THE UNIVERSITY
of EDINBURGH

Bacterial host attribution and bioinformatic characterisation of enteric bacteria *Salmonella enterica* and *Escherichia coli* from different hosts and environments

Nadejda Lupolova

Doctor of Philosophy

The University of Edinburgh

2018

Signed declaration

I, Nadejda Lupolova, confirm that the work presented in this thesis was composed by me and is my own. I confirm that the work has not be submitted for any other degree of professional qualification. Where information has been derived from other sources, confirm that this have been indicated in the thesis. The included publications in chapters 3.2, 3.3 are my own work. I also declare that publication that included in chapter 3.1, is jointly-authored publication and my contribution explicitly indicated below.

The work presented in Chapter 3.1 was previously published in Scientific reports as 'Phylogenomic approaches to determine the zoonotic potential of Shiga toxin-producing *Escherichia coli* (STEC) isolated from Zambian dairy cattle' by Geoffrey Mainda, Nadejda Lupolova, Linda Sikakwa, Paul R. Bessell, John B. Muma, Deborah V. Hoyle, Sean P. McAteer, Kirsty Gibbs, Nicola J. Williams, Samuel K. Sheppard, Roberto M. La Ragione, Guido Cordoni, Sally A. Argyle, Sam Wagner, Margo E. Chase-Topping, Timothy J. Dallman, Mark P. Stevens, Barend M. C. Bronsvoort and David L. Gally. N. Lupolova is an equal contribution first author with G. Mainda and the author of this declaration and David L. Gally is a supervisor of this thesis. I carried out all the bioinformatics analyses starting from short read quality assessment, all the way through to

reference genome mapping, development of bioinformatics pipelines for phylo- and serotype, phylogenetic tree building, interpretation and visualisation of results as well as together with other authors writing and editing the paper.

Nadejda Lupolova

September 2017

Acknowledgements

I would like to thank my incredibly inspiring and infinitely patient supervisor Prof. David Gally, for his trust, guidance and support during all stages of this work.

I thank Dr. Andy Law, Dr. Konrad Rawlik, Dr. Bryan Wee and Sharif Shaaban for their time over bioinformatics discussions.

We are grateful to all who helped gather our collection and provide us with sequences: Dr. Tim Dallman, Public Health England, Prof. Nicola Williams, University of Liverpool, Prof. Roberto Laragione, University of Surrey and Geoffrey Mainda, The Roslin Institute.

Finally, I would like to thank Dr. Nicola Holden, Prof. Ross Fitzgerald, Prof. Mark Stevens and Dr. Ross Houston for agreeing to be on my Thesis Committee and their comments to improve the research being performed.

Andre and Arina, you are my main driving force, no words can describe my gratitude for your understanding over these times!

Contents

Glossary

- AHC: Agglomerative Hierarchical Clustering
- ANI: Average Nucleotide Identity
- BIGSdb: The Bacterial Isolate Genome Sequence Database
- CGE: Center for Genomic Epidemiology
- CLIMB: Cloud Infrastructure for Microbial Bioinformatics
- DDH: DNA-DNA Hybridization
- DHC: Divisive Hierarchical Clustering
- DNA: Deoxyribonucleic acid
- DRT: Dimensionality Reduction Techniques
- EU: European Union
- GB: Gigabytes
- GEBA: Genomic Encyclopedia of Bacteria and Archaea
- GOLD: Genomes OnLine Database
- HAP: Host Associated Proteins
- HC: Hierarchical Clustering
- HIV: Human Immunodeficiency Virus
- HGT: Horizontal Gene Transfer
- LDA: Latent Dirichlet Allocation

- LPS: Lipopolysaccharide
- MDS: Multidimensional Scaling
- ML: Machine Learning
- MLEE: Multi-Locus Enzyme Electrophoresis
- MLST: Multilocus Sequence Typing
- NCBI: National Center for Biotechnology Information
- NN: Neural Network
- ONT: Oxforfd Nanopore technologies
- PacBio: Pacific Biosciences
- PCA: Principal Component Analysis
- PCR: Polymerase Chain Reaction
- rRNA: ribosomal Ribonucleic Acid
- RDP: Ribosomal Database Project
- RF: Random Forest
- rRNA: ribosomal Ribonucleic Acid
- sML: supervised Machine Learning
- SMRTs: Single Molecule Real Time sequencing
- ST: Sequence Type
- STEC: Shiga Toxin Producing *Escherichia coli*
- STm: *Salmonella enterica* serovar Typhimurium
- SVM: Support Vector Machine
- t-SNE: t-distributed Stochastic Neighbor Embedding

- uML: Unsupervised Machine Learning
- WGS: Whole Genome Sequencing

Lay Summary

There is an extensive history of research aimed at identifying virulent vs harmless bacteria and this has led to a 'pathotype' classification usually based around specific virulence genes important for infection or damage. Over the last twenty years it has become apparent that these virulence determinants are often horizontally-acquired on large regions of DNA, usually as integrated bacteriophage (prophage) regions or plasmids. With the advent of relatively low cost whole genome sequencing (WGS) techniques and advances in computing, it is now possible to obtain sequences from large numbers of bacterial strains and interrogate these in relation to both their core and accessory genomes. While there are some bacterial species with preferred hosts, especially in terms of disease, there has been no real systematic investigation of host and niche specificity associated with strains of *Escherichia coli* and *Salmonella* Typhimurium despite the fact that these bacteria can be isolated from many different host species and environments. The main aim of this project was to determine if host and/or niche-specific proteins could be

identified for 'multi-host adapted' bacteria such as *E. coli* and *Salmonella* Typhimurium and this information used to predict both the 'origin' of a strain and its potential to infect humans only using its genome sequence. For this research, two datasets of 'multihost' bacterial sequences were analysed: 1203 *S.* Typhimurium isolates from 4 hosts (avian, bovine, human, swine) and *E. coli* from 6 host (avian, bovine, canine, environmental, human, swine). Classical core genome analysis which included core phylogeny, multilocus sequence typing and phylogrouping found no clear way to predict the host from which the bacterium was isolated. Moreover, some of the methods were found impractical for analysis of large datasets (i.e. multiple sequence alignment) or the methods lacked sufficient resolution (MLST). The accessory genome was then investigated, and accessory host associated proteins (HAP) were found for each of the bacteria/host groups. The threshold for protein extraction could be changed, thus different numbers of HAPs could be extracted based on this threshold. At the setting used for further analysis, the average number of proteins associated with a host group was 648 for E. coli ranging from a minimum of 73 proteins for the human group (number of isolates = 409) to a maximum of 1311 for the canine group. It was interesting to note that a small number of HAPs was not necessarily a reflection of a reduced number of isolates in the group (compare: canine isolates = 36, number of HAP = 1311; bovine isolates = 703, number of HAP = 1090). For the more balanced *S.* Typhimurium dataset there was an average of 300 isolates per host with the minimum number of

187 HAPs in the human group and maximum of 685 in avian group. These proteins were used to build a machine learning classifier (support vector machine (SVM) to predict the isolation host of the bacterial isolates. The majority of isolates from both species were predicted correctly with prediction accuracy ranging from 67% to 90%. For both species the most challenging were bovine and swine host groups as these two had many features in common. The approach allowed both 'generalist' and 'specialist' strains from each host group to be estimated as well identifying genes potentially required for successful transmission between species, although these would have to be verified experimentally. This work has shown that the *E. coli* or *S. enterica* isolate sequences can be used as a baseline for prediction and quantification of human zoonotic potential as was demonstrated using *E. coli* O157 and *Salmonella* Typhi as examples. Overall this part of the research showed marked host restriction for both *S. enterica* and *E. coli*, with only a limited subset of isolates exhibiting host promiscuity by analysis of predicted protein content. Machine learning can be successfully applied to interrogate source attribution of bacterial isolates and has the capacity to predict zoonotic potential. Using the same machine learning approach another question was asked about how alike are known zoonotic pathogens. In the work described above all *E. coli* O157, independent of the isolation host, were scored as potentially zoonotic when compared to the wider *E. coli* population. However when only *E. coli* O157 isolates were studied then the approach identified a small subset of cattle strains with

predicted protein content associated with human strains and so were considered (and scored) more likely to be a serious threat to human health. This approach was verified with *E. coli* O157 human outbreak strains traced back to food sources. All of the outbreak strains independent of food or animal origin were scored as a 'human'. This finding has profound implications for public health management of disease because interventions in cattle, such a vaccination, could be targeted at herds carrying strains of high zoonotic potential. The final part of the research compared how effective different techniques and machine learning algorithms were at predicting the isolation host using the *S.* Typhimurium dataset. Dimensionality reduction techniques as well as unsupervised and supervised machine learning methods were applied to HAPs. All 3 supervised ML classifiers resulted in very comparable high levels of prediction (over 95%) while other methods were unsuccessful. Thus choice of which supervised classifier to use for host prediction should be based on the knowledge of the user and any requirements for downstream analysis. To conclude, supervised machine learning methods can be used successfully to predict the potential source of a bacterial isolate and quantify its infection threat to humans. The methods described here can be applied more broadly and have implications for monitoring, identification and targeted interventions applied to potentially zoonotic infections. The success of these approaches is dependent on high numbers of sequences with accurate metadata including the origin of the isolates.

Abstract

With the advent of relatively low cost whole genome sequencing (WGS), it is now possible to obtain sequences from large numbers of bacterial strains and interrogate their core and accessory genomes in relation to associated metadata. While there are some bacterial species with preferred hosts, especially in terms of disease, there has been no real systematic genomic investigation of host and niche specificity of 'generalist' bacteria, i.e., those that can be isolated from multiple hosts and environments.

The main aim of this research was to determine if host and/or niche-specific proteins can be identified for 'multi-host adapted' bacteria such as *E. coli* and *Salmonella* Typhimurium (STm) in order to predict the 'origin' of a strain and its zoonotic potential from its sequence.

Two datasets of 'multi-host' bacteria were analysed: 1,203 STm isolates from 4

hosts (avian, bovine, human and swine) and *E. coli* from 6 hosts (avian, bovine, canine, environmental, human and swine). Based on classical core genome analysis such as core phylogeny, multilocus sequence typing and phylogrouping, no strong correlations with host were identified.

The accessory genome was also investigated for host-based associations, and accessory host associated proteins (HAP) were identified for each of the bacteria/host groups. These proteins were used to build a machine learning (ML) classifier - support vector machine (SVM) - to predict the isolation host of the bacterial isolates. The majority of the isolates from both species were predicted correctly with prediction accuracy ranging from 67% to 90%. For both bacterial species the most challenging were bovine and swine host groups as these two had many features in common. The approach allowed not only prediction of host based on WGS but also an assessment of how much the genome of particular isolates resembled the features of the genomes of the same species isolated from other hosts. This allowed 'generalist' and 'specialist' strains from each host group to be estimated as well as the sequences that indicate successful transmission potential between hosts. This work also showed that diverse collections of *E. coli* or STm can be used as a baseline for prediction and quantification of zoonotic potential as was demonstrated with *E. coli* O157 and *Salmonella* serovar Typhi. Overall this part of the research indicated marked host restriction for both STm and *E. coli*, with only limited

isolate subsets exhibiting host promiscuity based on predicted protein content. ML can be successfully applied to interrogate source attribution of bacterial isolates and has the capacity to predict zoonotic potential.

Using the same ML approach, another question was asked about how similar are the known zoonotic pathogens. When studied apart, *E. coli* O157 can be classified further into human and bovine isolates with only a small proportion of bovine isolates predicted as 'human', pointing to the specific cattle strains that are potentially a more serious threat to human health. This approach was tested with 2 independent sets of O157 human outbreak strains with traced-back isolates from animals and food. The outbreak strains independent of the origin were scored as 'human'. This finding has profound implications for public health management of disease because interventions in cattle, such a vaccination, could be targeted at herds carrying strains of high zoonotic potential.

The final section the thesis research was based on the STm dataset and compared different ML approaches to test which algorithm performed best for host prediction. Dimensionality reduction techniques as well as unsupervised and supervised ML were applied to HAP. Dimensionality reduction techniques and unsupervised ML were not able to split the dataset by host and produced different results which could be challenging to interpret correctly in terms of bio-

logical significance of the factors that influenced clustering. On the other hand, all three supervised classifiers resulted in very comparable high levels of prediction (over 95%). Thus, the choice of supervised classifier for host prediction should be based on the knowledge of the end-user as well as on requirements for any further analysis.

To conclude, accessory genomes were successfully used for extraction of host associated proteins as well as for prediction of source host and quantification of zoonotic potential for bacteria species that can be isolated from multiple hosts. The methods described here can be applied to other bacteria and overall have implications for monitoring, identification and targeted interventions associated with potentially zoonotic infections. The results are completely dependent on the dataset quality which should be as large and diverse as possible. The research highlights the predictive potential of such algorithms but also the need for bacterial sequences to be gathered with as much useful metadata as possible, including isolation host.

Objectives

The overall objective for this work was to investigate if host/niche adaptation signals can be identified from whole genome sequences of enteric bacteria *Escherichia coli* and *Salmonella enterica* serovar Typhimurium.

- Study the relationships between isolates with known provenance.
- Explore similarities and differences in core and accessory genomes.
- Identify genetic features that are associated with host/niche adaptation.
- Determine if specific isolates pose an increased zoonotic risk due to their capacity to thrive in different environments; i.e. have evolved as generalists.
- Develop algorithms for prediction of strain origin.
- Compare different algorithms for prediction of strain origin.

List of Figures

List of Tables

Chapter 1

Introduction

This thesis describes bioinformatics approaches to investigate bacterial genomes. Both bioinformatics and bacterial genomics are relatively young fields of research and are expanding quickly, thus many approaches that were innovative at the beginning of this PhD have become obsolete or inappropriate over the course of this study. It is fascinating 'building up a head of steam' however such work in new areas requires caution and careful consideration with interpretation as there are very few examples to compare the work with.

The boom that we have witnessed in the past 10 years in microbial genomics is due mainly to one reason: the democratisation of sequencing. Illumina short read sequencing has become a quick and relatively affordable solution for many. Increased sequence dataset sizes and their analysis have lead to our realisation about the genetic complexity of many bacterial populations and resulted in the introduction of new terms such as 'pangenome'; ideas like the 'bacterial genome continuum' and projects like the Human Microbiome demonstrate weaknesses of a reductionist approach and inabilities of small datasets to correctly reflect reality. They often also point to a need for a more holistic view of problems. All of the above urge the development of new ways to analyse data, that are also helped immensely by another revolution, in computing, that has lead to the availability of clusters and clouds where all these large sequencing projects can be curated and stored as well as analysed.

This chapter introduces bacterial genomics as a discipline, gives an overview of the species that were investigated during this work, *Salmonella enterica* and *Escherichia coli*, briefly touches on the evolution of sequencing and its consequences for bacterial genomics as well as importance for public health. Moreover, this chapter provides brief reviews of the methods and concepts that were used in bacterial genomics before the advent of whole genome sequencing (WGS) and those that have now become the main players, dealing with larger data sets empowered by modern computing capacity.

1.1 Bacteria and bacterial classification

The main work was carried out with datasets from two bacterial species: *E. coli* and *Salmonella enterica*. Both are rod-shaped, Gram negative, facultatively anaerobic bacteria. *E. coli* is a common but low abundance commensal of the gastrointestinal tract of many mammals [1] and has been associated with a wide range of infections in both humans and animals with certain strains able to cause life threatening zoonotic infections. A long record of association with human and animal disease means that certain strains represent a health threat with significant costs to society [2] [3] [4] [5]. It is evident that *E. coli* can thrive in a wide range of hosts and ecological niches and while it is one of the first species of bacteria to colonise the human gut [6] [7] [8], it can exist, at least temporarily, outside of its 'primary' host habitat in soil, water, sediments and

plant tissues [9] [10]. Moreover, *Escherichia coli* is a model organism; strains K12 and B and their derivatives have helped advance our understanding of molecular biology, genetics and gene engineering.

E. coli seems to succeed in different environments with diverse conditions including variations in temperature and pH, and in the face of challenges such as an immune responses or antibiotic therapy. Some strains appear highly adapted to a particular niche, for example it is now appreciated that Shigella strains can be considered as part of the *Escherichia coli* species and these, to date, have only been found in humans and primates [11] and can cause diarrhoea in humans [2]. Other strains such as *E. coli* O157 seem to be well adapted not only to a subset of hosts like cattle (and perhaps sheep), but to a specific niche in the intestine of these ruminants [12]. Other strains, for example porcine, bovine and human enterotoxigenic *E. coli* are considered relatively host specific but this has been attributed to specific combinations of adhesins, although there is no evidence to suggest they could switch hosts solely based on exchange of these colonization factors.

The first stark indication of the diversity in *E. coli* followed the sequencing of a strain of *E. coli* O157:H7 and its comparison with *E. coli* K12. From this, 1,387 'new' genes were identified with the finding that *E. coli* O157 had a sig-

nificantly larger genome (5.5 Mb compared to 4.2 Mb) than the K12 strain. The *E. coli* O157:H7 strain possessed specific virulence factors as well as different metabolic capacities [13]. Although conversely, some studies show that there is much less diversity among certain *E. coli* pathovars than was previously anticipated [14]. Overall, it is now clear that the *E. coli* genome can exhibit incredible plasticity associated with horizontal gene transfer by bacteriophages and plasmids. As a consequence, these changes do not necessarily need many generations to consolidate and therefore can be associated with the rapid emergence of different and sometimes virulent strains, such as occurred in the 2011 atypical enterohaemorrhagic *E. coli* outbreak in Northern Germany [15]. An *E. coli* genome can perhaps be altered endlessly leading to appearance of new strains and/or rapid specialisation of existing strains. This same plasticity is important when considering adaptation to human interventions such as antibiotic treatment. Such selective pressures help to develop, maintain and potentially combine resistance and virulence traits.

Taxonomy of *Salmonella* is complex, with two main species *S. enterica* and *S. bongori*; with the latter species sometimes called the 'Salmonella of lizards', as this was first isolated from a lizard and for many years it was thought to be host restricted. Nevertheless, recently it has been isolated from dogs, birds and in some cases humans. [16] [17] [18].

S. enterica can be further divided into six subspecies *enterica, salamae, arizonae, diarizonae, indica*, and *houtenae* more often called by Roman numerals I, II, IIIa, IIIb, IV, and VI respectively. Both species, *S. enterica* and *S. bongori* can cause gastrointestinal disease - salmonellosis, however subspecies I (S. *enterica* subsp. *enterica*) is responsible for the vast majority of infections and is one of the most common bacterial infections in humans and animals worldwide. Diseases caused by *S. enterica* vary from self-limiting enterocolitis with very mild symptoms to severe systemic infections, as with the example of typhoid fever. The most common manifestation of salmonellosis is diarrhoea with fever, abdominal cramps, and vomiting.

Species of Salmonella are closely related to *E. coli* and they are estimated to have diverged from a common ancestor 100 million years ago; their genomes still display significant similarity, hence many functional identities. Many of the genes which are unique to *Salmonella* serovars, compared to *E. coli*, are found on large discrete genomic islands such as *Salmonella* pathogenicity islands (SPIs). These Salmonella-'specific' determinants include many genes associated with virulence and characterise the divergence of *S. enterica* from *S. bongori*. For instance, the SPI-2 genes which encode a type III secretion system present in *S. enterica* are absent in *S. bongori*.

Apart from acting as a blueprint for an organism, DNA contains information that allows us to understand phylogeny, the evolutionary path, antimicrobial resistance and environmental adaptation. Billions of years of evolution have allowed, in most cases, tremendous genomic diversity to be generated and selected with subsequent complexity in bacterial metabolic and anabolic pathways. There are more differences between certain bacterial species than between all eukaryotes combined see Figure 1.1). As a scientific discipline bacterial genomics is concerned with all of the hereditary information that can be found in a bacterial genome. Recently it has become possible to analyse and compare whole genomes sequences of thousands of bacteria. These recent advances indicate that we have underestimated bacterial diversity.

For many decades bacterial species were classified by a variety of phenotypic characteristics such as morphological and anatomical features, culture characteristics, reaction to stains, differences in nutrition, with analysis of a series of biochemical reactions and the presence and absence of different antigens and their structure. Moreover sometimes there is a need to distinguish bacteria of the same species, for example to identify bacteria associated with a disease outbreak. In this case, schemes such as serotyping, enzyme typing, identification of toxins or other virulence factors, or characterization of plasmids have become the methods of choice.

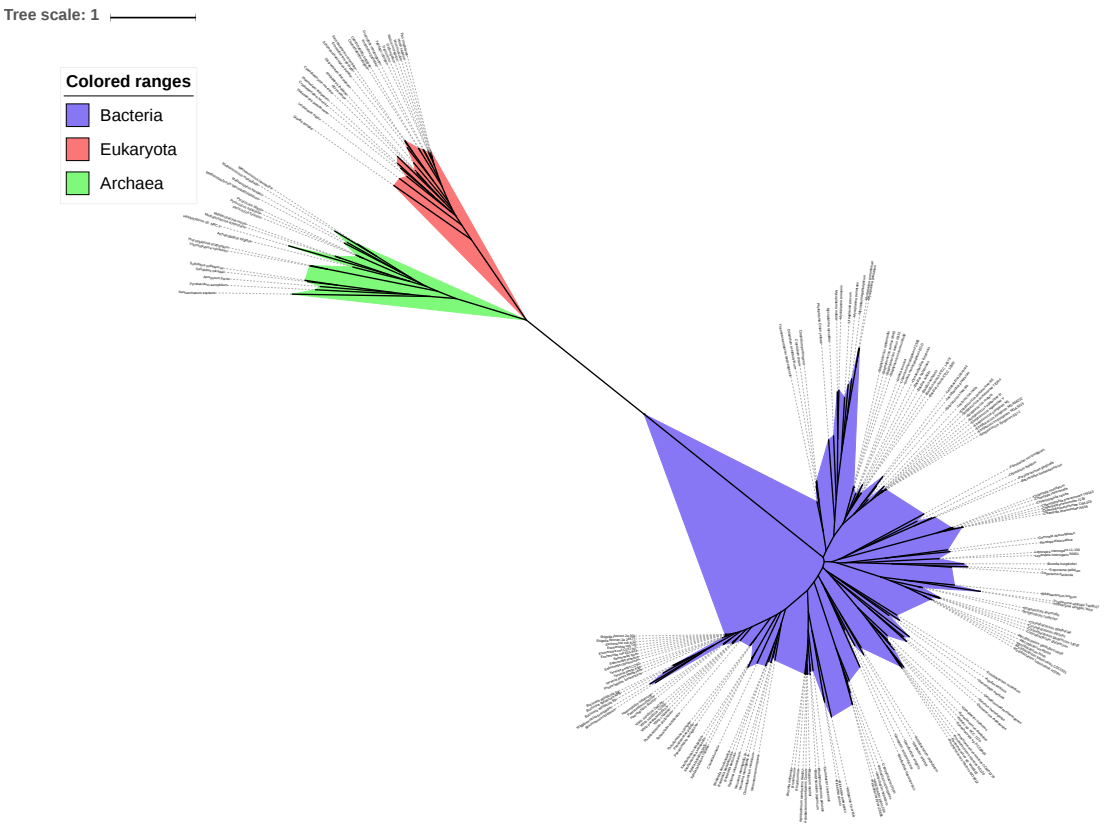


Figure 1.1: Tree of life adopted from iTOL. Relative diversity of three domains of live Bacteria (purple), Eukaryota (red) and Archaea (green) is shown. The phylogenetic tree built from concatenated alignment of 31 universal protein families and covers 191 species whose genomes have been fully sequenced. The tree scale represent substitutions per cite [19].

The first step towards genetic classification was based on DNA-DNA hybridization (DDH), a method that become a 'gold standard' for over 50 years. [20] [21]. The method measures the kinetics of re-association of the two strands of DNA from two bacteria, with the intuition that less similar DNA binds less efficiently, and therefore re-associates more quickly. The threshold of 70% was considered to be enough to call two bacterial isolates the same species, but in 2014 a new threshold with 79% similarity was proposed [22]. Even though DDH was used as the 'gold standard' for a long time, the method had its drawbacks as it was laborious and time consuming and there were difficulties in comparability and reproducibility.

Another successful method to establish relationships between isolates within the same species is the use of specific genes that are present in all sequences; these are then adopted into 'typing schemes'. From early work on 16S rRNA [23] and *gyrB* based phylogenetics [24] to development of new schemes such as MLEE, MLST, MLVA and PFGE [25] [26] [27] [28], these kinds of methods facilitate sharing and comparison of results between researchers, public health surveillance and outbreak investigations and serve as a means to reveal substructures of *E. coli* populations that sometimes can explain prototypical variation or shed light into the origin of strains.

The standard microbiological tests cannot distinguish from patients stool sample commensal, pathogenic and diarrhea causing *E. coli*. Advances in understanding of biology has led to multiple sub-qualifications of *E. coli* pathotypes that vary in pathogenic mechanism [29].

Summarised below are basic descriptions of the most common typing schemes that have been used in bacteria such as classification based on 16S rRNA, ANI, MLST, phylogroups and serogroups as well as some of the pros and cons of these methods. Arguably, the advent of whole genome sequencing (WGS) overcomes the majority of the pitfalls presented in these predecessor methods, although this information is often still extracted from WGS to allow comparison of strains, for example when only a subset has been subject to WGS.

1.1.0.1 16S rRNA

Over 30 years, ribosomal RNA operons (rRNA), specifically the 16S rRNA genes, have been used for taxonomic assignment and building of phylogenetic trees to determine microbial diversity and identify the phylogenetic position of novel isolates. 16S rRNA is so popular because at least one copy of the gene can be found in each bacterial species, the gene has not changed its function over time and contains both similar and variable regions that allow for easy PCR amplification and sequencing; followed by classification of the bacteria

into bacterial family, genus and species in the majority of the cases. Moreover the 16s RNA region is long enough (1,500 bp), so when bioinformatically searching for this region in more complex data, statistically weak hits can easily be removed.

16s rRNA is a part of 30S small subunit of the prokaryotic ribosome, and these genes have very slow mutational rates. Nevertheless the bacterial 16S gene contains regions of hyper-variability (V1-V10) by comparison of which the taxonomic classes can be detected. There is evidence that comparison of some of these regions (V4, V6) are more reliable and some (V2, V8) are less reliable for differentiating bacterial genera and species [30].

There is no clearly defined threshold for species identification based on 16S rRNA. The majority of cases can be resolved into species based on 97% similarity. On the other hand there is some evidence that even bacteria that have achieved this threshold of similarity can belong to different species [31] and conversely, isolates from bacteria of the same species, for example *E.coli*, can possess 16S rRNA regions when similarity is below than 97% (Chapter 3.1).

One of three separate methods for 16S rRNA analysis can be used: the first is to align short reads to a reference 16S RNA sequence; the second is to assem-

ble short reads; and third is to compare reads from the one of the most reliable hyper-variable regions. Each of these methods can introduce bias, however a recent study showed that an effective classifier can be achieved when just Illumina short reads from the V4 region, with an optimal size of 100-120bp, are compared between isolates [32]. The analysis of clinical specimens shows that the majority of isolates (83%) can be identified to species level. Those that could not be identified were due to limitations of the databases [33] and/or due to sequencing errors.

Overall, the 16S rRNA method can be used for bacterial classification to species level where the 16S RNA region is diverse enough to distinguish between species and highly similar within species. However for some bacteria such as *E. coli* (too diverse) or *Edwardsiella* species (too similar) 16S rRNA classification has its limitations. Nevertheless, 16S rRNA greatly aided the beginning of microbiome research when targeted 16S rRNA barcoding allowed whole microbial communities to be determined.

1.1.0.2 MLST

Multilocus sequence typing (MLST) is a technique by which bacterial isolates can be characterised based on comparison of DNA sequences from multiple internal fragments of common genes. Different alleles of these genes can

be found within populations, once a new allele is found and verified it will be assigned a number and stored in a database for that genus/species. MLST typing results in a sequence type (ST) which is a combination of the numbers obtained from allele assignments of each gene in a particular scheme. The first MLST scheme was developed for *Neisseria meningitidis* in 1997 [34] and consisted of 6 gene fragments. Nowadays there are 143 MLST schemes [35] with ST profiles ranging from 4 for *Salmonella* Typhi to a staggering 13,443 *Streptococcus pneumoniae* fragments.

Thus even though MLST was designed to simplify classification of bacteria into a few STs, it is clear that the diversity of the bacteria was underestimated and as the number of analysed bacterial sequences increased, so did the number of STs (see Figure 1.2). Over time ribosomal bacteriophage and plasmid MLST have been developed as well as nine eukaryotes MLST schemes [35].

There are two main methods to analyse MLST data. The first is to use software such as eBURST [37] to construct a minimum spanning tree or dendogram based on the pairwise differences between allelic profiles (i.e. numbers) assigned to each of the genes in a scheme [38]. Therefore, no meaningful phylogenetic tree can be constructed as just numbers of ST are assigned and compared. A second method uses the nucleotide sequences of these alleles

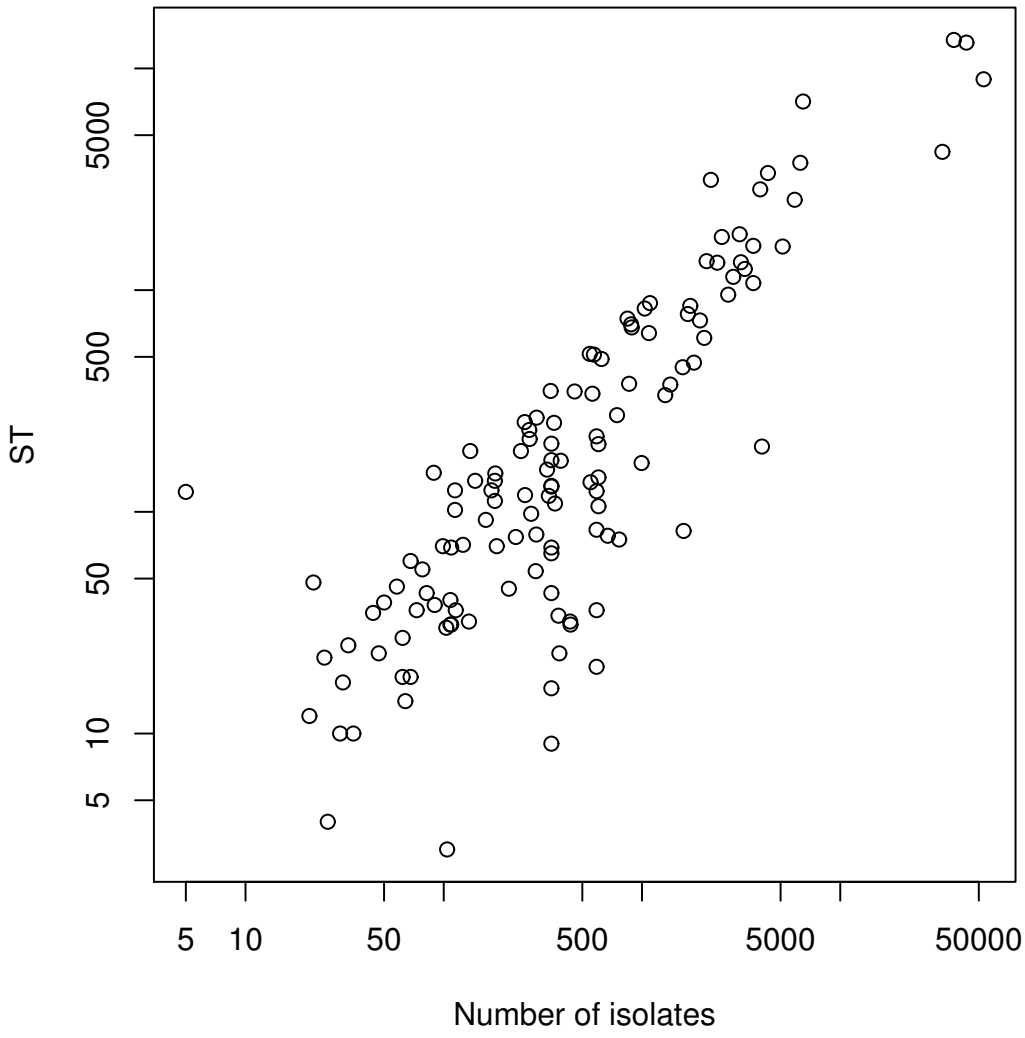


Figure 1.2: Sequence types (ST) vs number of analysed isolates for 133 existing MLST schemes of different bacterial species. The graph was generated from the data obtained from PubMLST database [36] on 24 July 2017.

rather than their allelic numbers. These alleles can be concatenated and their phylogeny inferred using a variety of schemes for SNPs substitution analysis. It seems that this method is more appropriate for bacteria that undergo clonal expansion than for bacteria with high rates of recombination [39].

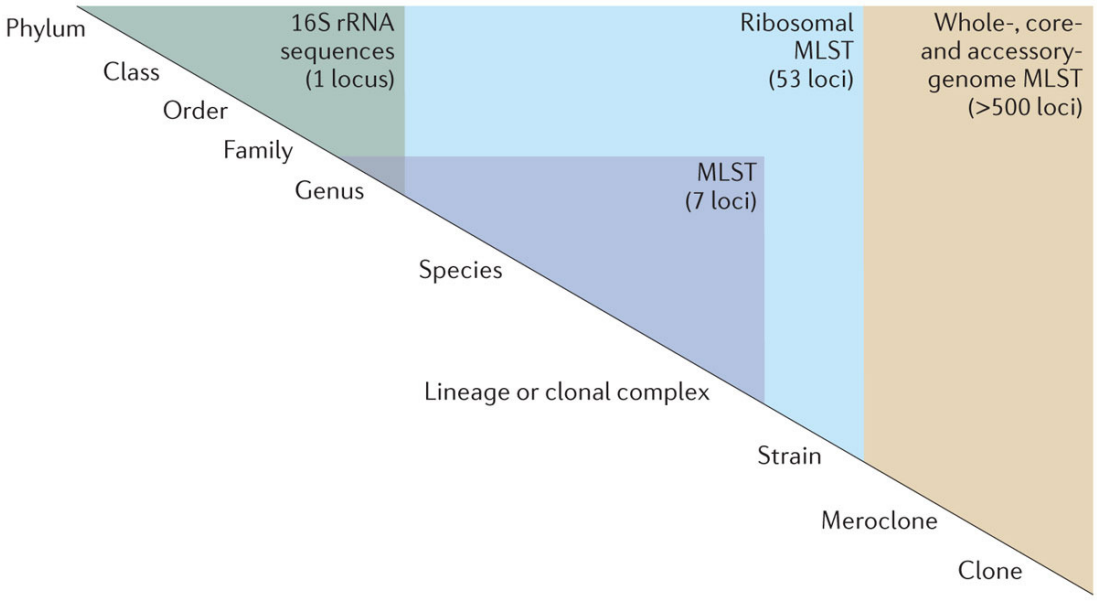


Figure 1.3: Level of resolution for different MLST schemes depends on the number of alleles in the scheme. [40]

MLST is based on the assumption that accumulation of mutations in housekeeping genes is a relatively slow process, thus allelic profiles continue to be stable enough over time allowing for species wide global epidemiological comparisons. It is very interesting that in some cases (as for *E. coli* and *S. enterica* in this work) MLST based phylogeny is comparable with core and whole genome analysis, showing quite remarkable similarities in terms of assigning isolates to the same clusters. However, MLST may lack resolution for certain

requirements such as outbreak investigations. To address these issues, the latest proposed MLST schemes generally extend to the whole core genome. The level of the resolution that such schemes can achieve can be seen in Figure 1.3.

Overall, MLST is a widely used, highly comparable and reproducible method that is often used for epidemiological investigations, although caution should be used when applied to highly recombinant bacteria. The schemes rely on well curated databases for their survival and nowadays that is sometimes based only on the pure enthusiasm and efforts of particular researchers. Moreover, any core gene MLST will need an adjustment when used for an individual collection that will require at least basic bioinformatic knowledge of the end-user. Also there is strong evidence that pangenome provides better resolution than MLST [41] and so the community needs to consider the resolution required and the value and costs of maintaining such schemes.

1.1.0.3 ANI

As introduced, DNA-DNA hybridisation methods have been used since the 1960s to indicate sequence similarities, however due to the laborious workflow involved and difficulties with comparison of the data obtained, [42] a new method based on Average Nucleotide Identity (ANI) [43] of sequences was

developed. The method compares genomes pairwise, each in turn serves as a reference while another is chopped to 'windows' of a desired size. Then these smaller sequences are mapped to the reference, the best BLAST score is noted and then the average nucleotide similarity is calculated. It is expected that ANI between sequences of the same species will be above 98-99%. Researchers also compared their method with DDH at 70% and found that this cut-off corresponds to 94% ANI and 65% or above of the conserved DNA.

Moreover, authors point out that even though the majority of species can be distinguished at 95-96 % of ANI, there are situations when the interspecies boundaries are as high as 99% based on ANI. One such example is *Bordetella* species that are found to be 99% identical by ANI definition however one of its species, *B. bronchiseptica*, possess 600 more genes compared with *B. pertussis* or *B. parapertussis*. Overall ANI is an easy to use, quick and reproducible method that can substitute DDH. However, this method uncovers a new problem in species classification; definition of the species should include not only genetic identity but also the ecological niche that will influence the number of extra genes present in an 'ecotype' genome.

Furthermore, if the two populations have almost identical genomes but do not compete with each other it can be considered as strong evidence that they are

2 different species. [44]. To conclude, there are some examples of successful use of ANI to delineate a species i.e. *Pseudomonas* species [45] and successful examples of use of ANI coupled with agglomerative clustering algorithm with an different cutoffs [46]. However, the exceptions list for ANI use even for already known species is too long, thus applying ANI to determine new species (or different sub-populations in vast ecological niches for the same species) is most likely a very unpractical choice.

1.1.0.4 Serotype

Serovars or serotypes indicate distinct variation within a species of bacteria based on their cell surface antigens. Serogroups are groups within serovars that allow within species classification. Classification is based on a set of unique reactions of cell surface antigens, usually based on antibody staining. The need for beyond species classification arises from the fact that even within a species bacteria are often quite diverse and exist with many different subgroups than can have quite different clinical manifestations. Thus, in clinical settings, serotyping can have a crucial role in identification and planning treatment and intervention options. Common antigens include: the 'O' antigen which is the outermost portion of the Lipopolysaccharide (LPS) and differ based on their chemical 'makeup'; the 'H' antigen which is based on the flagellar antigens which differ by protein content; K or capsular antigens and

F or fimbrial antigens [47]. *E. coli* can be classified into 150 - 200 serotypes while for *S. enterica* over 2,000 serovars can be distinguished. In *Salmonella* serovars can be grouped by relation to host: there are host adapted (that infect only one host like serovar Typhi - human), host restricted (that are restricted to a few hosts only like serovar Dublin (human and bovine) [48] and those that are found in multiple hosts such as serovar Typhimurium. Pathogenic *E. coli* are classified into pathotypes based on the production of broad classes of virulence factors and on the mechanisms by which they cause disease. Within each pathotype, strains are classified into virotypes or virulence gene profiles, based on the presence of combinations of virulence genes. Strains of a particular pathotype belong to a restricted number of serotypes.

1.1.0.5 Phylogeny

It is estimated that *Salmonella* Typhimurium and *E. coli* diverged from the common ancestor around 120 million years ago and since then have accumulated numerous point mutations as well as harvested and lost, by horizontal gene transfer (HGT), various parts of the chromosome [49].

The evidence for the genetic substructure of *E. coli* accumulated in the 20th century has led to development of a method that can easily and inexpensively

assign *E. coli* isolates to a certain phylogroup. Phylogroups are quite stable units in the population structure of *E. coli* and there were initially 4 main phylogroups that can give a indication of whether an isolate was commensal (phylogroups A and B1) or pathogenic (associated with phylogroups B2 and D). Initially, the method was based on the detection of two genes and one genetic fragment dividing all strains into 4 groups; a later method that analyses 5 genes and one genetic fragment could separate *E. coli* into 7 phylogroups and a cryptic clade allowing to gather under this classification all other *E. coli* that were very different in nucleotide composition but were very similar to all other *E. coli* phenotypically [50] [51] [52] [53]. The downside of the method when done not in silico is that it is still relatively time and labour consuming as well as subject to PCR related errors.

Salmonella enetrica subspecies *enterica* population structure for a long time was described as clonal, however recent studies show that five stable lineages with different estimated HGT levels can be delimited [54]. Moreover, STm population structure also can be defined by few lineages from which lineage I and II contain highly related sequences and lineages 3I and IV gather isolates with quite diverse sequences. Notably, some highly invasive sub-Saharan STm are estimated to have emerged independently and quite recently (around 50 years ago) due to spread of HIV and antibiotic treatment [55].

To conclude on classifications methods, overall, different schemes provide different perspectives about a bacterial population. Better resolution can be achieved when more information, such as gene content are included, although computational resources can be a limiting factor when analysing big datasets. Nevertheless, bacterial genomics has already moved from approaches reliant on either single or a few genes to whole genome models and these whole genome analyses are becoming routine.

1.1.1 The pan and core genome concept

Recent studies show that the diversity and number of the genes that have been estimated in nature are much larger than previously calculated [56] [57] [58] [59]. This reflects the billions of years of evolution over which bacteria have confronted Earth's ever changing environment, selecting the genes for all possible conditions. Mechanisms of horizontal gene transfer then allowed bacteria to exchange these genes and to reflect environmental changes by altering the genome. This is a comparatively rapid solution for adaptation compared to other 'gene-evolving' ways such as mutations, insertions and deletion. There are three main ways to acquire new genetic information by horizontal gene transfer:

- conjugation: DNA is directly exchanged as plasmids between bacterial

cells

- transformation: genetic material can be directly taken up from the environment
- transduction: the DNA is delivered by a virus (bacteriophages)

A few notes on the third 'method' of horizontal gene transfer as it becomes important later on (see Chapter 3.3). During the transduction process a bacteriophage may incorporate not only its own genome into the host bacterial genome but also can transfer genes of a previously infected bacterium by generalised or specialised transduction. The global population of bacteriophages is much higher than that of bacteria. There are 10^{23} phage invasion events per second [60], thus given the vast gene diversity already presented in bacteria and the capacity to transfer these genes in different combinations it has become evident that genome variation is staggering. Bacterial diversity could be seen at various levels, both between different taxonomic groups as shown on the tree of life 1.1 as well as within some bacterial species such as *E. coli* (3.1, 3.3 and STm 3.2. The currently sequenced 10 thousand unique bacterial species [61] as well as near 100 thousand different *E. coli* and STm isolates [62] are likely far from covering the whole bacterial population diversity both on a kingdom level as well as at species level.

To control HGT principally from bacteriophages, bacteria have defence systems that can occupy up to 10% of the bacterial DNA. Recent studies of bac-

terial defence systems classify them in a similar way to innate and adaptive immunity systems of higher organisms. So bacteria can modify its own DNA by methylation leading to restriction-modification systems, such that any DNA without these modifications can be recognised as foreign and dealt with this would be an example of an innate and quite general defence that is not specific in terms of recognition of the invader. By contrast, an example of adaptive immunity would be CRISPR-Cas systems that use phage genome fragments as transcripts to guide RNA to promote enzymatic cleavage of a 'remembered' invading phage genome. [63].

The gene collection within bacteria even of the same species can be quite diverse. Definition of a 'pan-genome' first appeared following analysis of 8 *S. agalactiae* genomes from which it was found that 80% of genes were shared amongst all of them, while 20% were only partially shared between genomes and some of these genes were strain specific [59]. Thus a pan-genome is all genes found within a group of organisms, usually of the same species. However pangenome, as a type of analysis can be applied to any groupings, one can try identify pangenome of Enteobacteria or look closely to just a serovar specific pangenome. Genes from pangenome can be sub-classified further: core genes present in all genomes in a group, and accessory genes, sometimes called non-essential genes, which would be present only in a fraction of genomes.

The core and pangenome concepts were established in 2005 by Tettelin in his work 'Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome' [59]. Core genes are found amongst all or the majority of strains (cut-off can vary) of the same species in a given dataset and the accessory genome (or variable genome or dispensable genes) are the genes that are present in only a proportion of the strains. Some authors divide the classification of the pangenome even further and distinguish singleton or strain specific genes these can be found only in a small proportion of the population.

The paper marked the beginning of new era in bacterial genomics and a new view of bacterial species; it was realised that there are species with limited diversity and therefore a closed pangenome, meaning that at some point it doesn't matter how many new isolate genomes are added into the collection, the number of new genes will not increase further (*B. anthracis*). On the other side there are bacterial species with an open pangenome, where new genes are constantly being detected as more isolates are sequenced as for example(*S. agalactiae*) [64]. Bacteria studied in this thesis have an open pangenome as in the case of *Escherichia coli* and the intermediate pangenome for *Salmonella enterica* (see Chapter 3.2).

Core genes usually are the genes that perform basic functions like fundamental growth and survival and some are responsible for species specific phenotypic traits. However, the number of core genes for a species with open pangenome would heavily depend on the number of isolates and sampling, i.e. bacteria from the same ST would share much more genes. Moreover, it is important to consider that with an average sequencing coverage of 30X (majority of isolates described in this thesis) there is a chance that some coding sequences can be missed. Another factor is that prediction/annotation software could miss some genes. Prodigal, the main underling annotation software for bacteria, estimates 5% of the error rate, mainly due to the stochastic nature of the algorithm. Also some small differences in the genetic sequence can lead to a frame shift and two almost identical sequences can be annotated differently, especially when predicting open reading frames and putative proteins.

It is important to note that core genes are not the same as essential genes. Firstly, because size of the core would heavily depend on the size and the diversity of the sub-population analysed. Secondly, the essential genes for each bacteria are dependant on the environment; therefore, for mixed populations, the core genome would contain some housekeeping genes but not in all of them.

Pan-genome analysis could also change the ways of how we classify different species. For example, it was found that *B. anthracis* differ from *B. cereus* only by two virulence-associated plasmids [64]. However, for species with an open pangenome like *E. coli*, definition of 'species' almost becomes meaningless with a genome-continuum paradigm where frontiers between recently called 'species' as an example *E. coli* and *Shigella* and *E. alberti* and *E. fergusonii* become blurred; and the more we sequence the more we find intermediate bacteria that can be classified by previous 'old' methods to either species.

The pangenome concept has lead to development of such new terms as pan-metabolism [65] and pan-regulon and these should incorporate all metabolic or regulation reactions respectively. The authors [65] concluded that open pan-genome is not reflected in the pan-metabolome, as a study of 29 *E. coli* found just over 1,500 total metabolic pathways from which almost 900 were core reactions. However, it is early days and metabolic pathway databases are still in their infancy, thus metabolic pathway databases like MetaCys and KEGG [66] [67] are far from complete. It is important to continue to fill such databases as in the future it could be used for such applications as predictions of metabolic pathways for the biodegradation of environmental toxins or biosynthesis of required speciality compounds.

Genome structure is quite variable in the bacterial world as well. So the proportion of proteins per genome (average coding density) is 85-90%. However, for some symbionts and pathogens, this number can be as low as 40% [68]. On the other hand, some genomes are quite redundant in terms of duplicated genes; pseudogenes and prophage related content also add to diversity of the bacterial DNA. On the surface, these additions often seem to have little evolutionary impact. Nevertheless from these immense levels of 'accidental' change, successful combinations emerge that contribute to better/faster adaptations to changes in the complex real-world environments encountered by many bacteria. It is this complexity in genomic information that should be amenable to investigation by machine-learning methods.

Accessory genes and in particular those which are rare singletons do not carry housekeeping function but contribute to species diversity and often can provide advantages when selective pressure is applied. Hypothetical proteins as well as phage- and transposon-associated genes often form the main bulk of the accessory genome in species like *E. coli* and seem not to be responsible for immediate survival as they are sparse. However, what are they, what are their functions? Traces of random gene exchange from the bacteria's past that easily can be lost, or do they contribute to bacterial fitness in specific envi-

ronments? In many cases species' evolution can be associated as much with gene loss as with acquisition [69] [70].

Of serious concern for human and animal kind is that bacterial pathogenesis and resistance are often linked to genes encoded in the accessory genome: these include a variety of virulence factors such as toxins and adhesins, antibiotic resistance genes, mobile elements such as plasmids, transposons, insertion sequences and phages [71]. In fact the pathogenicity of strains is often described by the distribution and expression of specific virulence factors such as toxins, adhesins, invasins and others that are encoded by either chromosomal or plasmid genes. Regions of chromosomal DNA that encode multiple genes linked to virulence, and are often horizontally required, are termed pathogenicity islands [70]. Pathogenic *E. coli* strains do not have a single evolutionary origin but may have arisen many times [26]. There is also the possibility that any *E. coli* strain can acquire appropriate virulence factors and give rise to a pathogenic form. However, it is debatable whether any strain can acquire any virulence factor as this also makes huge assumptions about the regulatory and other networks required to control the acquired genes apparently to become a successful pathogen.

From a bioinformatics point of view, note that even though the term 'pan-

genome' is used, blast search, alignment and comparison of sequences are usually carried out using translated amino acid sequences. There is some advantages of doing so. First, a redundant codon will count as a mutation. However, there is most likely no significant evolutionary pressure to prevent a silent mutation, so as long as the protein sequence is not altered these changes do not have to be calculated against the homology score. Second, statistical significance of an alignment will be more easily achieved when comparing alignments based on 20 different letters than alignment based on four letters. Third, some amino acid changes will not alter a protein dramatically, as for example isoleucine to valine, both similarly hydrophobic, these mutations can be accounted for in an amino acid alignment, while in DNA alignment this region will be treated as any other misalignment or substitution.

The number of genes (proteins) detected in any sequence can be influenced by various factors:

- Sequencing errors.
- Assembly errors: some of the genes will be lost as long as the assembly stays in a draft format, therefore contigs can be broken in the middle of a gene, so that gene is not detected.
- Annotation errors: Some software relies on annotation, nowadays annotation is the weakest link in many whole genome analysis steps, as the

amount of information increases dramatically while accuracy and verification of this information is often poor and slow. [72].

1.1.2 Zoonosis

Any disease or infection that is naturally transmissible from vertebrate animals to humans and vice-versa is classified as a zoonosis [73]. This is complex issue that involves interactions between animals, humans and their ecosystems and needs multidisciplinary collaboration and communication between public health and animal health professionals, environmentalist, microbiologist and many others to provide solutions that can benefit public health. There have been over 200 zoonotic diseases identified; listed below are the top priorities for the World Health Organisation (WHO).

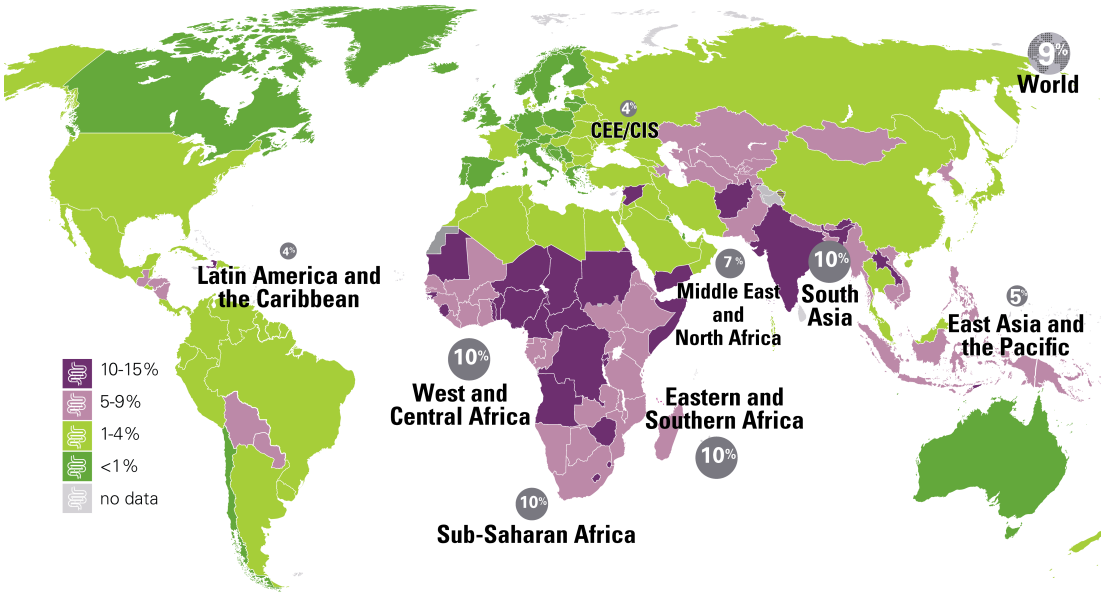


Figure 1.4: Map of incidence of diarrhoeal diseases worldwide in children under 5 years old. Figure adopted from WHO [73]

Main zoonotic diseases

- Anthrax
- Animal influenza
- Bovine Spongiform Encephalopathy (BSE)
- Food-borne zoonoses
- Haemorrhagic fevers
- Leptospirosis
- Prion diseases
- Tularaemia
- Variant Creutzfeldt-Jakob disease (vCJD)

The food-borne zoonotic diseases are estimated to kill 525,000 children and cause 1.7 billion incidences of childhood diarrhoea every year; they are the the major cause of mortality and morbidity for children under 5 worldwide (Figure 1.4). Diarrhoea is the result of acquired infection, with *Rotavirus* and *Escherichia coli* as the most common etiological agents that can be transmitted through water contaminated by human or animal faeces; food is another major source. The majority of these cases are preventable. However many general behaviours and practices need to be changed to achieve this. Currently, more money is spent on treatment of HIV than on these infections, even though less people die from HIV each year than from diarrhoeal diseases.

The main zoonotic pathogens identified by WHO are listed below.

Main zoonotic pathogens

- Salmonella
- Campylobacter
- Anthrax
- Brucellosis
- Verotoxigenic Escherichia coli
- Leptospirosis
- Plague
- Q fever
- Shigellosis
- Tularaemia

The human - animal - ecosystem interface (HAEI) includes all types of contacts between human, animals and the environment; direct and indirect and all types of pathogens are transmitted through these contacts. It is clear that the zoonotic problems cannot be treated only by one sector and seen only as, for example, a 'medical problem'. Thus, prevention and control measurements should be addressed at all levels.

Public health authorities and epidemiologists are starting to use genome sequencing as day-to-day tools to detect and warn populations about bacterial outbreaks as well as to rapidly identify pathogens. WGS has opened new

horizons in diagnostics and surveillance of outbreaks due to its high resolution. There are no limits in the areas of surveillance as it can be used for human clinical diagnostics in the same way it can be applied for detection of microbes in veterinary contexts or from environmental samples. The challenge remains in centralising such analysis, i.e holding the central database where uploaded sequences can be compared with other sequences, assembled, sequence typed, AMR detected, etc. There are many databases that are partially doing this, for example Enterobase [62] or the Center for Genomic Epidemiology [74]. However, the need for centralisation is obvious if it is to be really beneficial on a global scale.

There are multiple risk factors associated with the emergence of zoonosis, so it could be argued that pathogens with taxonomic and ecological broad host ranges are more likely to become zoonotic. The capacity to be transmissible between humans is also a risk factor, as well as a predisposition for the acquisition of DNA and evolution based on a high mutation rate. As a consequence, certain viruses may be more likely to become zoonotic compared to certain bacteria, and in turn this is more likely than zoonotic emergence of, for example, certain helminths and other eukaryotic microbial pathogens. Of course, many parasites have multi-host life-cycles which complicates this simplified view.

Other factors associated with emergence of zoonotic diseases include land use and agricultural/industry changes, international travel and commerce, dietary and medical practices [75]: 70% of emerging zoonoses originate in wildlife and this is increasing over time [76]. Novel and more intensive interactions between humans and livestock increases the risk of 'spillover' [77], and disruption of natural ecosystems such as the loss of native forest leads to loss of buffer zones. Overall habitat fragmentation leads to loss of biodiversity with distortion of population densities leading to an increased risk of zoonoses emerging. Human behaviour also increases the risk of zoonoses. For example, illegal animal trade is the 4th largest global illegal industry that can be one origin of zoonotic spillovers as with the bush meat trade. Overall, this acts by enhancing the contact between wildlife species and humans.

Another group of factors that can lead to an increased risk of zoonotic diseases are those associated with changes in the host. Susceptibility to infection varies with age, underlying disease and immune status. By 2040, 25% of EU population will be 65 and older with an age-reduced immune competence. Moreover, obesity is constantly increasing in the industrialised world, thus leading to chronic conditions like type 2 diabetes and a compromised immune system. In general, with improving medical technology many more people will live with immunosuppressive chronic diseases (diabetes, cancer, HIV, transplants) but be at an increased risk of infectious diseases, including zoonoses. Major socio-

economic factors should also be included such as poverty, war and famine. Currently 60 million people are forcibly displaced and social vulnerability reduces disease resistance.

Finally, changes in the pathogen or to a pathogen are important causes of emerging zoonoses. Microorganisms are constantly evolving to allow for survival in new and changing environments. Bacteria can generate variation that enables them to infect a new host in which they can then potentially specialise with further adaptation. In parallel, there is always the possibility of evolution towards novel more virulent strains; in turn these can acquire resistance to antimicrobials generating some of the multi-drug resistant pathogens that are now a major threat to human health.

While epidemiologists can track and monitor outbreaks, there is a need for prevention, prediction and targeted interventions to tackle zoonotic pathogens. These have become more realistic objectives now that WGS of thousands of bacterial isolates can be achieved and big scale population studies can be computed, all due to advances in both sequence technology as well as computational hardware and software.

1.2 Challenges and advances in bacterial genomics

Hundreds of thousands of isolates from over 50 different bacterial phyla have been sequenced since 1995 when the first two bacterial genomes of *Haemophilus influenzae* and *Mycoplasma genitalium* were sequenced [78] [79]. A 'genome sequencing' search in PubMed [61] returned 94,119 matches, demonstrating increase in use of these words in publications over the years (see Figure 1.5). Adding the word 'bacteria' decreases the number of hits almost 4-fold to 24,229, reflecting a predominance of work on human genome in search of genetic markers for health and disease.

Nevertheless, bacterial genomics is catching up, with the same upward trajectory over the past decade (see Figure 1.5), also reflecting how humankind is prioritising based on existing knowledge. In 2010, after establishing a human gut microbial gene catalogue [80] the number of publications in bacterial genomics is slowly but steadily increasing, doubling its number with over 2,300 publications in 2016. All this data requires computational 'warehouses' that enable storage, processing and backup.

As up to date (November 2018) there were 187,229 genomes available at NCBI genbak ftp site `ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria`.

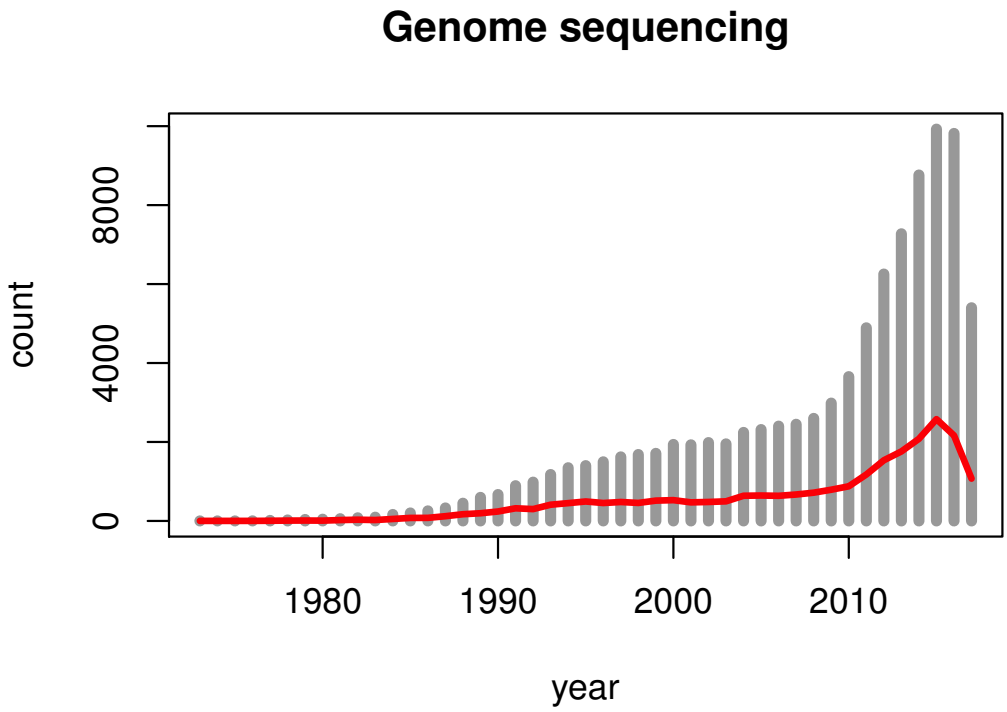


Figure 1.5: Number of sequencing related publications per annum. Number of publications (y axis) including words 'genome sequencing' that appeared in PubMed from 1970 to present (x axis). The red line represents the number of these publications which also contain the word 'bacteria'.

There is a number of publicly available databases that enable storage and often some analysis: National Center for Biotechnology Information NCBI [61], Genomes OnLine Database GOLD [81], Enterobase [62], the Genomic Encyclopedia of Bacteria and Archaea (GEBA) [82], Ribosomal Database Project, RDP [83], The Bacterial Isolate Genome Sequence Database BIGSdb with mostly MLST data [35]. Built-in tools for interaction with the sequences and their metadata often includes filtering and ordering based on metadata, sequence typing, clustering by similarity and some phylogeny.

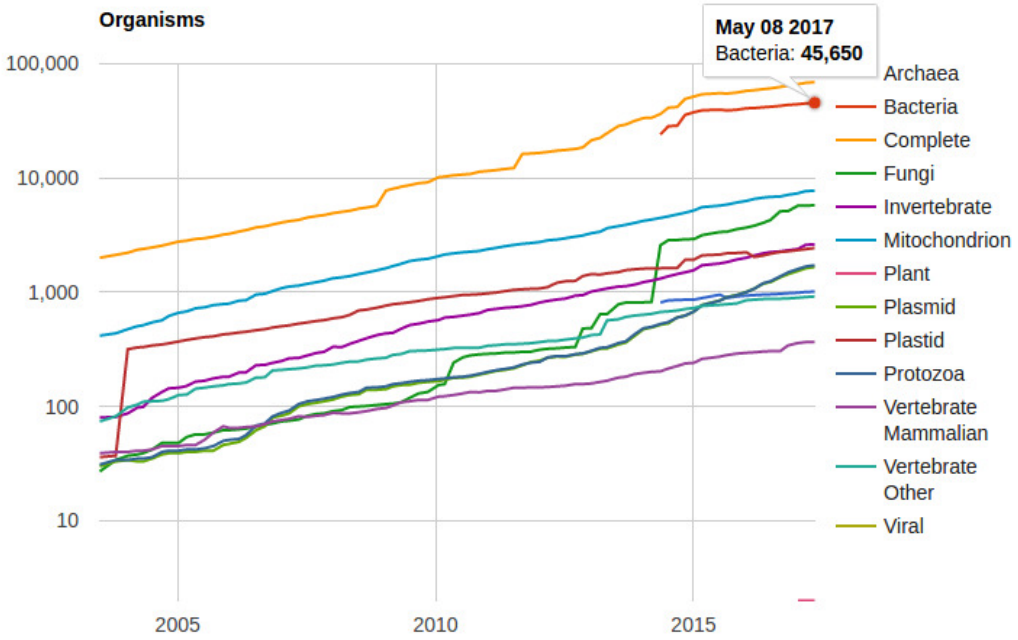


Figure 1.6: Growth in the number of genomes present RefSeq database from 2000 to present. It was 45,650 bacterial species on database as to 8-05-2017.

With regard to meta-data, there is still a huge debate around what information should be stored alongside the sequences. It is frustrating when one realises how little can be done handling only a sequence of the genome; all analysis becomes limited to comparison studies, and this is only if one is lucky enough to be working with a sequence which is similar to something that already exists in a database. Thus, some basic information is already mandatory for all of the above databases, however what is considered 'basic' can vary dramatically. Very often information can be deduced from the genetic sequence itself relatively easily: for example, sequence type (ST), serotype, phylogroup or antimicrobial resistnce (AMR) genes or virulence factor profiles. Nevertheless, most public databases already have in place 'minimal meta-data check lists' required for uploading raw sequence data and these check lists most likely

would include info such as: source of isolation, location and time as well as details about sequence manipulation such as sequencing platform or assembler, if applicable [84]. Thus, more recent submissions are improving the situation and are more likely to have appropriate information stored. However good metadata is still relatively sparse.

Over the time of this PhD research, new databases with new genomes have become available. One of these is EnteroBase, [62] the biggest database of enteric bacteria that contains 4 separate databases for S*almonella, Escherichia and Shigella, Yersinia and Moraxella*. EnteroBase aims to establish 'a world-class, one-stop, user-friendly, backwards-compatible but forward-looking genome database', coupled with a set of web-based tools such as EnteroBase Backend Pipeline, to enable bacteriologists to identify, analyse, quantify and visualize genomic variation principally within genera. The *Salmonella* subsection is the most populated database in Enterobase and contain 103,364 *Salmonella* genomes. The database curators at EnteroBase are trying hard to include metadata, so the sequences can be reused by others and can be used to address questions beyond those for which the original uploads were intended.

However, to illustrate a common state of affairs with metadata and the bias

towards human-related data, lets look at EnteroBase [62] which is one of the largest collections for enteric bacteria as well as one of the newest databases and therefore potentially have the best metadata. As of today (August, 5, 2017), there was just slightly over 100 thousand Salmonella genomes but for over 40 thousand of them there is no source of isolation information. One third of those with host information are isolated from humans. Moreover, human sampling is dominated by human pathogens: there are 5,625 sequenced genomes from *Salmonella* Typhi alone (specific human pathogen) while from *Salmonella* Typhimurium which can be found in different hosts there are 8,228 isolates with host information, 4,486 from human hosts and the remaining are unequally split between 32 different hosts. The biggest reservoirs for zoonotic split, such hosts as avian, bovine porcine each contain just slightly above 500 isolates). Therefore when planning future studies, principally those that require data acquisition, it would be very useful, where possible, to look for the opportunities to close existing gaps and try not to duplicate existing data.

Growth in the numbers of bacterial sequences in the NCBI RefSeq database are illustrated in Figure 1.6 with over 45,500 bacterial genomes in it. However, as expected reference genome collection is also heavily skewed towards model organisms and major pathogens, plus almost half of all the genomes sequenced are from the Proteobacteria phylum. Figure 1.7 shows bacteria whose number of sequenced genomes are above 1,000, with the 3 major play-

ers: *Streptococcus pneumoniae* with 8,267 genomes, *Staphylococcus aureus* - 7,997 genomes, *Salmonella enterica* - 7,426 genomes.

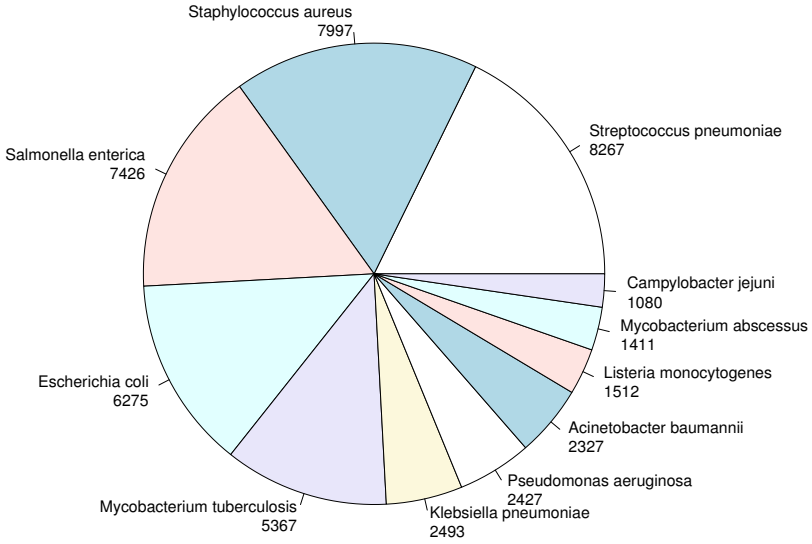


Figure 1.7: Bacterial species with more than 1,000 genomes sequenced in RefSeq. The pie chart showing the relative proportions of bacteria. Bacterial species names and numbers of sequenced genomes are shown.

In a genome sequencing project, the DNA of the target organism is broken up into millions of small pieces and processed on a sequencing machine. The resulting pieces of genomic informations are called reads, and depending on the sequencing technology can vary in length from few tens of bp to hundreds of thousands bp. Genome assembly is a process by which a large number of these reads are assembled back together in order to create a representation of the chromosome from which the DNA originated. In contrast to mapping, genome assembly assumes not prior knowledge of the genome structure and

composition.

The goal of a sequence assembler is to find overlaps and produce long contiguous pieces of sequence (contigs) from these reads. There are different ways to approach assembly, one of which is de Bruijn graph are described below as this strategy is implemented in the assembler used throughout this work (Spades [85]). The de Bruijn graph method of sequence assembly has its roots in Pevzner theoretical work from the late 1980s that studied the problem of reconstructing a genome sequence when only its set of constituent k-mers is known [86].

To construct the de Bruijn graph, each genomic read is broken into a sequence of overlapping k-mers. The distinct k-mers are added as vertices to the graph, and k-mers that originate from adjacent positions in a read are linked by an edge (see Figure 1.8). Then a path through the graph that visits each edge in the graph once - an Eulerian path - needs to be found. In practice, sequencing errors and sampling biases obscure the graph, so a complete Eulerian tour through the entire graph is typically not sought [**?**] [87]. Even when an Eulerian path through the entire graph can be found, it is unlikely to reflect an accurate sequence of the genome because of the presence of repeats, as there are a potentially exponential number of Eulerian traversals of the graph, only one of

which is correct [88]. In most instances, the assembler attempts to construct contigs consisting of the unambiguous, unbranching regions of the graph.

The choice of length of kmears affects the construction of the de Bruijn graph. Smaller values of k collapse more repeats together, making the graph more tangled. Larger values of k may fail to detect overlaps between reads, particularly in low coverage regions, making the graph more fragmented. Ideally, one should use smaller values of k in low-coverage regions (to reduce fragmentation) and larger values of k in high-coverage regions (to reduce repeat collapsing). Spades uses multi sized de Bruijn graph that minimise the above mentioned problems however not fully overcome them, and as a result the assembly for E coli and STm are fragmented with average number of contigs per genome 80. See 3.1.

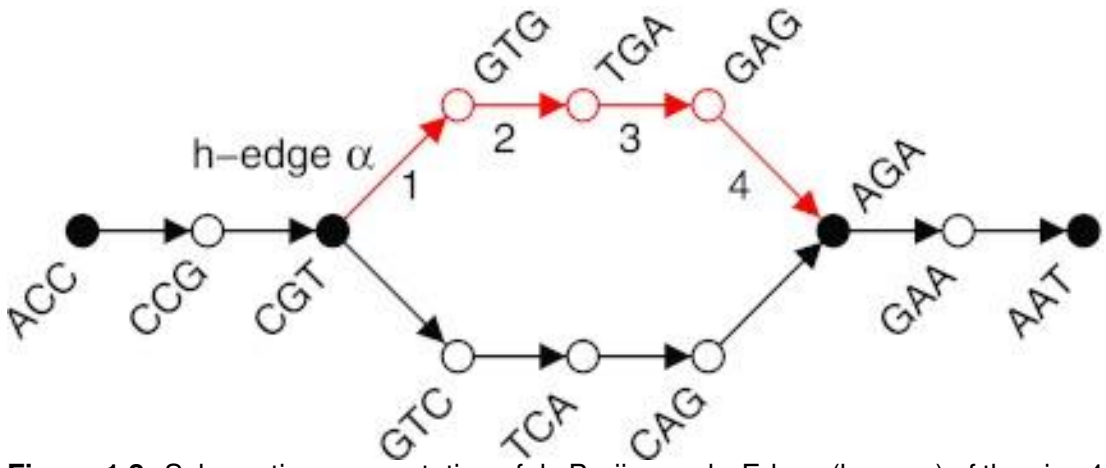


Figure 1.8: Schematic representation of de Bruijn graph. Edges (kmears) of the size 4 form vertexes of the size 3. The path can be read both ways (denoted by arrows) Adapted from [85]

A solution to the higly fragmented assembly is to produce longer reads. Third generation sequencing also known as single molecule real time sequencing (SMRT) done by companies such as Pacific Biosciences (PacBio) [89] and Oxford Nanopore Technologies (ONT) can produce much longer reads, however with much higher error rate. Compare PacBio: mean length = 10 kilobases, errors = 10%-15%; ONT: mean length = 2 kilobases, errors 65%-88%, Illumina: mean length = 150 bases, errors $<$ 2%. However, PacBio errors are distributed randomly, and as the technology now allows to sequence each strand of DNA multiple times (a pass), then when the number of passes is higher than 15, PacBio can show a extraordinary sequencing accuracy of 99.9% that is even greater than the 'Gold standard' Sanger technology. However, the cost of the genome sequencing by PacBio remains relatively high, limiting its use on a massive scale [90]. Nevertheless, long read technology was successfully used to define and compare integrated phages in EHEC [91]. All these sequencing advances are due to our better understanding of DNA (chemical and functional).

With the rapid increase in the number of sequenced genomes analysed, it has become evident that many approaches that worked for smaller datasets are inappropriate for large datasets. For example, whole genome multiple sequence alignments or building of core genome phylogenetic trees becomes very time and resources consuming, and therefore there is a need for develop-

ment of new better approaches and/or algorithms as well as further advances in computing capacities. Our computers can now store 10,000 times more information compared to when the first bacterial genome was sequenced; they can perform a quadrillion (10^{15}) calculations each second and we are building even bigger computers that aim to operate fifty times faster than is currently achieved https://exascaleproject.org/.

Nevertheless, bacterial sequencing technologies are improving faster than computational capabilities and today the reality is that analysing 20,000 bacterial genomes with approximately 5,000 genes per genome, in an all vs all protein comparison would take 4 months at the rate of a billion billion comparisons per second [92]. Building of core SNPs Maximum Likelihood tree from this work (see Figure 3.4) took 27 days at the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) [93] instance with 64 GB RAM and 12 CPUs.

The quantity of data also brings new requirements for visualisation: static ways to illustrate work are often no longer able to convey the complexity and granular detail of the data and fail to visualise the details necessary. To illustrate this point, one can compare a phylogenetic tree with 10 leaves vs 959 leaves (Figure 3.4), both printed on A4 format paper. Visualisation tools need to be quick and scalable in many directions: allow the user to quickly zoom in and

out; for example from a whole bacterial genome view to a specific 20 bp region. Ideally, the user could then zoom out for a view of thousands of genomes as well as then focus into a specific one from the dataset. Many such programs including visualising raw sequence data and visualisation, browsing and comparison of whole genome sequences have been developed recently [94].

A few new programs that allow for interactive visualisation and integration of different types of metadata are also becoming available:

- iTOL: Interactive tree of life

  https://itol.embl.de/
- Microreact: a Hierarchical and Geographical Analysis Tool

  https://microreact.org/showcase
- Phandango: an Interactive visualisation of genome phylogenies

  https://jameshadfield.github.io/phandango/

Even with such wonderful tools, a challenge still exists, for example, visualisation of the results from a 'thousand genomes project' for a publication is an act of pure creativity that involves not only out of ordinary thinking but excellent programming skills that allow implementation and the execution of the ideas.

Another logical and ongoing change is how we publish our research. Many

journals still expect that authors will supply their data in the form of spreadsheets, which over submission will be converted into pdfs. This ridiculous practice should stop, as such spreadsheets nowadays contain gigabytes (GB) of data that becomes virtually unusable as soon as they are converted into PDF. It also should be accepted that some details of research will be visualised not as a static pictures but with links provided to external resources, such as those from the previous paragraph.

In summary, over the last 20 years we have witnessed extraordinary changes in sequencing, computing, data management and data analysis. An exponential increase of the sequenced data has led to quick turnover of new software, as well as methods and tools in order to find new ways to interact and analyse complex multidimensional data.

1.3 Machine learning

Machine learning (ML) is a broad term referring to a class of algorithms that allows computers to 'learn' from experience, essentially enabling an algorithm to improve its predictive ability as it becomes exposed to data. In 1959, Arthur Samuel [95] defined machine learning as 'the ability to learn without being explicitly programmed'. ML applications in everyday life are diverse, from spam

filters to face recognition in airport security systems and targeted advertisement in your web-browser. ML shows promise not only in social and economic disciplines, but in the science and medicine arenas as well. ML is used in annotation pipelines such as Prodigal [96] as well as in skin cancer classification [97]. The Royal Society have recently pointed out importance of open frameworks, data and polices for successful advances of ML in the UK. [98].

An exponential increase of available datasets due to democratisation of sequencing has added to the captured complexity and understood natural heterogeneity of biological systems, which cannot be fully analysed with simplistic models and require: a) robust and easy to access databases that allow rapid data assessment and data feature retrieval; b) Development of machine learning models that can handle multi-dimensionallity. In biology, some models have already been implemented, e.g, for predicting macromolecule structure [99], tumour classification [100] [97], reconstruction of gene networks [101] and virtual drug discovery [102]. In bacterial genomics they have been applied to predict antibiotic resistance [103], solubility of recombinant proteins [104], and in clarification of taxonomic issues [105].

ML uses predictive analysis to identify patterns and hidden information on observed data without being explicitly programmed to do so. ML algorithms can

be divided into two main categories: supervised learning and unsupervised learning. uML attempts to automatically assign data to a number of groups, while supervised learning requires data that is already labelled in two or more groups. Unsupervised learning algorithms look for similarities/differences in data that were not previously labelled, trying to automatically classify them in independent groups. It is often used for initial data exploration, as no prior knowledge of the data structure is required. It is also easier to implement, as there are generally less parameters available to choose from. Another point in favour of unsupervised learning is that labelled data required for supervised learning is 'expensive' and often difficult to obtain, which is not required to run an unsupervised learning algorithm. For example, in public databases containing microbial genome sequences, information on the host of isolation is relatively rare and sometimes unspecific or incoherent (e.g., the label 'livestock' is sometimes employed, when a more detailed label would be necessary, requiring to manually relabel the data). Notwithstanding, the results of unsupervised ML can sometimes be hard to interpret, as the clustering can include features of little relevance for the subject of study.

With a number of clusters that should be defined a priori, uML would separate data starting with more obvious, dominant features. It is an excellent approach to explore data, however it could be an impractical choice when searching for particular answers that are based on subtle differences.

When the data is labelled, it is possible to use a class of more powerful algorithms based on supervised ML. The process by which a supervised algorithm is able to build a prediction is known as 'training' (see Figure 1.9). Once an algorithm has been trained on data, it is possible to apply the prediction to new, unseen data. The quality of a supervised algorithm depends on the type and amount of data, the number of features and the algorithm itself. If a learning algorithm is unable to perform a good prediction, this may be due to the lack of features to fit the data (underfitting), meaning that the model lacks complexity to describe the underlying data (low bias). The opposite may happen as well, when the model is overly complex and is thus able to describe very accurately the training data. Despite having a very small error on the training dataset, predictions performed on new data can be poor, and the algorithm is said to be overfitting.

The main algorithms that proved to work well with bacterial data to answer host quetions are Support Vector Machines (SVM) [106], Random Forest (RF) [107] and Neural Network (NN) [108]. SVM is a supervised learning algorithm that works by maximizing the margin (separation) between two sets of labelled data. It can naturally prevent overfitting, as the algorithm works my maximizing the margin between the two sets. SVM works well in the presence of many fea-

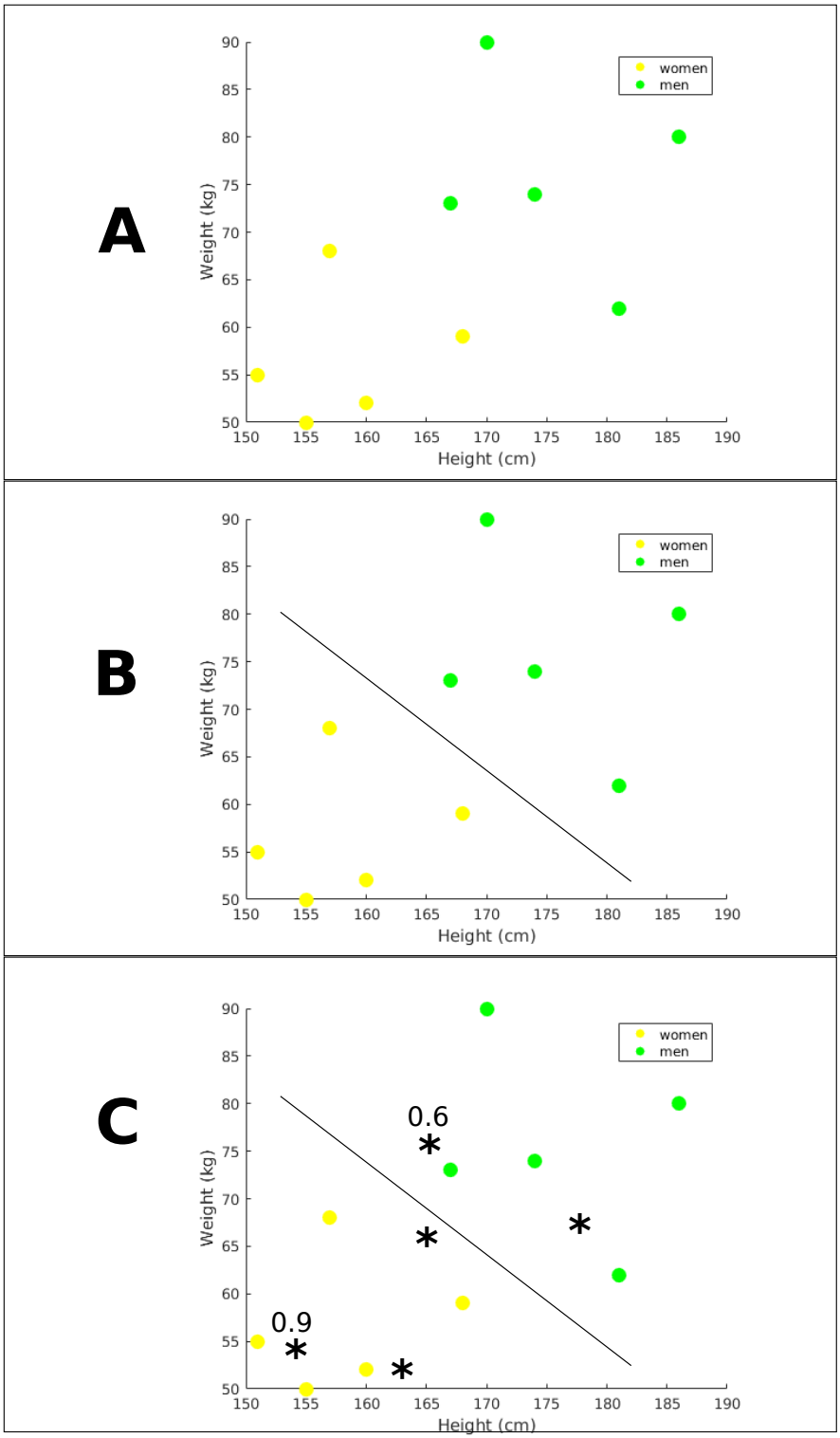


Figure 1.9: Simplified, linear ML classification example. **A** The simple dataset that consist of two classes and can be described by only 2 features represent here a training dataset. **B** During the training the rules of classification is calculated and decision boundary, here represented by line, are drown between 2 classes. **C** At the next stage, the new unseen before data can be introduced (stars) and predicted with higher or lower probabilities to belong to one of the classes.

tures, but training process can be computationally expensive. Once the training is done, it is easy to calculate predictions for new data. Some advantages of SVMs: High accuracy, nice theoretical guarantees regarding overfitting, and with an appropriate kernel that can work well even if the data is not linearly separable in the base feature space. Especially popular in text classification problems where very high-dimensional spaces are the norm; memory-intense, hard to interpret, and it may be difficult to execute and tune.

RF works by using decision trees generated with a random number of features, circumventing the overfitting that is usually seen when using regular decision trees. RF averages the results of these distinct decision trees, thus reducing the variance at the cost of bias.

Neural networks use an array (or layer) of 'neurons' which are able to apply and combine operations between a number of features with different weights. It is possible to create and model several of these layers in order to improve the accuracy of the algorithm at the expense of computational power. The algorithm is trained by finding the weights for each layer that maximize the prediction of the training dataset. Neural networks usually require large datasets and can be hard to find the right architecture to tackle a problem, but allow to solve very complex problems, for example in computer vision and robotics.

The major part of this thesis relies on machine learning classification that has been applied to identify 'host of isolation' for *Salmonella* serovar Typhimirium and *Escherichia coli* in Chapter 3.2, quantify zoonotic potential of a specific bacterial subgroup - *E. coli* serovar O157 in Chapter 3.3 and to see what patterns become dominant when different ML algorithms are applied to the same dataset in Chapter 3.4.

Chapter 2

Methods

Data

To distinguish between sub-datasets the following abbreviations are used: A - avian, B - bovine, C - canine, E - environmental, H - human, S - swine. Words 'sub-group' and 'sub-dataset' in this work mean a group of strains related by a host. The *Escherichia coli* dataset from multiple hosts was composed of all sequenced isolates in the possession of the group from previous projects and through collaboration, thus it is unbalanced in terms of host sources. The *Escherichia coli* dataset is publicly available as stated in [109]. The O157 *Escherichia coli* dataset is composed from UK and USA isolates from bovine host as well as clinical human isolates and is publicly available [110]. The *Escherichia coli* data were received in Illumina 1.9, paired-end, fastq format [111]. Raw reads vary in length from 32 to 251 bp. FastQC [112] was used to perform quality control, revealing that majority of data were good quality and did not need any further trimming. All datasets were checked for the presence of adaptors and some of the datasets (UK O157) were trimmed with cutadatp [113]. The *Salmonella enterica* dataset was downloaded from Enterobase [62] in form of draft genomes with number of contigs varied from 51 to 364. *Salmonella* dataset IDs and other metadata are published [114].

Assemblies

SPAdes [85] was used to assemble Illumina read sequences. After benchmarking a variety of options it was clear that better results are produced when a built-in error corrector is used instead of the recommended QUAKE [115] error corrector. To control mismatches and indels 'careful' flag proved to produce better results. Quality of assemblies were evaluated with QUAST software [116] - Quality Assessment Tool for Genome Assemblies. A wide range of statistical data from QUAST output for all assemblies were compiled into a spreadsheet describing assemblies in different parameters such as length of assembly, number and length of contigs, GC%, N metrics (N25, N50, N75), misassemblies report, number of Ns per 100kbp, gene statistics (unique, duplicated, genes larger than certain threshold).

Annotation

Annotation was carried out with PROKKA [117] - a prokaryotic genome annotation software. To achieve a better quality of annotation, a database with trusted proteins were gathered from reference quality *E. coli* genomes from NCBI. Use of custome database dramatically improved annotation and decreased number of 'hypothetical proteins'. For details refer to appendix 4.

Phylogenetic analysis

Maximum likelihood (ML) trees were constructed using RAxML [118]. To optimise the best tree search program was run with 500 rapid bootstrap (BS) followed by a slow Maximum Likelihood (ML) search under the GTRGAMMA model of heterogeneity.

To reconstruct phylogenetic relationships two separate approaches were used: one is based on the core genes and other is based on the pangenome. For core phylogeny core proteins present in 95% isolates were extracted from all sequences, aligned with MUSCLE, translated to nucleic acid and submitted to RAxML with parameters described above. The pangenome trees were extracted from ROARY output [119].

Typing schemes

The phylotyping described by [120] was used as a starting point to develop a small program that assigns each strain to the one of the 4 original possible phylogroups (A, B1, B2, D) based on the presence or absence of 4 genes: chuA, yjaA, TspE4.C2, arpA. To further distinguish between groups and assign strains to an additional 4 phylogroups (C, E, F, cryptic clade I) it was necessary to check for the presence of a fifth gene trpA and/or distinguish the specific alleles for the above genes. After performing all steps each *E. coli* sequence

were assigned to one of eight possible phylogroups.

To extract gene fragments from the genomes, a database that includes all sequences from the collection were build with BLAST+ [121]. The reference quality sequences for querying genes were downloaded from the NCBI website and blasted against *E. coli* dataset with *-max_target_seqs 1* parameter. The hit was considered positive when sequence length were covered < 90% with sequence similarity 99%.

MLST analysis was performed with SRST2 software [122] when short reads were available (*E. coli* dataset) and with MLST software developed by Torsten Seemann (unpublished, `https://github.com/tseemann/mlst`)
when analysing *Salmonella* dataset.

Serotyping was performed using command line blast [121] with curated alleles databse from SerotypeFinder gene databases [123].

Multiple gene alignment as for MLST, phylogroups, 16S rRNA phylogeny was carried out using MUSCLE [124]. Alignments were visualised with Geneious [125].

Protein clustering

First clustering results were obtained from Get_homologues [126] software. However, Get_homologues could not be scaled up for a larger dataset. See details in the Appendix 4. New software for clustering, Roary [119] was identified later on, and demonstrated promising speed and accuracy. The software takes as its input protein sequences extracted from .gff files from a PROKKA output. Different settings were trailed and for this work cut off for the assignment to the same cluster was set 95% of similarity at the amino acid level. To assign paralogs Roary uses synteny information of the nearest 5 genes both side from the gene in question. Due to draft quality of the assemblies it was decided not to split paralogs as this option originates many false positive new clusters, that is due to fact that the genes being on the ends of the contigs.

The computation for this work was performed on the HPC 'Eddie' facility from The University of Edinburgh and on CLIMB [93]. Data visualisation was done using R [127] and iTOL [19].

Other methods

Benchmarking of different methods and pipelines are described in the first year report 4. Any other methods are included in relevant chapter's methods sections or within papers.

Chapter 3

Results

3.1 Core genes phylogeny and analysis based on in silico typing schemes.

3.1.1 Introduction

There are three separate components in this chapter: 1) 1st year report in the appendix 4 ; 2) a published paper [109] (bounded); and 'core' analysis described below on a Chapter 3.2.

The first year report (see appendix 4) describes the work that was carried out during 1st year of this PhD and mainly addresses problems such as quality assessment of short Illumina reads, improvement of assemblies and annotation, outliers and in silico species classification based on sequence similarity: ANI calculator, 16S phylogeny, core genes similarity and some preliminary work on pangenome. Moreover, at that time only the *E. coli* dataset was available. As such, the work only superficially touched the main host question, therefore to maintain coherence of the thesis it was decided to omit that work from the main body of text.

Another work that forms part of the first chapter is a paper that is published as joined first authorship with my colleague, former PhD student Geoffrey Mainda [109]. My contribution to the paper 'Phylogenomic approaches to determine

the zoonotic potential of Shiga toxin-producing *Escherichia coli* (STEC) isolated from Zambian dairy cattle' was all the bioinformatics analyses starting from short read quality assessment, all the way through to reference genome mapping, development of bioinformatics pipelines for phylo- and serotype, phylogenetic tree building, interpretation and visualisation of results as well as writing and editing the paper.

The third component of this chapter describes and quantifies host associations that can be found based on typical typing schemes (MLST, serotype and phylogroup (phylogroup only for *E. coli*.) as well as core SNPs phylogeny. Strictly speaking, only MLST and core SNP phylogeny are core-genome analyses, while sero- and phylo-typing are looking not only to variation between specific genetics segments but also take into account presence or absence of some of the fragments, therefore not a strictly 'core'. These methods can still be regarded as 'typical' microbial genomics analysis [128].

Classical serological typing of *E. coli* is based on the O, H, and K surface antigens, first described by Kauffman in the 1940s and developed by Frits and Orskov, who made seminal contributions to *E. coli* typing. The O antigen forms part of the *E. coli* lipopolysaccharide (LPS) with over 180 different types; usually *fliC* gene encoding antigen of type H, which is the major part of flagella (53 antigenic types) and K relating to the capsular polysaccharide (CPS) which is

a thick, mucous-like layer of polysaccharide that surrounds some E. coli (>80 types). These antigens can be present in any combination, leading to an enormous number of strains that differ in their immunological profile. No strong association between serotype and host can be found except Shigella, that is known to be a specific human pathogen and to date was isolated only from humans and primates. Serogroup O55:H7 also seems to be restricted to human, however the number of isolates is still small: so in Enterobase there is 67 *E. coli* O55:H7 isolates and all from human host, 14 other O55 marked as environmental or food, however these contain different H antigenes. Nevertheless, it is known that some *E. coli* are more often associated with human disease (i.e shiga-toxin producing *E. coli* O157, O26:H11, O145:H28, O131).

Salmonella enterica can also be serotyped using the above antigens, as combinations of antigens differ between serovars. STm would contain O and H, while *S.* serovar Typhi and *S.* serovar Dublin would in addition contain capsular antigen. Even though by knowing *Salmonella enterica* serovar one can make some conclusions about its host (i.e. host restricted and host adapted serovars), combinations of antigens in *Salmonella* are not restricted to host. For example, serovar Typhi (only human pathogen) and Dublin (bovine, ovine), both contain all three antigens.

The major argument in favour of the theory that not any strain will acquire any virulence factor is the evidence of division of *E. coli* into phylogroups. Based on multi-locus enzyme electrophoresis (MLEE), Ochman and Selander [49] gathered a collection of 72 isolates from different mammalian hosts to represent the diversity of *E. coli* [25]. They noted that a small number of core genes can be used to organise *E. coli* into phylogroups. Importantly, certain phylogroups are known to contain the majority of strains that are pathogenic in humans (B2 and D) while others are known to contain many animal and human commensals (A and B1). This indicates that there are core lineage differences in the evolution of virulent strains arguing against the idea that acquisition of key virulence factors into 'any' background can produce a pathogen. As such it is a much more complex issue trying to determine which strains might emerge as significant pathogens by simple acquisition.

There is very limited evidence that phylogroup is associated with the host except a recent publication [129] in which a collection of 391 *E.coli* isolates from 3 hosts was phylogrouped and there was a statistically significant association of phylogroups A and F with poultry, B1 and E were associated with cattle and B2 and D with water buffalo. Study of water [130] isolates showed that these were of mixed origin but mostly from A and B1 phylogroups, with ration changes between these two phylogroups between dry and wet periods [130]. Stable number of isolates from B1 phylogrops over the year can indicate water

adaptation of B1 phylogroup.

Rather than focusing solely on virulence, the primary aim of the research proposed here is to ascertain if it is possible to predict the likely 'source' of an *E. coli* and STm isolate based on its WGS. The hypothesis underlying this research is that bacterial strains have evolved to replicate optimally in a specific environment. This does not preclude replicating and being 'successful' in multiple environments, but that each strain has an optimum niche. Extrapolating from this, it is likely that certain strains may have a more generic capacity to succeed in multiple environments than others based on their 'primary' habitat. These generalists may have an increased capacity to transfer between animal species and, depending on the factors they express, pose a zoonotic threat.

3.1.2 Results

The isolate sequence collection grew during the thesis research. The final collection is summarised in Figure 3.1. The STm dataset is more balanced than the *E. coli*, with an average of 300 isolates from each host (avian, bovine, human and swine). The *E. coli* dataset was more diverse in terms of host (avian, bovine, canine, environmental, human and swine), however it was dominated by human and bovine isolates. The environmental strains were isolated from

plants and plant roots from a specific field with no known livestock passage over the last 10 years. The metadata available for both datasets can be viewed at http://itol.embl.de/shared/nlupolova.

In this chapter, 'classical' core analysis was applied to *Salmonella* and *E. coli* isolate sequences, mainly focusing on one primary question: whether any typing schemes could be associated with a host?

3.1.2.1 Phylogenetic analysis

The core genes that are present in 95% of isolates were extracted, concatenated and used to reconstruct phylogenetic relationship for both *E. coli* and STm. STm phylogeny is presented in Figure 3.3 and Figure 3.5. Overall the tree is divided into 2 main mixed clusters (see vertical tree Figure 3.5 I and II or top and bottom clusters respectively). The diversity of the strains in the top cluster is much higher than in any other part of the tree (note branch length). This top cluster is dominated by very diverse human isolates (n=119, 35% of all STm human isolates) and a much less diverse tight cluster of bovine isolates (72 isolates 32% of all STm bovine isolates). The middle section of the top cluster is occupied by mixed isolates from bovine (n=32), avian (n=32) and swine (n=17) hosts.

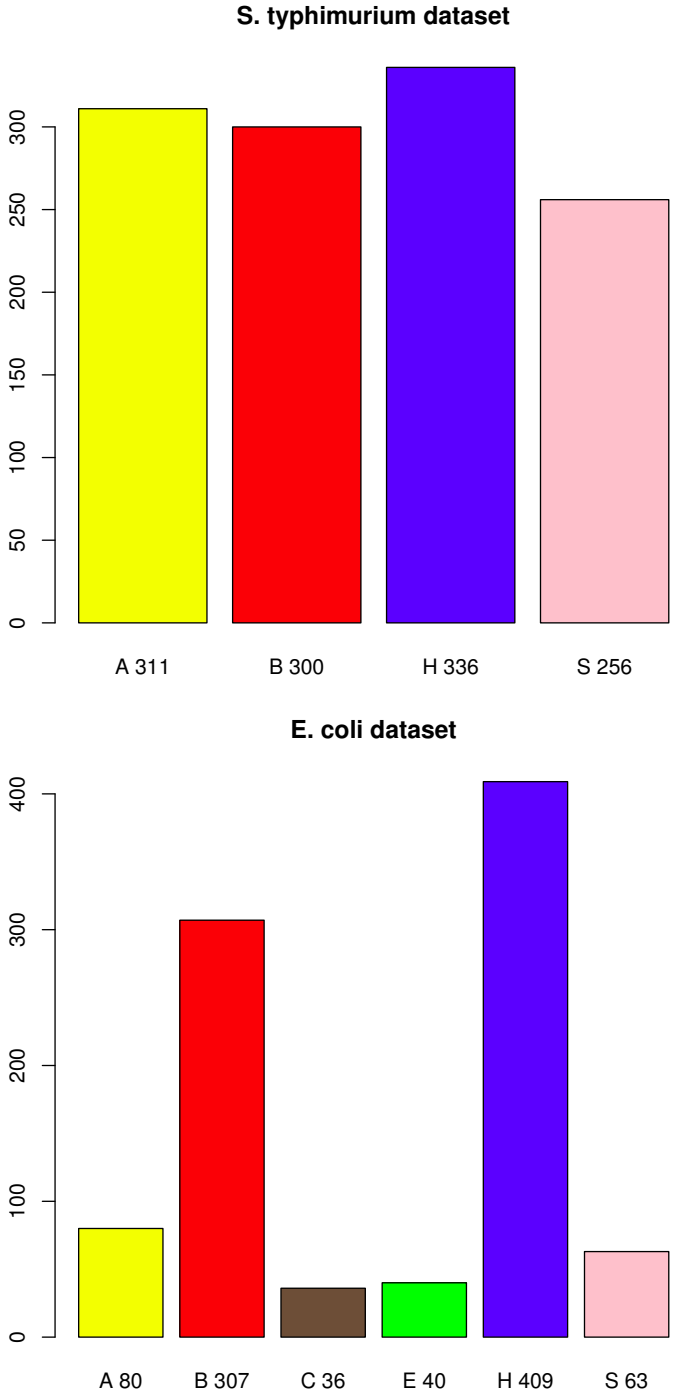


Figure 3.1: Host distribution in *S.* Typhimurium (top) *E. coli* (bottom) datasets. The number of genomes (y-axis) isolated from a specific host are shown. *S.* Typhimurium was isolated from 4 hosts: avian (yellow), bovine (red), human (blue), swine(pink). *E. coli* was isolated from 6 hosts/sources avian (yellow), bovine (red), human (blue), swine (pink) canine (brown) and environmental (green).

The II (bottom) cluster is formed from 2 sub-clusters, one composed almost exclusively from 2 'pure' host clusters human (115 isolates, 34%) and the biggest on the tree is an avian cluster with 225 isolates, 72% of all avian isolates. The remaining section of the tree is composed of clusters of mixed origin with the majority of the swine isolates (212, 83%), bovine (183, 61%), some human 66 isolates (20%) and only 9 avian isolates.

The *E. coli* phylogenetic tree (Figures 3.1,3.6) is much more diverse than STm, in part relating to the species vs serovar difference noted earlier and that the *E. coli* collection is composed of isolates from more hosts/sources. The majority of the isolates are found in separate branches with considerable length. Comparing an average branch length from both genera: for the STm tree this was 0.0002 and for *E. coli* this was 0.013 (100 times longer). The *E. coli* phylogenetic tree therefore demonstrates the much greater diversity than STm.

There were only 13 *E. coli* sequences with identical patterns and only some of the isolates formed clonal clusters. One of these was the O157 cluster (25 isolates) which was expected to cluster apart. Other clusters were 5 'pure' almost clonal human clusters that were situated in different parts of the tree and contained from 14 to 28 isolates. There were 3 tight bovine clusters, two of which contained 20 isolates and one 39. However 2 of the bovine clusters contained isolates from other hosts (3 human, 1 swine and 1 environmental) but the branch length for other hosts was the same as for the bovine isolates

in these clusters.

Isolates apart from the subsets from human and bovine hosts did not cluster together and are spread all over the tree.

3.1.2.2 Phylogroups

Phylotyping was performed in silico for all *E. coli* isolates. 31 sequences failed to be assigned to any phylogroup. The failure is likely to be due to assembly failures as the specific genes were searched from de novo assemblies. Nevertheless, the proportion of bioinformatically detected phylogroups was still higher than the proportion of assigned strains using PCR. For 40 Zambian strains on which PCR based typing was performed, 2 were untypable, while in silico all of these strains were assigned to one or other phylogroup. There is also high consistency between PCR and in silico phylotyping.

There are 16 possible combinations of variants that allow the assignment to 8 different phylogroups. Only 12 combinations were detected previously [120]. In our collection, all 16 combinations were detected.

Phylogroups fit well with core SNP based cluster divisions (Figure 3.2). However, there were some inconsistencies: the most interspersed were the A and C groups, both belonging to the same big cluster. A small number (21 isolates

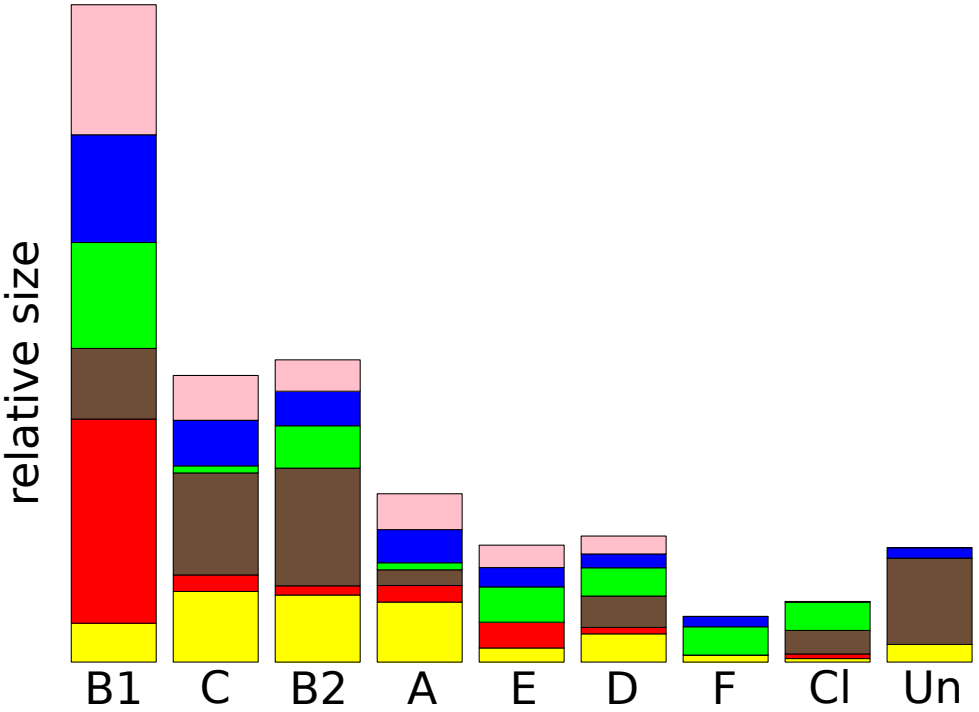


Figure 3.2: Phylogroups distribution of *E. coli* isolates. Colours represent hosts/sources: avian (yellow), bovine (red), human (blue), swine (pink), canine (brown) and environmental (green).

of the mixed origin) of phylogroup C were found in the B1 cluster. B1 is the largest phylogroup containing 442 isolates (A = 11, B = 222, C = 9, E = 15, H = 156, S = 29), 128 isolates were in group C (A = 20, B = 18, C = 13, E = 1, H = 66, S = 10). Phylogroup B2 had 109 isolates (A = 19, B = 10, C = 15, E = 6, H = 52, S = 7) and phylogroup A 99 isolates (A = 17, B = 18 , C = 2, E = 1, H = 53 , S = 8). Phylogroup E had 83 isolates (A = 4, B = 35, C = 0 , E = 5, H = 34, S = 5) and phylogroup D 36 isolates (A = 7, B = 6, C = 3, E = 3, H = 14, S = 3). All others isolates formed three small groups: phylogroup F had 21 isolates (A = 2, B = 0, C = 0, E = 4, H = 15, S = 0), cryptic clades (total 13 isolates A = 1, B = 4, C = 3, E = 4, H = 1, S = 0) and unassigned 31 (A = 5, B = 0 , C = 11, E = 0, H = 15, S = 0).

3.1.2.3 MLST

The MLST analyses were based on the 7 gene MLST scheme for both species. The vast majority of the STm isolates were successfully assigned to a sequence type (ST) 1,196 out of 1,203. In total, 35 different ST types were found in the STm dataset with heavy dominance of ST19 (n = 984, 82% of the dataset) and the second and third most frequent were ST34 (n = 122, 10%) and ST36 (n = 28, 2%) respectively. Distribution of ST by host was as follows: 11 STs in avian isolates, 15 in bovine, 14 in human, 14 in swine. ST19, as expected, was the dominant ST in all host subgroups but with significantly lower

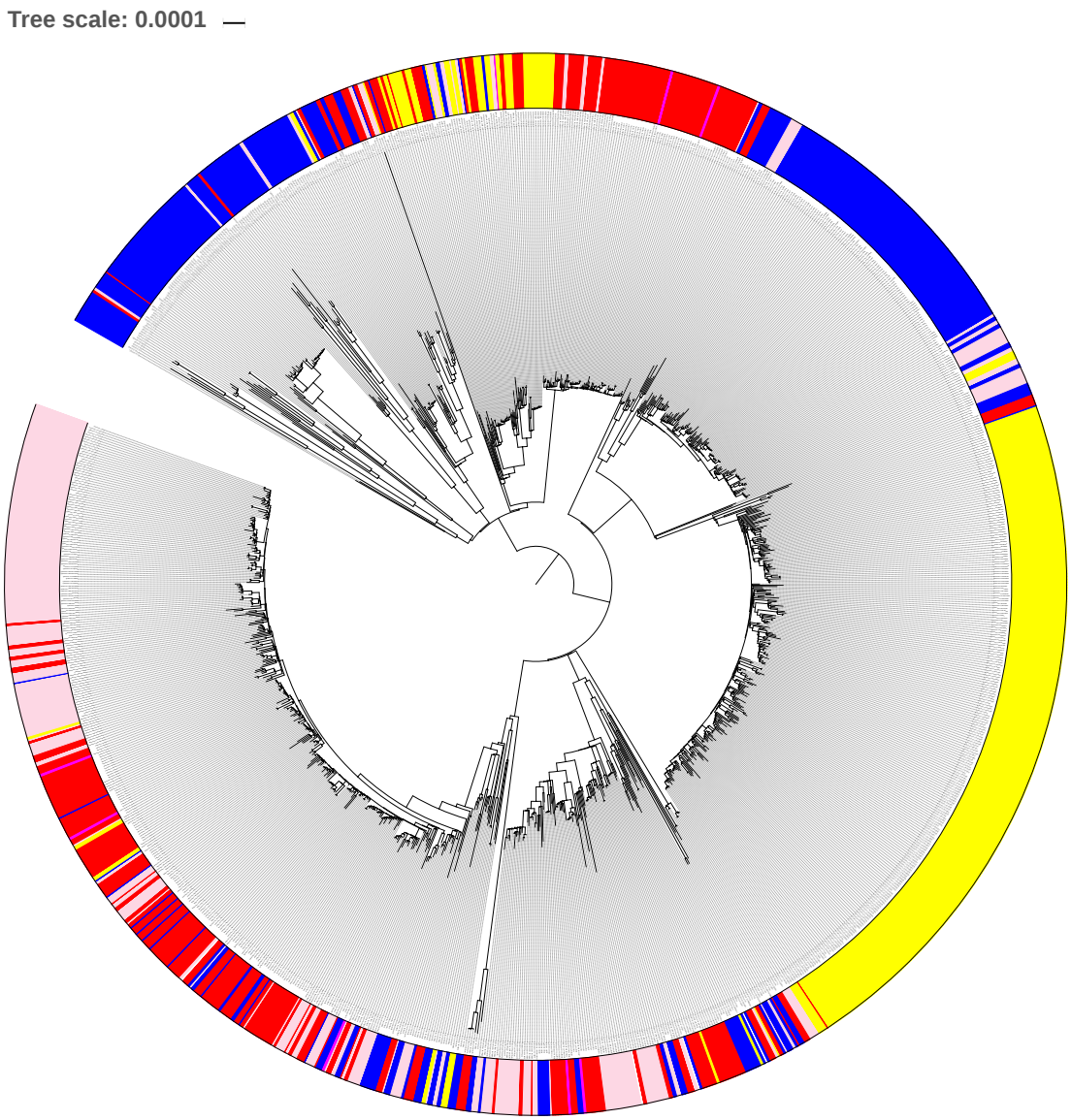


Figure 3.3: Maximum likelihood STm tree. The tree is based on core SNPs from all STm isolates (n = 1,203). Outer ring represents the host distribution: avian (yellow), bovine (red), human (blue), swine (pink).

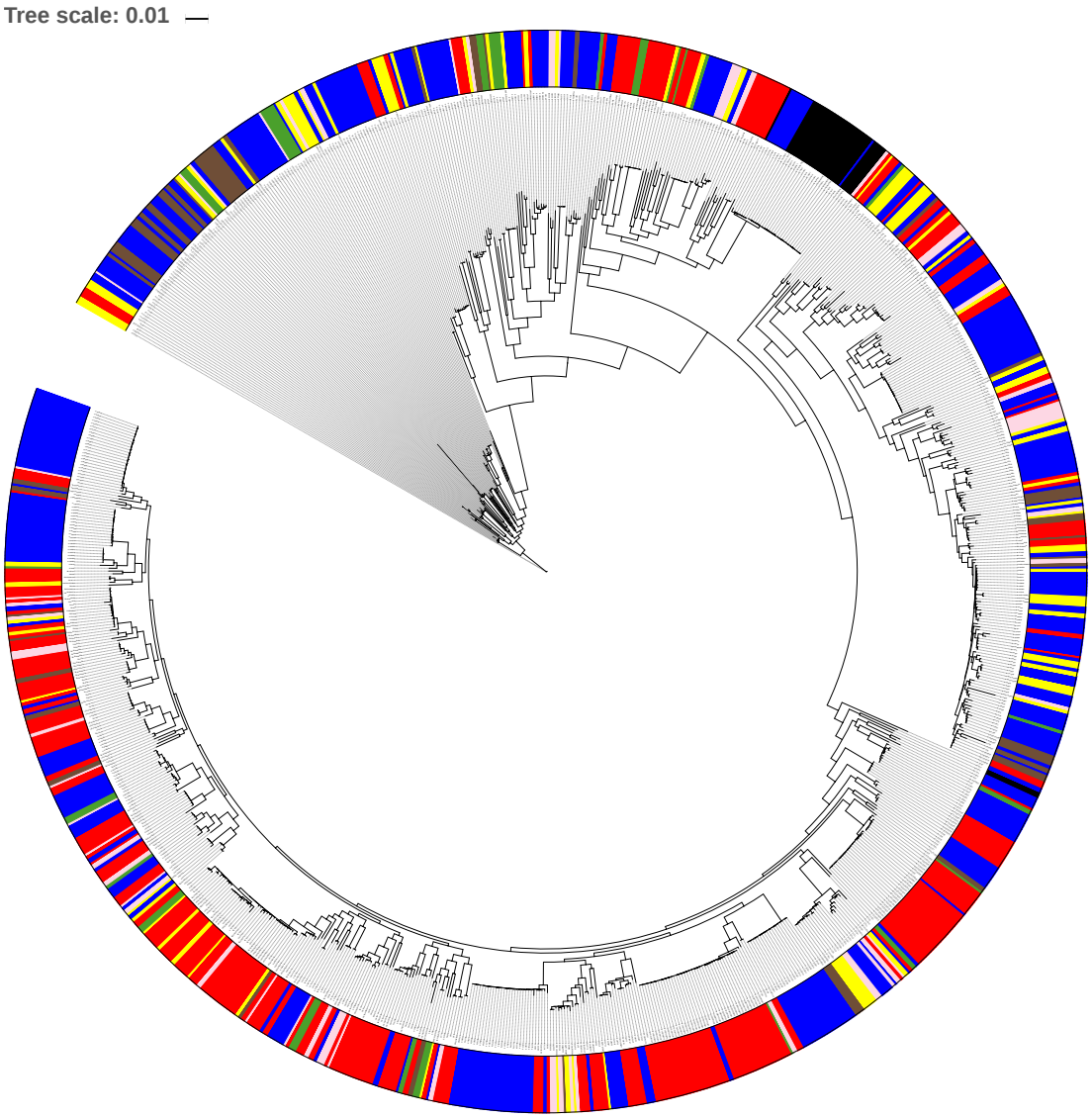


Figure 3.4: Maximum likelihood *E. coli* tree. The tree is based on core SNPs from all *E. coli* isolates (n = 967). Outer ring represents the host distribution: avian (yellow), bovine (red), human (blue), swine (pink), canine brown, environmental (green).

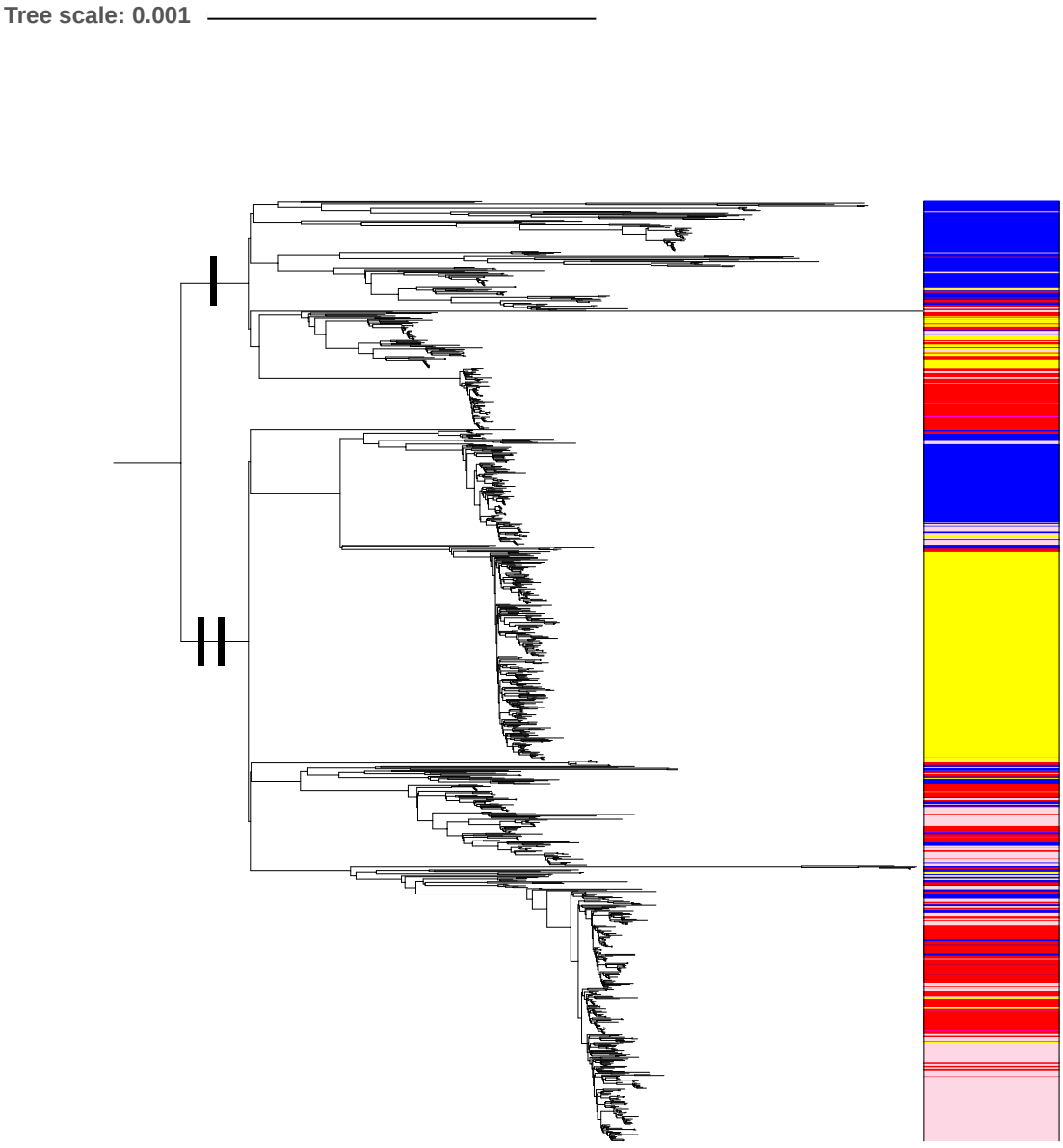


Figure 3.5: STm core ML tree, vertical, scaled. The tree is based on core SNPs from all STm isolates (n = 1,203). Coloured bar represents the host distribution: avian (yellow), bovine (red), human (blue), swine (pink).

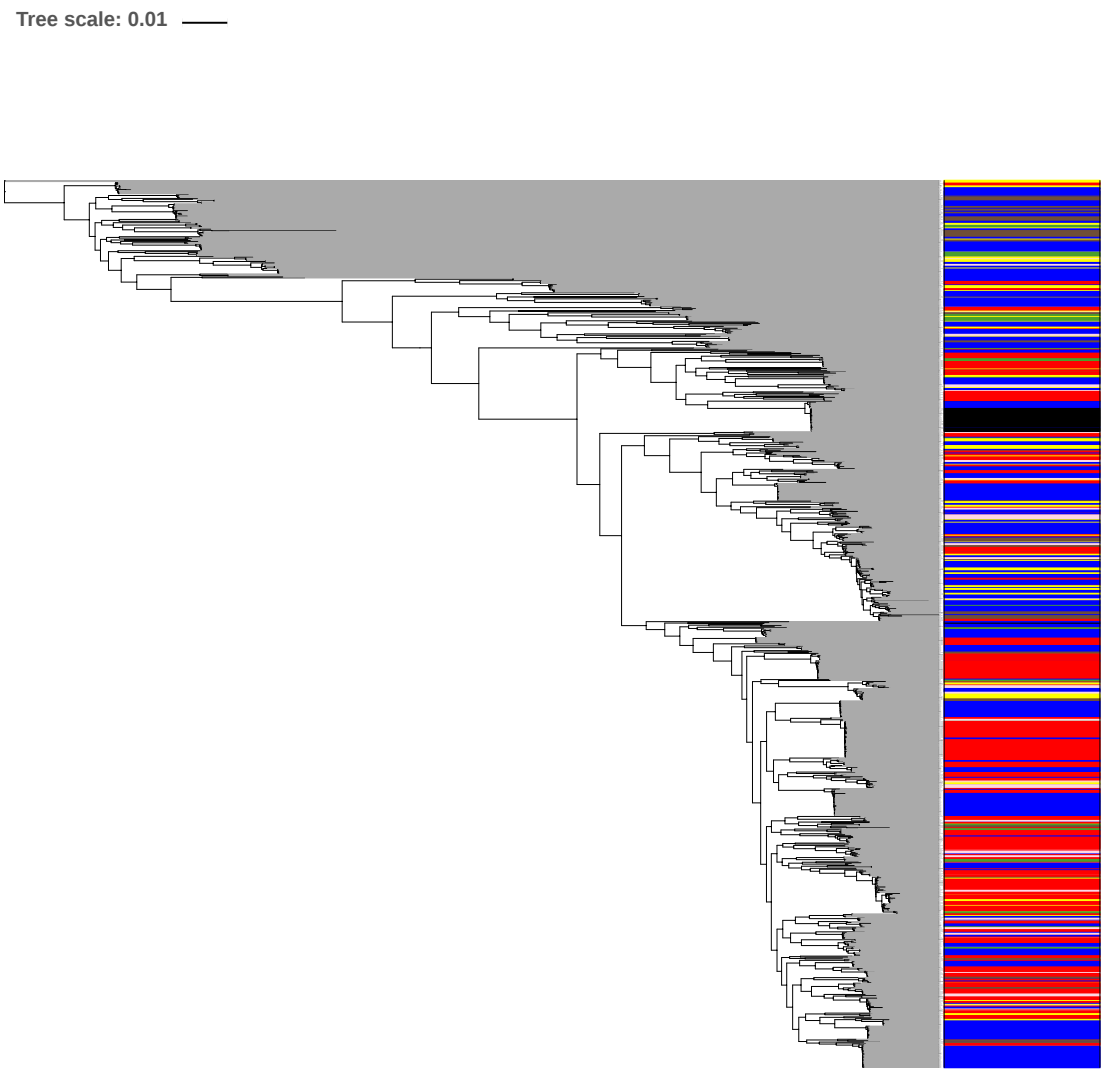


Figure 3.6: *E. coli* core ML tree, vertical, scaled. Coloured bar represents the host distribution: avian (yellow), bovine (red), human (blue), swine (pink), canine (brown), environmental (green).

prevalence in the human subgroup and slightly lower in swine (A = 95%, B = 95%, H = 55%, S = 86%) Apart from ST19, there were STs that were found across multiple sub host groups.

- ST19: A, B, H, S
- ST34: A, H, S
- ST11: A, H, S
- ST32: A, B
- ST321: A, B
- ST98: A, H
- ST138: A, S
- ST13: B, H
- ST302: B, S
- ST213: H, S

The entries in Enterobase for ST19 were examined, which turned up 11,024 isolates in total of which 2,694 were Human isolates, 1,627 Avian, 779 Bovine, 564 Swine; the remainder had no designation. The next most abundant was ST34 which was assigned to 3,832 isolates, broken down by host as follows: 29 Bovine, 44 Avian, 247 Swine, 2,074 Human. It is difficult to draw any conclusion from this number as database collection is biased by different studies.

Assigning ST for *E. coli* produced much more diverse results with the total number of STs from 963 sequences being 260, 139 sequences were not as-

signed to any ST. Only the top 5 STs had more than 20 isolates assigned to them: ST10 = 58, ST17 = 28, ST33 = 27, ST442 = 24 and ST11 = 23. It is difficult to deduce if any ST is associated with host as the groups are unequal and small. For example, for the largest ST in the whole dataset which was ST10, (n = 58) the distribution was: A = 12, 15%, B = 8, 3% E = 1, 3%, H = 31, 7.6% C = 5, 14% S = 1, 1.6%.

S

ST analysis was performed with all *E. coli* (47,590) sequences in Enterobase, which were distributed into a total of 3,023 STs. However, almost half of these had no information about the host assigned to them. After filtering sequences without host and without ST information there were 18,823 sequences that were assigned to a total of 1,902 STs (Figure 3.7, 3.8). Our collection only partially reflected ST distribution in Enterobase where most abundant STs were ST11 (the ST of the Sakai O157 strain), ST10 (ST of K12 strain), ST131 [131], ST21, and ST17. The majority of the STs were skewed towards human sampling, with some exceptions including ST10 for which the majority are laboratory strains. Overall any given ST has a mixed host population of isolates; human isolates can be found in any of these and it is extremely challenging to come to the conclusion that any particular ST is associated with a particular host. As with the example of ST131 for which the vast majority of the isolates are of human origin but is it due to the fact that these bacteria have evolved for a human host or due to the history of the research with this clone which is

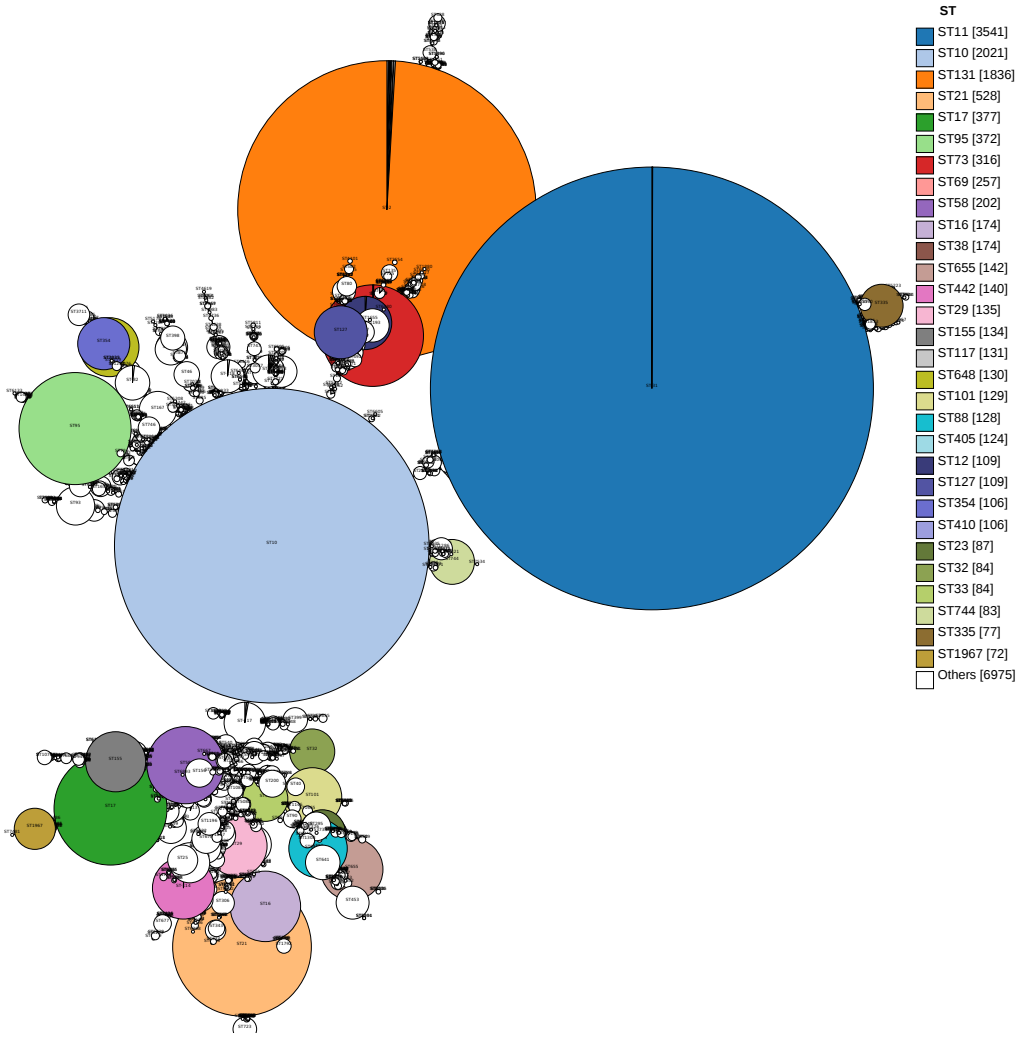


Figure 3.7: ST distribution of *E. coli* sequences form Eterobase. Size of the clusters are reflecting number of isolates. The Enetrobase is dominated by 3 ST: ST11 (blue), ST10 (light blue), ST131 (orange).

mostly focused on human studies? [131]

3.1.2.4 Pan-core proportions

Core proteins were extracted by clustering all proteins from all isolates using 95% sequence similarity and 95% inclusion (i.e a protein should be present

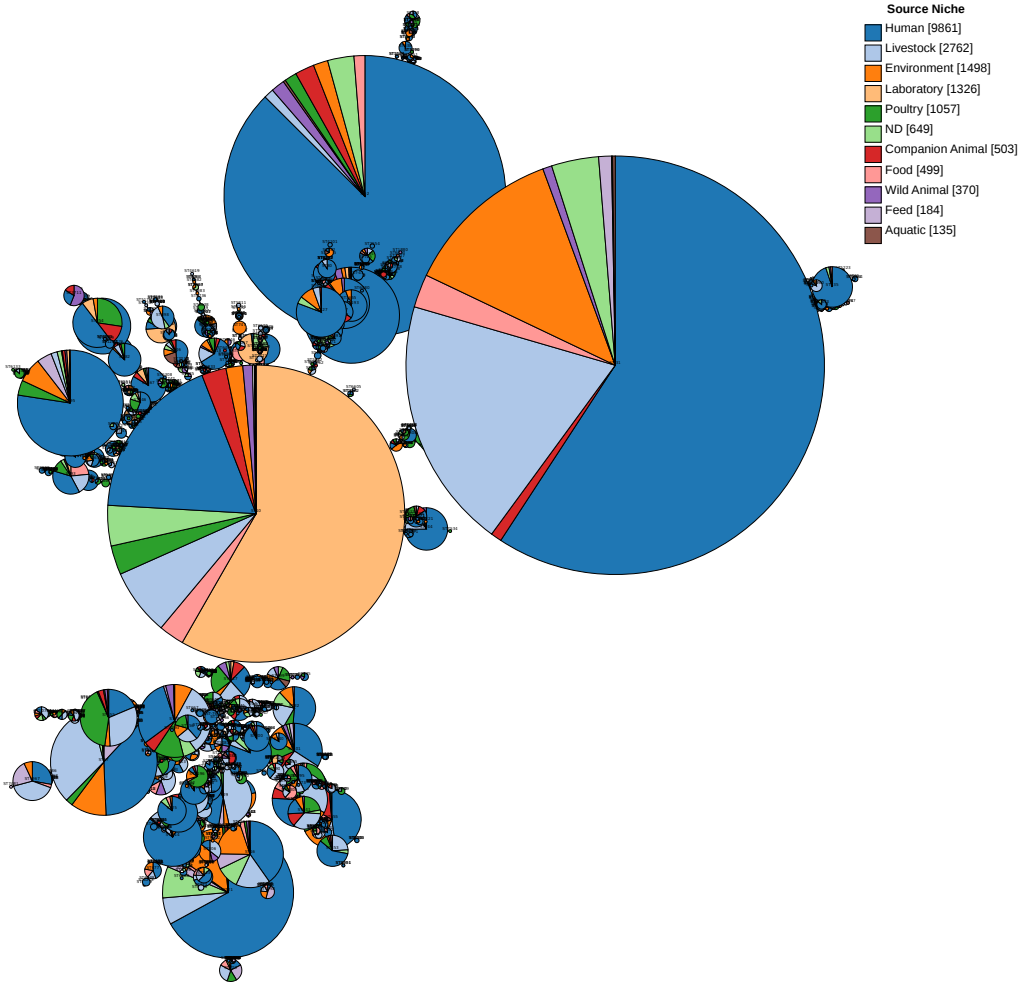


Figure 3.8: Host distribution within the the most abundant ST of *E. coli* sequences form Enterobase. The tree is clustered the same way as in the Figure **??**, but coloured by host. Human isolates (blue) can be found in all ST, but in different proportions.

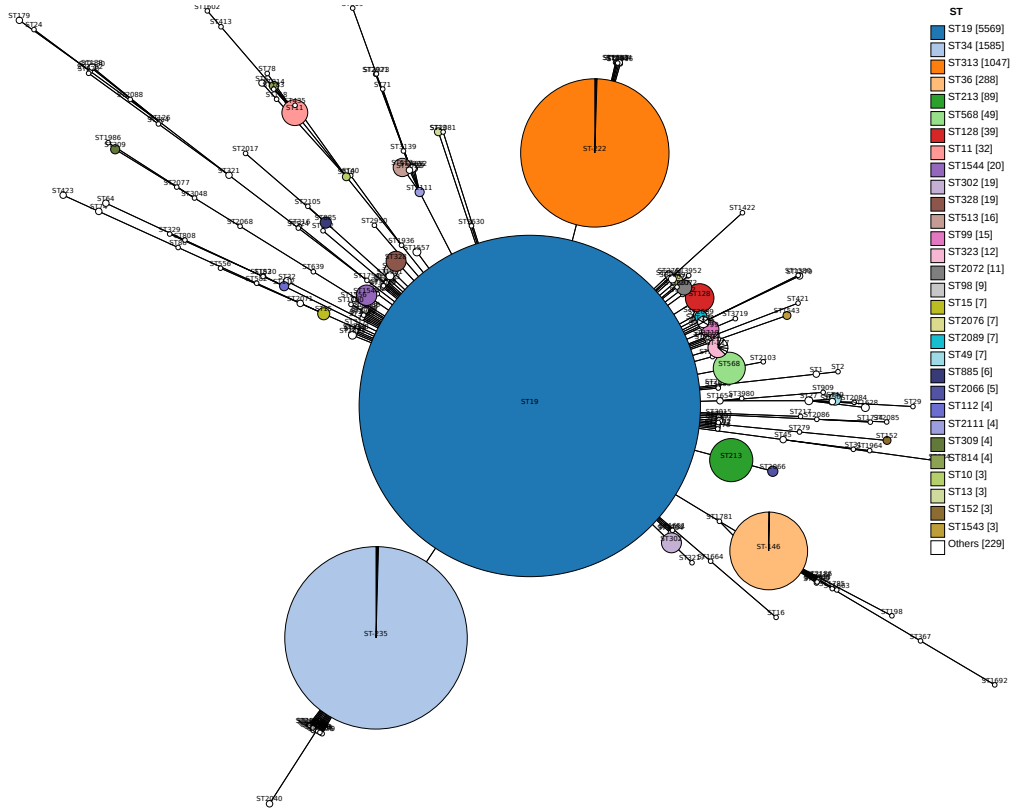


Figure 3.9: ST of all *Salmonella* Typhimurium isolates from Enterobase. Size of the clusters are reflecting number of isolates. The STm isolates from Enetrobase is dominated by 3 ST: ST19 (blue), ST222 (orange), ST235 (light blue), ST146 (beige).

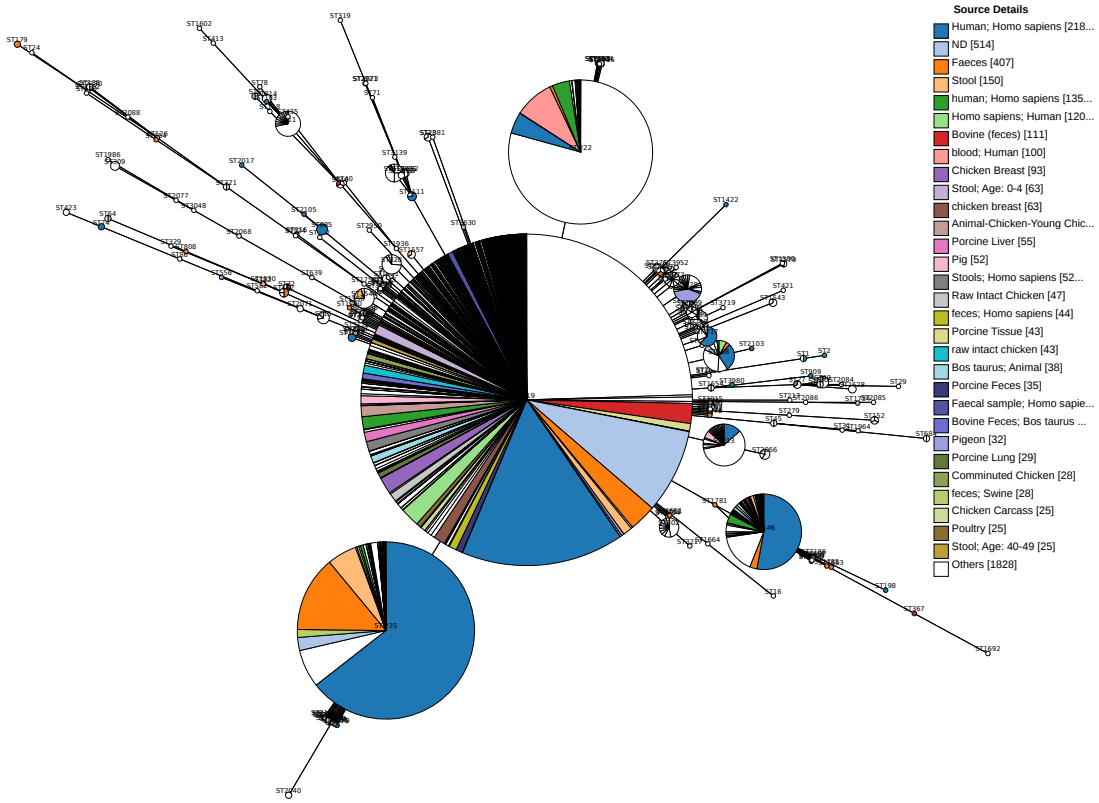


Figure 3.10: Host distribution by ST of all *Salmonella* Typhimurium isolates from Enterobase. The tree is clustered the same way as in the Figure 3.9, but coloured by host. Human isolates (blue) can be found in all ST, however in different proportions.

in at least 95% isolates) as the threshold. As these predicted proteins were translated from predicted genes, the terms genes and proteins are used interchangeably in this work as these were obtained by the same method (PROKKA annotations). Adjustment of 'core' inclusion strictness can be illustrated by the following example: when core is 'to be present in 100% of isolates' there were 1,991 core genes in STm dataset; with inclusion of 95% of isolates core = 3,991 genes. The average number of genes per STm genome was 4,626 gene per isolate. For *E. coli*, only 446 genes were found to be present in 100% of isolates but this number rapidly increased if the threshold was less strict, so at 95% inclusion there were 2,179 predicted core genes. The average number of genes in an *E. coli* genome was 4,773 (see Figure 3.11).

The number of predicted core proteins in STm avian isolates was significantly lower (z-test, $p<0.0001$) compared to all other STm hosts: A = 2,218, B = 3,054, H = 3,056, S = 3,065. On the other hand, the mean number of proteins in sequences per host showed little variation: A = 4,632, B = 4,645, H = 4,573, S = 4,636, but lowest in the human group. Thus, the proportion of shared proteins was the smallest in the avian group (48%) and varied from 66 to 68% in all others host subgroups.

For *E. coli*, the number of core proteins was: A = 1,615, B = 1,031, C = 1,433, E

= 1,857, H = 803, S = 1,715, and the mean number of proteins per host group was: A = 4,801, B = 4,743, C = 4,998, E = 4,656, H = 4,829, S = 4,563. As the size of the groups varied dramatically, the analysis was also carried out with a random set of 36 isolates to match the number of canine isolates as this was the smallest group. The sampling procedure was repeated three times and the results averaged. Even though by normalising we reduced the complexity of the dataset, it gives an estimate of the core diversity by host group with core proteins distributed as: A = 1,749, B = 2,130, C = 1,433, E = 1,870, H = 1,956, S = 1,934 and the mean total proteins was A = 4,752, B = 4,797, C = 4,998, E = 4,674, H = 4,893, S = 4,527. Therefore, subsampling also indicated that the canine group is significantly different, with a much smaller core than all other subgroups (only 28% of all genes was shared between all canine isolates), followed by avian (36%), and then all others with core proportion variation being from 40% in human to 44% in bovine.

Overall, the proportions of core proteins was lower for *E. coli*, but it must be noted that the *Salmonella* isolates where represented by a single serovar, while the *E. coli* dataset is quite diverse.

3.1.3 Published work: Phylogenomic approaches to determine the zoonotic potential of Shiga toxin-producing Escherichia coli (STEC) isolated from Zambian dairy cattle

SCIENTIFIC REPORTS

OPEN

Received: 20 January 2016
Accepted: 27 April 2016
Published: 25 May 2016

Phylogenomic approaches to determine the zoonotic potential of Shiga toxin-producing *Escherichia coli* (STEC) isolated from Zambian dairy cattle

Geoffrey Mainda[1,2,*], Nadejda Lupolova[1,*], Linda Sikakwa[3], Paul R. Bessell[1], John B. Muma[3], Deborah V. Hoyle[1], Sean P. McAteer[1], Kirsty Gibbs[4], Nicola J. Williams[4], Samuel K. Sheppard[5], Roberto M. La Ragione[6], Guido Cordoni[6], Sally A. Argyle[1], Sam Wagner[1], Margo E. Chase-Topping[7], Timothy J. Dallman[8], Mark P. Stevens[1], Barend M. deC. Bronsvoort[1] & David L. Gally[1]

This study assessed the prevalence and zoonotic potential of Shiga toxin-producing *Escherichia coli* (STEC) sampled from 104 dairy units in the central region of Zambia and compared these with isolates from patients presenting with diarrhoea in the same region. A subset of 297 *E. coli* strains were sequenced allowing *in silico* analyses of phylo- and sero-groups. The majority of the bovine strains clustered in the B1 'commensal' phylogroup (67%) and included a diverse array of serogroups. 11% (41/371) of the isolates from Zambian dairy cattle contained Shiga toxin genes (*stx*) while none (0/73) of the human isolates were positive. While the toxicity of a subset of these isolates was demonstrated, none of the randomly selected STEC belonged to key serogroups associated with human disease and none encoded a type 3 secretion system synonymous with typical enterohaemorrhagic strains. Positive selection for *E. coli* O157:H7 across the farms identified only one positive isolate again indicating this serotype is rare in these animals. In summary, while Stx-encoding *E. coli* strains are common in this dairy population, the majority of these strains are unlikely to cause disease in humans. However, the threat remains of the emergence of strains virulent to humans from this reservoir.

Shiga toxigenic *Escherichia coli* (STEC) are emerging pathogens of public health concern worldwide, including in Europe, North and South America and Asia[1,2]. Ruminants, in particular cattle, have been identified as the predominant reservoir of STEC[3,4], indicating that the bacteriophage-encoded Shiga toxins (Stx) are likely to confer an advantage to *E. coli* in these host animals. In Africa there is little information on the epidemiology of STEC in livestock systems and their impact on human health[1]. It is evident that only a subset of STEC are a serious threat to human health, these enterohaemorrhagic *E. coli* (EHEC) are associated with specific serogroups in particular the seven that have been defined as adulterants in beef production in the USA, O157, O26, O111, O45, O145, O103, O121[5,6]. Similar serotypes, especially O157 & O26 are also an issue in Europe. Typical EHEC strains can be further characterised by possession of a type 3 secretion system (T3SS) that enables colonisation of the gastrointestinal tract[7]. EHEC infections in humans are associated with diarrhoea and bloody diarrhoea, with the more serious sequelae of kidney and brain damage due to activity of Stx on the microvasculature in these organs[4,8].

The cost of whole genome sequencing (WGS) has drastically reduced and it is now possible to sequence large numbers of isolates and use bioinformatics approaches to extract strain relatedness and gene carriage data. For

[1]Roslin Institute and Royal (Dick) School of Veterinary Studies, Edinburgh, UK. [2]Ministry Livestock and Fisheries, Kabwe, Zambia. [3]University of Zambia, Lusaka, Zambia. [4]University of Liverpool, Liverpool, UK. [5]Swansea University Medical School, Swansea, UK. [6]University of Surrey, Surrey, UK. [7]University of Edinburgh, Edinburgh, UK. [8]Public Health England, London, UK. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to D.L.G. (email: dgally@ed.ac.uk)

Farm type	Estimate	95% CI	P
Commercial	1	–	–
Medium	7.05	1.76–28.28	0.007
Small	4.05	1.20–13.64	0.002

Table 1. Farming type STEC risk analysis.

E. coli, strains have classically been assigned into phylogroups that provide a good correlation with commensal versus pathogenic strains[9]. The phylogroups are based on particular combinations of specific genes and can be assigned from the WGS. Whole genome core SNP analysis to define strain relatedness is now commonly used and provides greater resolution than MLST[10]. In addition the serogroup, the O-chain of LPS, can also be inferred from their genetic determinants using WGS. The WGS of any strain collection is therefore a valuable resource allowing relatively rapid comparison of phylogeny and pathogenic potential.

Both small holding and large-scale dairy farming is important to the economic survival of communities in many developing nations, including Zambia[11,12]. As such, it is important to understand if practices on these units and their products may represent a threat to human health and where such risks exist suggest possible mitigation measures. A study has recently been carried out to sample *E. coli* strains from cattle across small, medium and large-scale (commercial) dairy farms in central Zambia, with the primary aim of understanding antibiotic use and antimicrobial resistance patterns in this sector[13].

These isolates have now been further analysed in the present study for *stx* prevalence and any association with the farming system. In addition, *E. coli* isolates from patients with diarrhoea were also screened and sequenced to determine evidence of relationship to the bovine strains.

Results

Detection of Shiga toxin alleles (*stx1* and *stx2*) in isolates from Zambian dairy cattle and humans. Eleven percent (41/371) of the bovine *E. coli* isolates were positive for the presence of Shiga toxin genes as defined by detection of appropriately sized PCR amplification products using an established *stx* multiplex assay[14]. Based on this, both *stx1* and *stx2* were detected in 54% (22/41) of the STEC, while 37% (15/41) had *stx2* only and 10% (4/41) had *stx1* only. Using this data, the overall adjusted prevalence of STEC across the different farming scales for the central Zambian study area can be estimated at 6% (95% CI: 2.5–10.2). The adjusted prevalence per farming scale was higher in medium-scale 17.1% (95% CI: 5.9–28.2) and small-scale 10.6% (95% CI: 6.6–14.5) farms when compared to the commercial farms 2.8% (95% CI: 0.3–6.0). Based on these ranges, there is a significant difference in estimated prevalence between the small and commercial scales. Logistic regression indicates that medium- and small-scale farming are significant risk factors for Shiga toxin producing *E. coli* (STEC) with commercial as a reference (Table 1). Out of the 73 *E. coli* isolates from human patients with diarrhoea for which good quality sequence information was generated, no Shiga toxin genes were detected.

As an additional investigation, the enrichment cultures for all the animals (n = 371) were streaked onto sorbitol MacConkey agar plates and any non-sorbitol fermenting colonies tested for O157 agglutination. Only one animal yielded a positive strain (ZB-2213N0194) and this was then added to the study.

Phylogenetics. In order to understand the genetic backgrounds of the STEC strains isolated in this study, including their potential threat to human health, their relationship to other human disease-associated EHEC were tested by phylogenetic methods. The WGS of 297 of the Zambian isolates (224 bovine and 73 human) were determined. This included 41 STEC, 37 of the 41 defined as *stx*+ by PCR from main study, three STEC strains from a pilot study and the single positively selected *E. coli* O157 strain. These were compared with 262 *E. coli* sequences from human, cattle, avian and canine hosts; one hundred and twenty nine strains in this second collection were human clinical STEC isolates (see Supplementary Table 1).

Alignment to a reference genome (*E. coli* O157:H7 str. Sakai, RefSeq assembly accession: GCF_000008865) resulted in 715,632 core positions with 68,327 single nucleotide polymorphism (SNPs) across all 559 sequences. A maximum likelihood phylogeny revealed the population structure of the *E. coli* strains (Fig. 1). While there was no clear clustering of the strains based on geographical location or host, there was, as anticipated, good correlation with established *E. coli* phylogroups, with only minor discordance. All possible phylogroups and cryptic clades were identified, however the majority of the *E. coli* strains (97%) were distributed across 5 phylogroups (Fig. 1).

The Zambian bovine strains (n = 224) predominately associated with the B1 'commensal' cluster (67%) with the remainder present as: A (9%); B2 (4%); C (8%); D (9%); other (3%). By contrast, the Zambian human strains (n = 73) had equivalent representation across the 5 main phylogroups: A (22%); B1 (19%); B2 (16%); C (22%); D (16%); other (5%). The Zambian cattle STEC strains were also predominately in the B1 phylogroup (27/41).

Serotyping. The majority of EHEC strains that are a threat to human health are associated with 7 specific serogroups. A bioinformatic approach was used to serotype the sequenced strains. H typing was possible for 550/559 strains and O-typing for 483/559 (summarized in Supplementary Tables 3 and 4). Failure to detect specific genes in some of the strains was most likely due to assembly issues with short read sequences.

With the exception of the positively selected *E. coli* O157 strain, none of the Zambian bovine STEC strains (0/40) were assigned to any of these seven serogroups. In fact, only 3 strains across the whole set of cattle isolates could be assigned within these serogroups (ZB-244; serogroup O45 and ZB-2213N0112; serogroup O111 and ZB-2213N0194; serogroup O157). Overall, the Zambian strains (cattle and human) exhibited an extensive array

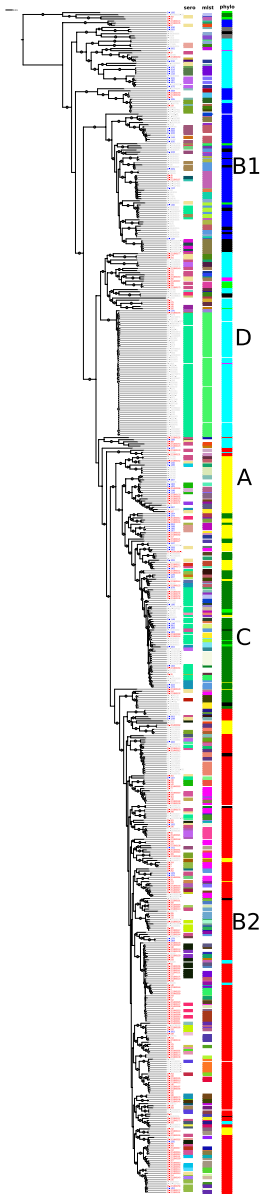


Figure 1. Phylogenetic context of Zambian isolates. The tree depicts the phylogenetic relationship of *E. coli* isolates from Zambia (bovine - red and human - blue) with other *E. coli* isolates (grey). The ML tree is based on core SNPs as described in Materials and Methods. The tree is un-rooted and grey circles on branches represent bootstrap values higher than 80. Vertical columns demonstrate: (1) Diversity of the sequence types (ST) based on MLST analysis where each colour represents a different ST; (2) Diversity of O-serogroups for which each colour represents a different group; (3) Phylogroups: A-yellow, B1-red, B2-blue, C-green, D-turquoise, E-pink, F-grey, cryptic clades-light green. The phylogroups are consistent with core SNP clustering with some minor discordance. White spaces on all columns indicate sequences that were untypable.

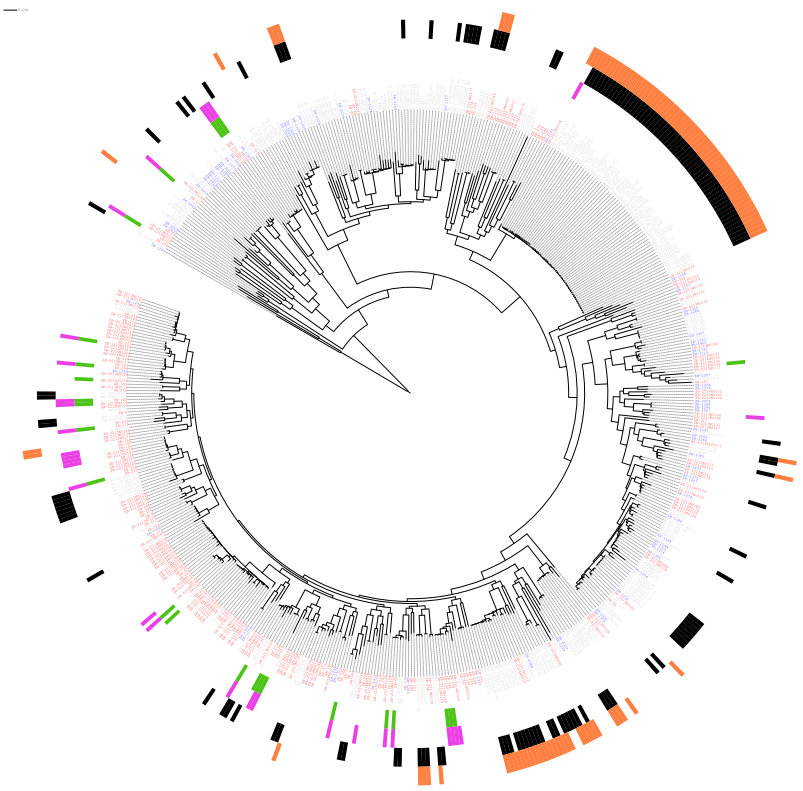


Figure 2. Phylogenetic relationship between STEC. The same ML core SNP tree as in Fig. 1 plotted in a circular manner to depict relationships between Shiga toxin encoding strains. The strain designations are Zambian bovine (red), Zambian human (blue), other *E. coli* (grey). For the Zambian strains, the coloured bars indicate the presence of Stx genes: *stx2* (purple) and *stx1* (green). Black blocks around the tree indicate non-Zambian *E. coli* encoding *stx* (1 or 2). Orange blocks highlight the presence of intimin (*eae*) and *sepL* indicating the possession of a type 3 secretion system. It is apparent that with the exception of one positively selected EHEC O157 (ZB-2213N0194), that the Zambian cattle STEC do not encode this system.

of serogroups (Fig. 1 and Supplementary Table 4) and H:O combinations were unique for each strain with little clustering or association with established human clinical isolates (Fig. 2).

Toxicity analysis of *stx*+ strains. To determine if the genotypically positive *stx* strains were able to express Stx, eighteen of the bovine STEC were examined for Vero cell cytotoxicity with and without mitomycin C (MMC) induction. 89% (16/18) of the MMC-induced STEC strains had a cytotoxic effect on Vero cells (Fig. 3). These samples were verified as Stx positive using a commercial ELISA (Fig. 3), with only one strain (4) exhibiting toxicity on Vero cells without any detection of Stx by ELISA.

Shiga toxin subtyping. Forty one STEC positive strains were included in the WGS analysis (Fig. 2) and from this *stx* alleles could be further subtyped using a published BLAST-based methodology[15] (Supplementary Table 5). It was evident that the most cytotoxic strains (Fig. 3) were those encoding Stx2a often in combination with Stx1a, in line with studies of cytotoxicity and pathology induced by enterohaemorrhagic strains with different Stx variants[16,17].

Stx association with type 3 secretion and enteroaggregative virulence factors. Typical enterohaemorrhagic *E. coli* strains are defined by the co-association of *stx* genes with a type 3 secretion system (T3SS)[7]. In the present study, the presence of a T3SS was determined by detection of both *eae* and *sepL*. Based on BLAST analysis, 3.6% (8/224) and 2.7% (2/73) of the Zambian bovine and human isolates respectively may encode a T3SS (Fig. 2). Excluding the positively selected O157 strain, neither intimin (*eae*) nor *sepL* were detected in the bovine

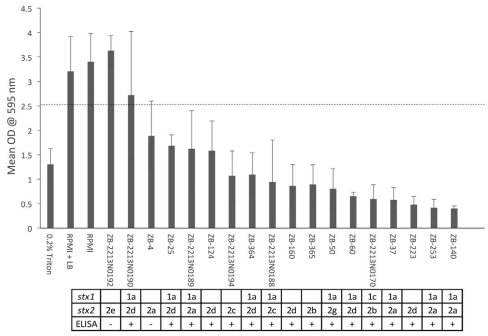


Figure 3. Shiga toxin activity and subtyping. The top panel graph indicates the cytotoxic effect of selected STEC strain supernatants on Vero cells. Increased cell survival results in higher values. 0.2% Triton X-100 was used as a positive control; RPMI + LB and RPMI alone were used as negative controls. Values below the dashed line indicate a cytotoxic effect on the cells. 89% (16/18) of the STEC supernatants tested demonstrated a cytotoxic effect. Supernatants were prepared as described in Materials and Methods. Stx subtypes are shown in the lower panel along with ELISA results for detection of Stx. Isolates with both *stx1a* and *stx2a* are associated with higher toxicity. Sample number ZB-4-stx contains *stx2a* and exhibited cytotoxicity on Vero cells but was negative by ELISA.

STEC. Other non-STEC but intimin positive strains were present within the Zambian strains analysed and some were in close proximity to clinical human STEC strains (Fig. 2). The cattle strains were also checked for the presence of the enteroaggregative *E. coli* (EAEC) adherence factors AggR and AA probe by PCR but all were negative indicating that these dairy cattle are not a common reservoir of enteroaggregative *E. coli* as associated with the atypical EHEC O104 outbreak in Northern Germany in 2011.

Discussion

Shiga toxins (Stx) can pose a serious threat to human health, and human infections are usually restricted to a subset of serogroups that can express Stx and a type 3 secretion system (T3SS) or other adherence mechanisms that can facilitate colonisation of the human gastrointestinal tract. In this study, *E. coli* isolates were obtained from cattle associated with dairy production in a region of central Zambia. A total of 371 isolates, each from an individual animal, covering 104 farms were tested for the presence of *stx* by PCR. Of these, 41 (11%) were positive. This gives an estimated prevalence (taking into account sampling and the study design effect) of 6% (95% CI: 2.5–10.2). To our knowledge, this is one of the first surveys to systematically analyse the proportion of random *E. coli* from a farm animal source that are positive for *stx* as other studies usually use positive selection methods from animals. Our survey does indicate that STEC are common in these dairy cattle. It was also evident that the small and medium sized production units had a higher prevalence of STEC than the commercial units sampled. This is of interest as it does indicate that management practices potentially influence the selection of STEC. On-going work will investigate these differences including the influence of breed which can differ between the farming scales[13] and/or diet which can change EHEC O157 prevalence[18].

Our study also examined 73 *E. coli* strains from human patients with diarrhoea presenting at the University Teaching Hospital (UTH) in Lusaka over the same time period. While the human sample numbers were relatively low, *stx* was not detected, indicating that STEC are unlikely to be common in the local human population. The overall phylogenetic analyses of the strains into phylogroups was as anticipated, with the majority of the Zambian bovine strains being present in the B1 group associated more with commensal strains, although this cluster contains non-O157 EHEC serotypes causing infections in humans that likely originate from cattle. A greater proportion of the human Zambian isolates clustered within the phylogroups associated with human disease, reflecting that the strains were collected from patients with diarrhoea and in some cases the strain may be the etiological agent.

Based on bioinformatic analyses of WGS, there was marked diversity of serotypes in the cattle and human sample populations (Fig. 1 and Supplementary Tables 3 and 4). None of the randomly selected bovine STEC were allocated to a serogroup commonly associated with human EHEC infections. Furthermore, none of these bovine STEC encoded a T3SS based on detection of intimin (*eae*) and *sepL* alleles (Fig. 2), as both genes are present on the locus of enterocyte effacement that encodes the system. An additional study to positively select for *E. coli* O157 isolates from the faecal pat enrichments only identified one positive sample from a farm. Taken together, this study indicates that while STEC are common in the Zambian dairy cattle these strains would not be classified as EHEC and are unlikely to be associated with serious human disease.

While it is encouraging that EHEC strains were extremely rare, many of the supernatants from the STEC strains were cytotoxic and appropriate backgrounds for EHEC emergence are present. As such, we should remain vigilant in case Stx-encoding prophages from this reservoir do emerge in other strain backgrounds that have a

higher capacity to cause disease in the human host. Continuing work on factors driving the maintenance of STEC strains in the bovine host will hopefully clarify approaches to reducing the threat from this emerging group of pathogens.

Methods

Bovine and human isolates from Zambia. Bovine *E. coli* isolates (n = 371) were collected as part of a previously published study investigating antimicrobial resistance[13]. In addition, a further 81 *E. coli* isolates from cattle were collected as part of a pilot study in 2013 in the same region[13]. Faecal sampling and animal handling of the farm animals was carried out in accordance with the approved guidelines issued by The Roslin Institute Animal Welfare and Ethical Review Body which approved this study[13]. In the main study, 376 dairy cattle from 104 farms representing about 20% of the dairy herds in the study area were randomly sampled and an *E. coli* was isolated from 371 animals (*E. coli* was not isolated from 5 animals) based on growth characteristics on both MacConkey agar and Bile-X-Glucuronide (TBX) plates (Oxoid, UK). Subsequent phylotyping indicated that 97% (361/371) could be allocated to established *E. coli* phylogroups[19]. In terms of subsequent studies the isolates were chosen as follows: From the main study all isolates (n = 371) were tested by the *stx* PCR to allow the prevalence to be estimated. 188 were sequenced but 186 were used in the phylogenetic analyses due to quality issues with 2 sets of reads. The sub-selected strains for sequencing were as follows (those with poor reads removed): (1) All strains showing phenotypic antibiotic resistance in the original study (n = 61)[13]; Strains positive for *stx* by PCR (37/41); the rest (n = 88) were randomly chosen from the remainder. From the pilot study 40 from 81 strains were sequenced including three that were *stx*+ as determined by PCR; 37/40 were included in the phylogenetic study as three had read quality issues. As a separate study, enrichments from all animals (n = 371) were plated onto sorbitol MacConkey and any non-sorbitol fermenting colonies (3 per plate) tested for O157 agglutination. Only one animal yielded a positive isolate (ZB-2213N0194) and this positively selected isolated was then sequenced and added into the phylogenetic analysis. In total there were 224 bovine *E. coli* good quality whole genome sequences that were analysed in this study.

E. coli isolates (n = 79) from patients presenting with diarrhoea were collected at Lusaka hospital between 4th December 2014 and 7th January 2015 as part of another project managed by Prof. J.B. Muma and generously supplied for sequencing. Informed consent was obtained from all subjects. Six of the isolate sequences were not analysed due to read quality leaving n = 73 for phylogenetic and virulence determinant analysis. Further strain and sequence details are provided in the Supplementary Table 1.

DNA extraction. DNA extraction was carried out using either a Wizard Genomic DNA Extraction Kit® or a Qiagen® DNA extraction kit from 1 ml of bacterial culture as defined in the manufacturers' protocols.

PCR detection of virulence determinants. All the bovine strains were screened by a published multiplex PCR for Shiga toxin genes and intimin[14]. The strains were also screened with a multiplex PCR for *aggR* and AA probe genes as markers for enteroaggregative *E. coli*[20]. The PCR products were visualised and captured using multi imaging software (Fluorchem HD2) following electrophoresis in 1.5% w/v agarose gel (Agarose, Melford, UK) and staining with Gelred®.

Verocytotoxicity assays. Established method[21,22], with these minor variations: Single colonies were selected from LB agar plates and suspended in 10 ml of LB broth for 24 h (overnight). 50 μl of overnight culture was added to 5 ml (1:100) of fresh LB broth and incubated for 60 min. Then 20 μl of 5 μg/ml mitomycin C (MMC) was added followed by overnight incubation.

Supernatant samples were screened for the presence or absence of Stx using a commercial ELISA kit (RIDASCREEN® Verotoxin ELISA (C2201), R-Biopharm AG, Darmstad, Germany) according to the manufacturer's instructions.

Statistical analysis. The adjustment of prevalence estimates per farming scale and the risk factor analysis were carried out using logistic regression in 'survey package'[23] in R software environment version 3.1.1 (http://cran.r-project.org/), $p < 0.05$ values were taken as statistically significant. The statistical analyses and more information on the definition of the different level farming systems were as described previously[13].

***E. coli* whole genome sequence analyses.** To better understand how the Zambian *E. coli* strains dataset compared with other *E. coli*, the Zambian strain genomic sequences were analysed with a larger strain collection that consisted of 559 *E. coli* genomes, including clinical and commensal isolates from 4 different broad categories of animal and human hosts (Supplementary Table 1). New short read sequence files have been uploaded to European Nucleotide Archive under the study accession number: PRJEB11782, PRJEB11950, PRJEB11956. Some genome sequences from the Zambian strain sets were removed due to poor read and/or assembly quality, resulting in 297 Zambian genome sequences (224 bovine and 73 human) available for analysis.

Sequencing analysis. All reads were generated by Illumina 1.9 paired-end read sequencing with read lengths from 36 to 251 bp. FASTQC[24] was used for quality assessment and where necessary trimming was done with cutadapt[25]. Short reads were aligned to a reference *E. coli* O157:H7 str. Sakai (RefSeq assembly accession: GCF_000008865) by combining BWA[26], SAMtools and SnpEff[27] in a custom-made python script. The consensus sequence for each alignment of 5,590,092 bp was produced using the majority rule.

Consensus sequences for each alignment were concatenated into one multifasta file that were then parsed to find core positions. Multifasta files of concatenated core nucleotides for each strain were used for recombination analysis with GUBBINS[28]. The recombinatorial regions were removed from the final alignment. The final

alignment was then used to construct a Maximum Likelihood (ML) tree with RAxML[29] under a GAMMA model of heterogeneity with 100 bootstrap replicates (BS). The trees were visualised with ITOL[30].

An established phylotyping scheme[31] was used as a starting point to develop a programme that assigned each strain to one of the 4 possible phylogroups (A, B1, B2, D) based on the presence or absence of one of 3 genes *chuA*, *yjaA*, *arpA* and one genetic fragment *TspE4.C2*. To further distinguish between groups and assign strains to an additional 4 phylogroups (C, E, F or cryptic clades), it was necessary to check for the presence of a fifth gene *trpA* and for the presence of specific alleles for the above genes. *arpA* alleles were used to distinguish between phylogroups D and E based on specific primer sequences described in[31].

To establish gene presence or absence a database that includes all sequences from the collection were built with BLAST+[15]. Query gene's sequences of intimin, *sepL*, *chuA*, *yjaA*, *arpA*, *trpA*, genetic fragment TspE4.C2 were downloaded from the NCBI website. Gene identifiers are presented in the (Supplementary Table 2). Query Shiga toxin sequences identified in[32] also were downloaded from the NCBI website. Gene's presence were established based on a E-value = 0 and similarity match at >90% coverage of the query sequence. For Shiga toxins blast results were filtered based on bit score above 1000, if multiple contigs were involved only the highest result was kept.

Serogroups were identified based on presence of one or several alleles from the following genes: for O-typing - *wzx*, *wzy wzm* and *wzt*; for H-typing the flagellin genes *fliC*, *flkA*, *fllA*, *flmA* and *flnA*. Databases were provided by Dr Flemming Scheutz and colleges[32]. Multi locus sequence type were identified using SRST2 software[33].

References

1. Islam, M. Z. *et al.* Regional variation in the prevalence of *E. coli* O157 in cattle: A meta-analysis and meta-regression. *PloS One* **9** 10.1371/journal.pone.0093299 (2014).
2. Chase-Topping, M. E. *et al.* Pathogenic potential to humans of bovine *Escherichia coli* O26, Scotland. *Emerging Infectious Diseases* **18,** 439–448 (2012).
3. Pruimboom-Brees, I. M. *et al.* Cattle lack vascular receptors for *Escherichia coli* O157: H7 Shiga toxins. *Proceedings of the National Academy of Sciences* **97,** 10325–10329 (2000).
4. Karch, H. *et al.* In *Zoonoses-Infections Affecting Humans and Animals* 235–248 (Springer, 2015). URL: http://link.springer.com/chapter/10.1007/978-94-017-9457-2_9#page-1. (Date of access: 15/11/2015).
5. Pearce, M. *et al.* Prevalence and virulence factors of *Escherichia coli* serogroups O26, O103, O111, and O145 shed by cattle in Scotland. *Applied and Environmental Microbiology* **72,** 653–659 (2006).
6. Friesema, I. *et al.* Emergence of *Escherichia coli* encoding Shiga toxin 2f in human Shiga toxin-producing *E. coli* (STEC) infections in the Netherlands, January 2008 to December 2011. *Euro Surveill* **19,** 26–32 (2014).
7. Kaper, J. B., Nataro, J. P. & Mobley, H. L. Pathogenic *Escherichia coli*. *Nature Reviews Microbiology* **2,** 123–140 (2004).
8. Fruth, A., Prager, R., Tietze, E., Rabsch, W. & Flieger, A. Molecular epidemiological view on Shiga toxin-producing *Escherichia coli* causing human disease in Germany: Diversity, prevalence, and outbreaks. *International Journal of Medical Microbiology* (2015). URL: http://www.sciencedirect.com/science/article/pii/S1438422115000831. (Date of access: 15/11/ 2015).
9. Clermont, O., Gordon, D. & Denamur, E. A guide to the various phylogenetic classification schemes for *Escherichia coli* and the correspondence among schemes. *Microbiology*, mic. 0.000063 (2015). URL: http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.000063. (Date of access: 15/11/2015).
10. Hasman, H. *et al.* Rapid whole genome sequencing for the detection and characterization of microorganisms directly from clinical samples. *Journal of Clinical Microbiology*, JCM. 02452–02413 (2013).
11. Mulemba, H. *The Livestock Sector in Zambia and Rising Food Prices Country Briefing–Zambia*. International Institute for Sustainable Development, Manitoba, Canada (2009). http://www.iisd.org/sites/default/files/pdf/ag_scenarios_south_africa_zambia.pdf. (Date of access: 15/11/ 2015).
12. Mumba, C. *Economic Analysis of the Viability of Small holder Dairy Farming in Zambia*, The University of Zambia, (2012). URL: http://dspace.unza.zm:8080/xmlui/handle/123456789/1804. (Date of access: 15/11/2015).
13. Mainda, G. *et al.* Prevalence and patterns of antimicrobial resistance among *Escherichia coli* isolated from Zambian dairy cattle across different production systems. *Scientific Reports* **5** (2015).
14. Paton, A. W. & Paton, J. C. Detection and Characterization of Shiga Toxigenic*Escherichia coli* by Using Multiplex PCR Assays forstx 1, stx 2, eaeA, Enterohemorrhagic E. coli hlyA, rfb O111, andrfb O157. *Journal of Clinical Microbiology* **36,** 598–602 (1998).
15. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10,** 421 (2009).
16. Zweifel, C., Cernela, N. & Stephan, R. Detection of the Emerging Shiga Toxin-Producing *Escherichia coli* O26: H11/H-ST29 in Human Patients and Healthy Cattle in Switzerland. *Applied and Environmental Microbiology*, AEM. 01728–01713 (2013). http://aem.asm.org/content/early/2013/06/24/AEM.01728-13.short. (Date of access: 15/11/2015).
17. Amézquita-López, B. *et al.* Detection of Shiga Toxin Variants, Virulence Genes and the Relationship to Cytotoxicity of Shiga Toxin-Producing *Escherichia coli* (STEC) from Domestic Farm Animals. in *Meeting Abstract.* URL: http://sistemanodalsinaloa.gob.mx/archivoscomprobatorios/_14_resumeneventoscientificos/694.pdf. (Date of access: 17/12/2015).
18. Fox, J., Depenbusch, B., Drouillard, J. & Nagaraja, T. Dry-rolled or steam-flaked grain-based diets and fecal shedding of O157 in feedlot cattle. *Journal of Animal Science* **85,** 1207–1212 (2007).
19. Clermont, O., Bonacorsi, S. & Bingen, E. Rapid and simple determination of the*Escherichia coli* phylogenetic group. *Applied and Environmental Microbiology* **66,** 4555–4558 (2000).
20. Cerna, J. F., Nataro, J. P. & Estrada-Garcia, T. Multiplex PCR for detection of three plasmid-borne genes of enteroaggregative *Escherichia coli* strains. *Journal of Clinical Microbiology* **41,** 2138–2140 (2003).
21. Padhye, V., Zhao, T. & Doyle, M. Production and characterisation of monoclonal antibodies to Verotoxins 1 and 2 from *Escherichia coli* of serotype O157: H7. *Journal of Medical Microbiology* **30,** 219–226 (1989).
22. Krüger, A., Lucchesi, P. M. & Parma, A. E. Verotoxins in bovine and meat verotoxin-producing *Escherichia coli* isolates: type, number of variants, and relationship to cytotoxicity. *Applied and Environmental Microbiology* **77,** 73–79 (2011).
23. Lumley, T. Analysis of complex survey samples. *Journal of Statistical Software* **9,** 1–19 (2004).
24. Andrews, S. (2011). URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Date of access: 20/11/2015).
25. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal* **17,** pp. 10–12 (2011).
26. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
27. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
28. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, gku1196 (2014). URL: http://nar.oxfordjournals.org/content/early/2014/11/20/nar.gku1196.short. (Date of access: 15/11/2015).
29. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

30. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23,** 127–128 (2007).
31. Clermont, O., Christenson, J. K., Denamur, E. & Gordon, D. M. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports* **5,** 58–65 (2013).
32. Scheutz, F. *et al.* Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. *Journal of Clinical Microbiology* **50,** 2951–2963 (2012).
33. Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* **6,** 90 (2014).

Acknowledgements
We are grateful to the Government of the Republic of Zambia (GRZ) and the University of Zambia for facilitating the field and laboratory research while in Zambia. We are indepted to the help of Geoffrey Kwenda and Sydney Malama to enable collection of the human *E. coli* isolates. We acknowledge access to pathogenic avian strains supplied by Zoetis. G.M. is grateful to the Commonwealth Scholarship Commission (CSC) for providing funding (ZMSC-2012-640). N.L. acknowledges support from a University of Edinburgh College studentship. D.L.G., S.P.M., M.P.S. and B.M.B. receive core strategic funding to The Roslin Institute from the Biotechnology & Biological Sciences Research Council (BB/J004227/1). The sequencing by PHE was funded by the National Institute for Health Research scientific research development fund (108601).

Author Contributions
G.M., D.L.G., M.P.S., J.B.M. and B.M.B. designed the project and developed the survey. G.M., D.L.G., L.S. and B.M.B. carried out the sampling work. S.P.M., J.G., J.B.M., L.S., D.L.G. and G.M. assisted with the microbiology design and laboratory work. K.G., N.J.W., S.K.S., R.L.R., G.C., S.A.A. and S.W. supplied unpublished strains and sequences for analysis. G.M., P.B., D.V.H., M.E.C.T. and B.M.B. carried out statistical analyses. N.L., G.M., T.J.D. and D.L.G. carried out bioinformatics analysis and drafted the manuscript, and all the authors read and helped edit the manuscript.

Additional Information
Accession codes: Short read sequence files have been uploaded to European Nucleotide Archive under the study accession number: PRJEB11782, PRJEB11950, PRJEB11956.

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Mainda, G. *et al.* Phylogenomic approaches to determine the zoonotic potential of Shiga toxin-producing *Escherichia coli* (STEC) isolated from Zambian dairy cattle. *Sci. Rep.* **6**, 26589; doi: 10.1038/srep26589 (2016).

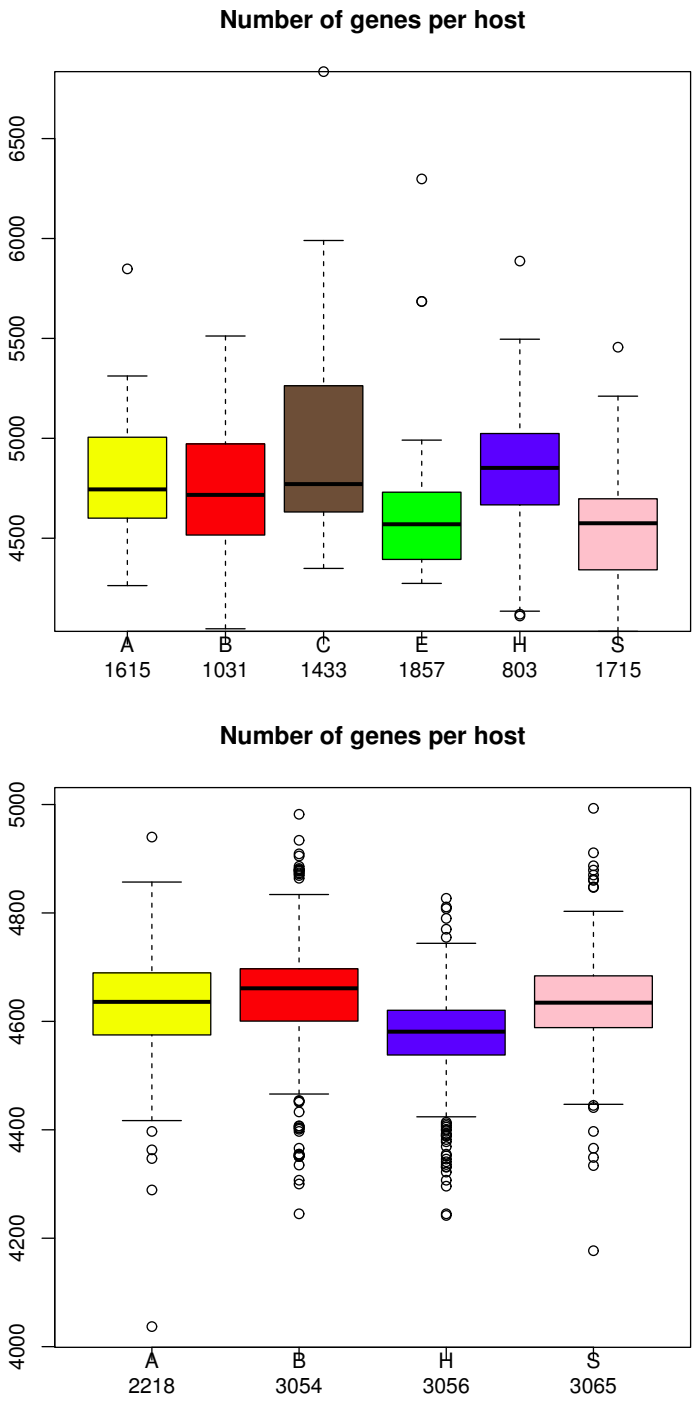


Figure 3.11: The figure illustrates the number of genes in core (x-axis) and pangenome (y-axis) in *E. coli* (top) and *S.* Typhimurium datasets (bottom). The host is represented by the following colors: avian (yellow), bovine (red), human (blue), swine (pink) canine (brown) and environmental (green).

Supplementary Information for:

Phylogenomic approaches to determine the zoonotic potential of Shiga toxin-producing Escherichia coli (STEC) isolated from Zambian dairy cattle.

Geoffrey Mainda
Nadejda Lupolova
Linda Sikakwa
Paul R. Bessell
John B. Muma
Deborah Hoyle
Sean P. McAteer
Kirsty Gibbs
Nicola J. Williams
Samuel K. Sheppard
Roberto La Ragione
Guido Cordoni
Sally A. Argyle
Sam Wagner
Margo E. Chase-Topping
Timothy J. Dallman
Mark P. Stevens
Barend M. deC. Bronsvoort
David L. Gally

Supplementary Table 1: Isolate sequences used in the study

Host	Source (country)/ Name prefix	Number/	Remark and source publication if relevant
Avian	Chicken (various) A- A-CH-	39	This study, two collections of avian isolates. (i) Disease associated isolates from the UK, Italy and Germany, courtesy of Zoetis Animal Health (ii) *E. coli* strains from the GIT of healthy birds (UK). Sequences released to public 31 January 2016 http://www.ebi.ac.uk/ena/data/view/PRJEB11956
	Turkey (Germany and Italy) A-T-	6	Isolates from diseased birds in Germany and Italy, courtesy of Zoetis Animal Health
	Duck (Germany) A-DK	4	Isolates from diseased birds courtesy of Zoetis Animal Health
Bovine	Cattle (Zambian) ZB-	224	Isolates from cattle fecal sampling in central Zambia (1). Sequences released to public 31 January 2016 http://www.ebi.ac.uk/ena/data/view/PRJEB11782
	Cattle (UK) W-	20	A subset of *E. coli* O157 strains isolated from UK cattle (2)
Canine	Canine (UK) C-	18	Multi-drug resistant strains isolated from dogs at the Edinburgh University Veterinary School (3) Sequences release to public 31 January 2016 http://www.ebi.ac.uk/ena/data/view/PRJEB11950
	Community (UK) C-	19	A subset of strains associated with community-acquired canine UTI (3) Sequences release to public 31 January 2016 http://www.ebi.ac.uk/ena/data/view/PRJEB11950
Human	Human (UK) HO- HS-	122	UK STEC strains as published (4, 5, 6) or this study
	Human (Zambia) ZH-	73	Isolated from patients exhibiting symptoms of diarrhea (this study). Sequences release to public 31 January 2016 http://www.ebi.ac.uk/ena/data/view/PRJEB11782
	Shigella isolates (UK) R-	3	NCBI: Ss046, Sb227, Sd197
	Reference genomes (various) R-	31	NCBI: H10407, REL606, HS, IAI1, E24377A, 55989, SE11, TW14359-O157, Sakai, Godstone, SMS, IAI39, CE10, 42, UMNO26, E2348-69, SE15, JJ1886, NA114, 536, S88, IHE3034, PMV, UTI89, UM146, LF82, 857C, 83972, CFT073, W3110, MG1655

Table 1 References:

1. Mainda G, Bessell PB, Muma JB, McAteer SP, Chase-Topping ME, Gibbons J, Stevens MP, Gally DL, Bronsvoort BM. 2015. Prevalence and patterns of antimicrobial resistance among *Escherichia coli* isolated from Zambian dairy cattle across different production systems. Scientific reports. 2015;5.

2. Dallman TG, Ashton PM, Byrne L, Perry NT, PetrovskaL, Ellis R, Allison L, Hanson M, Holmes A, Gunn GJ, Chase-Topping, Woolhouse MEJ, Grant KA, Gally DL, Wain J, Jenkins C. 2015. Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. Microbial Genomics, 10.1099/mgen.0.000029.

3. Wagner S, Gally DL, Argyle SA. 2014. Multidrug-resistant *Escherichia coli* from canine urinary tract infections tend to have commensal phylotypes, lower prevalence of virulence determinants and ampC-replicons. Vet Microbiol. 169:171-8. doi: 10.1016/j.vetmic.2014.01.003.

4. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovska L, Ellis RJ, Elson R, Underwood A, Green J, Hanage WP, Jenkins C, Grant K, Wain J. 2015. Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. Clin Infect Dis. 61:305-12. doi: 10.1093/cid/civ318.

5. Dallman TJ, Chattaway MA, Cowley LA, Doumith M, Tewolde R, Wooldridge DJ, Underwood A, Ready D, Wain J, Foster K, Grant KA, Jenkins C. 2014. An investigation of the diversity of strains of enteroaggregative *Escherichia coli* isolated from cases associated with a large multi-pathogen foodborne outbreak in the UK. PLoS One 9(5):e98103. doi: 10.1371/journal.pone.0098103.

Supplementary Table 2. Gene Identifiers

Gene	GI	Position
arpA	556503834	4222487-4220301
chuA	15829254	4391446-4389464
trpA	556503834	1317222-1316416
TspE4.C2	7330942	Not applicable
yjaA	556503834	4213234-4213617
SepL	NC_002695.1	4593776 - 4594831
eae	NC_002695.1	4596458 - 4599262

Supplementary Table 3: Most frequent H-types defined from whole genome sequence analysis

Collection (all including Zambian)		Zambian only	
H type (44 types identified)	Number of strains 550	H type (42 types identified)	Number of strains 297
H7	242	H21	38
H4	47	H8	26
H21	47	H7	25
H8	31	H4	22
H10	23	H10	16

Supplementary Table 4: Most frequent O types assigned from whole genome analysis

Collection (all including Zambian)		Zambian only	
O type (129 types identified)	Number of strains 483	O type (102 types identified)	Number of strains 235
O157	57	O8	17
O8	20	O25	9
O6	16	O102	9
O25	13	O6	6
O117	13	O150	6

Supplementary Table 5: Shiga toxin subtypes of bovine Zambian STEC

Shiga toxin	Subtype	Number of isolates	%
stx1 only	sxt1a	4	9.76
stx2 only	stx2a	3	7.32
	stx2b	1	2.44
	stx2d	7	17.07
	stx2e	1	2.44
	stx2c	2	4.88
Stx1 and stx2	stx1a, stx2d	9	21.95
	stx1a,stx2a	10	24.39
	stx1a,stx2g	1	2.44
	stx1c, stx2b	1	2.44
	stx1a,stx2c	2	4.88

3.1.4 Conclusions

Core SNP-based phylogeny and typing techniques were used to explore the core genome of the two bacterial datasets (Chapter 3.1) and assess their suitability for investigating bacterial host association. It is appreciated that the *S.* Typhimurium sequences are part of a better dataset as they are more balanced in terms of host of isolation. By contrast, the *E. coli* dataset was more limited although it did provide some insights into two extra host/niche source groups (environmental and canine).

The relative core and pangenome attribution for the species or serovar can vary dramatically depending on both the biology of the grouping (e.g., open vs closed genome species), as well as technical issues such as strictness of the analysis. As an example of this, the core genome for (*E. coli* can vary from 400 genes (at 100% inclusion) to 1,500 (at 95% inclusion). This difference is unlikely to be due to genuine underlying biology but instead reflects the draft quality of assemblies, for example when contigs break in different places disturbing downstream analysis including gene prediction. Inclusion in the core being based on presence in greater than 95% of the genomes seems to be reasonable and more realistic for comparative analysis purposes.

Overall, *E. coli* proportionally has a much smaller core genome and more diverse pangenome that STm. The significantly smaller canine core than that of other hosts can be partly explained by the fact that all the canine isolates were clinical isolates from UTIs with many associated with complicated medical histories and multiple antibiotic treatments. The set was geographically restricted and many of these isolates contain multiple diverse plasmids that aid bacterial survival in these conditions [132]. Taken together, this means the set lacks diversity but the isolates can have extended accessory genomes.

It is interesting that avian isolates for both *E. coli* and STm had much smaller core sizes compared to other host. For both these species, the smaller size of the avian core genome was not coupled with an increased size of its pangenome. Again, these isolates may lack diversity by comparison to some of the other host groups but further investigation on possible adaptation by gene decay should be investigated.

Some phylogroups for *E. coli* are known to be associated with human disease (B2, E), while others (B1, A, C) are mostly associated with commensal carriage. Moreover, a few previous studies have tried to quantify host-phylogroup relationships; however, the overall conclusion is that the distribution of isolates from specific sources being associated with different phylogroups is mainly due

to sampling biases (compare [133], [134], [135]). From our study, the main conclusion is that isolates from any host source can be found in any phylogroup, however with higher or lesser prevalence. So bovine and swine isolates have higher prevalence in the B1 phylogroup, while all other isolates from other hosts were relatively equally distributed amongst other phylogroups. Environmental isolates were quite diverse as a group and were distributed equally amongst the majority of phylogroups, however they had significantly lower prevalences in the A and C phylogroup. Overall, phylogroups for *E. coli* is an interesting concept as phylogrouping provides a stable clustering within the *E. coli* population with the vast majority of the isolates from the same phylogroup clustering together; for example in my current study 80 to 100% isolates from the same phylogroup were clustered together in either the core or pan trees. Nevertheless, it is not yet clear which biological processes make these structures so reliable, and whether core gene alterations in the different phylogroups are associated in any way with acquisition and loss of accessory genes. In other words, what is it about a particular core type that should lead to pathogen vs commensal differences?

As expected, phylogeny of the *E. coli* dataset is much more diverse than that of STm as the *E. coli* phylogeny 3.4 represents a species population diversity while STm 3.3 is an example of within serovar diversity. Core SNP-based phylogeny clustered the majority of avian, swine and human isolates by host,

although for human isolates there were 2 main clusters. On the other hand, nearly 20% of each host population (except bovine) fell into mixed clusters from different hosts. While all host-mixed clusters may contain 'true' generalist isolates, it is curious to investigate what led to formation of only two host clusters like avian-bovine and swine-human 3.3. No link to the same outbreak or same place or even the same submitter to the database was detected.

MLST classification based on house keeping genes has been long accepted as thought to be representative to the population diversity. Nevertheless, with WGS becoming more affordable, MLST schemes were criticized as lacking resolution, with core-MLST schemes starting to appear on the horizon. The current study provides pros and cons for the MLST debate demonstrating clearly that MLST based phylogeny can, within limits, represent population diversity. This study and others [53] [136], [137] demonstrate that phylogeny based on the seven MLST genes for *E. coli* produces a 'true' phylogeny that can be obtained from the whole core alignment. Thus, MLST is a useful approach as it can save time and computational resources when building phylogenies, principally in the modern reality when huge datasets may need to be analysed. On the other hand, there is no direct connection between sequence type and a phylogroup. The same is true for STm MLST based classification which maps well to the whole core phylogeny, but any ST can be found in any of the tree brunches. As an example, the majority of STm isolates belong to the same

ST19 which reflects the situation present on Enterobase where ST19 is by far the most dominate ST. ST19 isolates are distributed evenly across the tree in my study (Figure 3.9) and can also be found in all branches of the tree derived from 10,000 genomes of *Salmonella Typhimurium* from Enterobase 3.10, [62]. Moreover, ST19 is the most diverse ST in relation to host (Figures **??**). Therefore, at least for ST19, ST classification has little value, as these isolates do not cluster together.

In summary, STm demonstrates quite good correlation of phylogeny with host for the majority of host subpopulations, while only rare associations between typing schemes and host were found for this particular datasets for both STm and *E. coli*, the work of other researchers demonstrate quite promising result of using microbial sub-typing with integration of case-control data for source attribution in STm [138], [139] and by additional identification and understanding source specific risk factors for shiga-toxin producing *E. coli* [140].

3.2 Pangenome: host attribution

3.2.1 Introduction

Earliest microbial genomes comparisons made heavy use of reference strains and direct alignments to them, thus overlooking within-species heterogeneity due to additional content. Approaches described in Chapter 3.1 rely on this alignment to a reference genome or reference genes. The drawbacks of such method are obvious, for bacteria species with an open pangenome, *E. coli* for example, much of the information contained in genome would be lost or not considered as the core genes correspond less than a half of the genome content.

According to the idea of a genomic continuum [141], bacteria would absorb and discard genes dependent on particular circumstances and selective pressure. Thus, the accessory genome is more likely to contain traits of sequence adaptation. For a long time, variable gene content except for virulence factors has been left behind without proper analysis. The accessory genome provides new information that was difficult to access, analyse and quantify until now. Accessory genes can be unique to a particular strain or can be shared between few or many but may also not be present in all strains.

Another consideration is that with increasingly large datasets the analysis that can work with hundreds of sequences is computationally non-viable for thousands of genomes. For example, multiple genome sequence alignment or probability based phylogenetic trees from large alignments can be both time and resources consuming to compute and difficult to interpret.

Therefore, scalable methods that allow work with a whole genome, that account for sparse data and can also handle multi-factorial, multidimensional data, should be applied. Machine learning is one of the approaches that seems to be perform well with similar complex datasets in finance, social media and other sciences. One of the first studies using machine learning in bacterial populations looked to predict pathogenicity of bacteria. The study took into account whole genome sequences and identified yet uncharacterised proteins that can play roles in pathogenicity. They also built an algorithm that predicts if a new isolate is pathogenic or not [142]. Such methods are a step further from current typing schemes or phylogeny, as they allow to combine not only genetic information but phenotypic thus identifying strains that are more similar to these that are known to be a threat to human health. Moreover, ML provides a unique opportunity to assign a probability of any new unknown isolate to belong to one of the groups under analysis. Such capability of the algorithm can lead to development of new prognostic and diagnostic tools that can be useful for both monitoring as well as for targeted intervention.

This chapter is presented by the published paper 'Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*' published in Microbial Genomics 2017 [114]. In this study whole genomes of bacterial isolates from STm and *E. coli* dataset are analysed and any associations between accessory genome content and host are quantified.

3.2.2 Results: Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*

MICROBIAL GENOMICS

RESEARCH ARTICLE

Lupolova *et al.*, *Microbial Genomics* 2017;3

DOI 10.1099/mgen.0.000135

Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*

Nadejda Lupolova,[1] Tim J. Dallman,[2] Nicola J. Holden[3] and David L. Gally[4,*]

Abstract

Salmonella enterica and *Escherichia coli* are bacterial species that colonize different animal hosts with sub-types that can cause life-threatening infections in humans. Source attribution of zoonoses is an important goal for infection control as is identification of isolates in reservoir hosts that represent a threat to human health. In this study, host specificity and zoonotic potential were predicted using machine learning in which Support Vector Machine (SVM) classifiers were built based on predicted proteins from whole genome sequences. Analysis of over 1000 *S. enterica* genomes allowed the correct prediction (67–90 % accuracy) of the source host for *S.* Typhimurium isolates and the same classifier could then differentiate the source host for alternative serovars such as *S.* Dublin. A key finding from both phylogeny and SVM methods was that the majority of isolates were assigned to host-specific sub-clusters and had high host-specific SVM scores. Moreover, only a minor subset of isolates had high probability scores for multiple hosts, indicating generalists with genetic content that may facilitate transition between hosts. The same approach correctly identified human versus bovine *E. coli* isolates (83 % accuracy) and the potential of the classifier to predict a zoonotic threat was demonstrated using *E. coli* O157. This research indicates marked host restriction for both *S. enterica* and *E. coli*, with only limited isolate subsets exhibiting host promiscuity by gene content. Machine learning can be successfully applied to interrogate source attribution of bacterial isolates and has the capacity to predict zoonotic potential.

DATA SUMMARY

1. Data used for this work can be downloaded from https://figshare.com/s/7a3ededa8cedd95b9fb7. The files include isolate IDs, protein variants (PVs) and their annotations for *Salmonella enterica* and *Escherichia coli*.

2. Descriptive PVs for each model also can be found at https://figshare.com/s/7a3ededa8cedd95b9fb7. The name of the file describes the model for which these PVs were used. So salmonella_PV_30_AO_annotations.csv means these are the PVs that describe *Salmonella* Typhimurium Avian isolates vs all Other isolates.

3. Isolate metadata for both species (original host, predictions, place, year and multilocus sequence type) are visualized using pan genome trees and can be viewed on ITOL: http://itol.embl.de/shared/nlupolova.

INTRODUCTION

Salmonella enterica and *Escherichia coli* can be isolated from a large number of animal hosts, in particular birds and mammals. When isolated, *S. enterica* serovars are usually associated with disease whereas the majority of *E. coli* are commensals with only a subset considered overt pathogens [1, 2]. Infections caused by these two genera are a major burden on human morbidity and mortality and many of these infections are zoonotic, i.e. are transmitted from animals to humans. Host restriction or specificity has been a key area of research for *Salmonella*, and host-specific serovars such as *S.* Typhi and *S.* Gallinarum are responsible for more severe systemic disease in their primary host, whereas serovars with broader host ranges, such as *S.* Typhimurium (STm) and *S.* Enteritidis are often restricted to gastrointestinal disease in their different hosts. However, this differentiation is increasingly appearing simplistic with identification of invasive strains of STm, such as ST313, in humans [3–5]. The fundamental biology underlying host restriction is important to understand as it shows the barriers these bacteria need to overcome to successfully colonize and cause disease in a new host. From a public health perspective, the capacity to ascribe correctly the source of an infection is

Received 21 May 2017; Accepted 5 September 2017

Author affiliations: [1]University of Edinburgh, Edinburgh, UK; [2]Public Health England, England, UK; [3]James Hutton Institute, Dundee, UK; [4]Division of Immunity and Infection, The Roslin Institute, University of Edinburgh, Easter Bush, Edinburgh EH25 9RG, UK.

***Correspondence:** David L. Gally, dgally@ed.ac.uk

Keywords: host specificity; machine learning; Support Vector Machine; *Salmonella*; *E. coli*; zoonosis.

Abbreviations: MLST, multilocus sequence typing; PV, protein variant; ΔPV30, subtractive difference equal to 30 or less in proportions of a PV between two classes; ST, sequence type; STm, *Salmonella* Typhimurium; stx, *Shiga* toxin alleles; SVM, Support Vector Machine.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Twelve supplementary figures are available with the online Supplementary Material.

important as it can inform ways to intervene and limit human infection from animal and food sources.

Compared to *S. enterica*, the host-specificity of *E. coli* has been less well investigated. Classically the species has been divided into phylogroups (A, B1, B2, C, D, E and F) which are based on possession of a small number of specific alleles [6]. This classification was considered to have only a weak association with isolation host but has the advantage that commensal and pathogenic isolates are often assigned to separate types [7]. The genetic relatedness demonstrated by a reduced allele methods such as multilocus sequence typing (MLST) and phylogroup agree well with high-resolution core genome SNP typing [8, 9]. More recent studies that also take into account accessory genome information do provide examples of host specialization for *E. coli* [10, 11], which should also include *Shigella* species as a type of enteroinvasive *E. coli* [2, 12]. Fundamentally, *Escherichia* and *Salmonella* share the same gene acquisition, mutation and recombination systems as well as overall physiology. As such both should have the same genetic potential for plasticity that could result in host adaptation or host promiscuity. In the last few years short read sequences from thousands of bacterial genomes have been deposited in databases, although often their use is limited by lack of associated metadata. Where the isolation host is known, this now provides an opportunity to interrogate such sequence data for genetic signals and predictors of host specificity and determine how these map onto the phylogeny of the different species. Recently, a machine-learning algorithm, Support Vector Machine (SVM), was used to analyse sequence data from *E. coli* O157 isolated from cattle and humans, so as to determine if all cattle isolates had the same genetic potential to cause detectable infections in humans [13]. There were isolates from cattle that had genetic information allied more closely to isolates associated with human infection, indicative of strains with increased zoonotic potential. In the current study, we have combined a machine learning approach with pan genome analyses of both *S. enterica* and *E. coli* to investigate the relatedness of isolates from different hosts. The primary aim of this study was to demonstrate the potential of machine learning, in this case SVM, to predict the source of an isolate, and indicate its potential to transfer between hosts, including the zoonotic threat to humans.

IMPACT STATEMENT

Both *Salmonella enterica* and *Escherichia coli* are bacterial species with a broad animal host range with strains that can colonize humans and in some cases cause lethal infections. Both species have large accessory genomes and it is established for certain subtypes (serovars) of *S. enterica* that these can be host-restricted with both gene acquisition and loss contributing to the degree of host specificity. The extent of host restriction for *E. coli* and *Salmonella* serovar Typhimurium is not known and the capacity to predict the source of human infections with these bacteria is important to understand the origin of zoonoses and aid public health interventions. The work in this study has successfully applied a machine learning algorithm, Support Vector Machine, to attribute the source animal or environment of these bacteria based on their genome content. The work should have value to allow the sources of zoonotic outbreaks to be identified and also to assign sources to environmental and water pollution events. The research will also help identify the genes and pathways that lead to host restriction and are therefore required for infection in different animal hosts.

METHODS

Genome analysis

Illumina short read sequences were assembled with SPAdes [14] and annotated with Prokka [15]. Sequence type was assigned using MLST v. 2.4 [16]. Pan genomes were clustered with Roary [17], paralogues were split and the threshold for sequence similarity was set to 95 % at the amino-acid level. The core SNP trees were built with RAxML [18] based on aligned core genes (Table 1). Accessory trees based on the presence or absence of accessory genes were extracted from the Roary output (Table 1). Shiga-toxin (*stx*)-positive isolates were detected using a BLASTN search with an *stx1* query (NC_004913.3, coordintates: 33251–31917) and *stx2a* query (NC_002695.1, coordinates: 1266960–1267928).

Support vector machine analysis

SVM implementation in R package e1071 [19] was used to build classifiers with radial kernel, weighted classes, and 'gamma' and 'cost' parameters adjusted after tuning for each host. Protein presence and absence output from Roary were used to identify features for each class of an SVM model. Proteins were clustered with high (95 %) similarity, and therefore related proteins (less than 95% similarity) were allocated into different clusters. For these the term 'protein variants (PVs)' was introduced to more precisely describe the Roary output. PVs that differentiate the two classes under test were chosen for the respective classifier. For example, the proportion of each PV found in the STm avian host group was compared with the proportion found

Table 1. Summary of gene content for *S. enterica* and *E. coli* isolates analysed in this study

Section	Description	*S. enterica*	*E. coli*
Number of isolates		1682	943
Core genes	99 %≤strains≤100 %	3175	1328
Accessory genes	0 %≤strains≤99 %	20 132	91 087
Soft core genes	95 %≤strains<99 %	236	815
Shell genes	15 %≤strains<95 %	2098	3516
Cloud genes	0 %≤strains<15 %	17 748	86 746
Total genes	0 %≤strains≤100 %	23 307	92 415

in the 'all others' STm host groups. PVs which differed by at least 30 % between the two groups (ΔPV30) were used as descriptive features for the SVM STm model. The higher the ΔPV values the more clearly distinct the groups are, although there is a trade-off between PV discrimination, the number of PVs and model accuracy (Figs S1 and S2, available in the online Supplementary Material).

All SVM classifiers in this study were based on comparing two groups of data. For example, for serovar Typhi vs Dublin this was straightforward and the two training classes were based on the predicted proteins extracted from these two specific serovars. For analysis of STm and for *E. coli* datasets a 'one against all' approach [20] was used. For each classifier, the differential PVs are defined by comparing PVs of the isolates from one specific host with those from isolates from the remaining hosts combined. This means that for STm four different classifiers were built (avian vs the rest; cattle vs the rest; human vs the rest; and porcine vs the rest) and for *E. coli* six different classifiers were used (human vs the rest; porcine vs the rest; canine vs the rest; avian vs the rest; environmental vs the rest; and cattle vs the rest).

For each classifier 10× cross-validation was performed, meaning the data were split randomly into 10 groups and trained on 90 % of the isolates and the remaining 10 % used for testing, and this process was then repeated with different gamma and cost values. Different approaches to sub-sampling for test sets were taken. The main method was to remove each isolate in turn from the training set; each time PVs were re-calculated, the model was tuned and parameters were adjusted. After this the removed isolate was tested. From this a probability attribution for each isolate and for each host was generated. The overall performance for each classifier was assessed by plotting true positive vs true negative rates and calculating an area under the curve for each of the classifiers (Fig. S3 for STm and Fig. S4 for *E. coli*).

The Typhi and Dublin datasets were analysed differently. (1) Initially 20 isolates were randomly selected from each Typhi-human and Dublin-bovine dataset to be used as the test groups. The remaining Typhi-human and Dublin-bovine isolates (training dataset) were labelled and differential PVs were calculated between them. ΔPV90 values were used as features for the model. To access model accuracy 10× cross-validation was performed on the training set and the best parameters of 'cost' and 'gamma' were extracted from the tuning step. (2) A second test was similar to the above except that for the test group we included four Dublin human isolates. (3) To test the Dublin serovar alone, one isolate was removed for each assessment cycle, and the SVM classifier was retrained on all other human and bovine Dublin isolates; this involved recalculating discriminatory PVs (with ΔPV50 for the model) and then testing the removed isolate. (4) The combination involved the STm human and bovine datasets as a training model with ΔPV30, and then all isolates from the Dublin-bovine, Dublin-human and the Typhi datasets were tested.

Significance of the results was described using *P* values, which were obtained using basic R functions as required: Student's *t*-test (t.test function) and Fisher's exact test (fisher.test function).

RESULTS

Salmonella enterica

In total, 1682 *Salmonella* sequences were obtained from Enterobase [21]. The collection included serovar Typhi (250 human isolates), serovar Dublin (187 bovine isolates and 40 human isolates), and serovar Typhimurium (STm; 336 human isolates, 300 bovine isolates, 311 avian isolates, 256 swine isolates) (defined in supplementary file 'Salmonella_-data.tgz'). The isolates were diverse in terms of their year of isolation, ranging from 1945 to 2016, and their geography, which covered all continents with the exception of Antarctica. We assigned sequence type (ST) based on the *S. enterica* MLST scheme and identified 52 different STs in the whole dataset, although for each serovar one or two STs were dominant: Typhi ST 1 (*n*=185); Dublin ST 10 (*n*=206); and Typhimurium ST 19 (*n*=992) and ST 34 (*n*=122).

Both core (3175 genes) and accessory (20 132 genes) genome relationships (Table 1) were plotted as trees based on information derived from the sequences. The clustering obtained for both trees (Figs 1 and S5) was quite similar with serovars Typhi, Dublin and STm on separate branches, illustrating a good correlation between core SNPs and accessory genome content at serovar level. Overall, both core and accessory trees show a large avian cluster that contained 82 % of all the avian STm isolates, a few different human clusters with the largest two in the accessory tree containing 38 and 20 % of all human STm isolates and one bovine cluster with 23 % of all bovine STm isolates (Fig. 1). Swine isolates had no sub-cluster that contained at least 20 % of the isolates together in the accessory tree but the core tree did contain such a cluster. The accessory genome therefore indicates some clustering by host for STm, especially for avian and human isolates, but many of the STm isolates from different hosts were interspersed within several branches containing isolates of mixed origin. *S.* Typhi and *S.* Dublin were included as 'established' host-restricted serovars and provided a framework for analysis of the host association of STm isolates.

SVM prediction of isolation host

To predict the isolation host of an *S. enterica* isolate, the SVM classifier was built using a 'one against all' approach, i.e. differentiating one host group from all other host sequences based on discriminatory PVs as described in the Methods. Initially, the classification and prediction method was applied to serovars Typhi and Dublin. *S.* Typhi is human host-specific while *S.* Dublin is generally associated with severe infections in cattle with some human cases. There were 752 ΔPV90 found almost exclusively in the Typhi isolates and similar numbers (*n*=746) describing the Dublin isolates. Randomly taking 20 isolates from each serovar for testing, and training on the remainder, it was

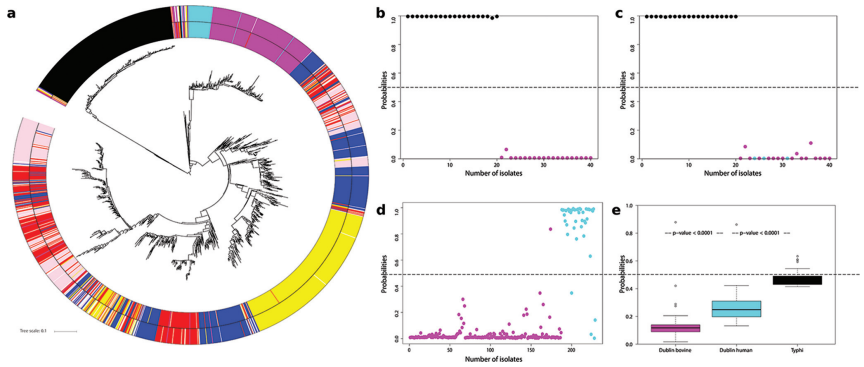


Fig. 1. Host association of *Salmonella enterica*. Colour scheme of serovars: Typhi (black); Dublin-bovine (magenta); Dublin-human (cyan); STm avian (yellow); STm bovine (red); STm human (blue); STm swine (pink). (a) Clustering of isolates based on accessory genome content (non-core): distinct branches are evident for Typhi and Dublin serovars. Inside of STm there is some clustering associated with host; the majority of avian isolates cluster together, 80 % of the human isolates cluster in three groups, while the bovine and swine isolates are mostly found in groups of mixed origin. The outer ring shows the SVM host prediction when >0.5 (see Methods) and is otherwise left blank. (b) SVM prediction of *Salmonella* Typhi (human) vs serovar Dublin (bovine). Twenty isolates were randomly taken from each serovar for testing, and the model was trained on the remaining sequences (230 Typhi-human, 167 Dublin-bovine). Prediction was 100 % accurate due to highly discriminatory PVs (ΔPV90=1349, ΔPV100=8). (c) The SVM classifier in (b) was applied to serovar Dublin isolates from both cattle (magenta) and humans (cyan): this primarily discriminates the serovar not the host as there is still complete separation between Typhi (black) and Dublin serovars (cyan and magenta). (d) If predictions were based on training with only Dublin human and bovine information then the Dublin isolates can be separated by this classification. (e) In this case STm bovine and human isolates were used as the training sets and testing was carried out on the distinct serovars: Dublin-bovine, Dublin-human and Typhi-human. Notably, the three groups can now be separated by the STm classifier in a logical trend based on isolation host.

found that these isolates could be separated with 100 % accuracy (Fig. 1b). This basic prediction was to be expected from the obvious differences in genetic content (ΔPV90=1349, ΔPV100=8) and their separation on both core and pan-genome trees. Note that in this situation any human vs bovine signal is masked by more significant serovar/phylogenetic differences. This is shown by adding Dublin human isolates into the analysis; in this case both Typhi and Dublin can still be accurately predicted (100 %), with separation not due to host as the human and bovine Dublin isolates receive the same prediction scores (Fig. 1c).

To try to exclude the serovar genetic signal, we tested SVM assignment of isolation host for a single serovar, STm, using the sequences of isolates collected from different sources (human, bovine, avian and swine). From 17 145 total PVs within the STm pan genome, the subtractive difference between the average presence of PVs in a host group versus all others isolates (ΔPV) can be ranked and binned by its discriminatory power (Fig. 2a). The aim was to predict the isolation host with as few PVs as possible while maintaining an acceptable accuracy for the model and its application across all host datasets. A series of test runs were carried out based on different ΔPV values, assessing the quality of prediction and trying to find a 'one solution fits all' for the datasets. ΔPV30 was eventually used as a features discriminator for SVM (Fig. 2b). However, it is clear that ΔPV30 is not the best option for all host groups and to further improve predictions for individual studies it would be advisable to choose the ΔPV value according to the dataset.

The main SVM analyses were then carried out by removing a single isolate and training the model with the remainder and then testing that isolate; this process was then repeated until all isolates had been tested. The distribution of discriminatory PVs differed in the four host groups (Fig. 2a): 684 ΔPV30 describe the avian group, 284 PVs for swine, 198 PVs for the bovine group and 182 PVs for the human isolates. Several highly discriminating PVs were identified for the avian group (45 ΔPV80) while for the other host groups there were only a limited number at the ΔPV50 level (bovine=2, human=13, swine=6).

SVM generates a probability, based on comparison of genetic content with the training set, of each test isolate belonging to a specific host group. As such, a logical starting point for our assessment of the methodology for host assignment was to determine how well test isolates could be assigned based on a prediction probability of >0.5 for a specific host. Using this threshold, the majority of the isolates could be classified in relation to their isolation host: 89 % of avian isolates (276 out of 311), 67 % of bovine isolates (202/300), 90 % of human isolates (301/336) and 75 % of swine isolates (192/256) (Fig. 2c, d). The distribution of probabilities was quite distinct for the different host groups. So while

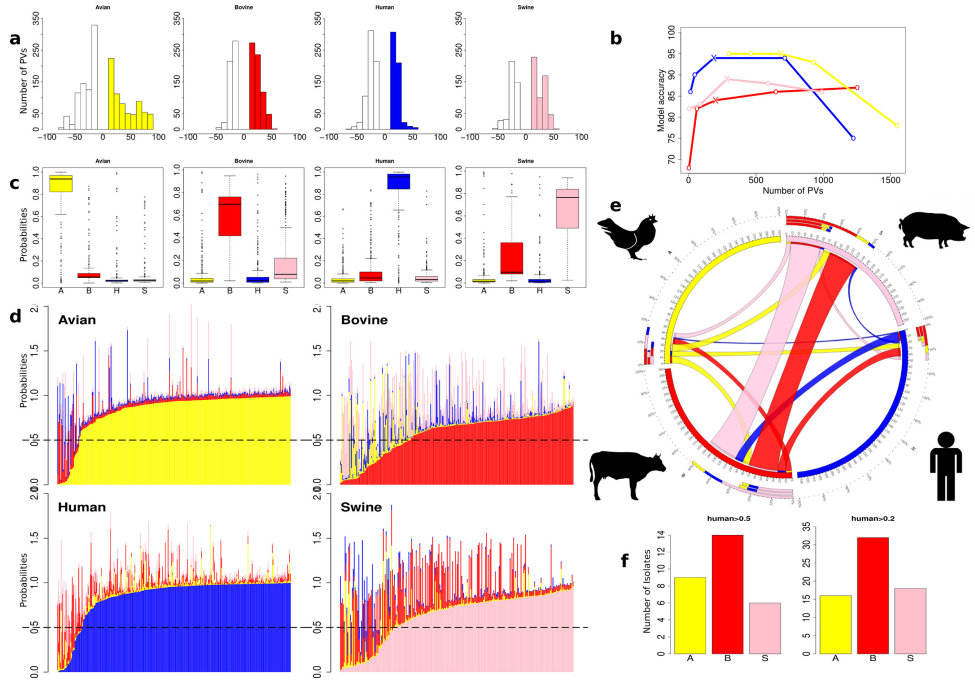


Fig. 2. Host prediction by SVM for STm. Colour scheme: STm-avian (yellow); STm-bovine (red); STm-human (blue); STm-swine (pink). (a) The number of and differential PVs on which predictions were based for each host. PVs that differed by less than 10 are not shown (see Methods). The coloured bars are the number of PVs that are present in higher levels in the specified host group, while white bars are the number of PVs more abundant in the 'all other' population. (b) Graph showing the relationship between the number of PVs and model accuracy for each host group. Individual points relate to the number of PVs at different ΔPV thresholds from ΔPV>10 to ΔPV>50, plotted from right to left. Crosses define the number of PVs and model accuracy at ΔPV>30, which was applied in the study. (c) Probability assignments of isolate genome content to each host. All STm isolates were tested for their score assignment to each host, expressed as a probability. The sources of the majority of the isolates were predicted correctly, although some hosts have isolates that were more likely to contain genetic information that overlapped with another host. (d) SVM-assigned probabilities for each host plotted for each isolate as a stacked bar. This allows a comparison of the level of host specificity for each isolate. (e) Circos plot depicting the proportion of STm sequence features from each host that can be found in another host. For example, 51 swine isolates with strong porcine prediction scores (>0.5) also had high (>0.5) scores for genetic features from bovine isolates and these are shown as a pink ribbon going from the swine host to bovine host. In total, 52 bovine isolates had a high (>0.5) swine signal and are depicted with a red ribbon going from cattle to swine. The outer ring plots these data as the percentage of isolates assigned other host scores for each specific host. (f) STm isolates scored as human from the different hosts. For each STm isolate the probability of belonging to the human training group was assessed. With a threshold probability of 0.5, there were: nine avian (3 %), 14 bovine (5 %) and six swine (2 %) isolates. When the threshold was set at 0.2, there were 16 avian (5 %), 32 bovine (11 %) and 18 swine (7 %). At this threshold the higher proportion of cattle isolates with human isolate features is significant (Fisher's exact test: P=0.035).

the majority of human and avian isolates had high host assignment scores (above 0.8), only a small proportion of isolates achieved such high scores in the bovine and swine groups. The strong SVM assignment for avian and human strains correlated well with their accessory genome clustering (compare Figs 1a and 2c, d). It was also evident that the majority of all isolates achieved a score higher than 0.5 for only one host (94 %), indicating dominant genomic characteristics for this host. However, there were isolates in each host group that scored highly for two or more hosts (total n=73), indicating that such isolates, termed 'generalists', already contain genetic information that could facilitate existence in at least one other host. Some host groups had low proportions of generalist strains, i.e. the avian and human groups, in which only a minority of isolates were assigned with second host probabilities, >0.5 (human n=6, avian n=11), while the proportions were much higher among the bovine (n=27) and swine (n=29) isolates (Fig. 2c, d).

At the left-hand side of each ranked host probability plot (Fig. 2d) are those isolates that have a low probability score for that specific isolation host and would be called as not significantly associated with that host. These isolates often have higher scores for other hosts. This may reflect: (1) the limitation of our strain sets in that we are failing to capture all genetic information relevant to a specific host; (2) incorrect metadata or methodology around collection; and (3) the isolate may be transient and may not persist in that host. It is notable that isolates that are not significantly associated with a particular host (the blanks in the outer ring of Fig. 1a) through the SVM classifier are present in grouped clusters in the pan-genome tree; this includes isolates with a range of host allocations. The implication of this is that particular STm sub-clusters may have a greater potential to switch between specific hosts based on analysis of their genome content.

SVM was used to assign scores to each isolate in relation to its host-relevant genetic content, creating a unique measure of host specificity, and indicates, from this collection of isolates, which animals may be more likely to exchange STm isolates (Fig. 2e). In line with the core and accessory genome trees, the analysis demonstrated a surprising level of host specificity for STm isolates, in particular with avian and human isolates, providing evidence that there may be more human-specific strains circulating beyond our current concerns with ST313. According to this approach, swine and bovine STm isolates can share significant genetic information (Fig. 2e, f).

In each of the three non-human host groups (avian, bovine and swine) there were isolates that achieved an isolation host probability of >0.5 but also reasonable scores for human association. At a 0.5 threshold for human isolate content there was no significant difference between the three hosts, although the highest numbers of such isolates were from the bovine host [avian (n=3, 0.9 %), bovine (n=9, 3 %), swine (n=0, 0 %)]. At a lower threshold (probability assignment >0.2) bovine isolates were significantly more likely to have genetic content associated with human STm isolates when compared with avian and swine isolates [avian (n=6, 1.9 %), bovine (n=20, 6.6 %), swine (n=2, 0.8 %)] (Fig. 2e, f).

We note that as with STm, SVM analysis of bovine and human isolates from within the Dublin serovar can also be predicted with high accuracy (Fig. 1d), as the classifier is again working within the same serovar and so presumably is not confounded by the serovar signal. We then investigated whether it is possible to predict isolation host across serovars, in this case by training on human and bovine STm isolates and testing on human and bovine Dublin isolates, as well as Typhi (Fig. 1e). It was evident that the Dublin isolates could be differentiated by their source even though the training was with STm genome content. Furthermore, *S.* Typhi isolates could be further differentiated in this model based on the STm classifier with significantly higher human association scores (average probability scores for Dublin bovine=0.15, Dublin human=0.27, Typhi=0.48) (Fig. 1e).

Escherichia coli

The *E. coli* dataset was composed of sequences from 943 isolates from six different sources: avian (n=87), bovine (n=308), canine (n=57), environmental (n=40), human (n=388) and swine (n=63). The analysis also included three *Shigella* isolates as these cluster genetically within the *E. coli* species and are considered human-specific. Clustering by relatedness of the accessory genomes is summarized in Table 1. While the number of *E. coli* isolates analysed is almost half that for *S. enterica*, this produced a pan genome that was four times larger than that of *S. enterica*, with more than 90 000 genes. The differences between these two bacterial species were also reflected in the size of their core genome, for which *S. enterica* had almost three-quarters of its genome content shared among the isolates examined while for *E. coli* only one-fifth was conserved across the sequences analysed (Fig. S6). In total, 279 different STs were attributed to *E. coli*, again indicating much greater diversification than for *S. enterica*. A direct comparison of the *E. coli* and *S. enterica* phylogenetic clusters at either accessory (Figs 1a and 3a) or core levels (Figs S5 and S7) indicates a less clear association by host for *E. coli*, although multiple clusters by host were present, especially for human and bovine isolates. Therefore, compared to what was observed for *S. enterica*, there is far more mixing of sub-clusters based on source association for *E. coli*, making prediction of host/habitat attribution more challenging from the accessory genome data presented in this format.

SVM prediction of isolation host for *E. coli*

The host/habitat association was predicted based on the SVM approach, in the same manner as for STm. In contrast to the STm analysis, only *E. coli* human and bovine datasets had equivalent isolate numbers so we first tested the impact of reducing dataset sizes on prediction accuracy by working with sub-samples of the larger datasets for both STm and *E. coli*. It was apparent that prediction capacity was substantially reduced when working with fewer than 100 isolate sequences (Figs S1 and S2). Therefore, while predicting the presence of both human and bovine genetic content is valid based on our group sizes, predictions for genetic content pertaining to avian, canine, environmental and swine isolates would require more isolates from these sources to be sequenced and made available. For the larger human and bovine datasets, the prediction capacity was equivalent to that for STm isolates: 72 % (223/308) of bovine *E. coli* isolates and 89 % (346/388) of human *E. coli* isolates were predicted correctly as originating from those hosts based on ΔPV30 (Fig. 3b) using a prediction probability of >0.5.

As with the STm analysis, this indicates a stronger genetic signal for human isolates and greater genetic diversity for bovine isolates. By reducing to ΔPV20 and based on analysis of PV distributions (Fig. S8), the prediction scores for avian, swine and canine *E. coli* isolates show patterns similar to human and bovine isolates when similar size training sets

were used (Figs S2, S4, S9 and S10). Therefore, we propose that an equivalent prediction capacity for these sources should be achievable when more isolate sequences are available to train the classifiers. Of note was the pattern for the environmental isolates, half of which showed a strong environmental score while the other half showed a negligible association. This may reflect two different populations of *E. coli* present in the environment, one that is plant/soil associated and the rest more directly related to animals. The ΔPV30 assignments were used to examine human and bovine genetic traits across all the *E. coli* isolates in the study (Fig. 3c). Only a minority of isolates outside of the same host had substantial genetic content associating them with bovine and human hosts (P>0.5), which may indicate that only a specific subset may be able to transfer and effectively colonize the two different hosts. Although based on small datasets, it was evident that the inter-relationship between bovine and swine genetic content as seen for STm was not apparent for *E. coli*. Zoonotic potential based on this content can be plotted as for STm. When using a threshold of P>0.5, there was a clear and statistically significant hierarchy working towards content in human isolates [environmental (n=0, 0 %), avian (n=5, 6 %), bovine (n=19, 6 %), canine (n=7, 12 %), swine (n=12, 19 %), Fisher's exact test, P=0.002216; Fig. S11)]. The numbers and percentages at a lower threshold of probability, >0.2, were: environmental (n=1, 2.5 %), avian (n=16, 18 %), bovine (n=40, 13 %), canine (n=16, 28 %) and swine (n=22, 35 %) (Fisher's exact test, P=1.023e-05). Independent of the threshold and based on the percentage of isolates (rather than actual number as group sizes varied), porcine isolates had the strongest association with human isolates. Overall, the data indicate that environmental *E. coli* isolates may be less likely to directly infect humans and that bovine, swine and canine isolates are much more likely to be a zoonotic threat than isolates from birds. While this assessment will be refined as more sequences become available, it does demonstrate the utility of the approach.

Testing the predictive capacity of SVM with an established bacterial zoonosis

As a proof of principle to support the SVM assignments in this study, we determined how machine learning would score sequences from a well-characterized zoonosis. We chose *E. coli* O157 as this clonal group colonizes cattle as an asymptomatic reservoir host and can cause potentially fatal disease in humans as an incidental host. To generate a baseline, all human and bovine isolates from the *E. coli* dataset, but excluding *E. coli* O157 (n=688, human=381, bovine=307), were used for SVM training with prediction of host source using ΔPV30 (n=139). In this case, training was carried out on 90 % of isolates and testing on 10 % until all isolates had been tested. Overall the source of 92 % of isolates was predicted correctly: 279 of 307 bovine isolates (91 %) and 352 of 381 human isolates (92 %). Most of the isolates were predicted with very high probabilities of originating from human or bovine hosts, with a mean probability of 0.8 (1st quartile of 0.95 and 3rd quartile of 0.98) for human assignments and a mean probability of 0.13 (1st quartile of 0.01 and 3rd quartile of 0.106) for bovine assignments (Fig. 4a). *E. coli* O157 isolates (n=25: 14 human, 11 bovine) were then tested in this context along with the three *Shigella* (human isolates) (Fig. 4a). The majority of the probabilities assigned for the O157 isolates were in the mid-range between high human and bovine scores (mean 0.58, 1st quartile 0.44, 3rd quartile 0.73), indicating that the *E. coli* O157 isolates contain ambiguity in their gene content that may allow association with both hosts.

One potential source of bias in the training dataset was the presence of *stx*-positive strains other than O157. Therefore, another analysis was carried out in which *stx*+ isolates were identified and removed from the training dataset (see Methods). The baseline was re-assessed and similar results were

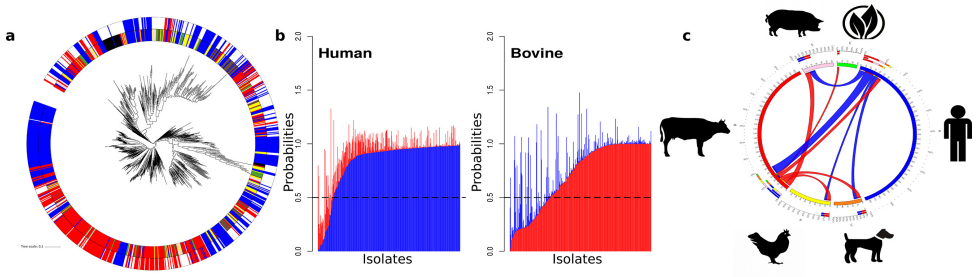


Fig. 3. Accessory genome analysis and host prediction by SVM for *E. coli*. Colour scheme: avian (yellow); bovine (red); human (blue); swine (pink). (a) Accessory genome tree based on PVs: some clustering by host for human and bovine isolates was evident. The outer ring indicates the position and isolation host of isolates incorrectly called as human by SVM analysis. (b) SVM host assignment probabilities for human and bovine hosts. The probabilities for each isolate are plotted as stacked bars. (c) The proportions of isolates from each host with human or bovine features.

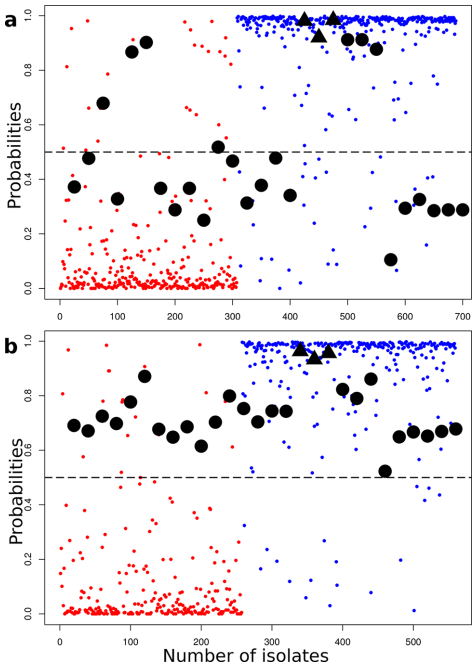


Fig. 4. Host assignment of an established bacterial zoonosis: *E. coli* O157. Colour scheme: human (blue) and bovine (red). *E. coli* isolates from both cattle and humans are plotted with their predicted host assignment probability. All these isolates were used as a training dataset to determine host assignment probabilities for O157 isolates (black circles) and three *Shigella* isolates (black triangles). (a) Training set containing *stx*+*E. coli* isolates but not serovar O157, and host assignment probability was then predicted for an O157 test group. (b) Training set with all *stx*-positive isolates removed and the host assigned for the same *E. coli* O157 test group. In both cases the *E. coli* O157 isolates, irrespective of their isolation host, score as containing mixed genetic information in relation to the training set of human and bovine *E. coli* isolates, indicating transmission/zoonotic potential.

obtained with 91.5 % (214 of 234) of bovine isolates and 93 % (266 of 285) human isolates predicted correctly. When probabilities were assigned to the O157 isolates, their distribution changed significantly [mean=0.7203, 1st quartile=0.6520, 3rd quartile=0.7842 (P=0.001)] when compared with the previous analysis that included *stx*+ isolates in the training set (Fig. 4b). It is interesting that there were different sets of differential PVs that describe the human and bovine *E. coli* populations depending on the presence (ΔPV30=136) or absence (ΔPV30=248) of *stx*+ isolates. Overall, this result provides strong support for the capacity of the SVM classifier to predict isolates with across-species transmission potential with *E. coli* O157 being assigned probabilities more indicative of human isolates despite cattle being their primary reservoir.

DISCUSSION

Public repositories of bacterial whole genome sequences, even with very limited metadata, allow new approaches to be tested that address fundamental biological questions such as host specificity and zoonotic potential. In this study we wanted to determine if a machine learning approach, specifically SVM, could assign the isolation host/habitat for both *S. enterica* and *E. coli* isolates based on analysis of differential predicted PVs. Moreover, we wanted to determine if the capacity for inter-species transmission was predictable from the gene content, including estimation of human zoonotic potential. The methods were first applied to *S. enterica* isolates, as serovars such as *S.* Typhi and *S.* Dublin exhibit host specificity and restriction, respectively. As was apparent by both core genome (SNP) and accessory genome analyses, including SVM, *Salmonella* serovars were distinct and easily assigned. By contrast, serovar Typhimurium can be isolated from many different hosts and can cause significant disease in humans with animals often considered the initial source of the infection. Both core and accessory genome clustering provided clear evidence for sub-clusters of STm and several of these were strongly host-associated, in particular for avian and human isolates. Although our analysis is restricted to only a small sample size, it does indicate that host-restricted lineages of STm may extend beyond those receiving attention in relation to their disease severity [22]. The SVM analysis supported these findings with strong host assignment scores for STm isolates. Conversely, only particular sub-clusters contained STm isolates from multiple hosts, and SVM calling of source host in these was much more challenging. However, this does indicate that particular clusters have genetic content that may be more associated with inter-species transmission, indicative of patchy promiscuity within the species.

Certain isolates from each animal host had more genomic content allied with human STm isolates potentially reflecting more of a capacity to infect humans. Overall the bovine STm isolates had the highest predicted 'human' scores, even compared with avian isolates. The fact that human STm infections may be more commonly associated with poultry [23] may reflect aspects of the food chain rather than the comparative infection threat of avian STm isolates. In fact, our analysis indicates that the majority of avian STm isolates analysed were quite host-specific and may not pose a public health threat. Support that the SVM classifier was using 'host-related' genetic information was provided by training on differential PVs from human and bovine STm isolates and testing on *S.* Dublin from humans and bovine as well as *S.* Typhi from humans. These sets were successfully discriminated by host (Fig. 1e), despite strong phylogenetic signals for the serovar. It is difficult to assess how the phylogeny impacts on the host assignment and in some cases the evolution of particular subtypes may have been driven by host association, in which case phylogenetic and

host signatures may overlap. When all isolates from a particular host are combined, the information coming from specific branches/sub-clusters, and therefore the phylogenetic signal, will be diluted and mixed with information from other isolates from other branches. When we then use PVs that describe the 'avian population' from these different branches, we decrease the importance of the tree structure. For example, we can predict avian strains from different regions of the tree despite there being a dominant avian isolate cluster. A primary driver for this publication is to demonstrate the potential for machine learning alongside phylogeny approaches, and the value and relationships between these will become apparent as sequences of more isolates from different sources become available.

E. coli, in comparison to *S.* Typhimurium, had more limited host-specific sub-clusters based on core and accessory genome analyses, although it was still possible to correctly call the host of origin for the more populated datasets of bovine and human isolates using the SVM classifier. We included isolates from other hosts/habitats to provide more discriminatory power in the 'one host vs all approach' but again prediction accuracy for *E. coli* from different sources will increase as more of these host/habitat-related sequences are made available (Fig. S2a, b). Even so, it was evident that environmental *E. coli* isolates had very little overlap with human isolates and that human infection may therefore be more likely from animal-adapted *E. coli* isolates. With the SVM approach, bovine, swine and canine isolates all had subsets that shared significant genetic content with human isolates. The analysis of *E. coli* O157 isolates provided validation that the SVM classifier, trained on bovine and human *E. coli* isolates, could identify isolates with increased zoonotic potential, as isolates of this established zoonotic clone produced intermediate scores reflecting mixed genetic assignment between other human and bovine isolates.

Both STm and *E. coli* isolates exhibited marked host restriction when genetic content was evaluated using a combination of phylogenetic and machine learning methods. We consider this is counter to a perception that these bacteria are 'generalists' capable of switching between hosts. Instead, our analyses indicate that only specific subsets of strains have 'mixed' genetic content, which we suggest indicates the capacity to transfer and succeed in different hosts, although this now needs to be tested using experimental approaches. We consider that machine learning has tremendous potential to interrogate complex sequence datasets and identify genes/sequences associated with host specificity. This will have value for source attribution in both a public health context and, for example, in ascribing the source of water pollution events if sequences of the bacteria are obtained.

Funding information
N. L. received a PhD scholarship from the college of Medicine and Veterinary Medicine, University of Edinburgh. D. L. G./N. L. are supported by a Roslin Institute strategic programme funded by the BBSRC (BB/P013740/1).

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
No ethics or consent approval was required for the research in this study.

Data bibliography
Figshare, https://figshare.com/s/7a3ededa8cedd95b9fb7 (2017).

References

1. **Kaper JB, Nataro JP, Mobley HL.** Pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2004;2:123–140.
2. **Chaudhuri RR, Henderson IR.** The evolution of the *Escherichia coli* phylogeny. *Infect Genet Evol* 2012;12:214–226.
3. **Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N *et al.*** Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. *Proc Natl Acad Sci USA* 2015;112:863–868.
4. **Bäumler A, Fang FC.** Host specificity of bacterial pathogens. *Cold Spring Harb Perspect Med* 2013;3:a010041.
5. **Okoro CK, Barquist L, Connor TR, Harris SR, Clare S *et al.*** Signatures of adaptation in human invasive *Salmonella* Typhimurium ST313 populations from sub-Saharan Africa. *PLoS Negl Trop Dis* 2015;9:e0003611.
6. **Clermont O, Christenson JK, Denamur E, Gordon DM.** The clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ Microbiol Rep* 2013;5:58–65.
7. **Clermont O, Olier M, Hoede C, Diancourt L, Brisse S *et al.*** Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. *Infect Genet Evol* 2011;11:654–662.
8. **von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A *et al.*** Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution. *Nat Genet* 2014;46:1321–1326.
9. **Mainda G, Lupolova N, Sikakwa L, Bessell PR, Muma JB *et al.*** Phylogenomic approaches to determine the zoonotic potential of Shiga toxin-producing *Escherichia coli* (STEC) isolated from Zambian dairy cattle. *Sci Rep* 2016;6:26589.
10. **White AP, Sibley KA, Sibley CD, Wasmuth JD, Schaefer R *et al.*** Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle adaptations but not host specificity. *Appl Environ Microbiol* 2011;77:7620–7632.
11. **Bauchart P, Germon P, Brée A, Oswald E, Hacker J *et al.*** Pathogenomic comparison of human extraintestinal and avian pathogenic *Escherichia coli*-search for factors involved in host specificity or zoonotic potential. *Microb Pathog* 2010;49:105–115.
12. **The HC, Thanh DP, Holt KE, Thomson NR, Baker S.** The genomic signatures of *Shigella* evolution, adaptation and geographical spread. *Nat Rev Microbiol* 2016;14:235–250.
13. **Lupolova N, Dallman TJ, Matthews L, Bono JL, Gally DL.** Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proc Natl Acad Sci USA* 2016;113:11312–11317.
14. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al.*** SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
15. **Seemann T.** Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
16. **Seemann T.** 2017. MLST. GitHub – tseemann/mlst: scan contig files against PubMLST typing schemes. https://github.com/tseemann/mlst [accessed 2017].
17. **Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S *et al.*** Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
18. **Stamatakis A.** RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.

19. **R package e1071, version 1.6-7.** 2015. Functions of the Department of Statistics, Probability Theory Group, TU Wien. https://cran.r-project.org/web/packages/e1071/.
20. **Cortes C, Vapnik V.** Support-vector networks. *Mach Learn* 1995; 20:273–297.
21. **Enterobase [online].** Enterobase. http://enterobase.warwick.ac.uk/ [Accessed 2016–2017].
22. **Parsons BN, Humphrey S, Salisbury AM, Mikoleit J, Hinton JC *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* 2013;7:e2487.
23. **Foodborne Outbreak Tracking and Reporting [Internet].** 2016. Centers for disease control and prevention. http://wwwn.cdc.gov/foodborneoutbreaks/.

Additional file 1

Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*

[2]

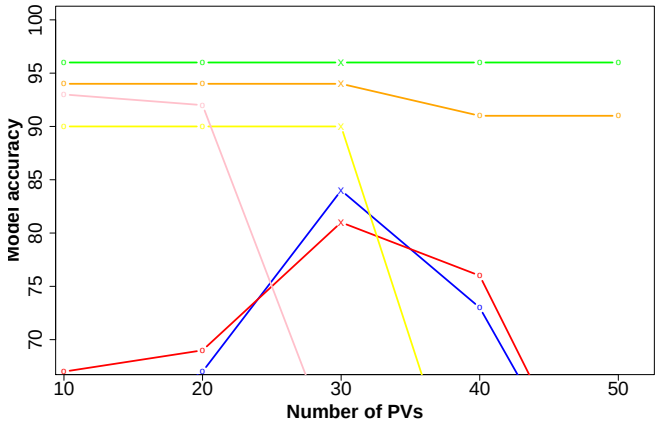


Figure S1: **Model accuracy vs. number of PVs for *E. coli*.** Each point from left to right indicates ΔPV50, ΔPV40, ΔPV30 (shown as crosses, these were chosen for the final model), ΔPV20, ΔPV10. The aim was to find a value that could be used for all the training models within the *E. coli* set, but it is clear that a "one fits all" is not the best strategy for this particular analysis. It is evident that the same threshold as applied to STm (ΔPV30) challenging to use for all *E. coli* sub datasets as in some of them (swine and avian) were too few PVs available. Similar to the *Salmonella* dataset, this analysis indicates that increasing the number of ΔPVs does not always lead to an increase in accuracy of the model.

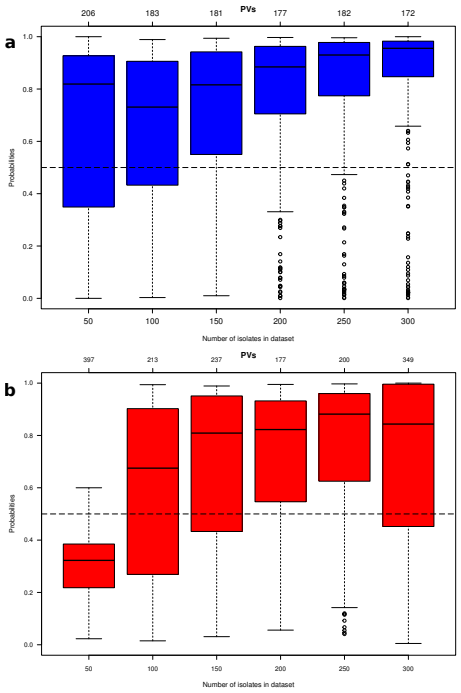


Figure S2: **Influence of dataset size on the number of PVs and prediction accuracy.** **(a)** Boxes represent predictions for gradually increasing number of *S.* Typhimurium human isolates, while the number of bovine isolates is kept constant. **(b)** The same as above with an increasing number of bovine *E. coli* bovine isolates and a constant number of human isolates. Increasing the number of isolates in the dataset mostly improves predictions.

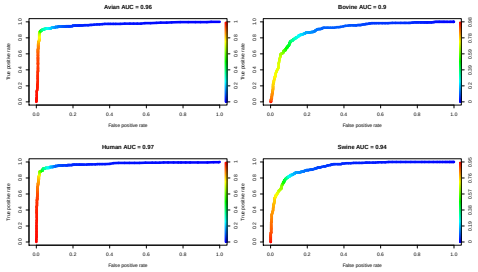


Figure S3: **Performance of SVM models for *S.* Typhimurium isolates.** Area under the curve illustrating performance of four classifiers for each host model for *S.* Typhimurium dataset.

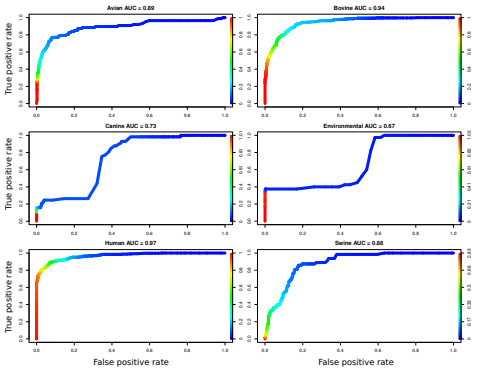


Figure S4: **Performance of SVM models for *E.coli.* isolates.** Area under the curve illustrating performance of six classifiers for each host model for *E. coli* dataset. As expected the best performance achieved for the datasets with highest number of isolates (human and bovine).

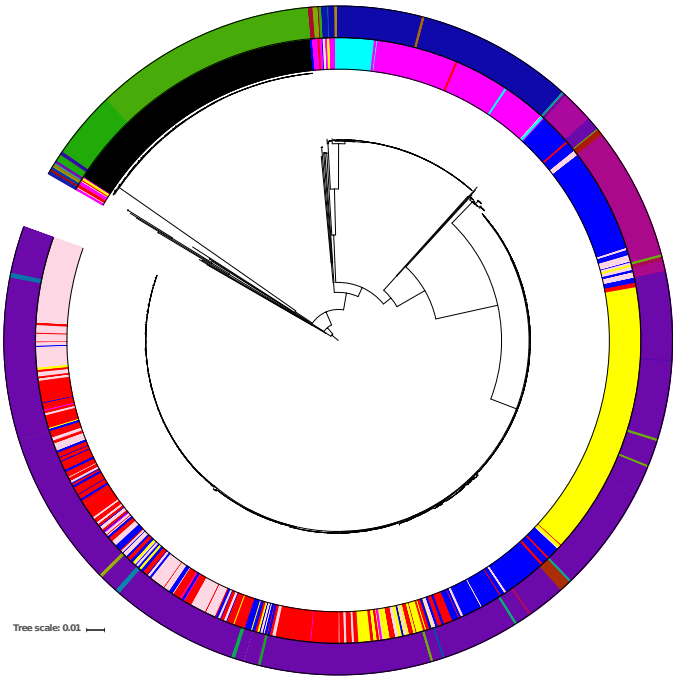


Figure S5: ***S. enterica* core genes tree**. Maximum likelihood core genes tree with host and serovar information shown in the inner circle (blue-human STm; yellow-avian STm; red-bovine STm; pink-porcine Stm; black-S.Typhi; dark pink-bovine *S.* Dublin; cyan-human *S.* Dublin) and MLST Sequence Type information in the outer circle.

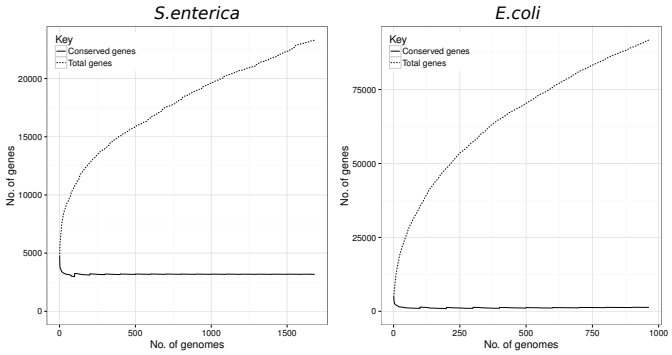


Figure S6: **Pan genome sizes of *S. enterica* and *E. coli*.** The figure illustrate the differences in pan-genome structures for *S. enterica* and *E. coli*. Even though almost only half as many isolates were analysed for *E. coli* (n = 943) compared to *S. enterica* including Typhi and Dublin (n = 1682), *E. coli* had a pan-genome that was 4 times the size of pan- genome of *S. enterica*

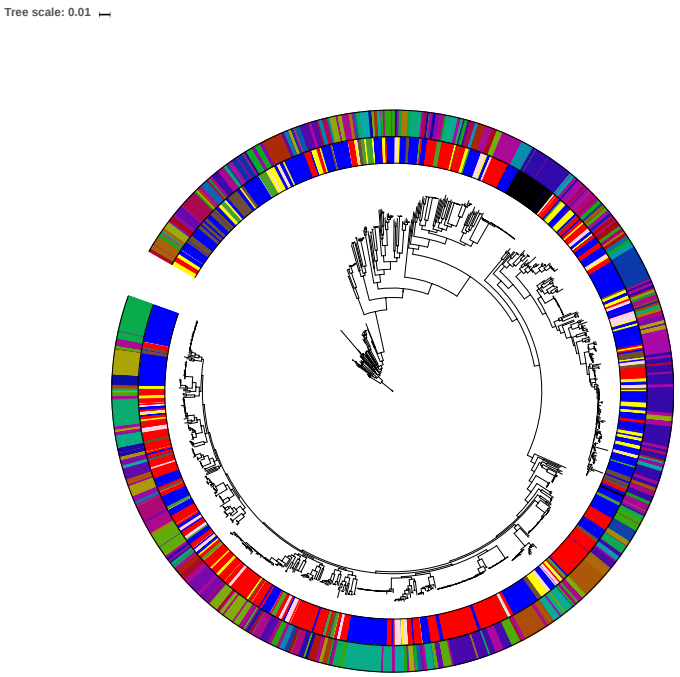


Figure S7: ***E. coli* core genes tree** with host information shown in the inner circle (blue-human; yellow-avian; red-bovine; pink-porcine; green-environmental; brown-canine) and Multi Locus Sequence Type-MLST information shown in the outer circle.

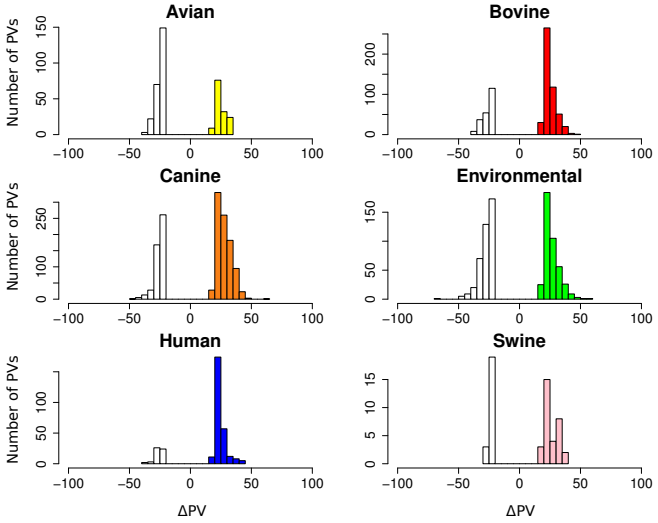


Figure S8: **Distribution of descriptive PVs for *E. coli***. The number of PVs is shown on the Y axis and the ΔPV range on the X axis with positive values indicating increased presence of the PV in the defined host group and negative values meaning increased presence of the PV in the remainder.

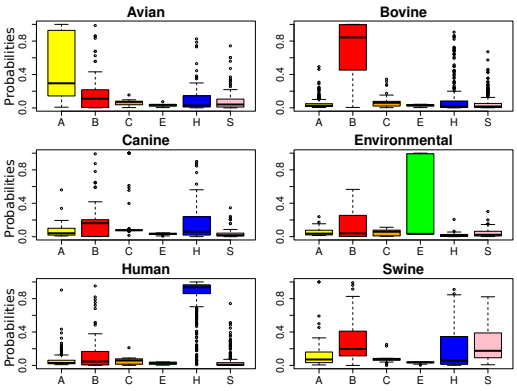


Figure S9: ***E. coli* boxplot predictions.** Distribution of probabilities of *E.coli* isolates plotted as a boxplot for each host. Color scheme: yellow - avian, red - bovine, orange - canine, green - environmental, blue - human, pink - swine.

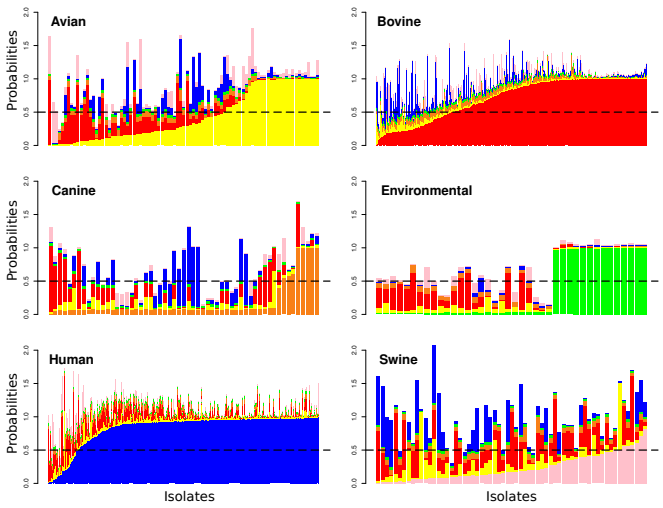


Figure S10: ***E. coli* prediction of host assignment plotted as stacked bar-plots.** As discussed in the main text, the lack of specific assignment for all hosts/environments other than bovine & human may be due to lack of isolate data and so care needs to be taken in interpreting these graphs. It is evident that the environmental group does have a very different structure and indicates a subset of *E.coli* with a strong environment-specific attribution.

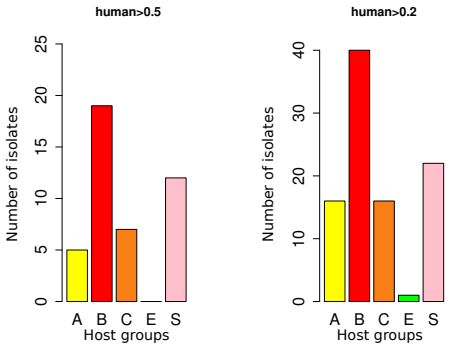


Figure S11: ***E. coli* isolates scored human**. X-axis host groups, Y-axis number of isolates. There was a clear and statistically significant hierarchy working towards content in human isolates (environmental(n=0, 0%), avian(n=5, 6%), bovine(n=19, 6%), canine(n=7, 12%), swine(n=12, 19%), Fisher's Exact Test, p-value = 0.002216. The relative numbers at the p> 0.2 threshold were: environmental(n=1, 2.5%), avian(n=16, 18%), bovine(n=40, 13%), canine(n=16, 28%), swine(n=22, 35%), Fisher's Exact Test: p-value = 1.023e-05.

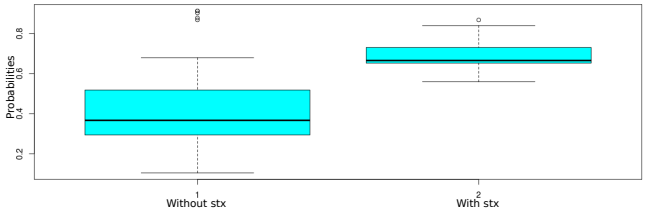


Figure S12: ***E. coli* O157 isolates predictions**. The figure illustrates how 'human isolate' predictions changed when 24 *E. coli* O157 isolates were tested on either all *E. coli* human and bovine isolates (with stx) as the training sets or with stx+ containing isolates removed from these two training sets (without stx).

3.2.3 Conclusions

In this section, the ML classifier, Support Vector Machine was applied to STm to predict the host of isolation. The majority of the isolates was predicted correctly. Despite the success of the prediction, this study highlights a few issues, including: (1) size of the training dataset; (2) the reliability of metadata and biological significance of identified genomic markers; (3) the thresholds that should be set.

It is very important to have enough labelled data to construct a training set in order to obtain reliable predictions. This principle is illustrated in the supplementary materials for my paper from the work in this chapter [114]) for which reducing the dataset to 50 isolates led to relatively meaningless random predictions. Current prediction accuracies are above 80% for the dataset of 300 isolates. It is important to try to increase and diversify the datasets to find out if the accuracy can be improved or whether there is a limit in learning that cannot be overcome even with larger datasets and within a particular model, such as SVM. If the 'error' rate remains the same, it could mean that data is mislabelled or that genetic content assessed here plays only a partial role in host adaptations and therefore other features should be taken into account, such as expression patterns that may be influenced by factors by epigenetic factors such as genome rearrangements and methylation profiles.

One the other hand, the error rate can be indicative of the true biological trends underlying the data, so the generalist and specialist dichotomy of STm is a biological reality and the STm population is not homogeneous. There are strains that are much more adapted to only one host while others have genetic content that could possibly facilitate colonisation in another animal host. Indications of specialist and generalist heterogeneity of STm populations can also be seen in the phylogeny where all host STm host subpopulations have at least one tight host cluster and 20% or more of the isolates are clustered in mixed clusters from multiple hosts.

Another difficulty with this type of study is to position the threshold of 'correct' prediction. In the study, it was placed midway (0.5), however, we don't know how much is enough to be successful in any particular host. It can also be that that threshold would change depending on the host and/or dataset analysed. These settings should be verified in competing experiments. Based on the outcome of grant applications, our future work will include competitive experiments to test the SVM predictions described in this chapter. It is also possible to look at correlations between differential protein variants identified in this study and genes considered important for colonisation of STm in different hosts using a genome-wide method such as TraDIS. We would then have much more confidence in alleles that were identified using the separate approaches and these could then be examined using traditional site-directed mutagenesis techniques and testing in vivo.

The study also indicates that the phylogenetic differences should be taken into account at least in questions such as host adaptation. Even though there was a trend from more bovine scores for *S*. Dublin to more human to *S.* Typhi when tested following training on the STm background it was clear that between-serovars differences are shadowing weaker between-host signals. Another illustration of this is *E. coli* O157 isolates that were all classified as having scores indicative of 'human' isolate content and perhaps indicating a human threat. This was the case whether from a bovine or human source when the testing is carried at the resolution of 'human' vs 'cattle' *E. coli* and delta PV40 cut-off. While at one level this is the case for *E. coli* O157 and supported by epidemiological data, we also know that not all *E. coli* O157 isolates have the same disease threat. O157 isolates appear as a clonal group in the context of the whole *E. coli* population, whether examined by SNP- or pangenome-based phylogenies. However, at a higher resolution, O157 isolates can be quite diverse in terms of both core SNPs and accessory content, which does then begin to identify variation that can influence pathogenic potential and zoonotic threat. In the same way, the next chapter uses finer margins in terms of differential PVs to compare host associations within the *E. coli* O157 'clonal' cluster and shows it is possible to then separate and predict variation in host proclivity for different *E. coli* O157 isolates.

3.3 Pangenome: zoom in. Zoonotic threat of *E. coli* O157

3.3.1 Introduction

In the previous chapter, predictions of the host of isolation were correct for the vast majority of STm and *E. coli* isolates. Interestingly, all 25 *E. coli* serotype O157 isolates were assigned high 'human' score, meaning they were more similar to human clinical isolates than to a wide range of bovine isolates. Therefore, these human scores were an indicator of the higher zoonotic potential for all of these isolates.

Are all of *E. coli* serovar O157 isolates are alike? Clonality of O157 can be seen in the phylogeny 3.1, 3.2), where independently of the source (MLST, core SNPs or WGS) which phylogeny was inferred, *E. coli* O157 isolates always seem to cluster together at the same branch with relatively short branches indicating reduced diversity between these strains. However, when zooming in, there is great diversity which can be noted in between these isolates. Some classifications of O157 serovar are based on phylogeny - there are three lineages (I, I/II, and II), on a phage sensitivity [28] - there are more then 80 phagetypes for O157, on pulsed-field gel electrophoresis (PFGE) there are above 20 profiles, and each of these can be devided further [25]. These three

classification schemes demonstrate how diverse O157 serovar can be.

It seems that the previous study while correctly identifying O157 as a zoonotic threat, was lacking resolution, and therefore each isolate from the O157 population was seen as similar to each other, almost undistinguishable. All of the O157 isolates achieved similar, around 70% probabilities. Nevertheless, when the biggest differences that account for diversity of all *E. coli* are removed and phylogeny inferred only from O157 isolates, it becomes clear that the strains in that subset are quite diverse: (from 185 isolates used for the analysis described below not a single strain shared the same SNPs pattern or had the same proteins presence and absence sequence.

Therefore, this section describes how the same approach (ML, SVM) was applied to predict host of isolation of only *E. coli* serotype O157:H7 isolates. *E. coli* O157 is in the category B NIH global food pathogen list [143]. Pathogens on this category are moderately easy to disseminate, they result in moderate morbidity rates and low mortality rates and require specific enhancements for diagnostic capacity and enhanced disease surveillance. *E. coli* O157 serotype usually has various prophage incorporated in their genome, some of which is shiga-toxin producing, thus these strains can cause further health complications and be lethal in some cases. Moreover, due to dynamic gene exchange

some combination of virulent factors can have summative effect as in the example of German seed sprout outbreak where already pathogenic and antibiotic resistance EAEC strains acquired stx2 from shiga-producing *E. coli* strains leading to a devastating effect in the human population.

3.3.2 Results: Support vector machine applied to predict the zoonotic potential of E. coli O157 cattle isolates

Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates

Nadejda Lupolova[a], Timothy J. Dallman[b], Louise Matthews[c], James L. Bono[d], and David L. Gally[a,1]

[a]Division of Immunity and Infection, The Roslin Institute and The Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian EH25 9RG, United Kingdom; [b]Public Health England, National Infection Service, London NW9 5EQ, United Kingdom; [c]Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom; and [d]US Meat Animal Research Center, Agricultural Research Service, United States Department of Agriculture, Clay Center, NE 68933

Edited by Roy Curtiss III, University of Florida, Gainesville, FL, and approved August 2, 2016 (received for review May 2, 2016)

Sequence analyses of pathogen genomes facilitate the tracking of disease outbreaks and allow relationships between strains to be reconstructed and virulence factors to be identified. However, these methods are generally used after an outbreak has happened. Here, we show that support vector machine analysis of bovine *E. coli* O157 isolate sequences can be applied to predict their zoonotic potential, identifying cattle strains more likely to be a serious threat to human health. Notably, only a minor subset (less than 10%) of bovine *E. coli* O157 isolates analyzed in our datasets were predicted to have the potential to cause human disease; this is despite the fact that the majority are within previously defined pathogenic lineages I or I/II and encode key virulence factors. The predictive capacity was retained when tested across datasets. The major differences between human and bovine *E. coli* O157 isolates were due to the relative abundances of hundreds of predicted prophage proteins. This finding has profound implications for public health management of disease because interventions in cattle, such a vaccination, can be targeted at herds carrying strains of high zoonotic potential. Machine-learning approaches should be applied broadly to further our understanding of pathogen biology.

machine learning | zoonosis | Shiga toxin | *E. coli* | cattle

For important global bacterial zoonoses such as *Salmonella*, enterohemorrhagic *Escherichia coli* (EHEC), and *Campylobacter*, tracking of disease outbreaks and identification of infection source are critical to limiting further disease. Whole-genome sequencing (WGS) has provided a revolution in our capacity to identify and trace outbreaks that would have been virtually impossible with more traditional techniques such as phage typing and pulsed-field gel electrophoresis (1, 2). Currently, most analyses rely on extraction of a core "shared" genome and isolate relationships are deduced based on SNPs in this core; conversely, accessory genome information is largely ignored due to its variability, although a number of approaches have recently been applied to interrogate pan-genome data (3).

EHEC infections, in particular by serogroups O157 and O26 (4), have emerged as a serious threat to human health in the last 30 y, driven by the integration of bacteriophages encoding Shiga toxin (Stx) into the genomes of specific *E. coli* strain backgrounds. Strains encoding Stx subtype 2a and a type 3 secretion system are often associated with the most severe human infections, which can lead to bloody diarrhea (hemorrhagic colitis) and kidney damage. Stx kills capillary endothelial cells and the host's attempt to repair this damage can result in red blood cell hemolysis in capillaries known as hemolytic uremic syndrome, which can be fatal (5–7). There has been extensive work to determine which strains in ruminants, in particular cattle, represent the most serious threat to human health (6, 8, 9). This led to the definition of lineages and clades for which lineage I or lineage I/II are more likely to be associated with human disease, whereas lineage II strains are more restricted to cattle (10, 11). Within these lineages certain clades predominate, so clade 8 within lineage I/II has been associated in the United States with more serious disease in humans (12). In the United Kingdom, a recent WGS analysis of over 1,000 EHEC O157 human and cattle isolates was used to determine their phylogeny based on core genome SNP analysis (13). The most serious disease in the United Kingdom is associated with lineage I strains and a specific phage type (PT) designated PT21/28; phage typing of UK strains is based on susceptibility testing with a collection of diagnostic bacteriophages (14). The United Kingdom has a high incidence of serious EHEC O157 infections, and the emergence of these infections in the 1990s coincided with the acquisition of the Stx 2a subtype into UK cattle strains already encoding a Stx2c subtype (13, 15).

Current core genome analysis of EHEC strains indicates complete mixing of human and bovine EHEC O157 isolates (Fig. 1*A* and Fig. S1). This fits with the concept that the majority of cattle strains within particular lineages and encoding Stx 2a are a serious threat to human health. In the present study, we aimed to determine whether a pan-genome analysis of EHEC O157 strains could distinguish between human isolates and isolates from cattle. In particular, we wanted to test whether machine-learning approaches such as support vector machine (SVM) (16) could be used to discriminate a subset of bovine strains that might represent a threat to human health and would allow more targeted interventions in cattle. SVM has been applied in many areas of bioinformatics, including prediction of protein function, prediction of transcription initiation site, and classification of gene expression data as well as cancer prediction and prognosis (17, 18).

Results and Discussion

UK Dataset. We initially analyzed an extensive UK dataset that consisted of WGS for 185 *E. coli* O157 strains isolated from

Significance

Zoonotic infections with enterohemorrhagic *Escherchia coli* O157 have emerged as a serious threat to human health. Conventional sequence-based analyses indicate that most human infections originate from particular pathogenic lineages. In this study, we apply a machine-learning approach to complex pangenome information and predict the human infection potential of cattle *E. coli* O157 isolates. We demonstrate that only a small subset of bovine strains is likely to cause human disease, even within previously defined pathogenic lineages. The approach was tested across isolates from the United Kingdom and United States and verified with food and cattle isolates from outbreak investigations. This finding has important implications for targeting of control strategies in herds.

Author contributions: N.L., T.J.D., L.M., and D.L.G. designed research; N.L. performed research; N.L., T.J.D., and J.L.B. contributed new reagents/analytic tools; N.L. and D.L.G. analyzed data; and N.L., T.J.D., J.L.B., and D.L.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: dgally@ed.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1606567113/-/DCSupplemental.

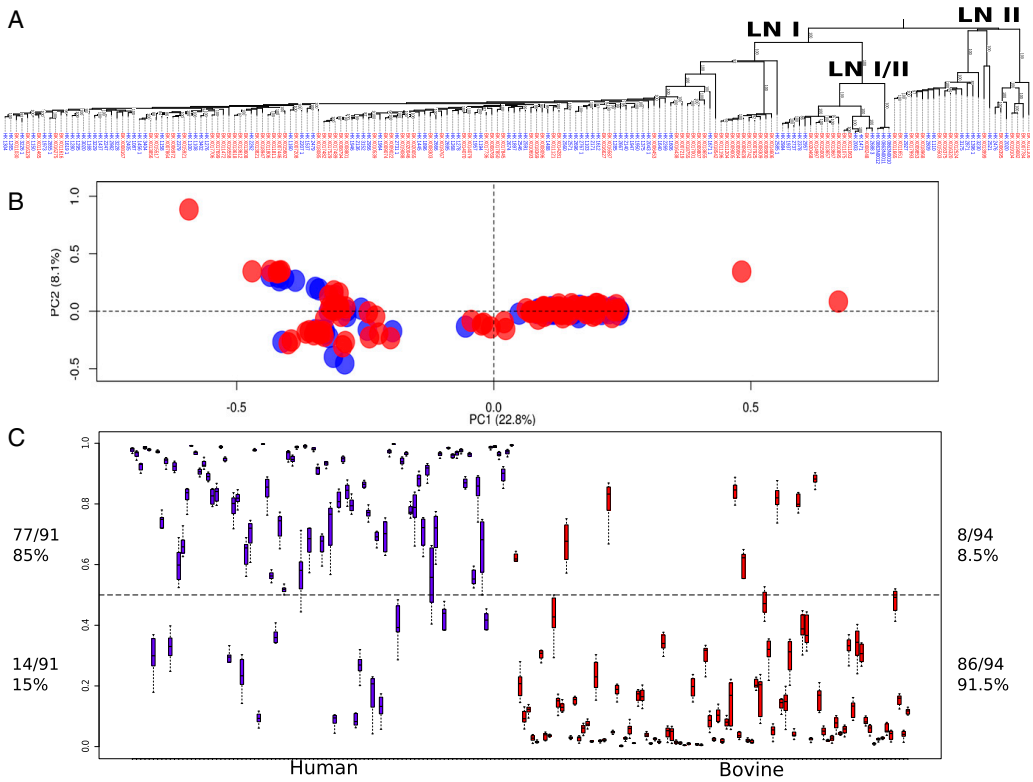


Fig. 1. UK Shiga toxin-producing *E. coli* (STEC) O157 dataset analysis. (*A*) Core SNP phylogenetic tree. The three main lineages (LN) are shown. The majority of the UK isolates are in lineage I (LN I) with bovine (red) and human (blue) isolates interspersed across the tree. Bootstrap values shown on branches. (*B*) MMDS plot with each isolate represented by a circle. The denser cluster on the right-hand side is composed primarily of LN I isolates with equivalent numbers of human isolates overlaid by bovine isolates. (*C*) SVM probability plot based on repeated testing of isolates in the different subsets. The probability of each isolate belonging to the human or bovine group was calculated over random repeated samples; median values are shown with interquartile ranges. The predicted "host" of the isolate is based on whether the mean probability is below 0.5 (bovine) or above (human). The percentages of isolates assigned to each host by the model are shown at the sides of the graph.

human patients in the United Kingdom ($n = 91$) and cattle ($n = 94$). The 185 strains share 4,737,622 core positions, which is equivalent to 85% of the reference *E. coli* Sakai strain genome (19). A maximum likelihood phylogenetic tree based on these positions clearly splits into distinctive branches, even within this relatively clonal serotype (Fig. 1*A*). The pattern for the UK O157 phylogenetic tree is consistent with previous studies (11, 13, 20) and represents a typical split for UK strains based on lineages: lineage I ($n = 140$, 70 bovine), II ($n = 25$, 15 bovine), and I/II ($n = 17$, 9 bovine). The average number of SNPs within two sequences of the same lineage was 1,859, 379, and 2,190 for lineages I, I/II, and II, respectively. The vast majority of the lineage I sequences were PT21/28 (101 out of 140) and the second most prevalent was PT32 (24 out of 140). The dominant PT in lineage II was PT8 (15 out of 25) and in lineage I/II was PT2 (14 out of 17). Based on phylogenetic analysis of core SNPs, it was not possible to detect any evidence of clustering by human or bovine host (Fig. 1*A*).

Determination of the accessory genome using the Roary pan-genome pipeline indicated that among 185 UK isolates there were 14,636 protein clusters assigned, based on 95% amino acid sequence similarity. Core proteins present in more than 95% of the sequences generated 4,369 clusters; 979 clusters originated from proteins predicted in 15–95% of sequences, leaving a high number of rare clusters (9,288) that were present in less than 15% of isolates. The majority of all protein clusters (10,653) were annotated (i.e., were similar to already-annotated proteins from a public database) and 3,983 were hypothetical. There were only 5,485 unique protein names across all of the genomes, and 3,807 of these produced single copy clusters. Due to these rules of cluster assignments, many homologous proteins generated multiple clusters. We have termed these protein variant (PV) clusters. An exceptionally high number of PVs were produced by phage-related proteins, confirming that phage sequences are highly variable (21).

An accepted way to analyze complex pan-genome data is to apply metric multidimensional scaling (MMDS) with different methods of matrix distance calculations. In the present study, methods of distance calculation had little effect on the final MMDS plots, and thus all MMDS plots presented in this paper are based on simple dissimilarity calculations (Fig. 1*B*). Dense clusters on the MMDS plot were highly correlated to the lineages shown on the phylogenetic tree. Thus, further clustering of UK isolates by *k*-means resulted in two clusters: one with all isolates having 100% support and originating from lineage I (128 isolates

MICROBIOLOGY

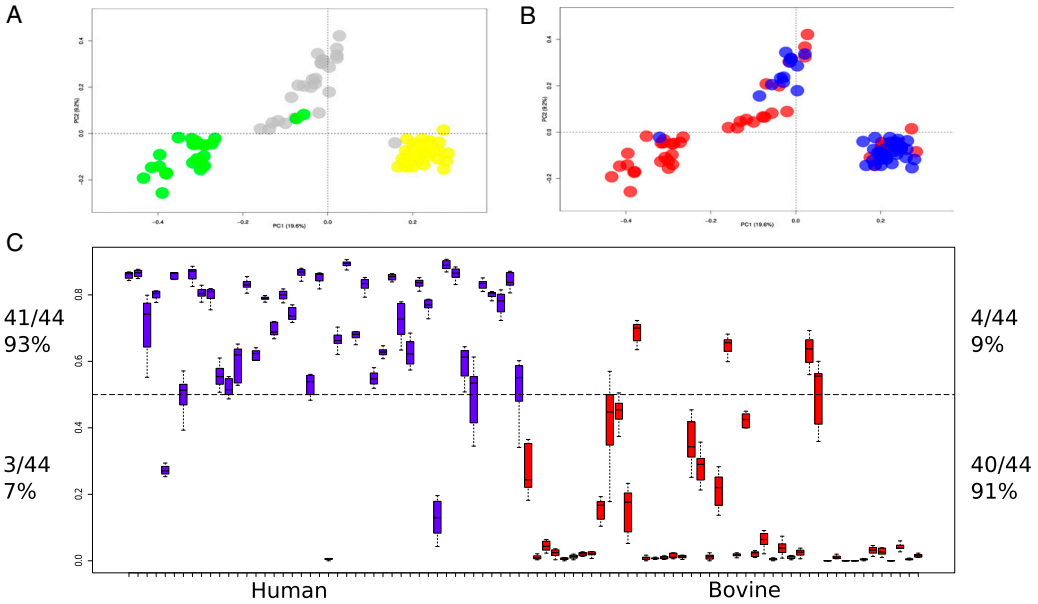


Fig. 2. US STEC O157 dataset analysis. MMDS analysis of US pan-genome dataset with each isolate represented by a circle. (*A*) MMDS clustering with isolates colored by lineage: lineage I in yellow, lineage I/II in gray, and lineage II in green. (*B*) MMDS clustering with isolates colored by host: red, bovine isolates and blue, human isolates. (*C*) SVM probability plot based on repeated testing of isolates in the different subsets. The probability of each isolate belonging to the human or bovine group was calculated over random repeated samples with median values and interquartile ranges shown. Red shading for bovine isolates; blue shading for human isolates. The predicted "host" of the isolate is based on whether the mean probability is below 0.5 (bovine) or above (human). The percentages of isolates assigned to each host by the model are shown at the sides of the graph.

out of all 140 lineage I isolates) and the second with isolates primarily from lineages II and I/II (support higher than 90%) and with only 11 isolates from lineage I (support between 85–90%). All isolates in the second cluster that belonged to lineage I were PT32 (Fig. S2). Overall, MMDS methods provided results similar to the phylogenetic analysis, namely, separation of lineage I and little capacity to distinguish between human and bovine isolates.

Machine learning methods have been routinely applied to investigate complex data in several areas of science, although, until now, it has not been used to analyze bacterial genomic data to predict phenotype from the genotype. Therefore we built an SVM classifier trained on *E. coli* isolates with known isolation host (human or cattle) and tested whether the classifier could predict the likely host origin (human/bovine) of isolates from their PV profile. To choose the features for the model, the proportions of each PV present in each host group were calculated separately. There were a total of 10,878 clusters with a different proportion of PVs between the two hosts (Table S1). To reduce the number of features introduced into our model, while preserving accuracy, we used only PVs with a subtractive difference between the two hosts of >10 (n = 638) and have defined the discriminatory PVs at >20 (n = 82) in Tables S2 and S3. The probability of each isolate being assigned to the human or bovine group was then calculated by random repeated sampling and the resulting probabilities plotted in Fig. 1*C*. Overall, using a probability of 0.5 as the separation value, 85% of human and 91% of bovine isolates were assigned in accordance with the host from which they were isolated, and the majority with high probabilities. This shows that it is possible to differentiate these isolates based on the isolation host, indicating that host-specific information for *E. coli* O157 can be derived from the sequence data alone. Because ruminants, in particular cattle, are a primary reservoir for EHEC O157 strains, there was an a priori assumption that it may not be possible to assign isolates to the two host groups because the majority of human isolates are likely to originate from cattle. However, this was not the case, and it is an important observation that a minor subset of isolates originating from cattle were classified into the human group (Fig. 1*C*). These same bovine isolates were persistently called as human, meaning that the model does find features in these isolates that make them more similar to those from the human population than from cattle. This finding indicates that not all bovine isolates have the same zoonotic potential; in fact, the majority of bovine *E. coli* O157 isolates were not predicted to be associated with human disease.

The majority of either bovine or human isolates did not change their assignment probabilities with multiple subtesting (the majority close to 0 or 1) and strains called distinct from their isolation host were called so consistently. Midrange isolates had more variable assignment probabilities (Fig. 1*C*) and this may indicate genomes with both human- and bovine-specific features. In addition, the bovine isolates called as "human" and the isolates called in the reverse direction cannot be explained by available metadata including lineage and PT; for example, the bovine isolates represent a mixed group of PTs: PT21/28 (n = 4) and one of each PT 31, 32, 33, and 49. Six of these isolates possessed *stx2a/2c*, one 2a, and one was negative for *stx*. We note that MMDS analysis of this differential PV subset did not separate strains by isolation host with clustering still tied to lineages and SNP core phylogeny (Fig. S3).

SVM models can be analyzed for accuracy and prediction capacity (Fig. S4), with accuracy calculations based on the level of "incorrect" assignments. However, there is an expectation that

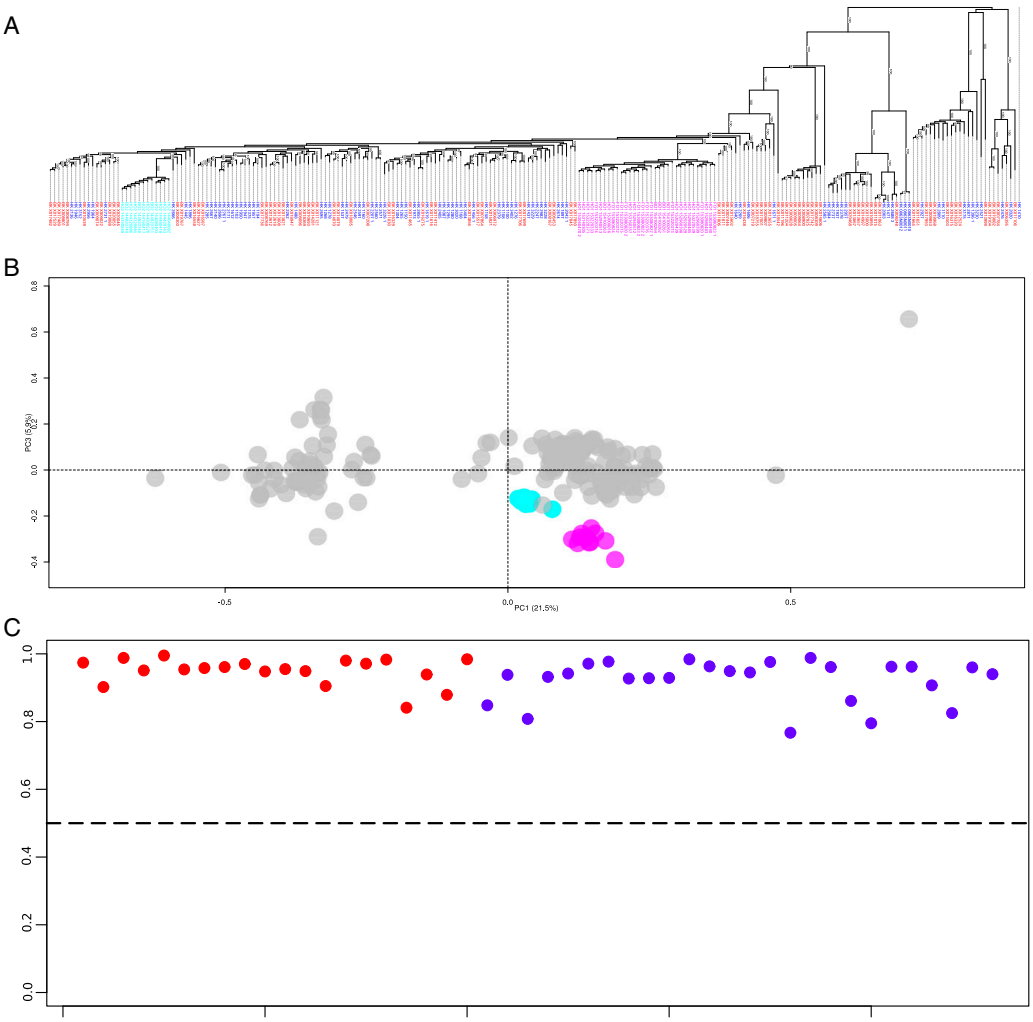


Fig. 3. Analysis of STEC O157 outbreak isolates. (*A*) Core SNP phylogenetic analysis of bovine UK (red) and human UK (blue) isolates and the two EHEC O157 outbreaks (magenta and cyan) showing that isolates from both outbreaks fall within lineage I and cluster tightly. (*B*) MMDS analysis of the two outbreaks (magenta and cyan) relative to the UK isolate subset (gray). The outbreak isolates form distinctive clusters although associated with lineage I. (*C*) SVM probability plot of each outbreak isolate without repeated sampling. Isolates from cattle, milk, or hamburger meat from both outbreaks are in red, and isolates from human hosts from both outbreaks are colored in blue. All isolates from both outbreaks (milk, hamburger, cattle, and human) were predicted to be "human."

our two host groups are not mutually exclusive, in other words that some isolates can colonize both hosts and therefore will contribute to model "inaccuracy." A logical extension of this point is that if the model were 100% accurate, then there would be no strain cross-over between the groups, indicating complete host adaptation or a very rare subset of cattle isolates with zoonotic potential. Therefore, accuracy estimations can reflect the underlying biology of the isolates and should be considered minimum estimates.

An important potential concern for data analysis by SVM is overfitting, for which the model is not using biologically relevant information to separate the groups. There are a number of ways to test for this; the most rigorous is to train the model on one dataset and then test it on a completely separate dataset. We apply this model successfully in the next section using isolate sequences from the United States. In addition, for our UK dataset we also tested whether we could train the model on two randomly labeled sets (containing both human and bovine isolates) and determined whether strains from these random groups could then be correctly assigned back to these groups. This was carried out in two ways. The first involved subsampling from our groups (random or bovine/human) with differential PVs (>10)

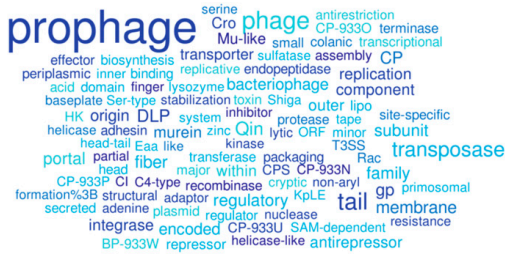


Fig. 4. Word cloud depiction of annotated PVs that were in higher proportions of human strains compared with cattle strains based on analysis of the UK dataset. Many of the PVs from human and cattle isolates are of bacteriophage origin.

determined for these subsamples. Forty isolates distinct to the training sets but within the assigned groups were then tested. This subsampling, PV determination, and testing was repeated 20 times. As expected, the isolates from the random groups had a normal distribution of probabilities reflecting their random assignment (Fig. S5*C*); in contrast, the bovine/human isolates had a different distribution with the majority having high probabilities of host assignment (i.e., 1 and 0) (Fig. S5*A*). Moreover, the repeated subsampling of the bovine/human groups yielded a significantly higher mean number of PVs (637.1, SD = 63, SE = 14) than subsampling the random groups (168, SD = 70, SE = 15) and individual PVs were more likely to be resampled from the host-related groups compared with the random groups (Fig. S5 *B* and *D*). Taken together, there is significantly more genetic information relevant to the bovine/human groupings compared with random groups. Second, when the PV selection used to assign the complete human and bovine groups (for Fig. 1*A*) was applied to randomly selected groups, the majority of probabilities were around 0.5 (Fig. S6). Both approaches give us confidence that our capacity to differentiate bovine and cattle isolates is not a result of chance and overfitting from a complex dataset.

US Dataset. We obtained 44 human and 44 bovine isolate sequences from the United States (Dataset S1, US isolates). The isolate distribution based on continental differences is apparent in the phylogenetic tree (Fig. S1). The US isolates occupy separate branches from UK strains, even within the same lineages, and show anticipated bias in host designation with lineage. There were 30 human and 8 bovine strains in lineage I, 13 human and 12 bovine strains in lineage I/II, and 1 human and 23 bovine strains in lineage II. Also, US strains share between them fewer "core" positions, covering only 79% of the Sakai genome. MMDS analysis showed results similar to the UK dataset: The isolates were separated predominantly by lineage (Fig. 2 *A* and *B*). Before testing our UK isolate model across to this dataset, we first built an SVM classifier based only on the US dataset, and the results were similar to UK isolates: The model accuracy was 91.3%, with 92% of the strains assigned correctly according to the host the isolate was from. Four out of 44 bovine isolates were called "human." Thus, even though the US isolates seemed to be distinct in terms of the human/bovine split on the phylogenetic tree and in an MMDS plot, the SVM analysis identified a small group of bovine strains (again just under 10%) that possessed genome features that can be found in the majority of disease-associated human isolates and therefore possibly have greater zoonotic potential. Also, as in the UK dataset, the predicted probabilities of most isolates had little variation, and therefore potentially contain strictly bovine or human features, whereas a smaller group exhibited much greater variability.

In the US dataset, there was a total of 10,590 PVs that varied between the two hosts, which is comparable with the UK dataset (10,878 PVs). However, the US dataset contained a much higher number of PVs with larger differences between hosts (Table S1). However, there was a relatively small overlap of discriminatory PVs (n = 197) between the two datasets. The US dataset was tested with the model trained on the UK dataset based on these 197 PVs. Despite the small number of overlapping PVs between the datasets, the model accuracy was 78%, with 38 out of 44 bovine isolates and 31 out of 44 human isolates assigned according to the host from which they were isolated. When trained on the US dataset and tested on the UK dataset, 86 out of 94 bovine isolates and 78 out of 91 human isolates were assigned to the isolation host. Therefore, even though there are considerable differences between the two datasets and a significant amount of continent-specific information has to be excluded, the same model can be applied, although with less accuracy, to a distinct dataset.

Despite the continental divergence between the UK and US isolates, we tried combining the two datasets for testing. Based on an MMDS analysis, human US isolates that belong to lineage I form a separate cluster far apart from other lineage I isolates (Fig. S7 *A* and *B*); however, the overall tendency is similar for the UK or US datasets alone, with lineage I isolates separated from all of the others. When the proportion of PVs was calculated for the sets combined, some descriptive features from one dataset become neutralized by the other dataset. The SVM model based on the combined dataset (Fig. S7) achieved 82% of model accuracy and predicted 84% of human isolates and 83% of bovine isolates correctly according to their isolation host. It was reassuring that among the bovine isolates that were called "human" were all of these that already were assigned "human" from the single-country models. The same applies to human isolates that were called "bovine." However, the subset of bovine strains called "human" in a mixed model increased potentially due to differences in PVs that define human/bovine separation in the United Kingdom and United States.

UK Outbreaks. Two main hypotheses can be generated from these findings, although they are not mutually exclusive: (*i*) Isolates associated with human infections represent a very specific subset of bovine isolates, in which case the majority of bovine *E. coli* O157 isolates that we have sequenced may be unlikely to cause human disease, and (*ii*) isolates change their genome content following transition into another host, so potentially they acquire phage/plasmid regions in the human host although they originate from cattle; the reverse transfer and adaptation is also possible. To address this question we analyzed EHEC O157 strains from two outbreak investigations. One outbreak was associated with hamburger consumption where both the meat and animal sources were identified (human n = 17, cattle n = 5, and hamburger n = 12). Another was associated with milk consumption (human n = 9 and milk n = 3). As anticipated, the individual outbreak strains closely relate to each other and in the phylogenetic tree formed individual tight clusters for each outbreak (Fig. S1). On an MMDS plot they clustered slightly separately from all other UK strains but in close proximity to lineage I PT 21/28 strains to which they belong (Fig. 3 *A* and *B*).

We trained the model on the all-UK dataset, excluding the outbreak isolates, and tested it on the outbreak isolates. From both outbreaks the "bovine" isolates (from milk, hamburger meat, and cattle) were classified as "human," with probabilities higher than 0.75 for any isolate (Fig. 3*C*). This supports the first hypothesis that the threat to human health originates primarily from a minor subset of strains and that the majority of bovine strains from both our UK and US datasets, despite their core SNP association and virulence gene content, are unlikely to be associated with disease in humans.

Descriptive Proteins. To assess what level of differences can be found at the core SNP versus PV level we selected four "pairs" of isolates that lie in close proximity to each other on the final branches of the phylogenetic tree but were isolated from different hosts and were predicted by the SVM model to be

associated with those hosts. These pairs had from 9 to 26 SNPs between them whereas the number of unique PVs ranged from 137 to 364, and the relative number of unique PVs between the pairs increased in line with the number of SNPs between the pairs (Table S4). This indicated that these PVs were being lost or acquired over relatively recent evolutionary time because the core mutation rate of *E. coli* has been estimated to be two to three SNPs per year (22).

We then summarized the differential PVs across the UK dataset based on their annotations, and for the complete UK dataset with ΔPV > 10 there were 292 PVs that had higher proportions in human compared with bovine isolates (summarized in Fig. 4; "hypothetical proteins" were not included in the figure). By comparison, 343 PVs (20% more) were present in higher proportions in the bovine isolates compared with the human. The main annotated proteins in both groups were similar and were predominately prophage-related proteins. Variation in prophage content therefore underpins the human/bovine classification demonstrated in this study. This accords with expectations about *E. coli* strain evolution being driven by prophage acquisition, rearrangement, and loss. Different prophage annotations do appear depending on the host (i.e., rac prophage with 3% for bovine isolates and dlp12 prophage with 3% for human), although work is now required to examine the biological impact of differential PVs and how these alter the potential of an isolate to infect or cause disease in humans.

Conclusions

This study has applied machine learning to predict the zoonotic potential of bacterial isolates. The analysis demonstrates that in the highly clonal *E. coli* O157 serogroup, host-specific information can be inferred from WGS analysis. Moreover, using an SVM classifier it was possible to generate a probability of host association that indicated that only a minor (<10%) subset of bovine strains were likely to have an impact on human health. In fact, none of the cattle isolates (apart from outbreak trace-back isolates) achieved very high human association probabilities (>0.9), potentially indicating that those posing a serious zoonotic threat are very rare. This finding has implications for public health management of this disease because it means that such strains can now potentially be identified in the ruminant reservoir and, if these are the exception, then targeted control strategies including vaccination or even eradication become a more realistic option to protect human health. The specific prophages that encode the differential PVs now need to be identified to progress our understanding of this zoonosis. A subset of isolates from humans were called as "bovine," and currently we do not know whether they differed in their disease severity, e.g., whether isolates from humans that had high bovine probabilities were more likely to be associated with asymptomatic infections (23). In summary, we consider that machine-learning approaches have tremendous potential to interrogate complex genome information for which specific attributes of the organism, such as disease or isolation host, are known.

Materials and Methods

UK and US datasets were previously studied (UK dataset in ref. 13 and US dataset in ref. 24). Illumina short read sequences were assembled with SPAdes (25) and annotated with Prokka (26). Maximum likelihood (ML) core SNPs trees were constructed with RAxML (27). MMDS was performed as described in ref. 28. Pan-genomes were constructed using Roary (29); the threshold was set to 95% of sequence similarity at the amino-acid level. A classifier based on an SVM algorithm was built using R package e1071 (30). The model was tuned and cost and gamma parameters were adjusted (f.ex to gamma = 1e-04 and cost = 100 for the UK dataset). No review board approval was required for the experiments described in this manuscript. Full details of methods are provided in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We would like to acknowledge the value of human and bovine *E. coli* O157 sequence data available from previous published studies, especially work from the Wellcome Trust IPRAVE consortium, Public Health England, and the Scottish *E. coli* reference laboratory. This work was supported by Food Standards Scotland and the Food Standards Agency Grant FS101055 (to D.L.G., T.J.D., and L.M.), which has allowed the continuation of significant EHEC O157 research in the UK. This research was also supported by a University of Edinburgh studentship (N.L.) and core Biotechnology and Biological Sciences Research Council strategic programme Grant BB/J004227/1 (to D.L.G.). T.J.D. was funded by the National Institute for Health Research Health Protection Research Unit in Gastrointestinal Infections at the University of Liverpool in partnership with Public Health England, University of East Anglia, University of Oxford, and the Institute of Food Research.

1. Quick J, et al. (2015) Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol* 16:114.
2. He M, et al. (2013) Emergence and global spread of epidemic healthcare-associated Clostridium difficile. *Nat Genet* 45(1):109–113.
3. Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. *Curr Opin Microbiol* 23:148–154.
4. Pearce MC, et al. (2006) Prevalence and virulence factors of Escherichia coli serogroups O26, O103, O111, and O145 shed by cattle in Scotland. *Appl Environ Microbiol* 72(1):653–659.
5. Gunzer F, et al. (1992) Molecular detection of sorbitol-fermenting Escherichia coli O157 in patients with hemolytic-uremic syndrome. *J Clin Microbiol* 30(7):1807–1810.
6. Griffin PM, Tauxe RV (1991) The epidemiology of infections caused by Escherichia coli O157:H7, other enterohemorrhagic E. coli, and the associated hemolytic uremic syndrome. *Epidemiol Rev* 13:60–98.
7. Mead PS, Griffin PM (1998) Escherichia coli O157:H7. *Lancet* 352(9135):1207–1212.
8. Wells JG, et al. (1991) Isolation of Escherichia coli serotype O157:H7 and other Shiga-like-toxin-producing E. coli from dairy cattle. *J Clin Microbiol* 29(5):985–989.
9. Borczyk AA, Karmali MA, Lior H, Duncan LM (1987) Bovine reservoir for verotoxin-producing Escherichia coli O157:H7. *Lancet* 1(8524):98.
10. Kim J, Nietfeldt J, Benson AK (1999) Octamer-based genome scanning distinguishes a unique subpopulation of Escherichia coli O157:H7 strains in cattle. *Proc Natl Acad Sci USA* 96(23):13288–13293.
11. Zhang Y, et al. (2007) Genome evolution in major Escherichia coli O157:H7 lineages. *BMC Genomics* 8:121.
12. Manning SD, et al. (2008) Variation in virulence among clades of Escherichia coli O157:H7 associated with disease outbreaks. *Proc Natl Acad Sci USA* 105(12):4868–4873.
13. Dallman TJ, et al. (2015) Applying phylogenomics to understand the emergence of Shiga-toxin-producing Escherichia coli O157:H7 strains causing severe human disease in the UK. *Microbial Genomics*, 10.1099/mgen.0.000029.
14. Pearce MC, et al. (2009) Temporal and spatial patterns of bovine Escherichia coli O157 prevalence and comparison of temporal changes in the patterns of phage types associated with bovine shedding and human E. coli O157 cases in Scotland between 1998-2000 and 2002-2004. *BMC Microbiol* 9:276.
15. Dowd SE, Williams JB (2008) Comparison of Shiga-like toxin II expression between two genetically diverse lineages of Escherichia coli O157:H7. *J Food Prot* 71(8):1673–1678.
16. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297.
17. Yang ZR (2004) Biological applications of support vector machines. *Brief Bioinform* 5(4):328–338.
18. Cruz JA, Wishart DS (2007) Applications of machine learning in cancer prediction and prognosis. *Cancer Inform* 2:59–77.
19. Hayashi T, et al. (2001) Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 8(1):11–22.
20. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA (2011) Genomic anatomy of Escherichia coli O157:H7 outbreaks. *Proc Natl Acad Sci USA* 108(50):20142–20147.
21. Ohnishi M, Kurokawa K, Hayashi T (2001) Diversification of Escherichia coli genomes: Are bacteriophages the major contributors? *Trends Microbiol* 9(10):481–485.
22. Lee H, Popodi E, Tang H, Foster PL (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc Natl Acad Sci USA* 109(41):E2774–E2783.
23. Silvestro L, et al. (2004) Asymptomatic carriage of verocytotoxin-producing Escherichia coli O157 in farm workers in Northern Italy. *Epidemiol Infect* 132(5):915–919.
24. Norman KN, Strockbine NA, Bono JL (2012) Association of nucleotide polymorphisms within the O-antigen gene cluster of Escherichia coli O26, O45, O103, O111, O121, and O145 with serogroups and genetic subtypes. *Appl Environ Microbiol* 78(18):6689–6703.
25. Bankevich A, et al. (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comp Biol* 19(5):455–477.
26. Seemann T (2014) Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068–2069.
27. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
28. Abdi H (2007) Metric multidimensional scaling (MDS): Analyzing distance matrices multidimensional scaling: Eigen-analysis of a distance matrix. *Encyclopedia of Measurement and Statistics*, ed Salkind NJ (Sage, Thousand Oaks, CA), Vol 2, pp 598–605.
29. Page AJ, et al. (2015) Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31(22):3691–3693.
30. Meyer D, et al. (2015) Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package e1071, version 1.6-7 (Vienna University of Technology, Vienna).

Supporting Information

Lupolova et al. 10.1073/pnas.1606567113

SI Materials and Methods

UK and US datasets were previously studied. The isolates used in this study, their isolation host, and lineage can be found in Dataset S1. For the convenience of this work prefixes were added to the original isolate names. So, UK human isolates have the prefix HK; UK bovine, BK; human US, HS; bovine US, BS; milk outbreak isolates from a human source are HO1 and from milk are BO1. Food outbreak isolates from humans are HO2 and from cattle are BO2.

To construct core SNPs trees short reads were aligned to a reference *E. coli* O157:H7 str. Sakai (RefSeq assembly accession no. GCF_000008865). Core positions from resulted consensus sequences were used to construct an ML tree with RAxML under a GAMMA model of heterogeneity with 500 bootstrap replicates.

Distance matrices were composed based on pairwise differences of the sequences calculated as number of changes divided by length of sequence and also as euclidean, manhattan, canberra, binary, and minkowski distance. However, in this study, methods of matrix calculation do not significantly change the final result. Calculations were done with "dist" function from R library statistics.

Pan-genome based on amino acid sequence similarity of 95% was constructed using Roary. Proportions of each PV were calculated separately for human and bovine hosts. PVs with differences between hosts higher than 10 were used to build a classifier based on an SVM algorithm from R package e1071. The model was tuned and cost and gamma parameters adjusted to gamma = 1e-04, cost = 100 for the UK dataset.

To determine accuracy of the model, datasets were divided by 6 and cross-validated over 100 runs. Accuracy (A) of the predictions were calculated as

$$A = \frac{TP + TN}{TP + FP + TN + FN},$$

where TP and TN are true positives and negatives and FP and FN are false positives and negatives. Specificity and sensitivity were defined as

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad \text{Sensitivity} = \frac{TP}{TP + FN}$$

and calculated separately for bovine and human predictions. Results were drawn as a bar plots. Receiver operating characteristic (ROC) curves were plotted for the each run using R library ROCR. To assess the influence of neighbors on probabilities calculated for each data point, 10 strains from each host were randomly sampled and combined into test datasets, then predictions were made and probabilities calculated. These steps were repeated to the point where each isolate from the dataset had been tested at least 10 times. Ranges of the resulting probabilities were graphed as bar plots.

We assessed the host-based model by comparing it to its performance with randomly allocated groups. For UK host groups, we randomly sampled 20 isolates from each host group to make a test set (20 + 20); ΔPV > 10 were then determined for the remaining human and bovine isolates on which the training was performed, and then the initial 40 isolates tested and probabilities assigned. This process was repeated 20 times. For randomly allocated groups, we divided the UK dataset into equal size groups then sampled 20 isolates from each of these, recalculated the PVs > 10 for the remainder on which SVM training was carried out, then used this model to tested the 40 strains. This process was also repeated 20 times. The distribution of assigned probabilities was compared between the groups as well as the frequency of PV selection.

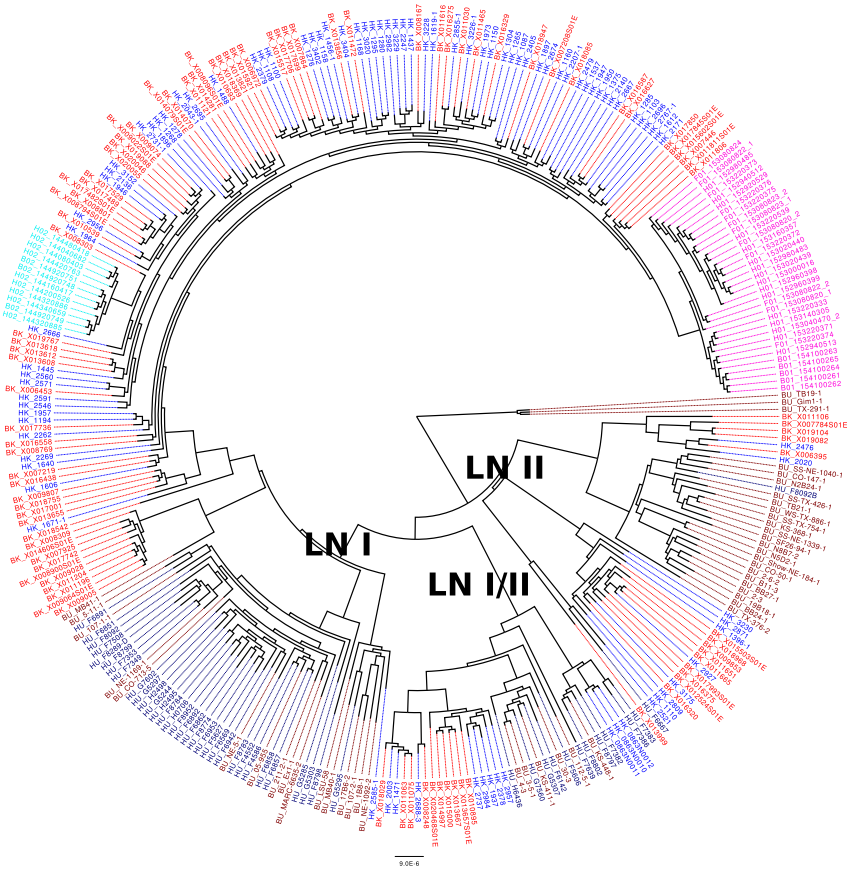


Fig. S1. ML phylogenetic tree of *E. coli* O157 strains, Roman numerals indicate lineages. Light red, UK bovine; dark red, US bovine; light blue, UK human; dark blue, US human isolates; cyan. milk outbreak isolates; and magenta, food outbreak isolates. The tree shows how the core SNP information aligns well with lineages and that UK and US isolates are generally separate, even within the same lineages. A minimum branch length is applied by the RAxML software, so very closely related isolate clusters seem more diverse than they are, but the overall relationship between isolates and clusters is correct.

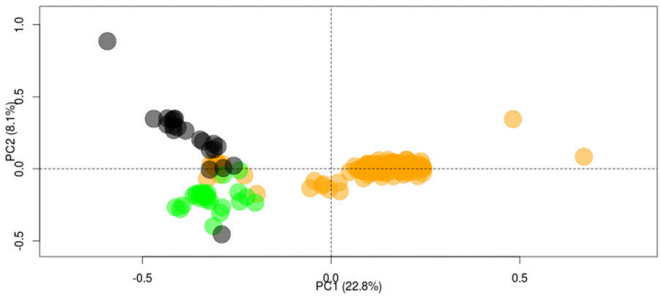


Fig. S2. UK isolates MMDS plot based on all PVs. Each circle represents a single isolate: orange, lineage I; dark gray, lineage I/II; and green, lineage II. The tight orange cluster on the right-hand side contains predominately lineage I PT21/28 strains. The second main cluster of lineage I isolates on the left-hand side (partially occluded) are predominately PT32.

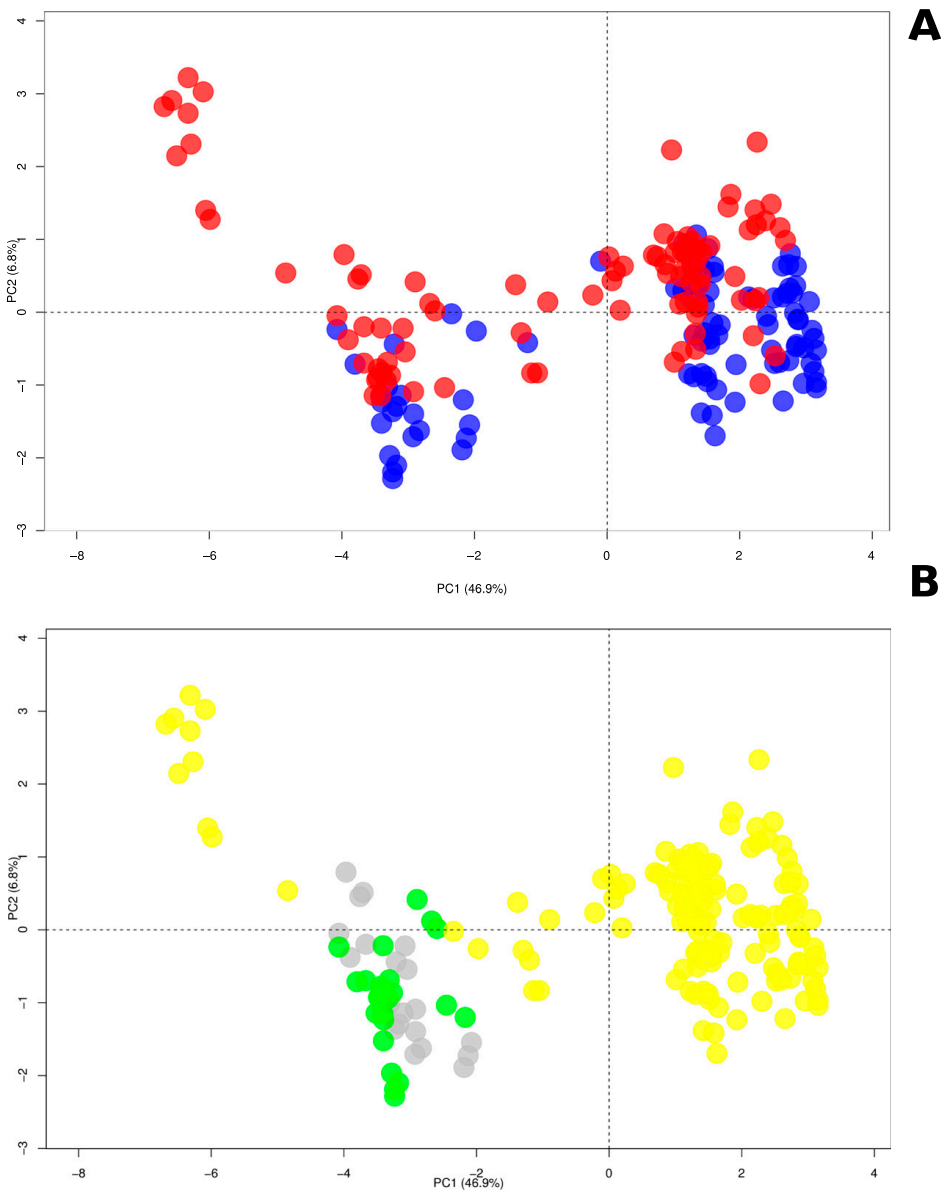


Fig. S3. UK isolates MMDS plot based on ΔPV > 10. Each circle represents a single isolate (*A*) colored by host: red, bovine and blue, human and (*B*) colored by lineage: yellow, lineage I; gray, lineage I/II; and green, lineage II.

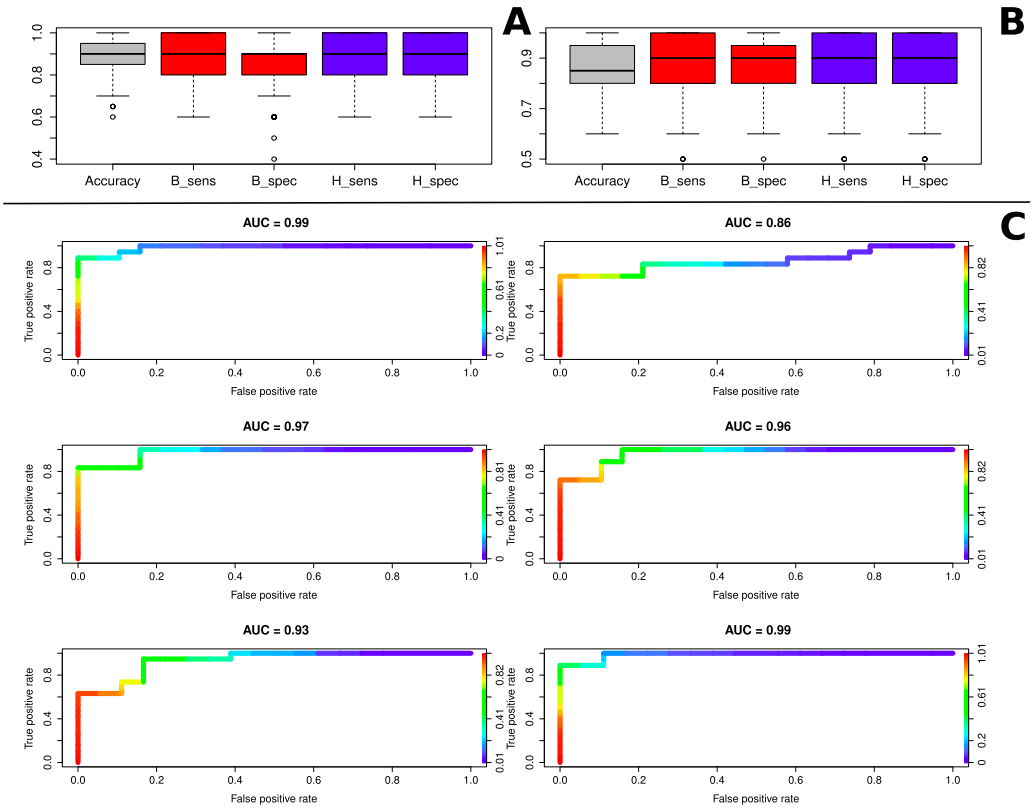


Fig. S4. Accuracy of the SVM model applied to the UK dataset. (*A*) Accuracy, specificity, and sensitivity of the UK model based on the ΔPV > 10 calculated over 100 cross-validation runs. Accuracy is the total accuracy of the model, which was 89%. (*B*) Accuracy, specificity, and sensitivity of the UK model based on the ΔPV > 20 calculated over 100 cross-validation runs. Accuracy is the total accuracy of the model, which was 86%. (*C*) ROC curves. The UK dataset was divided and each one-sixth was tested independently based on the training of the remaining five-sixths of the data. The area under the curve (AUC) value is plotted above each test subset.

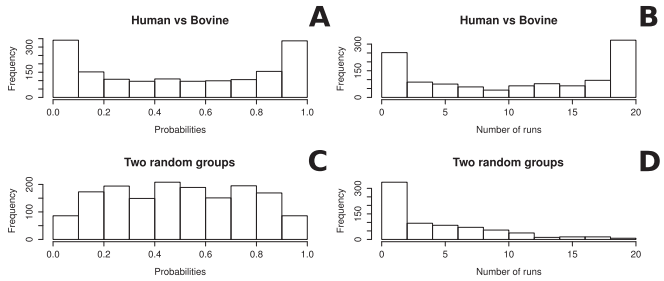


Fig. S5. Comparative analysis of random and host divided groups. (*A* and *C*) The probability distributions of isolates from 20 runs of 40 randomly sampled isolates from the study groups: (*A*) human/bovine groups and (*C*) two randomly divided groups. Differential PVs were recalculated for each run. (*B* and *D*) The number of runs in which the same discriminatory PVs were found. This distribution was markedly different between the bovine/human groups (*B*) and the random groups (*D*). For the human/bovine dataset there were more than 300 PVs that were repeatedly found in all 20 runs, whereas there were only five found for the random groups.

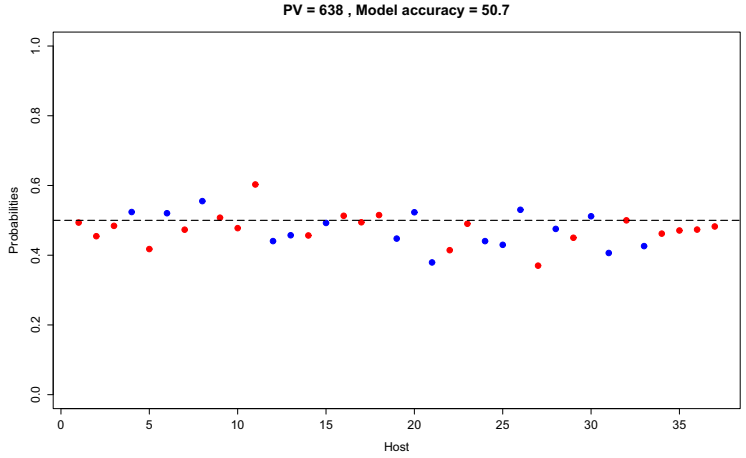


Fig. S6. Test of the SVM model for the UK dataset applying host-specific PVs to random groups. UK isolates (human and bovine) were randomly divided to generate a training dataset (80% of isolates) and a test dataset (20% of isolates). By applying the same algorithm to the model as used for the UK dataset alone, it was evident that the two groups cannot be distinguished. Prediction accuracy was 50% and all tested isolates achieved probabilities near the decision boundary of 0.5 (minimum 0.46 and maximum 0.52).

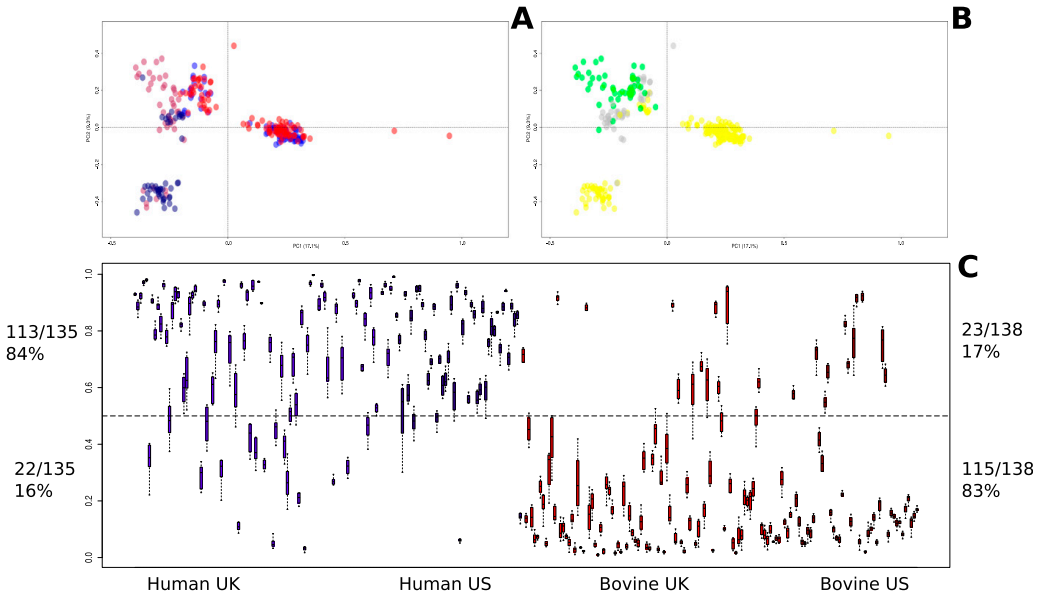


Fig. S7. Mixed-model analysis of combined UK and US datasets. (*A*) MMDS plot colored by host (red, bovine and blue, human) and (*B*) MMDS plot colored by lineage. (*C*) SVM analysis based on first combining the UK and US datasets and then multiple subtesting of different 20% subsets as described in *SI Materials and Methods*. All isolates that were called differently from their isolation host in the single-country model were also called this way in this combined model.

Table S1. Comparison of ΔPV by dataset

ΔPV	UK	US	Joined UK and US	Overlap UK and US
1. ΔPV > 5	1,326	2,429	1,448	556
2. ΔPV > 10	638	1,238	592	197
3. ΔPV > 20	82	532	84	12
4. ΔPV > 30	18	348	4	0

Number of differential proteins in the datasets shown in different thresholds. The proportion of each PV was calculated separately for human and bovine hosts.(i.e., PVs with a subtractive difference between the two hosts of greater than 5). Relatively few discriminatory PVs are shared between the UK and US dataset.

Table S2. ΔPV > 20 bovine

Cluster name	Bovine, %	Human, %	ID	Description
1. group_3598	93	62	NP_414783.1	CP4-6 prophage 3B uncharacterized protein
2. group_4316	93	66	PF12083.2	Hypothetical protein
3. group_22405	93	62	NP_415710.1	Murein tetrapeptide carboxypeptidase 3B LD-carboxypeptidase A
4. group_20911	93	60	NP_415709.2	Putative cation/proton antiporter
5. group_4846	79	56	NP_416090.2	Qin prophage 3B uncharacterized protein
6. group_4385	77	56	NP_416089.1	Qin prophage 3B uncharacterized protein
7. group_2374	90	70	NP_416552.1	Phosphomannomutase
8. group_22106	77	56	NP_309096.2	Regulatory protein
9. group_5053	90	69	NP_416550.1	Putative colanic acid exporter
10. group_8445	89	69	NP_416551.1	Colanic biosynthesis UDP-glucose lipid carrier transferase
11. group_10913	90	69	NP_416549.1	Colanic acid biosynthesis protein
12. group_8447	90	69	NP_416548.1	Putative glycosyl transferase
13. group_21878	86	51	NP_415711.2	Lytic murein endotransglycosylase E
14. intP_1	61	34	NP_288015.1	Integrase fragment 2C cryptic prophage CP-933P; Rac prophage 3B integrase
15. group_1710	90	69	YP_002271281.1	O antigen polymerase
16. group_2064	61	29	NP_417122.1	CP4-57 prophage 3B uncharacterized protein
17. group_2674	55	30	NP_417130.1	CP4-57 prophage 3B putative antirestriction protein ; CP4-6 prophage 3B
18. group_6602	55	34	YP_002402151.1	Phage protein
19. group_3028	81	44		Hypothetical protein
20. group_3128	69	47	YP_003234229.1	T3SS secreted effector NleG-like protein
21. group_1688	77	45	NP_417134.2 *(2)	Adhesin-like autotransporter
22. group_1685	80	44	NP_417134.2	Adhesin-like autotransporter
23. group_945	49	27	NP_415866.1	Rac prophage 3B exonuclease VIII 2C 5′ -_ 3′ specific dsDNA exonuclease
24. group_947	38	16	NP_415866.1	Rac prophage 3B exonuclease VIII 2C 5′ -_ 3′ specific dsDNA exonuclease
25. group_9649	51	27		Hypothetical protein
26. group_12348	50	26	NP_288007.1	Repressor protein encoded by cryptic prophage CP-933P
27. group_9656	50	26	YP_003233901.1	Putative antirepressor protein Cro
28. group_12345	48	25	YP_003078025.1	Plasmid stabilization system protein
29. group_2089	45	21	YP_002397033.1	Putative phage replication protein O (weak confidence)
30. group_3962	45	21	NP_415876.1	Rac prophage 3B uncharacterized protein ; phage regulatory protein
31. group_7348	62	15	NP_414651.1	Quinolinate phosphoribosyltransferase
32. group_2692	57	31	NP_417130.1	CP4-6 prophage 3B uncharacterized protein ; CP4-57 prophage 3B antirestriction protein
33. group_9542	47	22	YP_003222344.1	Antirepressor protein
34. group_4009	35	9	NP_287300.1	Tail assembly chaperone encoded by prophage cp-933n
35. group_658	61	16	YP_002272449.1	ISSfl4 ORF3
36. group_2261	37	15	NP_418704.1	IS30 transposase
37. group_1156	41	19	NP_416078.4	Qin prophage 3B uncharacterized protein
38. group_4925	40	18	NP_416352.4	Serine/threonine-specific protein phosphatase 1
39. group_9831	24	4		Hypothetical protein

ΔPVs > 20 that were present in higher proportions of bovine isolates than in human isolates. Highlighted in gray are proteins that formed multiple clusters (three different proteins formed two clusters each).

Table S3. ΔPV > 20 human

Cluster name	Bovine, %	Human, %	ID	Description
1. group_2494	70	91	YP_002268881.1	Tail fiber protein; tail fiber protein from prophage CP-933H
2. group_1918	38	78	YP_003228508.1	DNA biosynthesis protein (primosomal protein I); replication protein
3. group_2230	44	65	YP_002272118.1	DNA-binding protein
4. group_22597	38	67	NP_416507.1	CP4-44 prophage 3B uncharacterized protein
5. group_9655	46	70	YP_003233901.1	Putative antirepressor protein Cro
6. group_6372	56	77	YP_003224990.1	Phage repressor protein CI; e14 prophage 3B repressor protein phage e14
7. group_9543	35	67	YP_003222344.1	Antirepressor protein
8. group_12151	53	76		Hypothetical protein
9. group_12152	53	76		Hypothetical protein
10. group_9474	53	76	PF11225.2	Hypothetical protein
11. group_4840	35	64	YP_852520.1	Transcription regulatory protein
12. group_7349	33	74	NP_414651.1	Quinolinate phosphoribosyltransferase; molybdenum transport protein modd
13. group_22540	36	64	NP_416506.1	CP4-44 prophage 3B putative DNA repair protein
14. group_5954	50	70	YP_003233676.1	Outer membrane protein X; putative outer membrane protein Lom
15. ngoMIVR	48	71	P31032	Type-2 restriction enzyme NgoMIV
16. group_12291	48	71	P44068	Hypothetical protein
17. group_12364	48	70	YP_003235023.1	Putative antirepressor protein Cro
18. group_12365	48	70	YP_003235024.1	Putative phage repressor protein CI
19. group_3030	18	54		Hypothetical protein
20. group_12378	48	68		Hypothetical protein
21. group_2933	22	57		Hypothetical protein
22. group_896	16	64	NP_309418.1	BfpM-like protein; putative transporter
23. group_8446	10	31	NP_416548.1	Putative glycosyl transferase
24. group_5052	10	31	NP_416550.1	Putative colanic acid exporter
25. group_10912	10	31	NP_416549.1	Colanic acid biosynthesis protein
26. group_3959	10	40		Hypothetical protein
27. group_8444	10	31	NP_416551.1	Colanic biosynthesis UDP-glucose lipid carrier transferase
28. group_22105	23	45	NP_309096.2	Regulatory protein
29. group_2688	20	53	NP_417130.1	CP4-6 prophage 3B uncharacterized protein ; CP4-57 prophage 3B antirestriction protein
30. group_3185	26	56	NP_415861.1	Rac prophage 3B integrase
31. group_1711	10	31	YP_002271281.1	O antigen polymerase
32. group_12648	15	36		Hypothetical protein
33. group_654	12	32	YP_002272449.1	ISSfl4 ORF3; transposase; IS encoded protein within CP-933O; ORF 1 2C IS66 family
34. group_4378	18	44	NP_416089.1	Qin prophage 3B uncharacterized protein
35. group_4847	18	44	NP_416090.2	Qin prophage 3B uncharacterized protein
36. group_13920	9	32		Hypothetical protein
37. group_3600	7	36	NP_414783.1	CP4-6 prophage 3B uncharacterized protein
38. group_3844	7	29	NP_287540.1	Transposase within CP-933O 3B partial; transposase
39. group_22750	11	31	NP_286994.1	Tail fiber protein of bacteriophage BP-933W
40. group_21879	5	26	NP_415711.2	Lytic murein endotransglycosylase E
41. group_4380	7	31	NP_416089.1	Qin prophage 3B uncharacterized protein
42. cro	54	76	YP_002383192.1	Regulatory protein from bacteriophage origin
43. N	55	77	YP_002397117.1	Antitermination protein N

ΔPVs > 20 that were present in higher proportions of human isolates than in bovine isolates. Highlighted in gray are proteins that formed multiple clusters (three different proteins formed two clusters each).

Table S4. Pairwise comparison of human and bovine isolates

Pair	Bovine	Human	Unique PV	SNPs
1	BK_X011651	HK_2927	137	9
2	BK_X008248	HK_2688.3	143	17
3	BK_X010539	HK_2956	292	18
4	BK_X011616	HK_1619.1	364	26

The number of SNP and PV differences between human and bovine isolate pairs selected by relatedness from their core SNPs phylogenies (Fig. S1).

Other Supporting Information Files

Dataset S1 (XLS)

3.3.3 Conclusions

All isolates of the clonal group, *E. coli* serovar O157, in the previous chapter 3.2 were predicted to be of zoonotic threat. However, this separate O157 analysis when training and testing isolates are sampled from the the E. coli O157 'clonal' cluster confirms that differential host predictions can be achieved for this group. The resolution of the model is clearly reduced when phylogenetically distinct isolates are compared. General trends towards one or another host can be noted, but not the granularity of the sub-populations (see *S.* Dublin and *S.* Typhi examples tested on the STm-trained background 3.2 and compare *E. coli* O157 alone 3.3 with *E. coli* O157 tested on a wider, diverse scale 3.2compare). Distinct phylogeny was shadowing hidden patterns that eventually were discovered and allowed for separation by host even for such clonal group as *E. coli* O157. As such, the techniques can have more discriminatory power if you can focus down to specific sub-clusters and still have enough examples for training and testing.

A few issues were faced during this analysis and these were mostly due to clonal nature of O157 serogroup. First, it was noted that there were very few significantly differential PVs between human and bovine isolates. For the STm analysis, some PVs varied by 50% or even more; by contrast in the O157 study, no protein variants that differed by more than 20% were found. This presum-

ably reflects their clonal nature and that these isolates share quite a large core genome, above 4/5 of genome content was the same between isolates).

Another side of the analysis was the finding that the majority of the PVs on which the predictions were based were of phage origin. *E. coli* O157 isolates usually maintain from 10 to 20 prophage sequences incorporated in their genomes. Even though prophages can be quite diverse, they do have many related and sometime identical genes. To add further complexity to this situation, the phage content is usually further loaded with insertions sequences. These factors, similar regions and insertions, can heavily disrupt downstream analysis: assemblies usually break on phage regions, and due to insertions the same genes could be not be predicted correctly. This can negatively influence PV clustering which is based not only on sequence similarity but also on gene's neighbourhood. The high similarity of the clonal group and the complexity of phage content means that the O157 ML predictions were based on a large number of low differential value PVs: for example there were over 600 for the *E. coli* O157 analysis at PVdelta10 and around 100 used to discriminate the human bovine isolates at deltaPV40 for the wider *E. coli* population (Chapter 3.2). It is likely that some of these delta10 PVs will be erroneous, however at this point it would be impossible to distinguish between a true biological signal and noise introduced by the analysis. It is expected that noise could influence the predictions, however, the random testing showed significant differences

between a classifier based on the random data and well structured human vs bovine classifier outputs (see [110], supplementary material), reassuring that a 'host' signal can be extracted from the data.

To add to the mystery, our group's current work includes analysis of long sequence data (PacBio) and we have had a small subset *E. coli* O157 isolates sequenced including 12 phage type 21/28 isolates, some are already published[91]. When aligned, these sequences are extremely similar (above 90%). The most striking differences amongst these *E. coli* O157 genome assemblies are large chromosomal inversions that occur between highly homologous phage boundary regions. These potentially may explain some of the PVs identified, but conceptually such variation should have little impact on overall content identified. We still need to take account of possible plasmid-based differences as these have not yet been compared to PacBio data. In addition, the PacBio dataset is not yet big enough to consider a ML model to repeat an equivalent analysis from these long read assemblies, but this remains an objective. While phage and possibly plasmid based PVs will account for some of the differential PVs, we have cases of predictive score showing a differential host prediction for isolates from which the genome alignments (based on PacBio sequencing) show very little variation. There is concern that some of the variation may therefore be down to artefacts that appear during assembly and analysis, although we have also to consider that the Illumina method is from a broth culture and captures all read variants of an isolate's population

and the PacBio assembly is just one confirmation and form. We are seeing dramatic heterogeneity on some isolate cultures and this could also work its way through as a differential, for example IS and phage movements in sub populations. Such single isolate heterogeneity is a current focus in the laboratory. In summary, my confidence in the PVs reflecting true biological host differences increases with the differential score between the groups as it is less likely to be due to variation that can occur with assembly and other analysis issues.

Given the discussion in the last paragraph, this study, as with the previous chapter, would benefit from some experimental work, to try and confirm phenotype from genotype. However, for STm, performing competition experiments is quite realistic, for example competing 'chicken' and 'bovine' strains in a chicken gut, the equivalent for O157 would be more challenging to design and would require some human experiments which would be difficult to justify ethically. However, it may be possible to compete different isolates in cattle to see if differential 'bovine' scores have any differences in colonisation or excretion potential.

Despite the caveats and our lack of understanding of what may underpin the predictions, around 90% of isolates were classified correctly in both studies using SVM. Moreover, the *E. coli* O157 SVM correctly called the outbreak

isolates from meat and milk samples as 'human' giving us some confidence in this exciting approach to host attribution and my study was the first of its kind in bacteria. It is interesting how other ML methods would handle these datasets and if better predictions could be achieved. The next chapter explores this and considers what other ML methods could be valuable for bacterial host attribution.

3.4 Overview of ML methods for bacterial source attribution

3.4.1 Introduction

A single experiment can produce an extraordinary amount of complex data that is challenging to handle, analyse and interpret in order to identify and characterise features common to subsets before these can be understood in a biological context. While dataset is relatively small tabular format are an accepted practice to quickly sort or filter or plot data based on a few characteristics, nowadays these methods have reached their limitations. New ways to analyse and summarise a big data is needed in order to reveal any hidden structures that can then be evaluated by new experiments. Recently ML has become a popular choice for researchers in data rich subjects. However, it can be difficult for an inexperienced user to choose between the many algorithms available due to the complexity of both models and data.

This chapter aims to compare the capacity of different ML and other statistical approaches to analyse bacterial genome sequence data in order to determine if genetic signals relating to host 'specificity' or 'restriction' can be identified. A complication of this analysis is that while we know that certain serovars of *S. enterica* differ in their host restriction, i.e. that *S.* Typhi is a specialist human

pathogen, *S*. Dublin, initially considered restricted to cattle, has isolates that can cause infections in humans. Conversely, while *S.* Typhimurium is considered a generalist, as this serovar is isolated from multiple animal and human sources, the same as with *E. coli*, it may not mean that all isolates have the potential to infect multiple hosts.

It is anticipated that any such genetic signals of host-dependency will often be integrated with or masked by strong phylogenetic signals and at the moment we are reliant on classical approaches of sub-testing within datasets to validate the different methods being compared in this study. As such, this chapter is both a review and a research study into the different methods and their value to predict the host source of these particular bacterial species. It builds on the recent work described in the previous chapters predicting the zoonotic potential of *E. coli* and *S. enterica* using a single ML approach: Support Vector Machine in Chapters 3.2, 3.3 with the conclusion that supervised ML methods have significant potential to predict such complex phenotypes from bacterial WGS and this will be of value to public health authorities and others wanting to understand the zoonotic threat that surrounds us.

To explore dataset and classify bacterial sequences into 4 host categories variety of dimensionality reduction techniques (DRT) as well as supervised and

unsupervised machine learning (sML and uML) methods were used (See list below **??**). In short, DRT requires minimal prior knowledge of the dataset from the user, whereas for uML, the user should specify the number of classes by which the data should be separated. Some existing techniques to determine the optimal number of clusters for a given dataset were explored. By contrast for sML, an initial dataset (training dataset) with well defined subpopulations is required, thus the algorithm will deduce from it some specific characteristic of each class and learn to distinguish between them. Then, new data (test dataset) from unknown origin can be analysed and each new datapoint can then be assigned a probability to belong to one or another initial subpopulation based on previously learned characteristics of these subpopulations.

Dimensionality reduction techniques (DRT) are a group of methods that are often used as an initial step in exploratory analysis of data. Dimensionality or complexity of the data are reduced using an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of observations. This transformation is defined in such a way that the first principal component has the largest possible variance, i.e. it accounts for as much of the variability in the data as possible and each succeeding component, in turn, has the highest variance possible with the constraint that it is orthogonal to the preceding

components. The resulting vectors are an uncorrelated orthogonal basis set.

Moreover, DRT are the way to reduce the number of random variables under consideration by obtaining a set of principal variables. It is usefull for simplification of models, to make them easier to interpret, if computational resources is a bottleneck DRT can be applied prior training to shorten training times and also to reduce curse of dimensionality (the dimensionality increases, the available data become sparse, thus difficult to obtain statistically sound and reliable result. The amount of data needed to support the result often grows exponentially with the dimensionality. Reducing the number of random variables also would enhanced generalization. Overall, DRT can be divided into feature selection and feature extraction.

In simplified terms the supervised and unsupervised algorithms are described below.

k-means The first step when using k-means clustering is to indicate the number of clusters (k) that will be generated in the final solution. The algorithm starts by randomly selecting k objects from the data set to serve as the initial centers for the clusters. The selected objects are also known as cluster means or centroids.

Next, each of the remaining objects is assigned to itâĂŹs closest centroid,

where closest is defined using the Euclidean distance between the object and the cluster mean. This step is called 'cluster assignment step'.

After the assignment step, the algorithm computes the new mean value of each cluster. The term cluster 'centroid update' is used to design this step. Now that the centers have been recalculated, every observation is checked again to see if it might be closer to a different cluster. All the objects are reassigned again using the updated cluster means.

The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments stop changing (i.e until convergence is achieved). That is, the clusters formed in the current iteration are the same as those obtained in the previous iteration.

The **agglomerative clustering** is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named dendrogram. The **divisive clustering** starts by including all objects in a single large cluster. At each step of iteration, the most heterogeneous cluster is divided into two. The process is iterated until all objects are in their own cluster.

In natural language processing **Latent Dirichlet Allocation** is generative statistical model that facilitates the automatic discovery of themes in a collection of documents. Algorithm starts by going through each document, and randomly assign each word in the document to one of the K topics (similar to kmeans defined beforeahend). This step results in randomly assigned classes. To improve them for each document d go through each word w in d and for each topic t, compute two things: 1) p(topic t | document d) = the proportion of words in document d that are currently assigned to topic t, and 2) p(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w. Reassign w a new topic, where you choose topic t with probability p(topic t | document d) * p(word w | topic t) (according to our generative model, this is essentially the probability that topic t generated word w, so it makes sense that we resample the current wordâĂŹs topic with this probability). In other words, in this step, it is assumed that all topic assignments except for the current word in question are correct, and then updating the assignment of the current word using our model of how documents are generated. After repeating the previous step a large number of times, youâĂŹll eventually reach a roughly steady state where your assignments are pretty good. So use these assignments to estimate the topic mixtures of each document (by counting the proportion of words assigned to each topic within that document) and the words associated to each topic (by counting the proportion of words assigned to each topic overall).

Briefly, **SVM** works by identifying the optimal decision boundary that separates data points from different groups (or classes), and then predicts the class of new observations based on this separation boundary. Depending on the situations, the different groups might be separable by a linear straight line or by a non-linear boundary line.

Random forest in essence uses one feature at the time to separate data into classes. Combining random multiple features the decision tree is built. Combining multiple trees into a forest and calculating statistics on usefulness of each of the feature the final result is obtained.

A **neural network** is an interconnected group of layers of nodes, similar to the vast network of neurons in a brain. Each layer acts as a detection filter for the presence of specific patterns present in the original data (first layers detect most obvious patters and later layers detect smaller, more subtle patterns). Last layer combine all the data that was learned by the previous layers. In same cases (back propagation) the learning process does not end when last layer are reached, but after comparing obtained values with the labels provided will be disentangled backwards through network and adjusting learning to achieve better outcome.

All of the above mentioned techniques were used to address source attribution of *Salmonella enterica* serovar Typhimurium (see **??**). The results are focused on a few main points: (1) how the each method performs according to known

Table 3.1: Methods

	Dimensionality reduction methods	
1	Principal component analysis (PCA) [144]	*PCA tutorial*
2	Multidimensional scaling (MDS) [145]	*MDS tutorial*
3	t-Distributed Stochastic Neighbor Embedding (t-SNE) [146]	*t-SNE tutorial*
	Unsupervised machine learning	
1	K-means [147]	*k-means tutorial*
2	Divisive Hierarchical clustering (DHC) [148]	*DHS tutorial*
3	Agglomerative Hierarchical clustering (AHC) [148]	*AHC tutorial*
4	Latent Dirichlet Allocation (LDA) [149]	*LDA tutorial*
	Supervised machine learning	
1	Support vector Machines (SVM) [106]	*SVM tutorial*
2	Random Forest (RF) [107]	*RF tutorial*
3	Neural Network (NN) [108]	*NN tutorial*

host information about the dataset; (2) how user-friendly the techniques are; (3) the interpretability of the results and (4) any additional value of the technique. The 3 main methods are listed below. Explanation of the methods is out of the scope of this work, however references to original techniques, as well as good tutorials are provided.

3.4.2 Methods

For this study the dataset of 1203 *S.* Typhimurium genome sequences from the previous work (see chapter 3.2, [114]) were reused. Proteins variants associated with each host ($p > 0.05$) were calculated by pangenome GWAS software SCOARY [150] after 500 permutations. All calculations and visualisation presented in this chapter were carried out in R [127]. Significance tests were done using t.test{stats} and prop.test{stats}. Local polynomial regression fitting (loess{stats}) was used with 10% smoothing span and fitted with generic predict{stats} function. Diversity was explored with R package 'vegan' [151] from which following functions and methods were used: diversity() for Shannon index calculations, betadiver(x, 'z'), betadisper(), anova(), plot.betadisper(), permutest.betadisper(x, pairwise = T, permutations = 99) for beta diversity analysis.

Diversity indexes Dissimilarity matrix for all uML analysis was calculated based on euclidean distance. 5 fold cross-validation (sliding window) was used for all supervised models. To assign host scores one isolate at a time has been removed from the training, model has been trained and then the isolate were tested.

3.4.3 Results

3.4.3.1 Diversity

The same *Salmonella* Typhimurium dataset as in Chapter 3.2 was used in this study. The dataset is based on sequences from 4 hosts: 311 avian (A) isolates, 300 bovine (B), 336 human (H) and 256 swine (S). Pangenome matrix calculated by Roary was also reused from the previous study [114]. Pangenome contained 23,307 protein variants (PV) clusters. The mean number of predicted PVs for an isolate across the combined dataset was 4620 (min = 4037, max = 4993), while mean values of predicted PVs per each host were very similar with slightly less PVs in the human dataset (A = 4,632, B = 4,645, H = 4,573, S = 4,636). There was a significantly lower ($p < 0.001$, z.test, R [127]) number of core predicted proteins in the avian population (A = 2,218, B = 3,054, H = 3,056, S = 3,065) compared to the remainder. The total number of core protein clusters across all isolates was 1991, indicating only a partial overlap in proteins considered core in each host. Rare PVs, these that would be found in less then 15% of all isolates originated vast majority of the clusters (n = 18,203) and this number of rare PVs vary by host minimally (A = 18,354, B = 18,295, H = 18,272, S = 18,415).

One of the ways to examine biodiversity is to calculate diversity indices. Here

we used Shannon index [152] that quantifies the uncertainty in predicting the species identity of an individual that is taken at random from the dataset.The higher the index the more species are in the dataset. Moreover, Shannon index incorporates both aspects of diversity: richness (how many species in a site) and the evenness (how close in numbers each species). To translate to genomics, genes were treated as species and each host as a site. Shannon indexes were calculated across all PVs which proportions vary between hosts. The results demonstrated very similar levels of diversity with average value 7.88 with only slight decrease for human dataset (Shannon = 7.85). (see Figure 3.12 A). Another aspect of diversity can be described by beta diversity which defined as the differences in species composition among sites[153]. To analyse beta-diveristy, beta-dispersions were proposed [154] and used here to calculate an average dissimilarity from individual observation units(PVs) to their group(hosts) centroid in multivariate space. The clusters resulted from this analysis with average distances to the centroid A=0.09725, B=0.10578, H=0.12157, S=0.09975 were significantly different between hosts (Annova, p-value<0.0001) and in particular, between human and any other host (pairwise comparison with permutations, p-value <0.0001), see Figure 3.12 C). Avian and swine isolates were gathered into most tight clusters while bovine and human subpopulation were more disperse. Apart from the density, it can be noted, that some clusters overlap significantly as for example swine and bovine, while majority of avian isolates clustered quite distantly. Based on this

analysis human isolates can be divided into 2 subgroups, one was overlapped with bovine cluster and other was scattered between human and avian clusters and mixed with isolates from other hosts. Another interesting observation is that datapoints was not gradually dispersed from a host clusters, these isolates that are scattered away from the centroids usually appeared by more than one standard deviation away. (one SD from each centroid shown as an ellipse (see Figure 3.12 B).

To identify genetic content that associated with each host, all clusters that were present in 100% of isolates (i.e. core, n = 1,991) were removed, then the dataset was reduced further by removing all clusters that present in the equal proportions across different hosts, leaving only PVs clusters for which the proportions between hosts varied. Remaining 4,041 PVs shown on Figure 3.12 A. As a whole, no significant differences in distribution of PVs proportions in each host population were noted as shown by best fit lines (loess) (Figure 3.12 C).

Proportions of the majority of the 4,041 PVs varied just sligtly between hosts (Figure 3.12 B), however there were some PVs that were significantly associated with host groups as calculated by pangenome GWAS [150]. Furthermore, some of the PVs significantly associated with more than one host (i.e a PV was significantly overrepresented in one host and in the same time was be significantly under-represented in another host (Figure 3.12 D). There were

263 avian associated PVs and 113 (43%) of these were shared between other hosts, 78 PVs significantly associated with bovine host from which 30 (38.5%) were shared with other host groups, 197 human associated with 108 (55%) shared, 132 swine PVs with 67(51%) shared. As soe clusters were shared between hosts, therefore, the total number of unique significantly differential clusters was 495. Each host's differential PVs were plotted as black circles in Figure 3.12 C as well as in a Venn diagram to visualise the overlap of the PVs between each group (Figure 3.12 D). All subsequent analyses in this study were applied to the reduced matrix of 495 PVs that represent only those flagged by SCOARY as significant.

Dissimilarity matrix based on euclidian distances obtained from these 495 PVs showed some clustering (Hopkins statistics = 0.25), which were very similar to these that were obtained by phylogeny either SNPs or accessory genome based [114] (see Figure 3.12 E).

3.4.3.2 Dimensionality reduction techniques

Could dimensionality reduction by combining and transforming variouse features separate data into clear host related clusters? First two methods, PCA and MDS, show very similar results for which the majority of the avian isolates are separated by the first principal component from all other isolates. Moreover there is a small group of bovine and porcine isolates that are placed apart

from the bulk of other isolates as well as apart from the separated avian isolates indicating that these are seen as a different sub population apart from the remainder of bovine and porcine isolates. Both PCA and MDS show considerable overlap between the majority of all other hosts which are placed as a large cluster of mixed isolates, however some spread in range can be noted by PC 3 in the MDS analysis that place human isolates above bovine and porcine (See Figure 3.13 A, B).

All original features contributed in lesser or greater extent to separation of datapoints without any clear influencers. No meaningful biological information about what proteins are responsible for the separation can be extrapolated from this analysis. So the complexity of the data can be seen by percent of variation explained by principal components and for both PCA and MDS all 3 PC explained only 46% of variation (PC1 = 26%, PC2 = 12% PC3 = 8%)

Quite different results were obtained from the third dimensionality reduction technique, prbabilistic model t-SNE. Apart from the same differences observed with the majority of avian isolates, multiple tight groups were observed. It is interesting that the algorithm divided human isolates into 4 separate clusters, as well as divide bovine isolates into several of the most distant groups. One of the parameter that can be changed is a perplexity which is a guess how many neighbours each datapoint has. The plot shown (3.13 C) is done with

perplexity = 300 what reflects average number of isolates for each host group. Overall, any perplexity value between 30 to 350 resulted to the clustering similar to shown here **??** C, as expected decreasing perplexity produced badly separated one fuzzy group, while gradual increase above 350 produced one increasingly tighter group, both without any noticible structure.

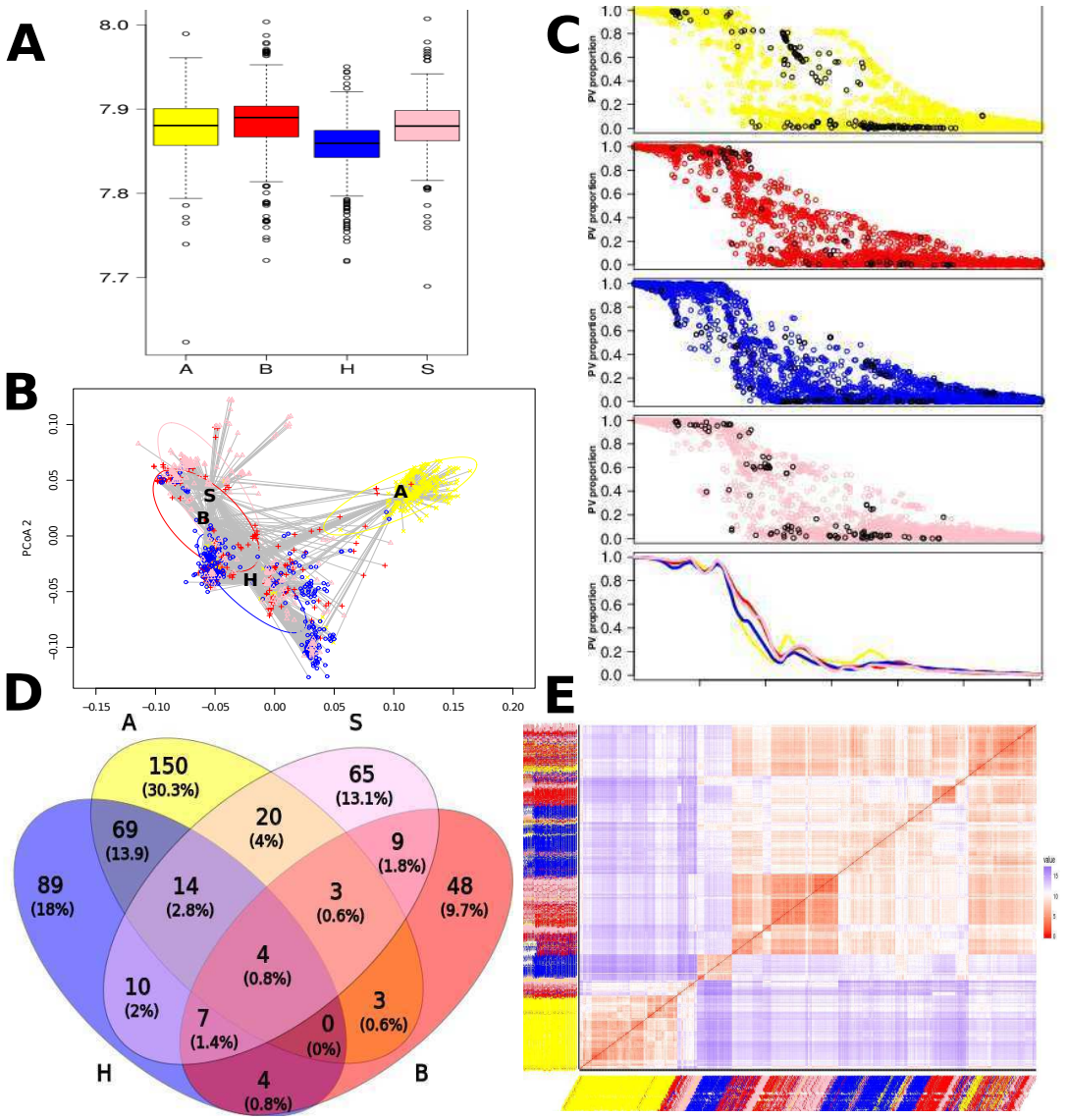


Figure 3.12: *S.* Typhimurium pangenome exploration. Colours represent host: avian (yellow), bovine (red), human (blue), swine (pink).**(A)** Shannon index calculated for each host on differential PVs.**(B)** Beta-dispersions shows multivariate homogeneity of isolates in each host. Non-euclidean distances between objectsare reduced to principal coordinates (x-axis and y-axis) Ellipses indicate one standard deviation from each host centroid marked as a letter.**(C)** 4,041 PV (x-axis), which proportions (y-axis) of presence vary between host (colours). Significantly associated with each host PVs (as calculated by pangenome GWAS) are plotted in black. Bottom panel shows best fit lines (Loess) for distribution of differential PVs from all host. **(D)** Numbers of PVs significantly associated with host and overlap of differential PVS by between different hosts. **(E)** Ordered dissimilarity matrix based on differential PVs. Heatmap colours: red (high) and blue (low) similarity. Labels are coloured by host.

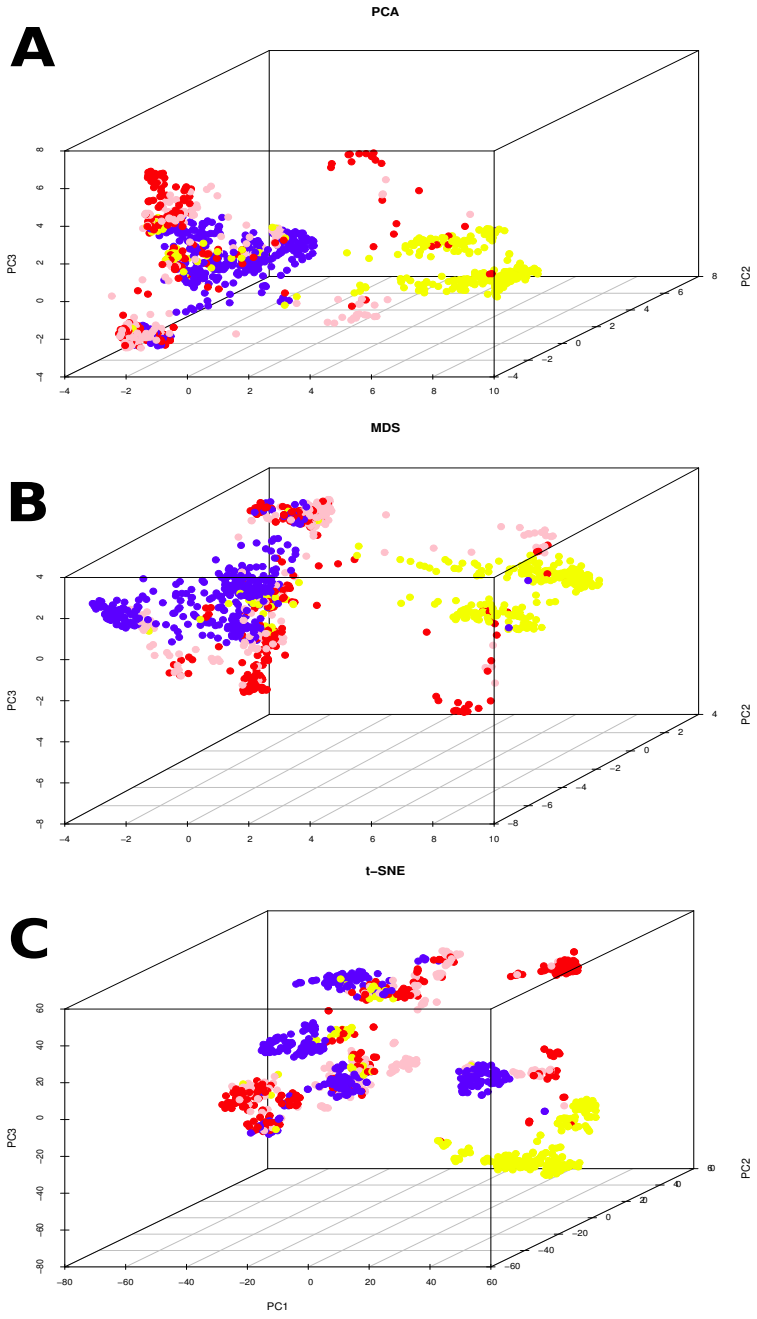


Figure 3.13: Dimensionality reduction techniques. Colors represent host: avian (yellow), bovine (red), human (blue), swine (pink) **A.** Principal component analysis (PCA), **B.** Multidimensional Scaling (MDS), **C.** t-Distributed Stochastic Neighbor Embedding (t-SNE).

3.4.3.3 Unsupervised ML

Even though the results of dimensionality reduction methods and unsupervised machine learning (clustering) seem to be similar as they both split datasets into smaller subgroups, these two methods differ in respect that DRT aims to compress features whereas clustering aims to compress datapoints. Also, for any of the clustering algorithm it is necessary to set a number of clusters that a user would like to obtain before running the analysis. Some techniques can indicate the optimal number of clusters for each of the uML algorithms: by computing different numbers of clusters and comparing within cluster 'sum of squares', sometimes called the elbow method, or average silhouette method, that computes how well each object lies within the cluster, or gap statistics methods which compare within intra-cluster variation with their expected values under null reference distribution. [155] [156].

All three cluster assessments were applied to demonstrate what number of clusters would be considered optimal by each of these methods, based on the k-means clustering. Figure 3.14 demonstrates that each method come to the different solution and number of the optimal clusters vary from 2 recommended by silhouette method to 10 by gap statistics. Moreover, according to the gap statistics graph, 10 is not yet the optimal number of clusters as the trend line has not yet reached a plateau. To sum up, there is no agreement between

cluster assessment methods, even though each of these produce stable solutions. In all subsequent uML analysis only 4 clusters were calculated as the objective was to access if it is possible to split the dataset by host and bacteria in the dataset were isolated from 4 hosts.

The above described cluster assessment methods can be used not only at the start, to guide the analysis, but also can at the end to assess how well clustering algorithms have performed. There are over thirty different indices that could be used and recent studies [157] indicate that some of these, including Silhouette, Davies-Bouldin and Calinski-Harabasz perform the best in a wide range of situations. Thus, Silhouette indices were used to measure how similar is a data point inside of a cluster compared to those in other clusters. Silhouette assigns a score to each data point and these scores range from -1 to 1, and the best clusters should have an average score near 1. If the average score is near 0 it could indicate that cluster members would be better separated into more, smaller clusters. When the value is negative it is an indication that the data points were wrongly placed into this cluster.

In addition, the results of the clustering were mapped to the previously obtained maximum likelihood phylogenetic tree see Figure 3.3, mainly because the phylogeny is a well established way to visualise bacterial datasets and provides a

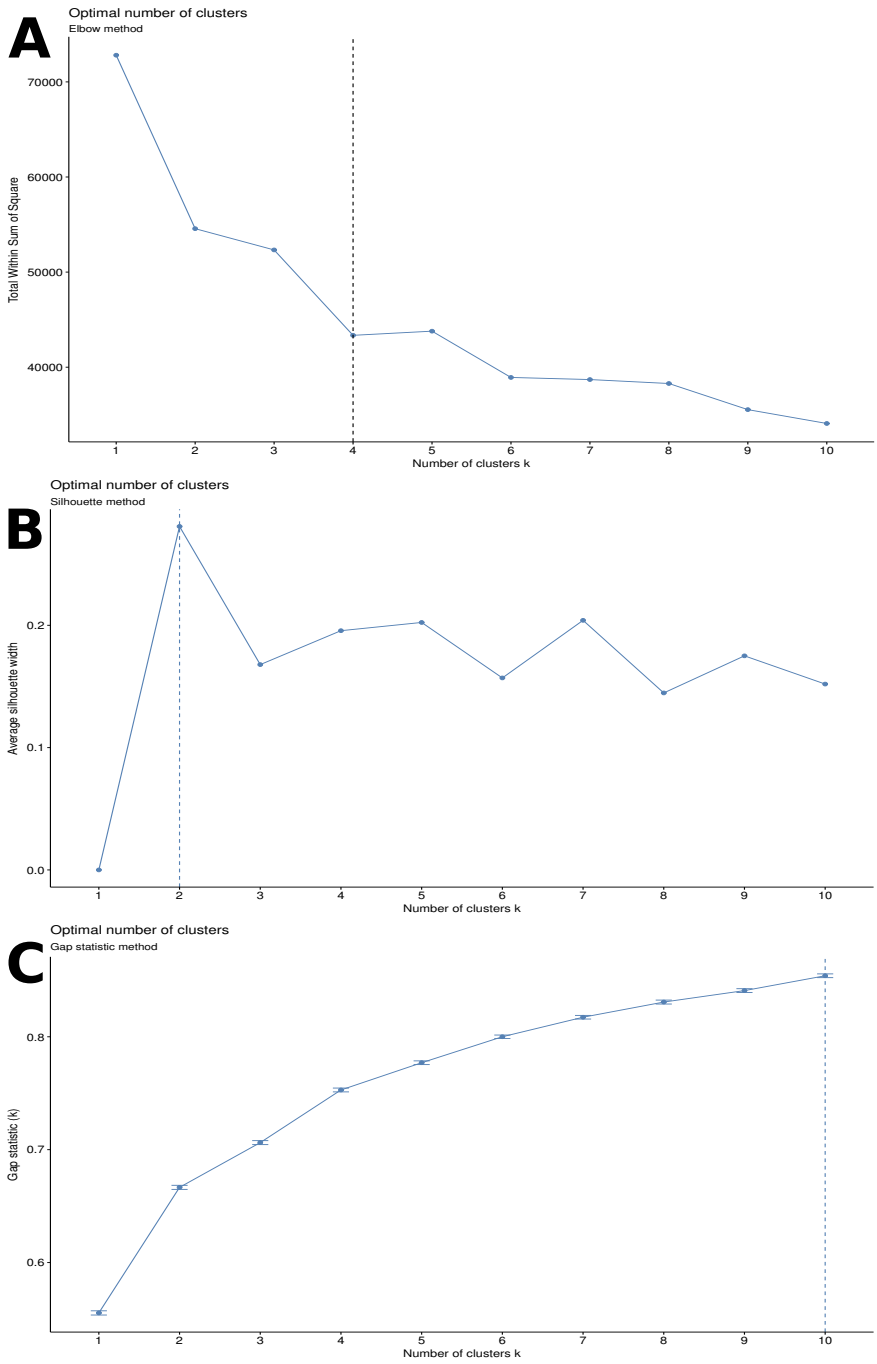


Figure 3.14: Optimal number of clusters as calculated by **(A)** Elbow method, **(B)** Silhouette method, **(C)** Gap statistic method. The methods that calculate optimal number of clusters in STm dataset were in disagreement and the recommended number of cluster ranged between 2 in 'silhouette' to above 10 in 'GAP statistics'

clear snapshot of the diversity of the bacteria in question as well as could infer relationships between particular isolates. In the previous sections, 3.1, 3.2, in

either the core or pan tree the more obviouse was avian cluster, however it incorporated only some of the isolates (80%) while the other 20% were spread across the tree and were found in close proximity to isolates from other hosts. Based on the phylogeny there is also a human cluster that contained 50% of isolates and a smaller bovine 30% group.

Overall uML agreed in allocation of the majority of the strains into particular clusters. (see Figure 3.15). So all the uML methods decided that the majority of the avian strains, also shown by phylogeny as related, should belong to the same cluster. Moreover, the human isolates were most of the time (kmeans HA, LDA) divided into 3 clusters, with one mainly human and two others of the mixed host origin. It is intriguing that all uML methods agreed that some of the phylogenetically close bovine strains (on the right side of the phylogenetic tree, Figure 3.15) were separated from the main bulk of bovine strains and allocated to the avian cluster. Comparing four different uML methods, it is evident that kmeans and hierarchical agglomerative clustering come to the almost same solution, with only 4.4% of the sequences allocated to different clusters, with main the disagreement about the human strains that by kmeans are allocated in the mostly human cluster, but by HA these strains are in the big mixed population cluster.

It is also very intriguing that almost all (except HD) methods not only joined

together the avian strains from the avian phylogenetic cluster but also added to that cluster bovine and swine isolates from the phylogenetic neighbourhood of the avian cluster.

Silhouette index for uML varied from 0.33 for LDA to 0.62 for HD, thus based on that measure HD had the most successful clustering strategy. Kmeans and HA both achieved very similar results and indexes of 0.46 and 0.43 respectively. However the number erroneously allocated isolates (silhouette index below 0) are higher with HA clustering. HD created the 'cleanest' avian cluster compared to all other uML (with only 2 human and 5 bovine included in that cluster), however HD also originated an enormous mix population cluster that was composed of 754 strains (67% of all strains).

The LDA method showed interesting result as it produced the most variable clusters. Also apart from the 'mostly' avian cluster its choice of the strains for a particular cluster, when compared with their phylogeny, seemed much more segmented, clearly indicating that this algorithm is finding a different and more granular pattern than other algorithms.

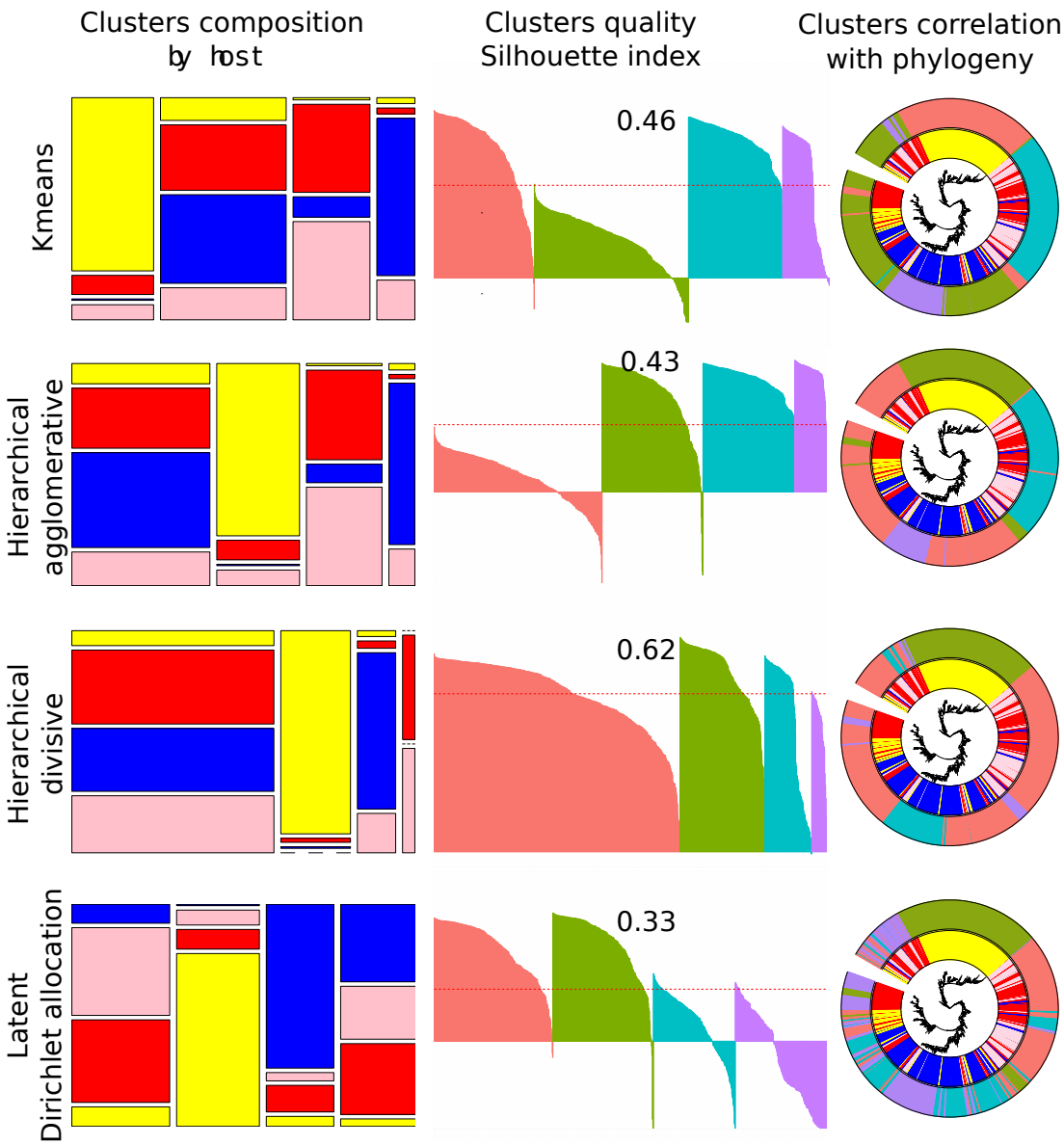


Figure 3.15: Unsupervised machine learning. The colours represent host: avian (yellow), bovine (red), human (blue), swine (pink). The first column of the figure shows the cluster's relative size and it composition by host, Second column demonstrates Silhouette index cluster assessment, where each of 4 clusters coloured differently and each isolate drawn as a bar with the silhouette index from (-1 to 1) allocated to it, and average of all individual indexes is plotted on the top of silhouette cluster and denoted as a red dotted line. The clusters are drawn in the same order as these from the first column. Third column illustrate cluster correlation with phylogeny (assessory genome tree) with inner ring depicting host and outer ring clusters

3.4.3.4 Supervised Machine Learning

SVM performance was demonstrated in the previous sections, thus this algorithm was rerun to verify that the results are stable and can be repeated, which was the case. Some variation during the training process were noted. So SVM and RF cross-validation model accuracy never were higher than 80-85% while for DL model cross validation accuracy could reach 100%. Nevertheless, when isolates were tested by 'leave one out' method, all algorithms showed very similar results with 85% $\pm$1% of overall accuracy with averaged accuracy by host (A - 90.3%, B - 78%, H - 92%, S - 75%). The performance of all algorithms indicated that avian and human host contain the easiest to learn patterns, while bovine and swine host have many features in common and therefore are difficult to distinguish. Figure Figure 3.16 shows the performance of each sML method with overall cluster composition and indication of assignment for each isolate. So the tendency for errors is as follows: all hosts except avian have a second preferred group in terms of probable assignments. For human isolates this is the bovine group, for bovine it is swine and vice-versa. For avian the erroneous assignments were spread equally between all other host categories. See Table 3.3, 3.2

Table 3.2: Host assignments as calculated by SVM. Predicted values are in columns and are equivalent to 'cluster composition by host' plotted on Figure 3.16.

Host	Ap	Bp	Hp	Sp	Total
A	278	16	9	8	311
B	13	234	15	38	300
H	1	25	309	1	336
S	7	45	12	192	256
Total	299	320	345	239	1203

Table 3.3: Host assignments as calculated by RF. Predicted values are in columns and are equivalent to 'cluster composition by host' plotted on Figure 3.16.

Host	Ap	Bp	Hp	Sp	Total
A	275	15	15	6	311
B	15	240	18	27	300
H	1	24	309	2	336
S	4	52	10	190	256
Total	295	331	352	225	1203

Table 3.4: Host assignments as calculated by DL. Predicted values are in columns and are equivalent to 'cluster composition by host' plotted on Figure 3.16.

Host	Ap	Bp	Hp	Sp	Total
A	281	13	11	6	311
B	14	226	19	41	300
H	1	19	305	11	336
S	6	47	11	192	256
Total	302	305	346	250	1203

High human scores for isolates from the non-human host group could indicate higher zoonotic potential for these particular isolates. Such scores occasionally assign one or another sML but here I am reporting only these isolates in which all 3 sML algoritms agrred in the assignment off to another, non human host. So it was 8 avian isolates that by all 3 sML were called 'human', 9 bovine, and 6 swine.

According to Silhouette index, the quality of clusters that were formed based on the prediction are of much worse quality than those from uML, with a Silhouette index 0,21 for SVM and RF and 0.19 for DL reflecting that the Silhouette index cannot capture the similarity of the patterns learned by sML. Moreover these strains that were allocated Sil index below 0 (n=365 for any of sML) all except 4 were of the same host as main population of the cluster.

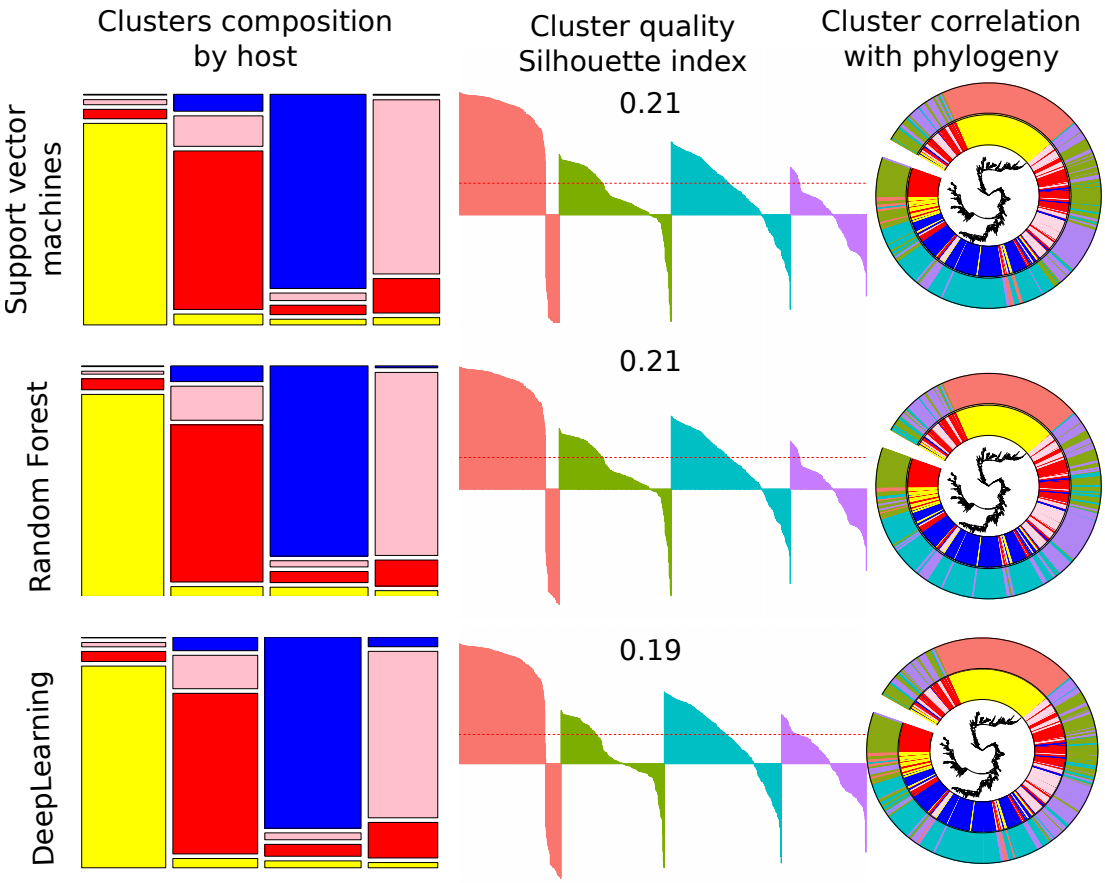


Figure 3.16: Supervised machine learning. The colours represent host: avian (yellow), bovine (red), human (blue), swine (pink). The first column of the figure shows the cluster's relative size and it composition by host, Second column demonstrates Silhouette index cluster assessment, where each of 4 clusters coloured differently and each isolate drawn as a bar with the silhouette index from (-1 to 1) allocated to it, and average of all individual indexes is plotted on the top of silhouette cluster and denoted as a red dotted line. The clusters are drawn in the same order as these from the first column. Third column illustrate cluster correlation with phylogeny (assessory genome tree) with inner ring depicting host and outer ring clusters. (also see Figure 3.16)

.

3.4.4 Discussion

In this chapter, different methods and their usefulness and capacity for bacterial source attribution were evaluated. Starting with the pangenome matrix of 1203 STm isolates from 4 hosts, the ratios of pan and core genes were analysed for each host as well as pan genome diversity calculated. The avian group had a significantly smaller core genome (2,218 compared to average 3,055 genes in all other hosts) and the human dataset had a slightly less diverse repertoire of accessory genes. STm have semi-open pan genome as number of genes didn't increased dramatically compared to other bacteria (compare 23)with After choosing only PVs significantly associated with each of the hosts, the pangenome matrix was reduced to only 493 features. Hopkins statistics indicated that the data is almost uniformly distributed (H = 0.25).
Ecological methods can be applied in genomics

Moreover, dimensionality reduction techniques all project data from multidimensional space int 2D or 3D, therefore it can be also challenging (or even incorrect to describe these results in terms of proximity/closeness).

Having too many features (dimensions) with little predictive power is a well-known problem in ML, as it can be computationally costly, and negatively affect the quality of the ML model. Several mitigating methods have been developed

to reduce the dimensionality of the data. These include, for example, principal component analysis (PCA) [144], multi-dimensional scaling (MDS) [145] and t-distributed stochastic neighbour embedding (t-SNE) [158].

PCA calculates the features that are able to represent the variance of the data, in the best way, eliminating dimensions that have low variance, effectively projecting those dimension into the remaining space. For example, if two features have high correlation between them, it is likely that both features are not necessary to effectively train the data, as only one of them would suffice to explain the variability. It is advisable to start any exploration analysis with visualisation by dimensionality reduction as well as for prediction use the simplest model possible as for example logistic regression, that in many cases could bring you more then half way closer to the answer. However in some cases dimensionality reduction would spot the most prominent features.

So dimensionality reduction techniques were applied to the 4 host dataset in order to check the ability of these methods to distinguish bacteria by host. Previously PCA and MDS have been applied to distinguish between 4 species of Enterobacteriaceae based on their biochemical profiles [159], demonstrating that PCA performed better than MDS, clustering the different species closer together. Our results demonstrate that both methods PCA and MDS were finding the same differentials in the dataset and both strongly agreed that the

majority of the avian dataset is quite different from all others. Both PCA and MDS collapsed the majority of the the data points into one cloud without any differentials between hosts.

On the other hand, the third technique t-SNE was able to separate the dataset into multiple tight clusters, revealing underlying structure of the dataset and differences between subgroups. Future work could explore the clusters that were obtained by t-SNE in order to find out what defines that structure, i.e. geography, time or environment related. However, all dimensionality reduction techniques map the multi-dimensional data to a lower dimensional space, thus the input features no longer exist in their original form as presence and absence values. Multiple features contribute to principal components, thus DRT becomes an impractical choice as a preliminary analysis of future wet lab experiments, analysis that could lead to further laboratory tests. So essentially it is mainly a data exploration and visualization technique, however DRT also can be used as a preliminary step in other supervised ML techniques where reduction of the data space is beneficial as for example when there are many irrelevant features, similar features or when computation becomes too costly.

Then I applied both uML and sML. The choice between uML and sML is not always straight forward. Usually, uML is quicker, more beginner friendly and

some times even the only choice because labelled data (metadata) is expensive and some times not even available. To use uML, the number of clusters should be decided beforehand. All of the three techniques that were used to decide the number of clusters showed different results, suggesting that the data is complex and no obvious clustering solution can be identified. On the other hand when all uML divided data into 4 clusters (aka 4 hosts), all uML come to a very similar solution, confirming that there is a stable underlying structure in the dataset, however this is not completely related to the host. Future work could explore uML further by trying to divide the dataset into a larger number of clusters, which could lead to multiple 'clean' clusters of the same host, similar to the human clusters identified in the t-SNE plot.

It is also very intriguing that the best host cluster, avian, had some bovine and swine isolates in it but ver minimal human ones. Most likely this indicate contamination by other host .

It is remarkable that phylogeny that takes into account not only molecular information (core SNPs) but also substitutional model (GTRGAMMA) [118] and probability distribution of all possible phylogenetic trees produced results that were quite similar by topology with the simple Neighbour Joining tree [160] that has completely different information as an input; (pangenome matrix of the presence and absence of protein variants written as 1s and 0s) which is, in essence a bottom-up (agglomerative) clustering.

sML uses an extra piece of the information, which is a label. sML is designed by giving the label the is giving the main 'weight', no matter the content of the whole vector of the features. Therefore, it is forced to find commonality between data points with the same label even though these commonalities are not the obvious ones. Thus new, previously undiscovered patterns can be revealed, however there is always a worry that the patterns would not be related to phenotype. Moreover sML uses many of the features in combination, thus the question becomes, are these learned patterns biologically relevant and how can we test this?

One of the directions to use when laboratory testing is not available due to complexity of the factors is to increase and diversify your dataset in order to blend out any of confounding signals. The one who is preparing the training dataset should ensure, where possible, that samples for any label come from different locations and in similar numbers, build a classical phylogeny of the collection and ensure that the isolates come from different phylogenetic branches, etc.

Another solution is to ensure at least a minimal number of samples with highly trustworthy metadata. These can be used in reinforcement ML methods when initial training is on a small sample, then unlabelled data is added and the algorithm trained and checks its learning against trustworthy samples.

To conclude, ML contains powerful methods that allow analysis of complex dataset, however, in order to use some of these techniques knowledge and caution are needed.

The different sML have different advantages. Advantages of SVMs: High accuracy, nice theoretical guarantees regarding overfitting, and with an appropriate kernel they can work well even if you're data isn't linearly separable in the base feature space. Especially popular in text classification problems where very high-dimensional spaces are the norm. Memory-intensive, need specialist knowledge to run and tune, hard to interpret, though. So I think random forests are starting to steal the crown.

Chapter 4

Final discussion

The detailed results of this study are described in the previous chapters and the main findings as well as some limitations are highlighted below. The objective of this work was to find out if genetic markers associated with different hosts can be identified from WGS of bacterial isolates. Core (Chapter 3.1) and accessory genome (Chapters 3.3, 3.2) were analysed and different methods that can be used for host identification were tested (Chapter 3.4). The algorithm to assess if some of the bacterial strains may represent an increased threat for zoonotic infection was also developed and tested (Chapters 3.3, 3.2).

Both *E. coli* and STm are enteric bacteria living in a wide variety of hosts and some strains are able to cause disease in humans. To date, both of these species were mainly considered as generalists capable of thriving in different hosts. Nevertheless, there are a few examples that demonstrate that some sequence types have become specialised human pathogens (i.e ST131 for *E. coli* and ST313 for STm), as well as some *E. coli* phylogroups (B1, A, C) which are more likely to be source of clinical infections, some *E. coli* serogroups such as O55 to date are usually only isolated from humans.

Although, with a few exceptions, belonging to specific sub-types of the primary typing schemes is not indicative of being associated with a particular animal host, these typing schemes are still valuable for a variety of analyses. For example, phylogeny based on MLST genes correlates well with whole core

phylogeny(at least for STm and *E. coli*) and therefore can be done in a fraction of the time and with fewer computational resources. Serotyping in Salmonella species is very indicative of phenotypic differences between serovars (i.e. Typhi vs Gallinarum vs Typhimurium), however within a particular serovar, in this case Typhimurium, a wide variety of strains could be found; those that differ by lifestyle i.e. isolated from different hosts, as well as genotypically (above 20,000 COG clusters from 1203 STm isolates). For *E. coli* some serogroups are indicative of pathogenity i.e O157, O26.

The main success of the core genome analysis was the translation of some molecular techniques (i.e. serotyping, sequence typing, phylogrouping) to an in-silico approach. When WGS is available, in silico typing provides, arguably, a more accurate result and (once the sequences are available) this takes less time and is less expensive. Over the time of this PhD, in silico typing has become widespread and multiple software and databases have become available to support thess types of analyses.

Most of the work during this thesis wa carried out by accessory gene analyses and at this stage it is difficult to compare these two parts and decide if core SNPs also could be used to predict the host of isolation. Therefore, a follow up study can be proposed to investigate further correlation of core gene changes/markers with the ability to thrive in a new host. Such a study can be

based on the following steps:

- GWAS on the core SNPs to extract SNPs associated with particular hosts.
- Build a machine learning classifier to find if the SNP patterns can be identified and bacterial host can be inferred from these.
- Compare core and accessory classification results.
- Check if core and accessory results can be combined for improved accuracy.

On the chapters 3.2, 3.3, using machine learning, differentiation by host was achieved, however, results still may not be valid. Supervised machine learning is a powerful technique for finding commonalities amongst same-label datapoints. To ensure that these commonalities are biologically relevant, validation of findings through biological tests are now needed. Some questionable areas of this research and possible ways to address these are highlighted below.

- In all models for this work predictions were based on multiple PVs (in different sub-datasets from 70 to 600). It will be challenging to design biological experiments to test all possible PVs identified, however, to some degree, validation of predictive genes can be achieved by using TraDIS (Transposon Directed Insertion Site Sequencing) data [161] [162]. Testing bacterial fitness by randomly perturbing genes as with TraDIS and then comparing TraDIS data with that obtained by ML would allow iden-

tification of important genes that enable bacteria to succeed in particular environments.

- It is very likely that not a single gene but a combination of genes are acting together to improve bacterial chances for survival in a particular environment. Pathway analysis could help to identify genes that are acting during different stages but impacting the same pathways, so shedding light on mechanisms of adaptation. However, pathway databases are still in their infancy, so at this stage such an analysis would be only partial.
- The threshold for success of predictions was placed at a 'midway' point between the extreme value probablities (0,1), however for success in a host that threshold could be different. It is also possible that the threshold would depend on what bacterial species is being studied and the host in question. Competition experiments for bacteria with different scores from different hosts could help to adjust a threshold in a ML model. Moreover, as discussed in the previous chapters higher and lower scores for bacteria from the same host could indicate a presence of specialist and generalist bacteria. To confirm this idea we could test isolates from 2 hosts with middle and high scores (i.e. 0.5, 0.9). We would expect isolates with 0.5 to colonise/survive equally well in any host, and isolates which scored 0.9 may survive well only in their original host. Moreover, specialised isolates should perform better in their original host than generalists.

- It should be remembered that these predictions were based only on genetic content, and we don't know at this point which of these genes are active and when. Therefore, an experiment where complex data (including genomic sequences, transcriptomes, proteomics) are gathered together would allow for a better understanding of adaptation/colonisation processes in bacteria.
- It is also not clear when (how quickly) adaptation related changes happen, so a resequencing experiment could shed light on this area. In essence one would choose a 'generalist' from cattle (or even 'generalist' from another host), sequence it, passage bacteria through a cow a few times and sequence excfreted isolates at differetn time points. These times series resequencing experiments should demonstrate changes that can happen while bacteria are trying to adapt. Moreover, this experiment could possibly clarify if the 'generalist' bacteria are a constant part of a host bacterial population or whether generalists are bacteria that have undergone a recent host jump and is are on their way to becoming a specialist in another environment/host.

In the last decade democratisation of sequencing and wider availability of high performance computing have enabled analysis of huge genomic datasets. We now better understand bacterial population dynamics as well as genomic structure and its modifications. There is also the growing realisation that reductionist approaches (aka one gene - one disease) often do not represent complex bio-

logical reality. Combining genomic and phenotypic data on a population scale and using machine learning to discover underlying patterns has shown promising results for bacterial host attribution. It is clear that machine learning with a combination of 'omics' and phenotypic data could help to elucidate many other questions in bacterial genomics. So this work and methods developed during this thesis can lead to development of various other projects.

One such project is prediction of pathogenicity of a bacterium. Urinary tract infections (UTIs) affect 150 million people worldwide; a minority of cases cannot be completely resolved by antibiotics and reoccurr. Moreover some of these UTIs progress and develop bacteraemia and sepsis. In England alone there are 5.000 *E. coli* bacteraemia related deaths per year. We (Prof. D. Gally group) are preparing a research grant application where we propose comparing thousands of *E. coli* from the broader phylogeny - commensal *E. coli* from a healthy population, E. coli from UTI cases and from bacteraemia and then build a classifier based on these groups to enable prediction of the potential of UTI strains to become bacteraemic. This would have value in informing which UTI infections to treat.

Another ongoing project for which I am developing a ML model, is a prediction of phage resistance and susceptibility of bacteria. The antimicrobial resistance crisis demands urgent and sustainable alternatives to antibiotics. One such solution is phage therapy. Most phages are strain specific and there-

fore should not disturb the patient microbiome in the way that antibiotics can. On the other hand that specificity is one of the main reasons why phage therapy still is not used worldwide - it is time-consuming and laborious process to find the right phage(s) for a particular infection. Therefore, if the big enough dataset of bacteria - phage interaction could be gathered, ML could be used to inform decisions about which phages to use. Moreover, incorporating into the model more phenotypic information should allow selection of different 'infection' mechanisms and therefore inform choice of the best combination of phages for bespoke cocktails that will avoid resistance.

To finish up, it is an exciting time to be a bioinformatician right now: affordable sequencing, available computing and powerful machine learning algorithms are aligned and ready to be applied by anybody who is keen to help humankind to leave it a safer place or even just for curiosity to uncover the wonders of biology.

Appendix

The Appendix includes a First year report submitted to the university of Edinburgh by Nadejda Lupolova. The report describes the work that was performed over the first year of this PhD and it mostly concerns finding the optimal parameters for running different software as well as the analysis of some isolates that were found to be outliers when compared to the main bulk of *E. coli* isolates. The original report is bound here.

Bioinformatic characterisation of *Escherichia coli* from different hosts and environments.

Nadejda Lupolova

9 month report submitted to The University of Edinburgh

July 14, 2015

Acknowledgements

I would like to thank my infinitely patient and understanding supervisor, Prof. David Gally, for his trust, guidance and support during all stages of this work.
I thank Dr Andy Law, Rodrigo Bacigalupe and Sharif Shaaban for their time over bioinformatics discussions.
We are grateful to all who helped gather our collection and provide us with sequences: Dr Tim Dallman, Public Health England, Prof Nicola Williams, University of Liverpool, Prof Roberto Laragione, University of Surrey and Geoffrey Mainda, The Roslin Institute.
Finally, I would like to thank Dr Nicola Holden, Prof Ross Fitzgerald, Prof Mark Stevens and Dr Ross Houston for agreeing to be on my Thesis Committee and their comments to improve the research being performed.

Contents

Chapter 1

Abstract

There is an extensive history of research aimed at differentiating pathogenic and commensal *Escherichia coli* strains and this has led to a 'pathotype' classification usually based around expression of specific virulence factors important for colonization or pathology. Over the last twenty years it has become apparent that these virulence determinants are often horizontally-acquired on large regions of DNA, usually as integrated bacteriophage (prophage) or plasmid regions. With the advent of relatively low cost whole genome sequencing (WGS) techniques, it is now possible to obtain sequences from large numbers of *E. coli* strains and interrogate these in relation to both their core and accessory genomes. While it is more or less accepted that a subset of Salmonella strains often have preferred hosts, especially in terms of disease, there has been no real systematic investigation of host and niche specificity associated with *Escherichia coli* strains, despite the fact that this species can be found to have adapted successfully to many host species and environments.

The main aim of this project is to determine if host and/or niche-specific genes can be identified in *E. coli*; from this it should then be possible to predict both the 'origin' of a strain and its zoonotic potential from its sequence. It is anticipated that any host restriction probably also requires appropriate gene regulation of these 'accessory factors', i.e. the correct genome context, with the hypothesis that this is why relatively few *E. coli* strains are truly zoonotic. In order to do this, it is necessary to first analyse the diversity present in *E. coli* genomes from different hosts and environments. This work focuses on *E. coli* strains from humans, dogs, birds (chickens, ducks, turkeys), plants/roots and cattle and uses Illumina short read sequences to interrogate both phylogenetic relationships and to identify genes that show significant associations with strain origin. Niche/host specific

gene expression by RNAseq will then be carried out to focus on how any host restriction may also relate to gene expression rather than just gene presence absence.

Chapter 2

Introduction

2.1 Biology

Extensive research has been carried out on *Escherichia coli* as a model organism; in particular K12 and B strains and their derivatives have helped advance our understanding of molecular biology, genetics and gene engineering. *E. coli* is a common commensal of the gastrointestinal tract in many mammals [1] and has been associated with a wide range of infections in both humans and animals with certain strains able to cause life threatening zoonotic infections. A long record of association with human and animal disease means that certain strains represent a health threat with significant costs to society. [2], [3], [4], [5]. It is evident that *E. coli* can thrive in a wide range of hosts and ecological niches and while it is one of the first species of bacteria to colonise the human gut [6], [7], [8], it can exist, at least temporarily, outside of its "primary" host habitat in soil, water, sediments, plant tissues [9], [10].

E. coli seems to succeed in different environments with diverse conditions including variation in temperature and pH, and in the face of challenges such as an immune or antibiotic therapy. Some strains appear highly adapted to a particular niche, for example it is now appreciated that Shigella strains can be considered as part of the *Escherichia coli* species and these, to date, have only been found in humans and primates [11] and can cause diarrhoea in humans [2]. Other strains such as *E. coli* O157 seem to be well adapted not only to a subset of hosts like cattle (and perhaps sheep), but to a specific niche in the intestine of these ruminants [12]. Other strains, for example porcine, bovine and human enterotoxigenic *E. coli* are considered relatively host specific but this has been attributed to specific combinations of adhesins, although there is no

evidence to suggest they could switch hosts solely based on exchange of these colonization factors.

The first stark indication of the diversity in *E. coli* followed the sequencing of a strain of *E. coli* O157:H7 and its comparison with *E. coli* K12. From this, 1,387 'new' genes were identified with the finding that *E. coli* O157 had a significantly larger genome (5.5 Mb compared to 4.2 Mb) than the K12 strain. The *E. coli* O157:H7 strain possessed specific virulence factors as well as different metabolic capacities [13]. On the other side, some studies show that there is much less diversity among certain *E. coli* pathovars than previously anticipated [14]. Overall, it is now clear that the *E. coli* genome can exhibit incredible plasticity associated with horizontal gene transfer by bacteriophages and plasmids. As a consequence, these changes do not necessarily need many generations to consolidate and therefore can be associated with the rapid emergence of a virulent strains, such as occurred in the 2011 enterohaemorrhagic *E. coli* outbreak in Northern Germany [15]. An *E. coli* genome can perhaps be altered endlessly leading to appearance of new strains and/or rapid specialisation of existing strains. This same plasticity is important when considering adaptation to human interventions such as antibiotic treatment. Such selective pressures help to develop, maintain and potentially combine resistance and virulence traits.

The pathogenicity of strains also can be described by distribution and expression of specific virulence factors as toxins adhesins, invasins and others that are encoded by either chromosomal genes, plasmids, or pathogenicity islands. Pathogenic *E. coli* strains do not have a single evolutionary origin but may have arisen many times [16]. There is also suggestion of the possibility that any *E. coli* strain acquiring the appropriate virulence factors may give rise to a pathogenic form. However, it is debatable if every strain can acquire any virulence factor and this also makes huge assumptions about the regulatory and other networks required to work in a co-ordinated way in a successful pathogen.

The major argument in favour of the theory that not any strain will acquire any virulence factor is the evidence of division of *E. coli* in to phylogroups. So, based on multi-locus enzyme electrophoresis (MLEE) Ochman and Selander gathered a collection of 72 isolates from different mammalian hosts to represent the diversity of *E. coli* [17]. They noted that a small number of core genes can be used to organise *E. coli* phylogroups. Importantly, certain phylogroups are known to contain the majority of strains that are pathogenic in humans (B2 and D) while others

are known to contain many animal and human commensals (A and E). This indicates that there are core lineage differences in the evolution of virulent strains arguing against the idea that acquisition of key virulence factors into 'any' background can produce a pathogen. As such it is a much more complex issue trying to determine which strains might emerge as significant pathogens by simple acquisition.

Rather than focusing solely on virulence, the primary aim of the research proposed here is to ascertain if it is possible to predict the likely 'source' of an *E. coli* isolate based on its WGS. The hypothesis underlying this research is that for any *E. coli* strain, it has evolved to replicate optimally in a specific environment. This does not preclude it replicating and being 'successful' in multiple environments, but that each strain has an optimum niche. Extrapolating from this, it is likely that certain strains may have a more generic capacity to succeed in multiple environments than others based on their 'primary' habitat. These generalists may have an increased capacity to transfer between animal speciesand, depending on the factors they express, pose a zoonotic threat.

It is important to understand the host restrictions that limit strain transmission and then the further issue of what is required to induce disease in a new host. A good example of this are *E. coli* strains that contain Shiga toxin encoding bacteriophages. There are multiple types of *E. coli* that contain these prophages in ruminants, but only a relatively small subset cause an issue to human health. It is proposed that this is because those strains require both appropriate colonization factors for the human gastrointestinal tract and signalling/expression systems to switch on the system.

Therefore it is proposed that any host restriction of *E. coli* will be a combination of the need for:

- Specific colonization and catabolic functions that are evolved to the host/tissue.
- Appropriate regulation of these factors by 'long term' adaptation of core regulatory networks.

2.2 Technology

This project aims to work with a large and varied set of strains to examine both genome content and expression differences that may be related to host restriction.

Advances in technology that improved our ability to sequence and analyse whole genomes opened new horizons for diagnostics and prevention of diseases. The major advance is a whole genome sequencing (WGS) that now is widely available for use in science and clinical microbiology. Modern technology is PCR-free (thus excluding PCR introduced biases) and enables the performance of millions of simultaneous reactions producing gigabites of data in a single run. The data consists of a short reads usually 150 - 300 nucleotides (nt) long that can be reassembled using known genome reference or assembled de-novo using advanced computational techniques.

The main problem that may appear while using short read sequencing is caused by the fact that the original position where a short read comes from is lost, so during the alignment as well as de-novo assembly the unique position where the read belong should be found. Thus, if a length of short sequence does not cover whole repetitive region there is a possibility that all similar sequences would pill up in the same place. On the other side it is counter productive to apply too strict rules in mapping as some mismatches should be deliberately allowed, to account for genetic variance between reference and aligned sequences. Moreover, sequence technology itself produces errors, that should be separated form natural variation later in analysis that sometimes can be impossible.

The newest technologies such as PacBio and MinION can produce much longer reads from 14,000 to 40,000 bases, although some factors limit usage of this technology. Thus, PacBio showed extraordinary sequencing accuracy of 99,9% that is even greater than a 'Gold standard' Sanger technology. However, cost of the genome sequencing by PACBio stays relatively high, limiting use a massive scale. MinION on the other side seems to be more accessible, but due to the fact that this is recent technology error rate of sequencing is still high. Nevertheless, advances not only in sequencing technologies but also in software development allow extract valuable information using short read Illumina sequences.

As a well studied, model organism *E. coli* has a large collection of previously sequenced

genomes many of which qualified to be a reference genome i.e. represent 'the highest quality dataset that is supported by curation by NCBI scientific staff and by collaborators'. Thus, one of the methods used in this work was reference based analysis when each newly sequenced genome is aligned to and compared with the reference, inferring relationship between strains based on sequence similarity of genome regions present in all.

Schemes that use genes that are present in all sequences have long been adapted for 'typing'. From early work on 16S rRNA [18] and gyrB based phylogenetics [19] to development of new schemes such as MLEE, MLST, MLVA and PFGE [17], [16], [20], [21], these kind of methods facilitate sharing and comparison of results between researchers, public health surveillance and outbreak investigations and serves as a means to reveal substructures of *E. coli* that sometimes can explain prototypical variation or shed light into the origin of the strains. However, only a miniscule amount of the genome information is used with techniques such as MLST and they leave behind valuable material that may contain crucial information about strain specific features.

So where one should look for adaptation traits?

It is not clear yet and should be investigated further to what extent core genes are involved in adaptation processes or associated with acquisition of virulence factors. Looking for another bacterial example such as *Staphylococcus aureus*, SNPs variation in a core region can be strongly associated with toxicity [22]. However, *E. coli* can undergo massive recombinations events [23], [24], that obscure phylogenetic relationships and association with ST or habitat. On the other hand we know that some segments of *E. coli* genome evolve clonally, the chromosome structures continue to be stable and insertions and deletions are mostly found in chromosomal hotspots [25], [26], [27].

According to the idea of genomic continuum [28] bacteria would absorb and discard genes dependent on particular circumstances and selective pressure. Thus the accessory genome is more likely to contain traits of sequence adaptation. For a long time, variable gene content except for virulence factors has been left behind without proper analysis. Accessory genome provides new information that was difficult to access, analyse and quantify until now. Accessory genes can be unique to a particular strain or can be shared between few or many but not present in all.

The differences between strains can lay not only in genome content but in how it it used. Therefore, it is planned to perform competition assays and/or forced evolution experiments with a subset of strains in order to quantify gene expression changes that can help to understand whether there is any host related associations. Apart from being an invaluable tool for analysis of bacteria with a large variation between strains, RNAseq also will improve delineation of untranslated regions and improve existing annotations.

2.3 Objectives

The list below describes steps that have been or will be taken in order to investigate host/niche adaptation of *Escherichia coli*.

1. Study the evolutionary relationships between strains with known provenance.
2. Explore similarities and differences in core and accessory genomes.
3. Test whether we can be confident about these findings. This will involve:
    - In silico verification of regions of interest from the accessory genome analyses.
    - Collaborative in vivo experiments examining short term acquisition and loss towards host adaptation.
4. Test impact of specific horizontally acquired regions in regulation of gene expression in *E. coli* by RNAseq.
5. Determine if a subset of strains pose an increased zoonotic risk due to their capacity to thrive in different environments; i.e. have evolved more as generalist.
6. Develop algorithms for prediction of strain origin.

Further strains may be added to the analysis, in particular UK bovine. The in silico methods established may also be applied to Salmonella enterica to determine if any overlap is detectable.

Chapter 3

Methods

3.1 Data

The current collection of *E. coli* strains are isolated from five different host/environments and consists of 564 strains. The list below summarise the collection.

- 1: Avian: chicken (39 isolates), turkey (6), duck (5)
- 2: Bovine: Zambian (135), UK (99)
- 3: Canine: Multidrug resistant (MDR) (18), community (19)
- 4: Environmental: grain (3), roots (5), soil (2), compost (3), slurry (2)
- 5: Human: UK (203), reference genomes (23)

To distinguish between sub-datasets the following abbreviation are used: A - avian, C - canine, E - environmental, R - human isolates downloaded form NCBI, H - all other human, W - UK bovine isolates of sequence type (ST) O157, Z - bovine isolates from Zambia. Words 'sub-group' and 'sub-dataset' in this work mean a group of strains related by a host.

The pathotype for some isolates is known and numbered in **Table 3.1**. Metadata gathering for the collection is still in progress. It is planned to identify the presence of resistant and virulence genes and report sequence type (ST) using SRST2 software [29].

Main steps that was taken to explore variation between strains in our collection are illustrated in **Figure 3.1**. This workflow is planned to be finished by the end of the first year.

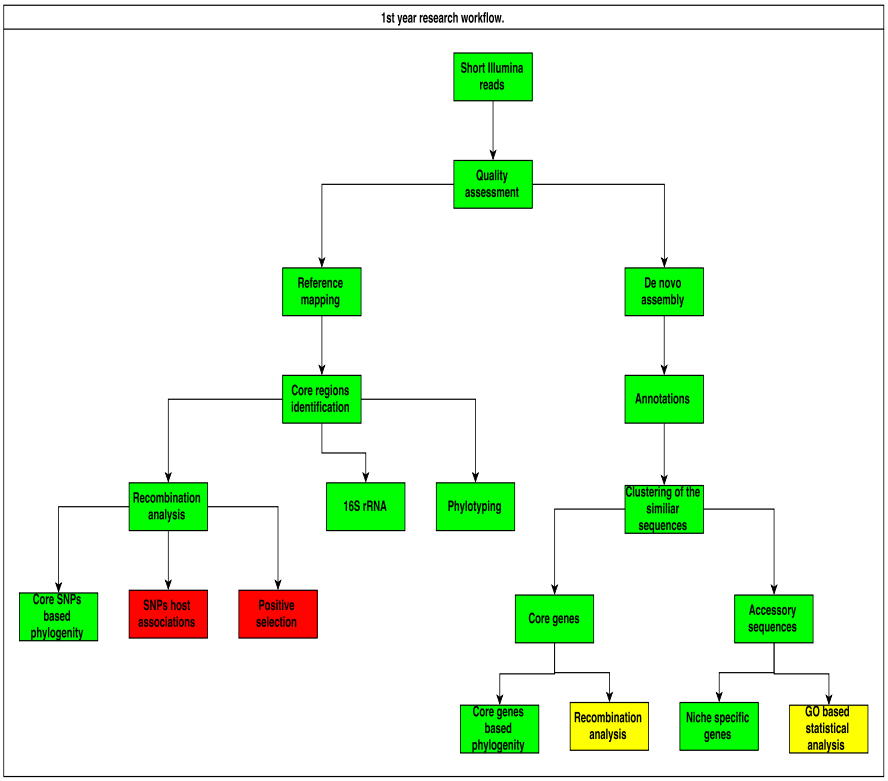


Figure 3.1: Steps taken over the first year of this work to explore diversity of the collection of *E. coli* genomes. Green squares indicates finished work, yellow - work is in progress, red - work is planned.

Table 3.1: Known Pathotypes

STEC	EAEC	ExPEC	UPEC	ETEC	AIEC	EPEC	Commensal
160(H)+97(B)	24(H)	30(A) + 4(H)	4(H)	3(H)	1(H)	1(H)	4(H)

3.2 Quality

The data were received in Illumina1.9, paired-end, .fastq format [30]. Raw reads vary in length from 32 to 251 bp. Fastqc [31] was used to perform quality control, revealing that majority of data were good quality and did not need any further processing. All datasets were checked for the presence of adaptors and some of the datasets (H- and W-) were trimmed with cutadatp [32].

3.3 Reference mapping

The reference quality genome of *E. coli* O157:H7 str. Sakai (RefSeq assembly accession: GCF_000008865) was chosen as reference genome for variety of methods used on this work.

Alignment of short reads to the reference genome were performed by combining BWA [33], [34], SAMtools [35], SnpEff [36] in a custom made python script. The script takes as an input paired-end short Illumina reads, aligns them to the reference genome and stores variants into the .vcf files. Resultant .bam files store short read multiple alignments. Then consensus sequence for each alignment is produced by taking the most common base from vcf and bam files at each position.

All resultant sequences .consensus.fa of the length 5 590 092 bp were concatenated into one multifasta file that served as an input to the custom made python script core_finder.py that checked presence or absence of a nucleotide (nt) for each sequence in each position. Starting from the beginning, if at the position one nt is present for each strain then this position called core position and is written to the separate file together with a nt at this position for each strains. If at any given position at least one nt is absent that position called non-core and thrown away.

The same algorithm was applied when core single nucleotide polymorphisms (SNPs) were calculated. At each given position if there is variation at least in one base then the position is

called a SNP and is written to another file.

3.4 Assemblies

SPAdes [37] was used to assemble short read sequences. After benchmarking a variety of options it was clear that better results are produced when builtin error corrector is used instead of the recommended QUAKE [38] error corrector. To control mismatches and indels 'careful' flag proved to produce better result. Quality of assemblies were evaluated with QUAST software [39] - Quality Assessment Tool for Genome Assemblies. A wide range of statistical data from QUAST output for all assemblies were compiled into a spreadsheet describing assemblies in different parameters such as length of assembly, number and length of contigs, GC%, N metrics(N25, N50, N75), misassemblies report, number of Ns per 100kbp, gene statistics (unique, duplicated, genes larger than certain threshold).

3.5 Annotations

Annotation was carried out with PROKKA [40] - prokariyotic genome annotation software. To achieve a better quality of annotation, a database with trusted proteins, in form of multifasta file, were gathered from reference quality *E. coli* genomes from NCBI.

3.6 Phylogenetic analysis

Maximum likelihood (ML) trees were constructed using RAxML [41].To optimise the best tree search program was run with 500 rapid bootstrap (BS) following with slow Maxumum Likelihood (ML) search under the GAMMA model of heterogeneity.

To reconstruct phylogenetic relationship two different approaches were used. An earlier attempt at core SNPs alignment described in the reference mapping section were used as a input to RAxML.

For the second approach core proteins presented in 100% isolates were extracted from all sequences, aligned with MUSCLE, translated to nucleic acid with EMBOSS Backtranseq [42]

and Pal2Nal [43], sorted using bash custom made script, concatenated and used as a input for Gubbins software [44] - Genealogies Unbiased By recomBinations In Nucleotide Sequences. With Gubbins recombination sites were excluded from the alignment, a tree obtained as an output from Gubbins was used as the Guide Tree for RAxML together with alignment of core proteins with filtered recombination sites. RAxML was run as described above with 500 bootstraps and following ML search. The trees were visualised with FigTree [45]

3.7 Phylotyping

The phylotyping scheme described by [46] was used as a starting point to develop a small program that assigns each strain to the one of the 4 possible phylogroups (A, B1, B2, D) based on the presence or absence of one of 4 genes chuA, yjaA, TspE4.C2, arpA. To further distinguish between groups and assign strains to an additional 4 phylogroups (C, E, F, cryptic clade I) it was necessary to check for the presence of a fifth gene trpA and/or for the presence of the specific alleles for the above genes. Performing all steps each *E. coli* sequence were assigned to one of eight possible phylogroups.

To extract genes fragments from the genomes, a database that includes all sequences from the collection were build with BLAST+ [47]. The reference quality sequences for querying genes were downloaded form NCBI website. Gene's identifiers are presented in a Table 3.2. Reference gene sequences were blasted against the database and blast output based on sequence length < 90% and E-value = 0 were filtered with a custom made python script which also maps filtered extracted genes back to the sequences they were obtained from and then assigned to phylogroup.

Table 3.2: Genes ID, for phylotyping

	arpA	chuA	trpA	TspE4.C2	yjaA
GI	556503834	15829254	556503834	7330942	556503834
Position	4222487-4220301	4391446-4389464	1317222-1316416		4213234-4213617

3.8 16S rRNA

rrsA 16S ribosomal RNA of the rrnA operon from *Escherichia coli* str. K-12 substr. MG1655 (gene ID: 948332, sequence NC_000913.3, coordinates 4035531..4037072) were downloaded from NCBI website to querying the database described in the previous section. Sequences with length more than 90% and E-value equal to zero were extracted from database and aligned with MUSCLE [48]. Alignments were visualised with Geneious [49]. Phylogenetic trees were build with RAxML, using the same parameters as described above.

3.9 Clustering

In this report some of the results were obtained from Get_homologues [50] software. However, Get_homologues cannot be used with a large dataset as this software struggles to be scalable. The amount of time and computer memory that the program requires means that calculation with over 200 sequences is not really possible.

New software for clustering Roary [51] was identified only a month ago and shows promising speed and accuracy. The software take as an input protein sequences extracted from .gff files from a PROKKA output. Cut off for the assignment to the same cluster was set 95% of similarity on amino acid level.

Visualisaton of pan-genome, calculation of genes per group and statistical analysis was done developing custom based scripts in MATLAB [52].

Chapter 4

Results and Discussion

4.1 Quality

Assess of raw and corrected reads by FastQC software demonstrated a good overall quality of reads. All reads passed N content sections, meaning there were less than 5% of uncalled bases. There were also no warnings in over-represented sequences, suggesting that these libraries, as expected, contain a diverse set of sequences; individually, each sequence makes no more than than 0.1% of the total pool. Warnings in these sections could indicate that the libraries are contaminated, but none were produced.

4.2 Reference mapping

The first attempt to map the short Illumina reads from 564 isolate sequences to the reference genome resulted in 134,743 core positions which represents only 2.41% out of 5,590,100 bp of the reference genome. Across the core positions 9,891 positions were found to be variable (SNPs). Based on SNPs patterns here, 167 sequences were absolutely identical. Similar sequences were identical to other sequences within the same sub-dataset, no sequence was detected with the similar SNP patterns across datasets. So there were 8 avian sequences out of 50 similar to other avian sequences, 11 Zambian out of 135, 132 O157 sequences out of 185 similiar to others O157 genomes, independent of host, 8 identical human non O157 STEC out of 98, 5 canine out of 37 and 2 reference genomes R-MG1655 and R-W3110 with identical SNPs pattern. The numbers of identical sequences per group confirm that O157 is a clonal

group with minimal dissimilarities and that Zambian sub-dataset is very diverse even at a core level. Also this indicates that the methodology works as the two K12-like strains R-MG1655 and R-W3110 were found to be identical in their core SNPs distribution.

The relatively small proportions of core positions could indicate either that there were mapping errors or that *E. coli* genomes are so variable that when large number are analysed together then it is difficult to find many similarities.

Due to an algorithm chosen to find core that removes any position that contains a gap, it is clear that gaps distribution is a key factor that influenced the core size. Surprisingly there were not a few outliers that dramatically influenced the core size but instead a trend of steadily increasing gap numbers depending on how far from reference any particular strain is **Figure** 4.1.

Genomes of the same ST as the reference (*E. coli*. str Sakai) posses around 5,000 gaps per sequence. Gradual increase from 20,000 gaps per sequence to 40.000 do not related to any particular sub-group of strains. The next increase in number of gaps from 40,000 to 54,000 mostly originated from NCBI reference sequences. The highest bars at **Figure** 4.1 originated by outliers described in section 4.5 (one avian, two environmental and one Zambian strain). Gap numbers for a strain were accessed independently of others strains, i.e. calculated only as bases that are absent in a strain in comparison to reference. We cannot be sure whether these strains with increased gaps numbers are outliers due to methodology error like sequencing, or they are from different sub population that we have not enough samples from.

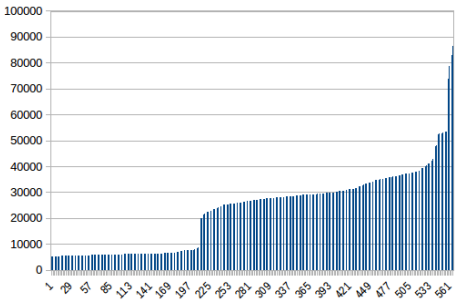


Figure 4.1: Number of gaps in consensus sequences; first 200 sequences are ST O157 which are similar to reference hence such small number of gaps were detected. Majority of the sequences with number of gaps between 40,000 and 50,000 are human isolated reference genomes.

When *E. fergusonii* is added to the analysis in order to use it as a root for phylogenetic inference the core proportions shrinking even more as each of the *E. fergusonii* .consensus.fasta files

bring additional gaps: 83457 for ATCC_35469 and 83211 for ECD227.

4.3 Phylogenetic analysis

First attempt to reconstruct phylogenetic relationships resulted in a tree presented in a **Figure** 4.2. The tree is based on alignment of the 564 sequences, that share 134,743 core positions and have 9,891 SNPs. 167 sequences were absolutely identical.

Approximately, the tree can be divided into six distinct clusters. Human isolates are spread across all clusters, however very few are presented in cluster number six. Tight cluster number three is formed by human and bovine isolates of ST O157. Most likely all the sequences in this cluster are of bovine origin as there is zoonotic spread from ruminants to humans. Cluster number five is composed of isolates from all host types, canine UTI strains lay in that cluster in close proximity with human pathogenic strains. Canine isolates can be also found in cluster number one with sequences that are genetically similar to K-12 strains. The majority of Zambian strains belong to cluster number two, however a few zambian bovine isolates can be found in each cluster. The cluster number four with long branches raised questions about species identification and lead to work described in part 4.5; this is formed by one avian A-DK8, two environmental E-5038, E-5088, 4 bovine and one human strain.

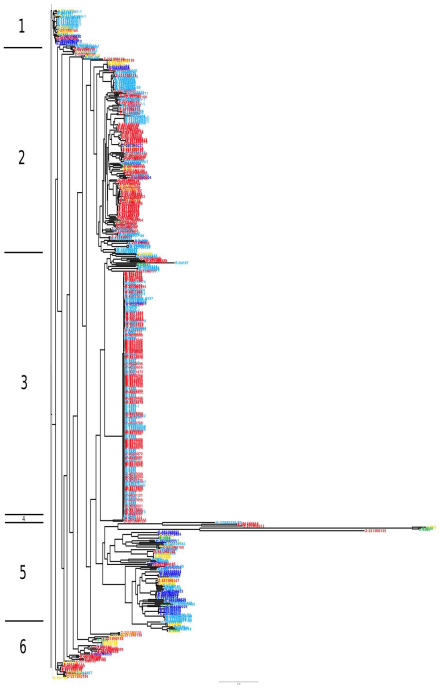


Figure 4.2: Core SNPs tree are coloured according of host: Avian - yellow, bovine - red, canine - blue, environmental - green, human - turquoise.

There were some evidence [53] that reconstructing phylogeny based only on concatenated SNPs may produce erroneous results as well as confuse the choice of the right substitution model and can produce misleading results, such as wrong topology and incorrect branch length. Thus for a second attempt invariant sites also were included into analysis. There was expected changes in a branch length detected but no differences in topology between the two runs. Moreover it was decided to exclude sequences with a number of gaps higher than 60,000, that increased the effective core size to 950,000 base pares with 83,522 variables sites. Again, no topology changes were detected. The tree is presented on **Figure** 4.3 Even after exclusion of strains with number of gaps higher than 60,000 the core obtained is still smaller than were descried in previous works estimating that sligtly less than half of the *E. coli* genome are expected to be a core [3], [54], [55].

4.4 Phylotyping

The evidence for genetic substructure of *E. coli* accumulated in the 20th century has lead to development of a method that can easily and inexpensively assign *E. coli* isolates to certain phylogroup. These phylogroups have been shown to have different characteristics in particular disease and commensal associations. Initially method was based on the detection of two genes and one genetic fragment dividing all strains into 4 groups, later method was improved allowing 7 phylogroups and one cryptic clade to be distinguished [27], [56], [57]. The downside of the method is that it is still relatively time and labour consuming as well as subject to PCR related errors. With advance of NGS and a wave of available whole genome sequences it was decided to write a small program that can check for presence of the key genes in order to even further accelerate phylotyping assignment of any strain.

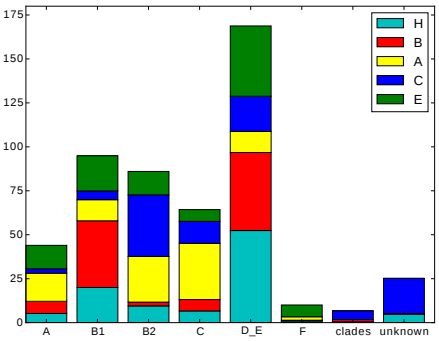


Figure 4.4: Percent distributions of hosts in phylogroups. Colours represent host: Avian-yellow, bovine-red, canine-blue, environmental-green, human-turquoise

The results of in silico performed phylotyping shows that only 8 out of 564 strains failed to be assigned to any phylogroup. These can be explained by assembly failures as a search for genes done in assembled genomes. Nevertheless, the proportion of bioinformatically detected phylogroups is still higher than the proportion of assigned strains using PCR. So for 40 Zambian strains on which PCR based typing was performed 2 were untypable, while in silico all of these were assigned to one or another phylogroup.

There are 16 possible combination of genes of which only 12 were detected previously [46]. In our collection 15 combinations were detected including 19 strains from previously 'unknown' phylotypes. Further scripting is required to distinguish between D and E phylogroup and between cryptic clades as their detection are based on allele differences.

Figure 4.3 and **Figure 4.4** demonstrate results of in silico phylotyping. Phylogroups fit well with core SNPs based cluster divisions. However, there are some inconsistencies. Thus the

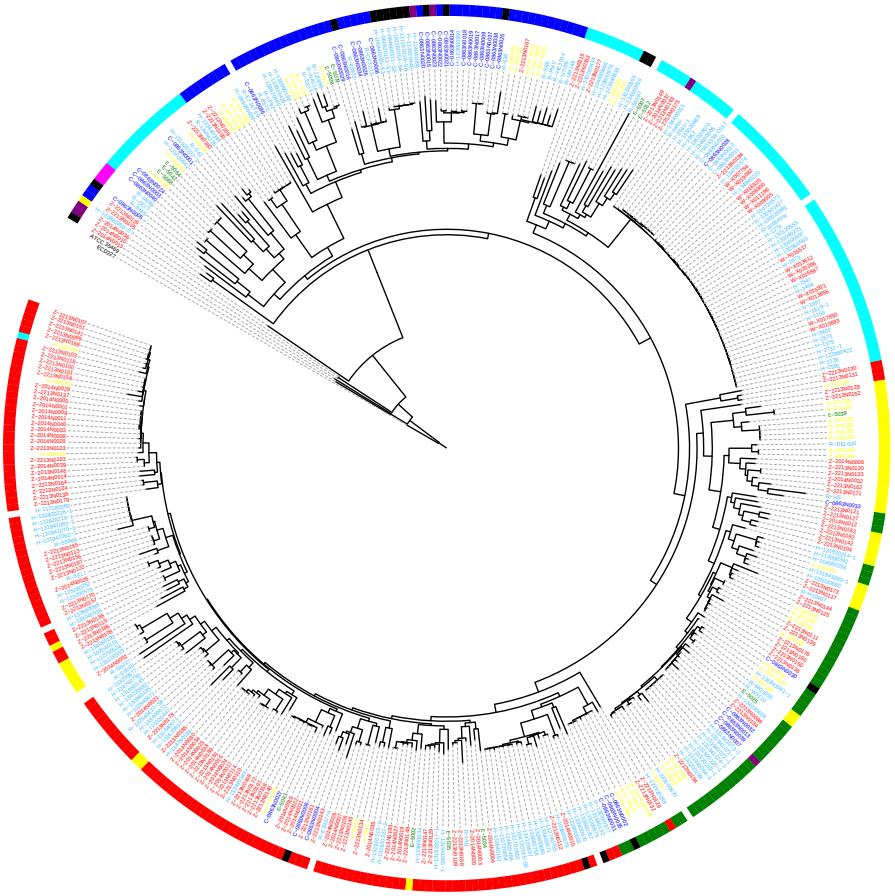


Figure 4.3: Core regions tree coloured according of host. Avian-yellow, bovine-red, canine-blue, environmental-green, human-turquoise. Outer circle represents phylogroups: A - yellow, B1 - red, B2 - blue, C - green, D and E - turquoise, F - pink, clades - violet. Strains failed to be identified (8) left white, those that belong to 'unknown' - black

most disperse is the A phylogroup for which strains appears in several clusters. Strains in the long brunched cluster 4 pretence to different phylogroups, which is expected as all strains in this clusters are very different between themselves as well as in comparison to other *E. coli* in the collection.Distribution by host shows that human isolates are present in all phylogroups except for cryptic clades with majority in phylogroup D_E, which is ex[[ected as a over one third of human isolates are of the ST O157. There is almost half as much human isolates in phylogroup B1 (prevalently commensal) compared to B2 (mostly pathogenic). 35% of dog and 26% of avian isolates are also assigned to phylogroup B2. In absolute numbers, phylogroup C is composed by equal part of human, bovine adn avian isolates, however in proportion it is the most popular phylogroup for avian isolates 40% of which assigned to this group. All bovine O157 is in phylogroup D_E, and two-thirds of bovine Zambian strains is in phylogroup B1 with one-third spread arounnd all other phylogroups. Previously not detected combination of genes are presented in a **Table 6.1** in the Supplementary material section

4.5 Outliers

All sequences in our collection were originally isolated as *E. coli* based on laboratory detected phenotypes (lactose +, citrate -, indole +), however, we have isolates from many different sources so I think we can expect variation at this level of laboratory testing. Raw short read sequences for strains that form cluster 4 in a core SNPs tree **Figure 4.2** were the same good quality as for all others sequences in the collection. Thus the question was addressed whether these sequences are from *E. coli* and fit within species limits or whether these are not *E. coli*. It was decided to perform a number of in silico experiments to see if it is possible to identify these sequences as *E. coli*.

4.5.1 ANI

DNA-DNA hybridisation methods have been used since the 1960s to indicate sequence similarities, however due to the laborious work-flow involved and difficulties with comparison of the data obtained [58] a new method based on Average Nucleotide Identity (ANI) of sequences has been designed. The method compares genomes pairwise, each in turn serves as a reference while another is chopped to "windows" of a desired size - in this analysis I have used 700nt 'windows', then these smaller sequences are mapped to the reference, the best blast score is noted and then average nucleotide similarity is calculated. It is expected that ANI between

sequences of the same species will be above 95%.

ANI comparisons were carried out between an 'outlier' duck strain A-DK8 and a 'normal' *E. coli* duck strain A-DK20; between two outliers A-DK8 vs environmental strain E-5038 and between *E. coli* and bacteria of a different species *E. fergusonii*. As a control two *E. coli* sequences were used that belong to the most distinct clusters from each other in the core SNPs tree (MG1655 and Sakai) **Figure 4.5**.

It is clear that the method works and that even between the most distinct sequences on the tree MG1655 and Sakai 98% sequence similarity was shown. From the analysis it could be concluded that the outliers are not *E. coli* species as they only have 90.5% similar, not related to *E. fergusonii* either with an ANI of 88.6%. Nevertheless, the outliers are more related to each other than other sequences of the collection with an ANI of 99.32%.

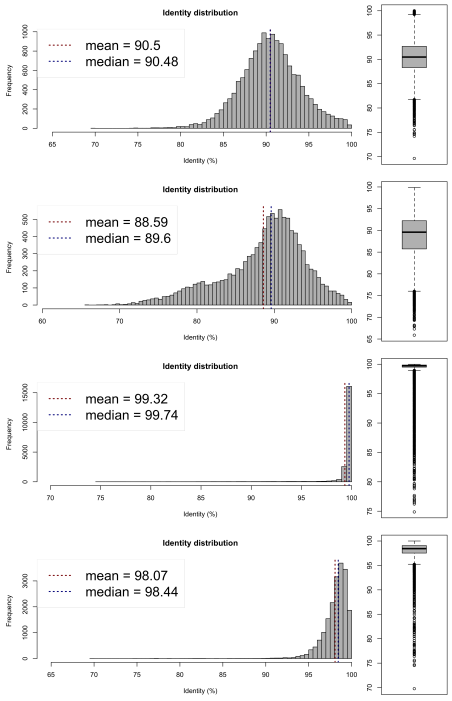


Figure 4.5: Average Nucleotide Identity (ANI) calculations performed for some outliers from cluster 4 in a **Figure 4.2**.
From top to bottom: (1) A-DK8 vs A-DK20, (2) A-DK8 vs *E. fergusonii*, (3) A-DK8 vs E-5038, (4) MG1655 vs Sakai.

4.5.2 16S rRNA

Another method to decide whether the outlier sequences are *E. coli* or not is to conduct a comparative analysis of the 16S. From time when Archaea was defined as a new domain of life using 16S ribosomal RNA phylogenetic taxonomy [59], this method became widely popular in phylogenetic due to slow rates of mutations in this region. Multiple sequences of 16S rRNA can be presented within a single bacterium (in case of *E. coli* there are 7).

Blast search resulted in 646 sequences between 1428 and 1551 nt in length. 506 sequences were exactly the same size as the query sequence. Aligned sequences produced a

consensus (1,570 bp) which contains 1,396 identical sites (88.9%) across all sequences. Pairwise sequence similarity between any 2 sequences was 99.5%, indicating that all sequences are *E. coli*.

ML tree based on 16S rRNA sequences demonstrate quite different results than core SNPs based tree. Outliers from the core SNPs tree do not seems to be outliers on a 16S rRNA tree **Figure 4.6**. They lay in between other 16S rRNA sequences and also share 99.5% of pairwise sequence identity with any other sequence from the analysis. 16S rRNA sequences from *E. fergusonii* str. ATCC 35469 and *E. fergusonii* str ECD227 (GenBank assembly accession: GCA_000026225.1 and GCA_000191665.1 respectively) were also added to analysis. Interestingly, these strains also lay amongst other *E. coli* strains with 99.5% pairwise sequence similarity **Figure 4.6**.

For many years 16S rRNA has been an attractive and easy method for species identification. However, the method can lack resolution when sequences share 99.5% 16S identity but have less then 50% sequence similarity over the whole genome as for Edwardsiella [60]. The method also has no power to distinguish between recently derived species [61]. Moreover, no well defined criterion for percent similarity cut off exists for 16S rRNA. In the majority of studies 99% of similarity used to establish species match. However an example provided by Janda et al describes Aeromonas veronii with a genome that contains multiple copies of 16S rRNA that vary among themselves by up to 1.5%. Clearly WGS will pose many challenging taxonomic issues including species definition.

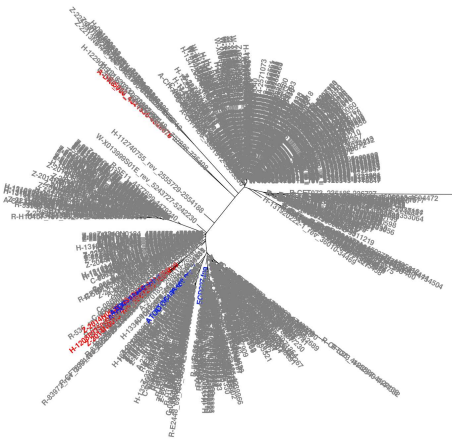


Figure 4.6: 16S rRNA tree, based on the alignment of 646 sequences of length between 1428 and 1551 nt; Red - the outliers from cluster 4 on the **Figure 4.3**, Blue - 16s rRNA from *E. fergusonii*

In this study the 7 copies found in K-12 MG1655 were compared between themselves to provide a baseline for comparison. So all seven genes code for 16S ribosomal RNA called rrsA,

rrsB, rrsC, rrsD, rrsE, rrsG, and rrsH were extracted, concatenated, aligned and visualised with Geneious. All sequences had 1542 bp length and formed a consensus of the length 1543 bp with 1522 identical sites across all sequences which is 98.6% of the consensus length. Pairwise identity between any two sequences was 99.5% percent.

Further, the 16S rRNA genes were extracted from 6 other *Escherichia* species

- *E.adecarboxylata* CIP JN175338; length 1527
- *E.fergusonii* ATCC 35469 AF530475; length 1473
- *E.hermannii* CIP 103176 JN175345; length 1478
- *E. coli* NBRC 102203 AB681728; length 1467
- *E.vulneris* NBRC 102420 AB681776; length 1465
- *E.blattae* CIP 104942 JN175333; length 1525
- *E.albertii* ICDDR 19982 AY696662; length 1501

and compared the same way. These 7 sequences produced a consensus of 1529 bp with 1,418 identical sites (93.0%) and pairwise identity of 96.8%. **Figure 4.7** shows the resulting tree.

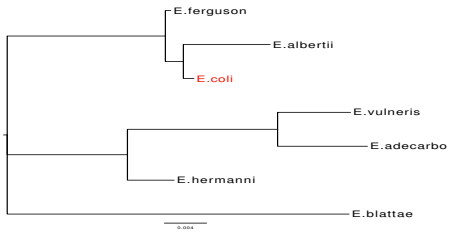


Figure 4.7: 16S rRNA tree based on Escherichia genera

This example indeed shows that some species inside of *Escherichia* genera can be defined, however not all of them. So if compare only *E. coli* and *E.fergusonii* they originate consensus of the length 1,542 bp identical at 1,466 positions which is 99.5% and pairwise identity 99.6% that is even higher than the pairwise identity across all homologues copies of 16S rRNA inside one K12 strain. Therefore it can be tricky to separate *E. coli* from *E. fergusonii* based only on 16S method. Based on this analysis it can be concluded that 16S rRNA is not an adequate method to determine species inside of *Escherichia* genera. Addressing outliers showed that two different methods provided opposite results and therefore should be interpreted with caution.

Apart from species identification, it should be noted that the 16S rRNA tree clearly separates into 2 distinct clusters. It can be investigated further why this division happens. To date this does no associate with any host or environment or ST, but does indicate a likely early divergence of the core genome in its evolution as a species.

4.6 Assemblies

All sequences were assembled with Spades with very good overall results that is summarised in **Figure 4.8**. Average numbers for assembly are as follow: length of assemblies - 5239617 bp, contigs per assembly - 224, N50 - 228762nt, unique genes - 5075 per sequence. There were some outliers: eleven out of 564 assemblies were longer then 6 Mbp, forty nine contained more than 300 contigs, and fourteen contained more than 6000 genes. Two of these characteristics are strongly correlated - longer assemblies means more genes detected, but not always longer assemblies contain more contigs.

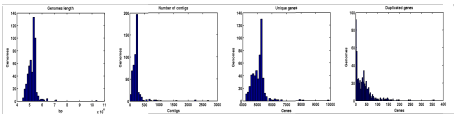


Figure 4.8: Assemblies statistic: lenght, contigs, genes, duplicated genes.

To assess if the longest assemblies are a biological entity or banal misassembly, some of the longest assemblies were compared with the same sequence, but resulted from sequencing by PacBio. The results show that these compared outliers are misassembled sequences. It is therefore predicted that all other assembly related outliers are also misassembled genomes. There is no any indication as to why these particular sequences are mis-assembled as they have the same short read lengths and quality. For future reference, these numbers could be used as a predictions for technical error when assembling a large number of genomes, thus based on the abnormally high numbers of genes detected 14 out of 564 genomes gives 2.5% of the erroneously assembled genomes. **Table 4.1** describes results obtained when comparing Illumina based assemblies with PACBio.

Moreover, knowing that the main cause of misassemblies are repeats and low complexity regions it can be recommended to include pre-assembly workflow that can address such problems. Repeats cause different types of errors in assembly: one of them is collapsed assembly, that happens when non adjacent copies of repeats can erroneously be joined together and 'or-

Table 4.1: Missasemblies

sequences	C-0863N0005	C-0863N0006	C-0863N0030
original length	5000721	5309743	5135913
assembly length	6142715	6146674	5985616
gaps in mapped	28 327	26 211	27 993

phan reads', that should be in the middle, left behind. These types of misassemblies increase the number of contigs but shorten the overall size of assembly. Another type are expansions when more than necessary repetitive copies joined together. Also repeats usually cause inversion, the most common type of misassemblies detected in all entries when comparing Illumina vs PacBio datasets. This type of assembly error when reads are wrongly joined in the opposite direction produces biased results in further downstream analyses. Therefore, future work will include RepeatMasker [62] in a pre-assembly workflow to detect and exclude, if necessary, such regions from the assembly.

4.7 Annotation

Running PROKKA with default parameters resulted in a large number of hypothetical proteins and less than 600 genes detected than expected. I decided to produce my own database with trusted annotated proteins from which annotations can be derived. **Table 4.2** summarises improvements in annotations.

Table 4.2: Annotations

	Total	Hypotheticsl
Sakai ref	5467	2136
Sakai default	4810	2275
Sakai with db	5408	1342

Difference between total numbers of gene products detected in the reference Sakai (5467) and the Sakai sequence annotated by PROKKA with own database (5408) can be explained by imperfections of annotation software that will not annotate anything smaller than 200 nt long, therefore leaving behind for instance tRNA. On the other hand it has achieved much better results regarding the number of hypothetical proteins. The differences most likely originated by reference status of Sakai genome. To obtain status of reference genome the annotation should be curated i.e. verified for evidence from literature or pass manually evaluated computational analysis, therefore proteins without such evidence will be called hypothetical. Our annotations extracted

inference based only on sequence similarity first from our own database and then from public databases that potentially full of errors. [63].

4.8 Homology Clustering

Clustering algorithms were developed and described well before advances in NGS technologies as a classification by patterns (by sequence similarity in our case) into groups. Even though clustering as a mathematical problem is not new, it is still 'a difficult problem combinatorially, and differences in assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur' [64].

In Bioinformatics clustering is used when dataset is large and other approaches such as multiple sequence alignment may fail or when very dissimilar genomes need to be compared. By tuning the percent of similarity one can decide how many and how similar sequences (genes or proteins) should be to become a cluster of homologues genes or proteins.

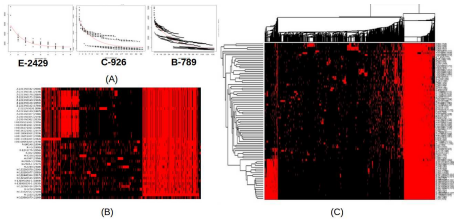


Figure 4.9: Get_Homologues results on smaller datasets. (A) Number of core genes in Environmental, Canine and Bovine datasets.(B) Human vs Bovine dataset: matrix of present (red) and absent (black) genes are plotted. (C) 135 sequences clustered based on presence and absence of genes.

Definition of 'pan-genome' has first arrived when analysing 8 *S. agalactiae* genomes it was found that 80% of genes were shared amongst all of them, 20% was only partially shared between genomes and some of the genes were strain specific [65]. Thus pan-genome is assembly of all genes found in a group of organisms of the same species. These genes can be sub-classified further: core genes are the genes present in all genomes in a group, and accessory genes, sometimes called non essential genes, that would be present only in a fraction of genomes.

To find what genes families are present in our collection, I used Get-Homologues that showed very promising results over pilot studies. However, the biggest dataset that the software was able to calculate was composed of only 135 sequences. Increasing further dataset led to unre-

alistic time for calculations needed as well as very high memory requirements. Some results from that analysis are presented in a **Figure 4.9**. It can be seen how quickly size of core genome shrinks when size of datasets as increasing from almost half of genome shared among 9 environmental sequences to only 789 genes when 135 bovine sequences were analysed. Some interesting blocks of genes were detected: some associated with bovine isolates as can be seen in a upper left corner on sub-figure B and genes associated with ST O157 that can be seen on a subfigure C at lower left corner.

Later, other software 'Roary' were identified and used for clustering. Note, that even though a word 'pan genome' is used, blast search, alignment and comparison of sequences are done using translated amino acid sequences. There is some advantages to do so. First, redundant codon will be count as a mutations, however there is most likely no significant evolutionary pressure to prevent a silent mutation, so as long as protein sequence not altered these changes do not have to be calculated against homology score. Second, statistical significance of an alignment will be easier achievable when comparing alignment based on 20 different letters than alignment based on four letters. Third, some amino acid changes will not alter protein dramatically, as for example isoleucine to valine, both similarly hydrophobic - these mutations can be accounted for in amino acid alignment, while in DNA alignment this region will be treated as any other misalignment.

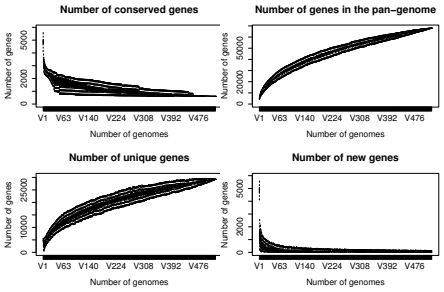


Figure 4.10: Pan genome statistics

Number of genes or proteins detected in any sequence can be influenced by various factors:

- Sequencing errors.
- Due to assembly some of the genes will be lost as long as assembly stays in a draft format, when contigs can be broken in a middle of a gene, so that gene is not detected.
- Some software relies on annotation, nowadays annotation is the weakest link in many whole genome analysis steps as amount of information increase dramatically while accuracy and verification of this information is often poor and slow [63].

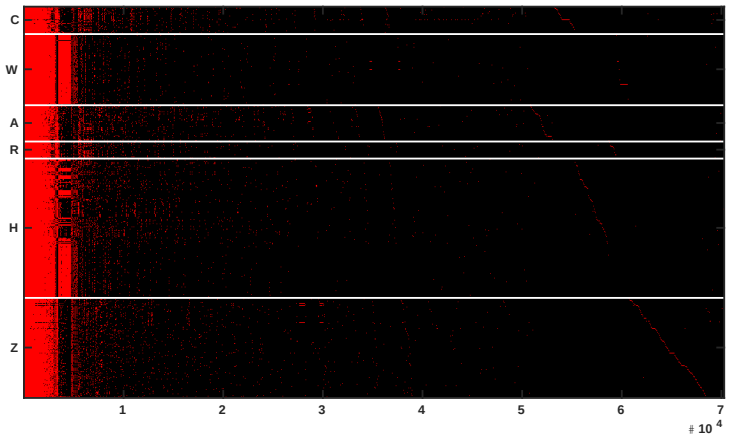
Thus it is expected a increasing rate of false positive errors in such analysis.

The results are visualised in a **Figure 4.11**. There are much bigger number of homologues proteins detected than in previous studies, however it is appreciated that blastp cut off for this work were set very high (95% of similarity), while in other studies protein sequences were joined into the same cluster if they are above 50% [55] or 80% [14], [66] of similarity. Nevertheless, there was 70,283 protein clusters produced, however that number decrease dramatically as soon as clusters that are composed by only one sequence (singletons, 30,908 clusters) are excluded.

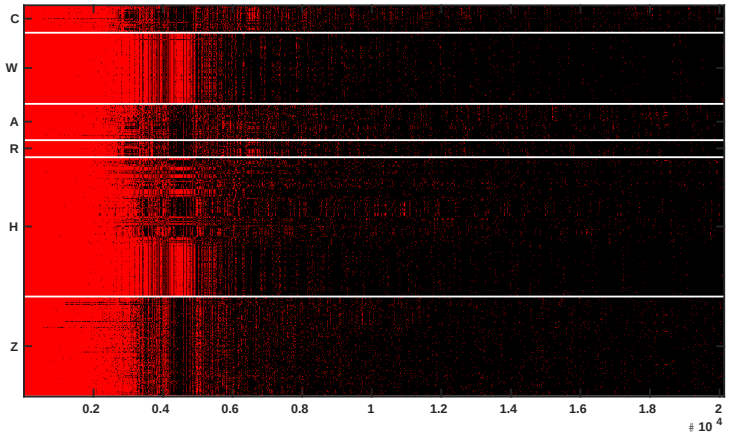
Table 4.3: Number of genes by host.

Dataset	Quantity	Pan	Core	Singletons
All	528	70283	599	30908
A	49	12954	1894	7480
C	37	22527	1447	9836
H	188	32565	1903	12496
R	23	14230	2388	5009
W	96	14104	3019	4730
Z	135	37705	947	16985

Size of pan genome for different subgroups is different, most likely due size of sub-datasets. **Table 4.3** present absolute numbers for pan-genomes, core genes and number of singleton clusters for each of 4 datasets. Preliminary analysis indicates that as expected some strains are influencing greatly in overall size of the core and pan genome. Such strains can be clearly seen in **Figure 4.11** when they appears as black doted lines in a core genome area, meaning no genes detected. However this problem should be analysed further if the fault is due to assembly or the genes of these particular strains are more divergent than 95% similarity. Therefore such situation will decrease number of core genes and increase pan genome size placing these proteins in a separate clusters.



(a)



(b)

Figure 4.11: The figure visualise a matrix of strains (rows) vs genes (columns) Red indicates presence of a gene, thus, the core genes can be seen at a left part of figure. (a) Pan-genome for 528 strains from the collection contains 70,283 clusters. (b) Pan-genome size is 20,136 clusters, if include clusters that contain proteins presented in at least 5 strains.

4.9 Core genes

Positions and numbers of core genes are demonstrated in a **Figure 4.12**. There are much less core genes in a canine - human and canine - avian datasets. What cause this situation should be analysed further.

Recombination analysis were performed on core genes using Gubbins. These genes resulted in an alignment of the length of 26,681 nt. Recombinations were detected and excluded, resultant alignment were 18,172 nt long, thus 8,509 nt were filtered. Number of SNPs inside and outside of recombination blocks were calculated and based on this ratio of recombination vs mutation will be calculated in future. The **Figure 6.1** in Supplementary materials illustrates preliminary 'map' for the recombination. There are clear hotspots of recombinations events that occur on the few genes. Work in recombination is ongoing, with plans to closely analyse the regions with massive recombinations and compose list of genes from these hotspots.

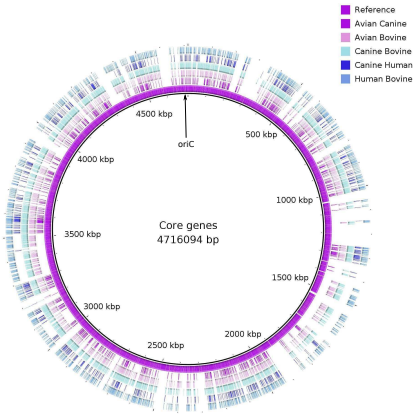


Figure 4.12: Core genes plotted against reference (inner violet circle) to demonstrate their position in a chromosome.

Conserved regions in *E. coli.* also include 1,060 conserved intergenic regions flanked by orthologues genes. These regions vary in size and place across genome. There are 12 regions that is bigger then 400 nt, with the biggest regions over 683 nt long. **Figure 6.2** in Suplementary materials shows numbers and length distributions of conserved intergenic regions across all genomes.

It is also appreciated that DNA alignment can provide invaluable information especially in detecting these silent mutations which does not matter for protein homology studies. The work of analysing DNA alignment of core genes is ongoing, aiming to determine the background drift level of mutation, to use it then to help quantify the amount of positive or negative selection.

4.10 Analysis of proteins associated with host or environment

Proteins presented in all strains were extracted and combined into table that provides percent informations for each host about how many strains contain each gene. **Table 4.4** bellow highlights some of the proteins that are found in a bigger numbers for particular host. The condition for highlighting were that protein should be presented in more than 80% of isolates for a particular host, and be in less than 60% in other host isolates.

The highest number of proteins associated for a particular dataset was found for a bovine groups: 50 for Zambian, and 42 for bovine, ST O157, UK. There are 10 proteins for the human reference dataset that are almost absent in other datasets, except for some proteins found in minority of avian strains. For avian isolates no proteins were detected that is presented in more than 80%, it was only one protein ygcG that is presented in more than 70% of the genomes while in other groups it is found in less than 60%, therefore it is unlikely that ygcG is an avian host specific protein.
This part of the work is ongoing and all findings should be verified further as well as should be statistically challenged.

Table 4.4: Proteins by host

	C	W	R	H	A	Z	Gene ID
ygcG	43.24	0.0	47.83	30.85	71.43	54.81	:gi|90111490|ref|NP_417258.4|
group_38327	94.59	0.0	0.0	2.13	0.0	0.0	
group_69352	0.0	0.0	86.96	0.0	0.0	0.0	
group_9061	0.0	0.0	86.96	1.6	0.0	0.0	:gi|145698331|ref|NP_418214.2|
group_69354	0.0	0.0	86.96	5.85	2.04	0.0	
group_69349	0.0	0.0	100.0	3.19	4.08	0.74	
group_28790	0.0	1.04	86.96	0.0	10.2	2.22	:gi|16131218|ref|NP_417798.1|
group_28787	0.0	1.04	100.0	0.0	40.82	5.19	:YP_671310.1
gadB	0.0	4.17	100.0	1.6	26.53	5.19	:gi|16131389|ref|NP_417974.1|
group_69348	2.7	0.0	86.96	1.6	4.08	0.74	
group_69355	2.7	0.0	86.96	3.72	4.08	0.0	
gadB_1	2.7	4.17	95.65	2.13	20.41	5.93	:gi|16129452|ref|NP_416010.1|
group_23656	0.0	80.21	0.0	47.34	0.0	0.0	
group_26057	0.0	81.25	0.0	44.68	2.04	0.0	:YP_003222976.1
group_524	0.0	82.29	0.0	42.02	18.37	3.7	:YP_001461404.1
group_21972	0.0	84.38	0.0	45.74	2.04	3.7	:gi|16129458|ref|NP_416016.1|
group_22453	0.0	84.38	0.0	45.74	2.04	3.7	:gi|16132094|ref|NP_418693.1|
group_35189	0.0	85.42	0.0	43.09	0.0	0.0	:gi|16130028|ref|NP_416593.1|
group_13881	0.0	86.46	0.0	39.36	4.08	0.74	:NP_311304.1
group_23340	0.0	86.46	0.0	45.21	0.0	0.74	:gi|16130540|ref|NP_417111.1|
group_21726	0.0	87.5	0.0	46.28	0.0	0.74	:YP_002268455.1
group_22584	0.0	87.5	0.0	46.81	0.0	0.74	:YP_002268454.1
group_30641	0.0	88.54	0.0	45.74	0.0	0.0	:gi|16131780|ref|NP_418377.1|
group_32213	0.0	88.54	0.0	46.81	0.0	2.96	
group_30017	0.0	88.54	0.0	47.34	0.0	0.0	:gi|16130295|ref|NP_416864.1|
group_9907	0.0	89.58	0.0	46.81	0.0	0.0	:gi|16130028|ref|NP_416593.1|
ycjA	0.0	89.58	13.04	35.64	0.0	3.7	:YP_002406020.1
group_25718	0.0	91.67	0.0	48.94	0.0	0.74	:gi|16129533|ref|NP_416092.1|
group_977	2.7	81.25	0.0	38.3	0.0	0.0	:gi|16131356|ref|NP_417941.1|
group_13320	2.7	81.25	0.0	45.74	0.0	0.74	:YP_003222327.1
group_38302	2.7	81.25	8.7	47.34	2.04	2.22	:NP_288358.1
group_8389	2.7	81.25	13.04	49.47	8.16	10.37	:YP_003229435.1
group_18648	2.7	81.25	17.39	46.28	14.29	15.56	:gi|16128532|ref|NP_415081.1|

	C	W	R	H	A	Z	Gene ID
group_35688	2.7	83.33	0.0	49.47	0.0	0.74	:NP_309096.2
group_26647	2.7	84.38	0.0	49.47	0.0	0.0	:gi|16131387|ref|NP_417972.1|
group_38694	2.7	87.5	0.0	45.74	0.0	0.74	
group_20520	2.7	87.5	0.0	46.28	6.12	3.7	:YP_001457045.1
group_28621	2.7	87.5	0.0	46.81	6.12	13.33	:YP_541946.1
group_2779	2.7	87.5	8.7	42.02	4.08	2.96	:gi|288551665|ref|NP_415949.2|
group_8403	2.7	89.58	0.0	47.87	6.12	14.81	:gi|16129121|ref|NP_415676.1|
group_15416	2.7	97.92	0.0	49.47	0.0	0.0	:YP_325651.1
group_35059	5.41	80.21	0.0	35.64	0.0	0.0	:YP_002402126.1
group_18949	5.41	81.25	8.7	47.87	6.12	8.15	:YP_539966.1
group_9393	5.41	89.58	17.39	32.98	2.04	3.7	protein motif:Pfam:PF12083.2
group_974	5.41	89.58	21.74	48.4	20.41	25.19	:gi|16131356|ref|NP_417941.1|
group_14858	8.11	81.25	0.0	40.43	22.45	11.11	:gi|90111276|ref|NP_415967.2|
cpsG_1	8.11	89.58	17.39	39.36	8.16	5.93	:gi|16129988|ref|NP_416552.1|
group_11599	13.51	81.25	0.0	39.89	22.45	10.37	:gi|226524714|ref|NP_415966.6|
group_5719	13.51	82.29	4.35	43.62	22.45	13.33	:gi|16129410|ref|NP_415968.1|
group_12919	13.51	82.29	8.7	44.68	22.45	13.33	:gi|16129413|ref|NP_415971.1|
group_11035	16.22	94.79	30.43	46.81	10.2	9.63	:YP_005280294.1
group_27425	18.92	82.29	0.0	47.34	12.24	8.89	:YP_539975.1
group_26809	18.92	94.79	39.13	49.47	26.53	11.11	
group_29374	21.62	81.25	13.04	48.94	4.08	9.63	:NP_310658.1
group_27410	21.62	82.29	13.04	49.47	12.24	10.37	:YP_539972.1
group_24640	24.32	86.46	17.39	49.47	0.0	6.67	:gi|16128526|ref|NP_415075.1|
group_21602	27.03	1.04	13.04	20.21	55.1	82.22	:gi|16129261|ref|NP_415816.1|
hpaH	27.03	1.04	13.04	21.81	55.1	82.22	:gi|90111116|ref|NP_414884.2|
group_26330	27.03	1.04	13.04	22.34	51.02	81.48	:gi|16129652|ref|NP_416211.1|
hpaI	27.03	1.04	13.04	22.34	55.1	81.48	:gi|16130180|ref|NP_416748.1|
group_25403	27.03	1.04	13.04	22.34	55.1	82.22	:YP_001726642.1
group_29037	27.03	1.04	13.04	22.34	55.1	82.22	:gi|90111202|ref|NP_415527.4|
hpaG	27.03	1.04	13.04	22.34	55.1	82.22	:gi|16129143|ref|NP_415698.1|
yiiF	27.03	1.04	26.09	31.38	53.06	88.89	:gi|90111663|ref|NP_418326.2|
yhaB	29.73	2.08	34.78	35.64	59.18	84.44	:gi|90111544|ref|NP_417590.2|
ybgP	32.43	1.04	26.09	25.0	51.02	82.22	:gi|16128692|ref|NP_415245.1|

	C	W	R	H	A	Z	Gene ID
maoC	32.43	1.04	26.09	28.19	48.98	81.48	:gi|16129348|ref|NP_415905.1|
setA	32.43	1.04	26.09	32.45	59.18	86.67	:gi|49175994|ref|YP_025293.1|
yafQ	32.43	1.04	34.78	27.13	44.9	87.41	:gi|16128211|ref|NP_414760.1|
gmr	32.43	1.04	34.78	35.11	59.18	88.15	:gi|16129246|ref|NP_415801.1|
feaR	32.43	1.04	39.13	33.51	59.18	86.67	:gi|16129345|ref|NP_415902.1|
group_17363	32.43	2.08	34.78	31.91	48.98	82.22	:gi|16128631|ref|NP_415181.1|
yddJ	32.43	9.38	21.74	35.64	53.06	81.48	:YP_002402676.1
group_7587	32.43	10.42	26.09	40.43	57.14	85.19	:gi|90111272|ref|NP_415950.4|
group_28945	35.14	1.04	26.09	32.45	59.18	85.93	:gi|16128833|ref|NP_415386.1|
group_20685	35.14	1.04	43.48	31.38	59.18	81.48	:gi|16132205|ref|NP_418805.1|
group_16872	35.14	2.08	0.0	26.06	6.12	82.96	:gi|16129175|ref|NP_415730.1|
group_15681	35.14	2.08	26.09	31.38	46.94	85.19	:gi|16130641|ref|NP_417214.1|
group_15935	37.84	1.04	34.78	32.45	59.18	83.7	:gi|16130946|ref|NP_417522.1|
group_23478	37.84	2.08	0.0	25.53	6.12	82.22	:gi|16129177|ref|NP_415732.1|
group_24524	37.84	2.08	0.0	25.53	6.12	82.22	:gi|16129176|ref|NP_415731.1|
group_14860	37.84	18.75	34.78	54.26	57.14	87.41	:gi|90111276|ref|NP_415967.2|
ycgY	40.54	1.04	26.09	32.45	59.18	85.93	:gi|16129159|ref|NP_415714.1|
group_69901	40.54	2.08	0.0	26.06	6.12	83.7	:gi|16129178|ref|NP_415733.1|
group_22187	43.24	1.04	21.74	35.11	55.1	88.15	:gi|16130924|ref|NP_417500.1|
feaB	43.24	1.04	26.09	31.91	53.06	81.48	:gi|90111264|ref|NP_415903.4|
tynA	43.24	1.04	26.09	33.51	53.06	82.96	:gi|162135920|ref|NP_415904.3|
paaJ	43.24	1.04	26.09	33.51	53.06	86.67	:gi|16129353|ref|NP_415910.1|
paaD	43.24	1.04	26.09	34.04	53.06	88.15	:gi|226524712|ref|NP_415909.4|
ygiL	43.24	1.04	30.43	32.45	55.1	83.7	:gi|16130939|ref|NP_417515.1|
group_25336	43.24	2.08	0.0	26.6	8.16	85.19	:gi|16129172|ref|NP_415727.1|

	C	W	R	H	A	Z	Gene ID
group_69217	43.24	2.08	0.0	26.6	8.16	85.93	:gi|16129174|ref|NP_415729.1|
paaH	45.95	1.04	26.09	30.85	51.02	88.89	:gi|16129356|ref|NP_415913.1|
paaI	45.95	1.04	26.09	32.45	53.06	89.63	:gi|16129357|ref|NP_415914.1|
paaG	45.95	1.04	26.09	32.98	53.06	88.89	:gi|16129355|ref|NP_415912.1|
paaF	45.95	1.04	26.09	34.04	53.06	87.41	:gi|16129354|ref|NP_415911.1|
paaA	45.95	1.04	26.09	34.57	53.06	86.67	:gi|16129349|ref|NP_415906.1|
yneK_2	45.95	1.04	34.78	32.98	48.98	82.22	:gi|16129486|ref|NP_416044.1|
paaY	48.65	1.04	34.78	35.11	59.18	94.07	:gi|16129361|ref|NP_415918.1|
yhiJ_2	51.35	7.29	34.78	39.89	59.18	86.67	:gi|16131360|ref|NP_417945.1|
paaC	54.05	1.04	26.09	34.57	53.06	88.15	:gi|16129351|ref|NP_415908.1|
group_5507	54.05	1.04	39.13	32.45	55.1	87.41	:gi|16129342|ref|NP_415899.1|
paaB	56.76	1.04	26.09	35.11	53.06	87.41	:gi|16129350|ref|NP_415907.1|
yrhB	56.76	1.04	34.78	32.98	55.1	86.67	:gi|16131318|ref|NP_417903.1|
ycbQ	56.76	2.08	30.43	30.32	55.1	84.44	:gi|90111190|ref|NP_415458.2|

Chapter 5

Conclusions and Future plans

Extracting all genes from more than 500 strains - provides robust ground to build powerful statistical tools that will help to detect and recognise the origin of a strain and therefore may be of invaluable help over epidemiological outbreaks and in general public health surveillance. Therefore it is planned to statistically challenge data in variety of dimensions.

The descriptive part of the work is not finished yet. It is planned to classify genes found by GO; detect virulence factors, insertion sequences, antimicrobial resistant genes and then combine all information in multi-step statistical pipeline that take as an input short Illumina reads and provides output file that describe the sequence with possible origin.

Next logical step will be to look closer to what actually are expressed and used. Thus some studies indicate that even though there is enormous diversity across *E. coli* genomes, its metabolome is much more stable and around 57% of metabolic reaction shared across all strains [67]. Mentioned study analysed 29 strains, suggesting that panmethabolome has reached plateau that so it is interesting if proportion maintain when applied to a bigger collection. Apart from strong dissection between cathabolic that are mostly core reaction, and anabolic processes that can be in accessory metabolome, analysis of metabolic reactions can provide clues about adaptation to specific live-stile whether pathogenicity or commensalism or adaptation to the different environments.

Also the idea that there is no single genome that represent a specie but genomic continuum with variability of possibilities to acquire new genes it would be interesting to include other species

into our collection and check for any overlaps.

PhD is a learning opportunity and I am trying to effectively manage my time to acquire new skills. Thus, over the first year I attended courses listed bellow. Over 2nd year of my PhD I am planning to learn R as I have limited experience with this language. I also will continue monitor major courses providers for a learning opportunities specific to my work.

Table 5.1: Courses attended

Course	Duration	Provider	
1	Bioinformatics Programming and System Management, level 11	1 semester	UoE
2	RNA-Seq Data Analysis workshop	2 day	Edinburgh Genomics
3	Hands-on Porting and Optimisation Workshop	1 day	ECDF, ARCHER
4	Software development for research	2 days	Software Carpentry, ARCHER
5	Manipulating data using linux/unix tools	1 day	The Roslin Institute
6	Computational Molecular Evolution	2 weeks	Welcome trust
7	Ensembl Browser Workshop	1 day	EBI
8	MATLAB programming	9 weeks	Coursera
9	Leadership and Entrepreneurship in the Biotech Industry	1 day	UoE
10	Life Sciences start-ups, spin outs and let downs	1 day	Innovation forum Edinburgh
11	Practical Project Management	1 daya	UoE

Chapter 6

Supplementary material

Table 6.1: 'Unknown' phylotypes

	arpA	chuA	yjaA	TspE4.C2
H-060520151-S1	-	-	+	+
H-100440290-S3	-	-	+	+
H-111160229-S6	-	-	+	+
H-113320446	-	-	+	+
H-084660371-S2	-	-	+	+
H-102260754-S4	-	-	+	+
H-133200678	+	-	+	+
H-122900502	+	-	+	+
C-0863N0015	+	+	+	+
C-0863N0016	+	+	+	+
C-0863N0002	+	+	+	+
C-0863N0006	+	+	+	+
C-0863N0005	+	+	+	+
H-132380155	+	+	+	+
H-134380810	+	+	+	+
C-0863N0030	+	+	+	+
C-0863N0025	+	+	+	+
C-0863N0021	+	+	+	+
Z-2014N0013	-	-	-	-

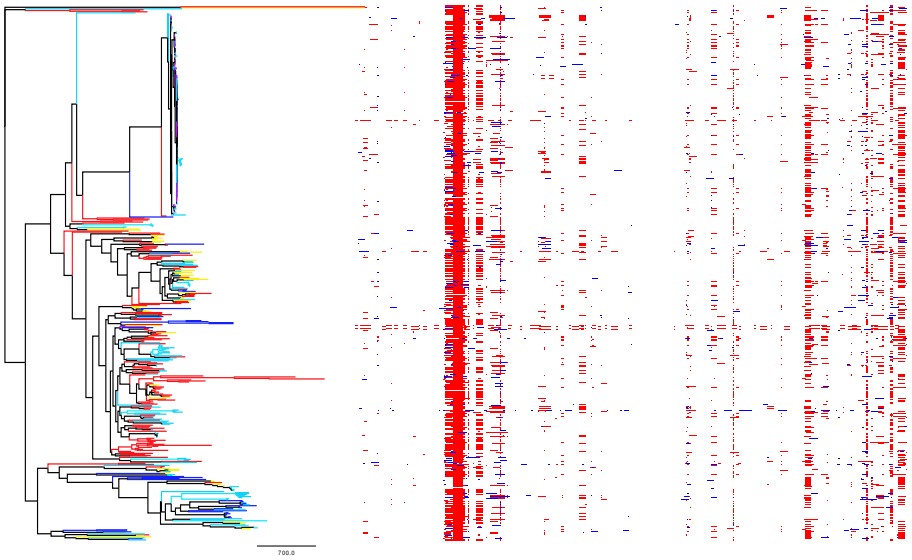


Figure 6.1: Recombinations detected in a core genes, plotted against core genes tree.

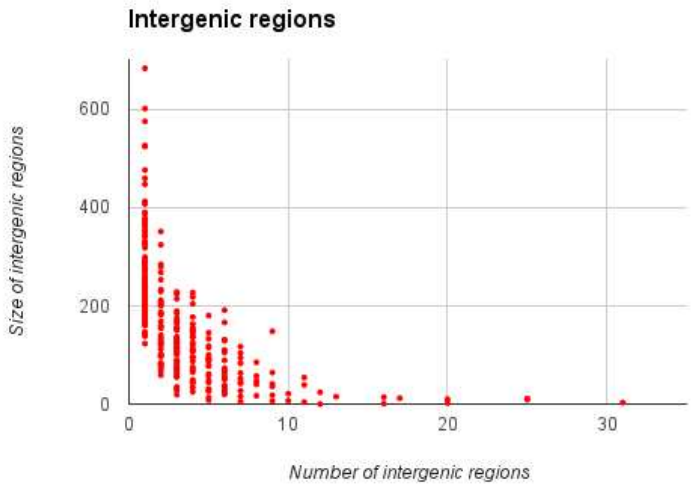


Figure 6.2: Size and number of conserved intergenic regions regions

References

[1] James B Kaper, James P Nataro, and Harry L Mobley. Pathogenic Escherichia coli. *Nature reviews. Microbiology*, 2(2):123–40, February 2004. ISSN 1740-1526. doi: 10.1038/nrmicro818. URL http://dx.doi.org/10.1038/nrmicro818.

[2] K L Kotloff, J P Winickoff, B Ivanoff, J D Clemens, D L Swerdlow, P J Sansonetti, G K Adak, and M M Levine. Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. *Bulletin of the World Health Organization*, 77(8):651–66, January 1999. ISSN 0042-9686. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2557719&tool=pmcentrez&rendertype=abstract.

[3] David A Rasko, Dale R Webster, Jason W Sahl, Ali Bashir, Nadia Boisen, Flemming Scheutz, Ellen E Paxinos, Robert Sebra, Chen-Shan Chin, Dimitris Iliopoulos, Aaron Klammer, Paul Peluso, Lawrence Lee, Andrey O Kislyuk, James Bullard, Andrew Kasarskis, Susanna Wang, John Eid, David Rank, Julia C Redman, Susan R Steyert, Jakob Frimodt-Mø ller, Carsten Struve, Andreas M Petersen, Karen A Krogfelt, James P Nataro, Eric E Schadt, and Matthew K Waldor. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. *The New England journal of medicine*, 365(8):709–17, August 2011. ISSN 1533-4406. doi: 10.1056/NEJMoa1106920. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3168948&tool=pmcentrez&rendertype=abstract.

[4] Thomas A Russo and James R Johnson. Medical and economic impact of extraintestinal infections due to Escherichia coli: focus on an increasingly important endemic problem. *Microbes and infection / Institut Pasteur*, 5(5):449–56, April 2003. ISSN 1286-4579. URL http://www.ncbi.nlm.nih.gov/pubmed/12738001.

[5] Paul D Frenzen, Alison Drake, and Frederick J Angulo. Economic cost of illness due to Escherichia coli O157 infections in the United States. *Jour-*

nal of food protection, 68(12):2623–30, December 2005. ISSN 0362-028X. URL http://www.ncbi.nlm.nih.gov/pubmed/16355834.

[6] J. H. Hewitt and Janet Rigby. Effect of various milk feeds on numbers of Escherichia coli and Bifidobacterium in the stools of new-born infants. *Journal of Hygiene*, 77(01):129, May 2009. ISSN 0022-1724. doi: 10.1017/S0022172400055601. URL http://journals.cambridge.org/abstract_S0022172400055601.

[7] M. F. MICHEL J. E. DEGENER, A. C. W. SMIT, H. A. VALKENBURG* MULLER, and L. FAECAL CARRIAGE OF AEROBIC GRAM-NEGATIVE BACILLI AND DRUG RESISTANCE OF ESCHERZCHZA COLZ IN DIFFERENT AGE-GROUPS IN DUTCH URBAN COMMUNITIES, 1983. URL http://jmm.sgmjournals.org/content/16/2/139.full.pdf.

[8] Forough Nowrouzian, Bill Hesselmar, Robert Saalman, Inga-Lisa Strannegard, Nils Aberg, Agnes E Wold, and Ingegerd Adlerberth. Escherichia coli in infants' intestinal microflora: colonization rate, strain turnover, and virulence gene carriage. *Pediatric research*, 54(1):8–14, July 2003. ISSN 0031-3998. doi: 10.1203/01.PDR.0000069843.20655.EE. URL http://www.nature.com/pr/journal/v54/n1/full/pr2003354a.html#bib1.

[9] Michael A. Savageau. Escherichia coli habitats, cell types, and molecular mechanisms of gene control. *The American Naturalist*, 122(6):pp. 732–744, 1983. ISSN 00030147. URL http://www.jstor.org/stable/2460914.

[10] Kathryn M Wright, Sean Chapman, Kara McGeachy, Sonia Humphris, Emma Campbell, Ian K Toth, and Nicola J Holden. The endophytic lifestyle of Escherichia coli O157:H7: quantification and internal localization in roots. *Phytopathology*, 103(4):333–40, April 2013. ISSN 0031-949X. doi: 10.1094/PHYTO-08-12-0209-FI. URL http://www.ncbi.nlm.nih.gov/pubmed/23506361.

[11] Thomas L. Hale and Gerald T. Keusch. *Shigella*. University of Texas Medical Branch at Galveston, 1996. doi: NBK8038[bookaccession]. URL http://www.ncbi.nlm.nih.gov/books/NBK8038/.

[12] David L Gally, Stuart W Naylor, J Christopher Low, George J Gunn, Barti A Synge, Michael C Pearce, William Donachie, and Thomas E Besser. Colonisation site of E coli O157 in cattle. *The Veterinary record*, 152(10):307, March 2003. ISSN 0042-4900. URL http://www.ncbi.nlm.nih.gov/pubmed/12650483.

[13] N T Perna, G Plunkett, V Burland, B Mau, J D Glasner, D J Rose, G F Mayhew, P S Evans, J Gregor, H A Kirkpatrick, G Pósfai, J Hackett, S Klink, A Boutin, Y Shao, L Miller, E J Grotbeck, N W Davis, A Lim, E T Dimalanta, K D Potamousis, J Apodaca, T S Anantharaman, J Lin, G Yen, D C Schwartz, R A Welch, and F R Blattner. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature*, 409(6819): 529–33, January 2001. ISSN 0028-0836. doi: 10.1038/35054089. URL http://dx.doi.org/10.1038/35054089.

[14] David A Rasko, M J Rosovitz, Garry S A Myers, Emmanuel F Mongodin, W Florian Fricke, Pawel Gajer, Jonathan Crabtree, Mohammed Sebaihia, Nicholas R Thomson, Roy Chaudhuri, Ian R Henderson, Vanessa Sperandio, and Jacques Ravel. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of bacteriology*, 190(20):6881–93, October 2008. ISSN 1098-5530. doi: 10.1128/JB.00619-08. URL http://jb.asm.org/content/190/20/6881.full.

[15] Outbreaks of E. coli O104:H4 infection. URL http://www.euro.who.int/en/health-topics/disease-prevention/food-safety/outbreaks-of-e.-coli-o104h4-infection.

[16] Maria Foti, Antonio Daidone, Aurora Aleo, Alessia Pizzimenti, Cristina Giacopello, and Caterina Mammina. Salmonella bongori 48:z ₃₅ :âĂŞ in Migratory Birds, Italy. *Emerging Infectious Diseases*, 15(3):502–503, mar 2009. ISSN 1080-6040. doi: 10.3201/eid1503.080039. URL http://wwwnc.cdc.gov/eid/article/15/3/08-0039{_}article.htm.

[17] Giovanni M Giammanco, Sarina Pignato, Caterina Mammina, Francine Grimont, Patrick A D Grimont, Antonino Nastasi, and Giuseppe Giammanco. Persistent endemicity of Salmonella bongori 48:z(35):–in Southern Italy: molecular characterization of human, animal, and environmental isolates. *Journal of clinical microbiology*, 40(9):3502–5, sep 2002. ISSN 0095-1137. doi: 10.1128/JCM.40.9.3502-3505.2002. URL http://www.ncbi.nlm.nih.gov/pubmed/12202604http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC130773.

[18] S Pignato, G Giammanco, C Santangelo, and G.M Giammanco. Endemic presence of Salmonella bongori 48:z35:-causing enteritis in children in Sicily. *Research in Microbiology*, 149(6):429–431, jun 1998. ISSN 09232508. doi: 10.1016/S0923-2508(98)80325-2. URL http://linkinghub.elsevier.com/retrieve/pii/S0923250898803252.

[19] I. Letunic and P. Bork. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1): 127–128, jan 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl529. URL http://www.ncbi.nlm.nih.gov/pubmed/17050570https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl529.

[20] B J MCCARTHY and E T BOLTON. An approach to the measurement of genetic relatedness among organisms. *Proceedings of the National Academy of Sciences of the United States of America*, 50(1):156–64, jul 1963. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/13932048http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC300669.

[21] C L SCHILDKRAUT, J MARMUR, and P DOTY. The formation of hybrid DNA molecules and their use in studies of DNA homologies. *Journal of molecular biology*, 3:595–617, oct 1961. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/14498380.

[22] Jan P Meier-Kolthoff, Richard L Hahnke, Jörn Petersen, Carmen Scheuner, Victoria Michael, Anne Fiebig, Christine Rohde, Manfred Rohde, Berthold Fartmann, Lynne A Goodwin, Olga Chertkov, T B K Reddy, Amrita Pati, Natalia N Ivanova, Victor Markowitz, Nikos C Kyrpides, Tanja Woyke, Markus Göker, and Hans-Peter Klenk. Complete genome sequence of DSM 30083T, the type strain (U5/41T) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. *Standards in Genomic Sciences*, 9(1):2, dec 2014. ISSN 1944-3277. doi: 10.1186/1944-3277-9-2. URL https://doi.org/10.1186/1944-3277-9-2.

[23] D J Lane, B Pace, G J Olsen, D A Stahl, M L Sogin, and N R Pace. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A*, 82: 6955–6959, 1985. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=2413450.

[24] Masao Fukushima, Kenichi Kakinuma, and Ryuji Kawaguchi. Phylogenetic analysis of Salmonella, Shigella, and Escherichia coli strains on the basis of the gyrB gene sequence. *Journal of clinical microbiology*, 40(8):2779–85, August 2002. ISSN 0095-1137. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=120687&tool=pmcentrez&rendertype=abstract.

[25] David C. Schwartz and Charles R. Cantor. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell*, 37(1):67–75, May 1984. ISSN 00928674. doi: 10.1016/

0092-8674(84)90301-5. URL http://www.sciencedirect.com/science/article/pii/0092867484903015.

[26] G M Pupo, D K Karaolis, R Lan, and P R Reeves. Evolutionary relationships among pathogenic and nonpathogenic Escherichia coli strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. *Infection and immunity*, 65(7):2685–92, July 1997. ISSN 0019-9567. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=175379&tool=pmcentrez&rendertype=abstract.

[27] Multilocus sequence typing (MLST) of Escherichia coli O78 strains. URL http://femsle.oxfordjournals.org/content/femsle/222/2/199.full.pdf.

[28] Phage-Typing Scheme for Escherichia coli 0157:H7. URL http://jid.oxfordjournals.org/content/155/4/806.full.pdf.

[29] Gerald T. Keusch, Olivier Fontaine, Alok Bhargava, Cynthia Boschi-Pinto, Zulfiqar A. Bhutta, Eduardo Gotuzzo, Juan Rivera, Jeffrey Chow, Sonbol Shahid-Salles, and Ramanan Laxminarayan. *Diarrheal Diseases*. The International Bank for Reconstruction and Development / The World Bank, 2006. ISBN 0821361791. URL http://www.ncbi.nlm.nih.gov/pubmed/21250340.

[30] Bo Yang, Yong Wang, and Pei-Yuan Qian. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC bioinformatics*, 17:135, mar 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-0992-y. URL http://www.ncbi.nlm.nih.gov/pubmed/27000765http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4802574.

[31] J Michael Janda and Sharon L Abbott. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9):2761–4, September 2007. ISSN 0095-1137. doi: 10.1128/JCM.01228-07. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2045242&tool=pmcentrez&rendertype=abstract.

[32] Orna Mizrahi-Man, Emily R. Davenport, Yoav Gilad, BA Chapman, and CJ Cox. Taxonomic Classification of Bacterial 16S rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs. *PLoS ONE*, 8(1):e53608, jan 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0053608. URL http://dx.plos.org/10.1371/journal.pone.0053608.

[33] S. Mignard and J.P. Flandrois. 16S rRNA sequencing in routine bacterial identification: A 30-month experiment. *Journal of Microbiological Methods*, 67(3):574–581, dec 2006. ISSN 01677012. doi:

10.1016/j.mimet.2006.05.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0167701206001655.

[34] Martin C. J. Maiden, Jane A. Bygraves, Edward Feil, Giovanna Morelli, Joanne E. Russell, Rachel Urwin, Qing Zhang, Jiaji Zhou, Kerstin Zurth, Dominique A. Caugant, Ian M. Feavers, Mark Achtman, and Brian G. Spratt. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenicâĂŁmicroorganisms. *Proceedings of the National Academy of Sciences*, 95(6):3140–3145, 1998. URL http://www.pnas.org/content/95/6/3140.abstract.

[35] Keith A Jolley and Martin C J Maiden. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, 11(1):595, dec 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-595. URL https://doi.org/10.1186/1471-2105-11-595.

[36] K A Jolley, M S Chan, and M C Maiden. mlstdbNet - distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics*, 5, 2004. doi: 10.1186/1471-2105-5-86. URL https://doi.org/10.1186/1471-2105-5-86.

[37] E J Feil, B C Li, D M Aanensen, W P Hanage, and B G Spratt. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol*, 186, 2004. doi: 10.1128/JB.186.5.1518-1530.2004. URL https://doi.org/10.1128/JB.186.5.1518-1530.2004.

[38] Brian G Spratt. Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the Internet. *Current Opinion in Microbiology*, 2(3):312–316, jun 1999. ISSN 13695274. doi: 10.1016/S1369-5274(99)80054-X. URL http://linkinghub.elsevier.com/retrieve/pii/S136952749980054X.

[39] D Gevers, F M Cohan, J G Lawrence, B G Spratt, T Coenye, E J Feil, E Stackebrandt, Y Van de Peer, P Vandamme, F L Thompson, and J Swings. Re-evaluating prokaryotic species. *Nat.Rev.Microbiol.*, 3 (1740-1526 (Print)):733–739, sep 2005. ISSN 1740-1526. doi: 10.1038/nrmicro1236. URL c:{%}5CKarsten{%}5CPDFs{%}5CGrundlagen-PDFs{%}5CGrund-2005{%}5CGeversetal.-Re-evaluatingprokaryoticspecies.pdf.

[40] Martin C J Maiden, Melissa J Jansen van Rensburg, James E Bray, Sarah G Earle, Suzanne A Ford, Keith A Jolley, and Noel D McCarthy. MLST revisited: the gene-by-gene approach to bacterial genomics. 2013. doi: 10.1038/nrmicro3093. URL https://www.nature.com/nrmicro/journal/v11/n10/pdf/nrmicro3093.pdf.

[41] Barry G Hall, Garth D Ehrlich, and Fen Z Hu. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. *Microbiology (Reading, England)*, 156(Pt 4):1060–8, apr 2010. ISSN 1465-2080. doi: 10.1099/mic.0.035188-0. URL http://www.ncbi.nlm.nih.gov/pubmed/20019077http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2889442.

[42] Johan Goris, Konstantinos T Konstantinidis, Joel A Klappenbach, Tom Coenye, Peter Vandamme, and James M Tiedje. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology*, 57(Pt 1):81–91, January 2007. ISSN 1466-5026. doi: 10.1099/ijs.0.64483-0. URL http://ijs.sgmjournals.org/content/57/1/81.long.

[43] Konstantinos T Konstantinidis and James M Tiedje. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7): 2567–72, feb 2005. ISSN 0027-8424. doi: 10.1073/pnas.0409727102. URL http://www.ncbi.nlm.nih.gov/pubmed/15701695http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC549018.

[44] Frederick M. Cohan. What are Bacterial Species? *Annual Review of Microbiology*, 56(1):457–487, oct 2002. ISSN 0066-4227. doi: 10.1146/annurev.micro.56.012302.160634. URL http://www.annualreviews.org/doi/10.1146/annurev.micro.56.012302.160634.

[45] Marco Scortichini, Simone Marcelletti, Patrizia Ferrante, Giuseppe Firrao, and M Shumway. A Genomic Redefinition of Pseudomonas avellanae species. *PLoS ONE*, 8(9):e75794, sep 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0075794. URL http://dx.plos.org/10.1371/journal.pone.0075794.

[46] Ramiro Logares, Stephane Audic, Sebastien Santini, Massimo C Pernice, Colomban De Vargas, and Ramon Massana. Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. *The ISME Journal*, doi(10), 2012. doi: 10.1038/ismej.2012.36. URL http://biomarks.eu/sites/default/files/pdf-refs/Logaresetal2012.pdf.

[47] Ellen Jo Baron. *Classification*. University of Texas Medical Branch at Galveston, 1996. ISBN 0963117211. URL http://www.ncbi.nlm.nih.gov/pubmed/21413329.

[48] S Uzzau, D J Brown#, T Wallis$, S Rubino, G Leori%, S Bernard&, J Casadesu ! S ', D J Platt, and ïĺąïĺőïĺď J E Olsen#. REVIEW

Host adapted serotypes of Salmonella enterica. *Epidemiol. Infect*, 125:229–255, 2000. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2869595/pdf/11117946.pdf.

[49] H Ochman and E A Groisman. The origin and evolution of species differences in Escherichia coli and Salmonella typhimurium. *EXS*, 69:479–93, 1994. ISSN 1023-294X. URL http://www.ncbi.nlm.nih.gov/pubmed/7994120.

[50] Olivier Tenaillon, David Skurnik, Bertrand Picard, and Erick Denamur. The population genetics of commensal Escherichia coli. *Nature reviews. Microbiology*, 8(3):207–17, March 2010. ISSN 1740-1534. doi: 10.1038/nrmicro2298. URL http://www.ncbi.nlm.nih.gov/pubmed/20157339.

[51] O Clermont, S Bonacorsi, and E Bingen. Rapid and simple determination of the Escherichia coli phylogenetic group. *Applied and environmental microbiology*, 66(10):4555–8, October 2000. ISSN 0099-2240. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=92342&tool=pmcentrez&rendertype=abstract.

[52] Sameer M Dixit, David M Gordon, Xi-Yang Wu, Toni Chapman, Kaila Kailasapathy, and James J-C Chin. Diversity analysis of commensal porcine Escherichia coli - associations between genotypes and habitat in the porcine gastrointestinal tract. *Microbiology (Reading, England)*, 150 (Pt 6):1735–40, June 2004. ISSN 1350-0872. doi: 10.1099/mic.0.26733-0. URL http://www.ncbi.nlm.nih.gov/pubmed/15184560.

[53] Erick Denamur, Olivier Clermont, and David Gordon. Guide to the various phylogenetic classification schemes for Escherichia coli and the correspondence among schemes. *Microbiology*, 161 (5):980–988, may 2015. ISSN 1350-0872. doi: 10.1099/mic.0.000063. URL http://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.000063.

[54] Xavier Didelot, Rory Bowden, Teresa Street, Tanya Golubchik, Chris Spencer, Gil McVean, Vartul Sangal, Muna F Anjum, Mark Achtman, Daniel Falush, and Peter Donnelly. Recombination and population structure in Salmonella enterica. *PLoS genetics*, 7(7):e1002191, jul 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002191. URL http://www.ncbi.nlm.nih.gov/pubmed/21829375http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3145606.

[55] Chinyere K Okoro, Robert A Kingsley, Thomas R Connor, Simon R Harris, Christopher M Parry, Manar N Al-Mashhadani, Samuel Kariuki, Chisomo L Msefula, Melita A Gordon, Elizabeth de Pinna, John Wain,

Robert S Heyderman, Stephen Obaro, Pedro L Alonso, Inacio Mandomando, Calman A MacLennan, Milagritos D Tapia, Myron M Levine, Sharon M Tennant, Julian Parkhill, and Gordon Dougan. Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. *Nature Genetics*, 44(11):1215–1221, sep 2012. ISSN 1061-4036. doi: 10.1038/ng.2423. URL http://www.nature.com/doifinder/10.1038/ng.2423.

[56] J Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson, William Nelson, Derrick E Fouts, Samuel Levy, Anthony H Knap, Michael W Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O Smith. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004. ISSN 0036-8075. doi: 10.1126/science.1093857. URL http://science.sciencemag.org/content/304/5667/66.

[57] Paul B Eckburg, Elisabeth M Bik, Charles N Bernstein, Elizabeth Purdom, Les Dethlefsen, Michael Sargent, Steven R Gill, Karen E Nelson, and David A Relman. Diversity of the Human Intestinal Microbial Flora. *Science*, 308(5728):1635–1638, 2005. ISSN 0036-8075. doi: 10.1126/science.1110591. URL http://science.sciencemag.org/content/308/5728/1635.

[58] Ruth-Anne Sandaa, Vigdis Torsvik, ÃÂŸIvind Enger, Frida Lise Daae, Tonje Castberg, Dittmar Hahn, Zeyer J., and Hahn D. Analysis of bacterial communities in heavy metal-contaminated soils at different levels of resolution. *FEMS Microbiology Ecology*, 30(3):237–251, nov 1999. ISSN 01686496. doi: 10.1111/j.1574-6941.1999.tb00652.x. URL https://academic.oup.com/femsec/article-lookup/doi/10.1111/j.1574-6941.1999.tb00652.x.

[59] Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, Robert T Deboy, Tanja M Davidsen, Marirosa Mora, Maria Scarselli, Immaculada Margarit y Ros, Jeremy D Peterson, Christopher R Hauser, Jaideep P Sundaram, William C Nelson, Ramana Madupu, Lauren M Brinkac, Robert J Dodson, Mary J Rosovitz, Steven A Sullivan, Sean C Daugherty, Daniel H Haft, Jeremy Selengut, Michelle L Gwinn, Liwei Zhou, Nikhat Zafar, Hoda Khouri, Diana Radune, George Dimitrov, Kisha Watkins, Kevin J B O'Connor, Shannon Smith, Teresa R Utterback, Owen White, Craig E Rubens, Guido Grandi, Lawrence C Madoff, Dennis L Kasper, John L Telford, Michael R Wessels, Rino Rappuoli, and Claire M Fraser. Genome

analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–5, September 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506758102. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1216834&tool=pmcentrez&rendertype=abstract.

[60] Roger W Hendrix. Bacteriophage genomics. *Current Opinion in Microbiology*, 6(5):506–511, oct 2003. ISSN 13695274. doi: 10.1016/j.mib.2003.09.004. URL http://linkinghub.elsevier.com/retrieve/pii/S1369527403001152.

[61] NCBI Resource NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 44(D1):D7–19, jan 2016. ISSN 1362-4962. doi: 10.1093/nar/gkv1290. URL http://www.ncbi.nlm.nih.gov/pubmed/26615191http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702911.

[62] Enterobase. URL http://enterobase.warwick.ac.uk/.

[63] Kira S Makarova, Yuri I Wolf, and Eugene V Koonin. Comparative genomics of defense systems in archaea and bacteria. *Nucleic acids research*, 41(8):4360–77, apr 2013. ISSN 1362-4962. doi: 10.1093/nar/gkt157. URL http://www.ncbi.nlm.nih.gov/pubmed/23470997http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3632139.

[64] Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6):589–594, 2005. ISSN 0959437X. doi: 10.1016/j.gde.2005.09.006. URL http://www.sciencedirect.com/science/article/pii/S0959437X05001759.

[65] Gilles Vieira, Victor Sabarly, Pierre-Yves Bourguignon, Maxime Durot, François Le Fèvre, Damien Mornico, David Vallenet, Odile Bouvet, Erick Denamur, Vincent Schachter, and Claudine Médigue. Core and panmetabolism in Escherichia coli. *Journal of bacteriology*, 193(6):1461–72, March 2011. ISSN 1098-5530. doi: 10.1128/JB.01192-10. URL http://jb.asm.org/content/193/6/1461.

[66] Ron Caspi, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, and Peter D. Karp. The

MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1):D471–D480, jan 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1164. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1164.

[67] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, jan 2000. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pubmed/10592173http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC102409.

[68] John P. McCutcheon and Nancy A. Moran. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10(1):13, nov 2011. ISSN 1740-1526. doi: 10.1038/nrmicro2670. URL http://www.nature.com/doifinder/10.1038/nrmicro2670.

[69] Kimberly A Bliven and Anthony T Maurelli. Antivirulence genes: insights into pathogen evolution through gene loss. *Infection and immunity*, 80 (12):4061–70, dec 2012. ISSN 1098-5522. doi: 10.1128/IAI.00740-12. URL http://www.ncbi.nlm.nih.gov/pubmed/23045475http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3497401.

[70] Ohad Gal-Mor and B. Brett Finlay. Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cellular Microbiology*, 8(11):1707–1719, nov 2006. ISSN 1462-5814. doi: 10.1111/j.1462-5822.2006.00794.x. URL http://doi.wiley.com/10.1111/j.1462-5822.2006.00794.x.

[71] L Rouli, V Merhej, P-E Fournier, and D Raoult. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New microbes and new infections*, 7:72–85, sep 2015. ISSN 2052-2975. doi: 10.1016/j.nmni.2015.06.005. URL http://www.ncbi.nlm.nih.gov/pubmed/26442149http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4552756.

[72] Alexandra M Schnoes, Shoshana D Brown, Igor Dodevski, and Patricia C Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*, 5(12):e1000605, December 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000605. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2781113&tool=pmcentrez&rendertype=abstract.

[73] WHO Zoonoses and the Human-Animal-Ecosystems Interface, 2013. URL http://www.who.int/zoonoses/en/.

[74] Center for Genomic Epidemiology. URL http://www.genomicepidemiology.org/.

[75] Elizabeth H Loh, Carlos Zambrana-Torrelio, Kevin J Olival, Tiffany L Bogich, Christine K Johnson, Jonna A K Mazet, William Karesh, and Peter Daszak. Targeting Transmission Pathways for Emerging Zoonotic Disease Surveillance and Control. *Vector borne and zoonotic diseases (Larchmont, N.Y.)*, 15(7):432–7, jul 2015. ISSN 1557-7759. doi: 10.1089/vbz.2013.1563. URL http://www.ncbi.nlm.nih.gov/pubmed/26186515http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4507309.

[76] Kate E. Jones, Nikkita G. Patel, Marc A. Levy, Adam Storeygard, Deborah Balk, John L. Gittleman, and Peter Daszak. Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993, feb 2008. ISSN 0028-0836. doi: 10.1038/nature06536. URL http://www.nature.com/doifinder/10.1038/nature06536.

[77] Andrew D Jones, Francis M Ngure, Gretel Pelto, and Sera L Young. What Are We Assessing When We Measure Food Security? A Compendium and Review of Current Metrics 1,2. *Adv. Nutr*, 4:481–505, 2013. doi: 10.3945/an.113.004119. URL http://www.fao.org/fileadmin/templates/ess/documents/meetings{_}and{_}workshops/cfs40/001{_}What{_}Are{_}We{_}Assessing{_}When{_}We{_}Measure{_}Food{_}Secu
pdf.

[78] R D Fleischmann, M D Adams, O White, R A Clayton, E F Kirkness, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty, and J M Merrick. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, 269(5223):496–512, jul 1995. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/7542800.

[79] C M Fraser, J D Gocayne, O White, M D Adams, R A Clayton, R D Fleischmann, C J Bult, A R Kerlavage, G Sutton, J M Kelley, R D Fritchman, J F Weidman, K V Small, M Sandusky, J Fuhrmann, D Nguyen, T R Utterback, D M Saudek, C A Phillips, J M Merrick, J F Tomb, B A Dougherty, K F Bott, P C Hu, T S Lucier, S N Peterson, H O Smith, C A Hutchison, and J C Venter. The minimal gene complement of Mycoplasma genitalium. *Science (New York, N.Y.)*, 270(5235):397–403, oct 1995. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/7569993.

[80] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen,

Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, MetaHIT MetaHIT Consortium, Peer Bork, S Dusko Ehrlich, and Jun Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285): 59–65, mar 2010. ISSN 1476-4687. doi: 10.1038/nature08821. URL http://www.ncbi.nlm.nih.gov/pubmed/20203603http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3779803.

[81] Supratim Mukherjee, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Olena Verezemska, Michelle Isbandi, Alex D. Thomas, Rida Ali, Kaushal Sharma, Nikos C. Kyrpides, and T. B. K. Reddy. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Research*, 45(D1):D446–D456, jan 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw992. URL http://www.ncbi.nlm.nih.gov/pubmed/27794040http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5210664https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw992.

[82] Nikos C Kyrpides, Philip Hugenholtz, Jonathan A Eisen, Tanja Woyke, Markus Göker, Charles T Parker, Rudolf Amann, Brian J Beck, Patrick S G Chain, Jongsik Chun, Rita R Colwell, Antoine Danchin, Peter Dawyndt, Tom Dedeurwaerdere, Edward F DeLong, John C Detter, Paul De Vos, Timothy J Donohue, Xiu-Zhu Dong, Dusko S Ehrlich, Claire Fraser, Richard Gibbs, Jack Gilbert, Paul Gilna, Frank Oliver Glöckner, Janet K Jansson, Jay D Keasling, Rob Knight, David Labeda, Alla Lapidus, Jung-Sook Lee, Wen-Jun Li, Juncai Ma, Victor Markowitz, Edward R B Moore, Mark Morrison, Folker Meyer, Karen E Nelson, Moriya Ohkuma, Christos A Ouzounis, Norman Pace, Julian Parkhill, Nan Qin, Ramon Rossello-Mora, Johannes Sikorski, David Smith, Mitch Sogin, Rick Stevens, Uli Stingl, Ken-Ichiro Suzuki, Dorothea Taylor, Jim M Tiedje, Brian Tindall, Michael Wagner, George Weinstock, Jean Weissenbach, Owen White, Jun Wang, Lixin Zhang, Yu-Guang Zhou, Dawn Field, William B Whitman, George M Garrity, and Hans-Peter Klenk. Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS biology*, 12(8):e1001920,

aug 2014. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001920. URL http://www.ncbi.nlm.nih.gov/pubmed/25093819http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4122341.

[83] James R. Cole, Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 42(D1):D633–D642, jan 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1244. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1244.

[84] D Field, G Garrity, T Gray, N Morrison, J Selengut, P Sterk, T Tatusova, N Thomson, M J Allen, S V Angiuoli, M Ashburner, N Axelrod, S Baldauf, S Ballard, J Boore, G Cochrane, J Cole, P Dawyndt, P De Vos, C de Pamphilis, R Edwards, N Faruque, R Feldman, J Gilbert, P Gilna, F O Glockner, P Goldstein, R Guralnick, D Haft, and D Hancock. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnol*, 26, 2008. doi: 10.1038/nbt1360. URL https://doi.org/10.1038/nbt1360.

[85] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, Alexey V Pyshkin, Alexander V Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A Alekseyev, and Pavel A Pevzner. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 19(5): 455–77, May 2012. ISSN 1557-8666. doi: 10.1089/cmb.2012.0021. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3342519&tool=pmcentrez&rendertype=abstract.

[86] Phillip E C Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11): 987–991, nov 2011. ISSN 1087-0156. doi: 10.1038/nbt.2023. URL http://www.nature.com/articles/nbt.2023.

[87] Daniel R Zerbino. Using the Velvet de novo assembler for short-read sequencing technologies. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 11:Unit 11.5, September 2010. ISSN 1934-340X. doi: 10.1002/0471250953.bi1105s31. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2952100&tool=pmcentrez&rendertype=abstract.

[88] Carl Kingsford, Michael C Schatz, and Mihai Pop. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformat-*

ics, 11(1):21, dec 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-21. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-21.

[89] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex DeWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910), 2009. URL http://science.sciencemag.org/content/323/5910/133.

[90] Anthony Rhoads and Kin Fai Au. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, 2015. ISSN 16720229. doi: 10.1016/j.gpb.2015.08.002. URL http://www.sciencedirect.com/science/article/pii/S1672022915001345.

[91] Sharif Shaaban, Lauren A. Cowley, Sean P. McAteer, Claire Jenkins, Timothy J. Dallman, James L. Bono, and David L. Gally. Evolution of a zoonotic pathogen: investigating prophage diversity in enterohaemorrhagic Escherichia coli O157 by long-read sequencing. *Microbial Genomics*, 2(12), dec 2016. ISSN 2057-5858. doi: 10.1099/mgen.0.000096. URL http://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000096.

[92] Miriam Land, Loren Hauser, Se-Ran Jun, Intawat Nookaew, Michael R Leuze, Tae-Hyuk Ahn, Tatiana Karpinets, Ole Lund, Guruprased Kora, Trudy Wassenaar, Suresh Poudel, and David W Ussery. Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, 15(2):141–61, mar 2015. ISSN 1438-7948. doi: 10.1007/s10142-015-0433-4. URL http://www.ncbi.nlm.nih.gov/pubmed/25722247http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4361730.

[93] Thomas R. Connor, Nicholas J. Loman, Simon Thompson, Andy Smith, Joel Southgate, Radoslaw Poplawski, Matthew J. Bull, Emily Richardson, Matthew Ismail, Simon Elwood Thompson, Christine Kitchen,

Martyn Guest, Marius Bakke, Samuel K. Sheppard, and Mark J. Pallen. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microbial Genomics*, 2(9), sep 2016. ISSN 2057-5858. doi: 10.1099/mgen.0.000086. URL http://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000086.

[94] Cydney B Nielsen, Michael Cantor, Inna Dubchak, David Gordon, and Ting Wang. Visualizing genomes: techniques and challenges. *Nature Methods*, 7(3s):S5–S15, mar 2010. ISSN 1548-7091. doi: 10.1038/nmeth.1422. URL http://www.nature.com/doifinder/10.1038/nmeth.1422.

[95] Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, 3:1–8, mar 2015. ISSN 2214-7535. doi: 10.1016/J.BDQ.2015.02.001. URL http://www.sciencedirect.com/science/article/pii/S2214753515000224.

[96] D Hyatt, G L Chen, P F LoCascio, M L Land, F W Larimer, and L J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11, 2010. doi: 10.1186/1471-2105-11-119. URL https://doi.org/10.1186/1471-2105-11-119.

[97] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, jan 2017. ISSN 0028-0836. doi: 10.1038/nature21056. URL http://www.nature.com/doifinder/10.1038/nature21056.

[98] MACHINE LEARNING: THE POWER AND PROMISE OF COMPUTERS THAT LEARN BY EXAMPLE. URL https://royalsociety.org/{~}/media/policy/projects/machine-learning/publications/machine-learning-report.pdf.

[99] Oliver Stegle, Linda Payet, Jean-Louis Mergny, David J. C. MacKay, and Julian Leon Huppert. Predicting and understanding the stability of G-quadruplexes. *Bioinformatics*, 25(12):i374–i1382, jun 2009. ISSN 1460-2059. doi: 10.1093/bioinformatics/btp210. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp210.

[100] Ying Liu. Active Learning with Support Vector Machine Applied to Gene Expression Data for Cancer Classification. *Journal of Chemical Information and Computer Sciences*, 44(6):1936–1941, nov 2004. ISSN 0095-2338. doi: 10.1021/ci049810a.

URL http://www.ncbi.nlm.nih.gov/pubmed/15554662http://pubs.acs.org/doi/abs/10.1021/ci049810a.

[101] I. Pournara and L. Wernisch. Reconstruction of gene networks using Bayesian learning and manipulation experiments. *Bioinformatics*, 20 (17):2934–2942, nov 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth337. URL http://www.ncbi.nlm.nih.gov/pubmed/15180938https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bth337.

[102] Yukiko Fujiwara, Yoshiko Yamashita, Tsutomu Osoda, Minoru Asogawa, Chiaki Fukushima, Masaaki Asao, Hideshi Shimadzu, Kazuya Nakao, and Ryo Shimizu. Virtual Screening System for Finding Structurally Diverse Hits by Active Learning. *Journal of Chemical Information and Modeling*, 48(4):930–940, apr 2008. ISSN 1549-9596. doi: 10.1021/ci700085q. URL http://www.ncbi.nlm.nih.gov/pubmed/18351729http://pubs.acs.org/doi/abs/10.1021/ci700085q.

[103] James J. Davis, Sébastien Boisvert, Thomas Brettin, Ronald W. Kenyon, Chunhong Mao, Robert Olson, Ross Overbeek, John Santerre, Maulik Shukla, Alice R. Wattam, Rebecca Will, Fangfang Xia, and Rick Stevens. Antimicrobial Resistance Prediction in PATRIC and RAST. *Scientific Reports*, 6(1):27930, sep 2016. ISSN 2045-2322. doi: 10.1038/srep27930. URL http://www.ncbi.nlm.nih.gov/pubmed/27297683http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4906388http://www.nature.com/articles/srep27930.

[104] Narjeskhatoon Habibi, Siti Z Mohd Hashim, Alireza Norouzi, and Mohammed Razip Samian. A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli. *BMC bioinformatics*, 15:134, may 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-134. URL http://www.ncbi.nlm.nih.gov/pubmed/24885721http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4098780.

[105] Jérôme Azé, Christophe Sola, Jian Zhang, Florian Lafosse-Marin, Memona Yasmin, Rubina Siddiqui, Kristin Kremer, Dick van Soolingen, and Guislaine Refrégier. Genomics and Machine Learning for Taxonomy Consensus: The Mycobacterium tuberculosis Complex Paradigm. *PLOS ONE*, 10(7):e0130912, jul 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130912. URL http://www.ncbi.nlm.nih.gov/pubmed/26154264http://www.pubmedcentral.

nih.gov/articlerender.fcgi?artid=PMC4496040http://dx.plos.org/10.1371/journal.pone.0130912.

[106] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[107] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[108] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[109] Geoffrey Mainda, Nadejda Lupolova, Linda Sikakwa, Paul R Bessell, John B Muma, Deborah V Hoyle, Sean P McAteer, Kirsty Gibbs, Nicola J Williams, Samuel K Sheppard, Roberto M La Ragione, Guido Cordoni, Sally A Argyle, Sam Wagner, Margo E Chase-Topping, Timothy J Dallman, Mark P Stevens, Barend M deC Bronsvoort, and David L Gally. Phylogenomic approaches to determine the zoonotic potential of Shiga toxin-producing Escherichia coli (STEC) isolated from Zambian dairy cattle. *Scientific reports*, 6: 26589, may 2016. ISSN 2045-2322. doi: 10.1038/srep26589. URL http://www.ncbi.nlm.nih.gov/pubmed/27220895http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4879551.

[110] Nadejda Lupolova, Timothy J Dallman, Louise Matthews, James L Bono, and David L Gally. Support vector machine applied to predict the zoonotic potential of E. coli O157 cattle isolates. *Proceedings of the National Academy of Sciences of the United States of America*, 113(40): 11312–11317, oct 2016. ISSN 1091-6490. doi: 10.1073/pnas.1606567113. URL http://www.ncbi.nlm.nih.gov/pubmed/27647883http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5056084.

[111] Peter J A Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–71, April 2010. ISSN 1362-4962. doi: 10.1093/nar/gkp1137. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2847217&tool=pmcentrez&rendertype=abstract.

[112] Simon Andrews. Fastqc high throughput sequence qc report version 0.10.0. URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[113] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, May 2011. doi: http://dx.doi.org/10.14806/ej.17.1.200. URL http://journal.embnet.org/index.php/embnetjournal/article/view/200.

[114] Nadejda Lupolova, Tim J. Dallman, Nicola J. Holden, and David L. Gally. Patchy promiscuity: machine learning applied to predict the host specificity of Salmonella enterica and Escherichia coli. *Microbial Genomics*, oct 2017. ISSN 2057-5858. doi: 10.1099/mgen.0.000135. URL http://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000135.v1.

[115] David R Kelley, Michael C Schatz, and Steven L Salzberg. Quake: quality-aware detection and correction of sequencing errors. *Genome biology*, 11(11):R116, January 2010. ISSN 1465-6914. doi: 10.1186/gb-2010-11-11-r116. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3156955&tool=pmcentrez&rendertype=abstract.

[116] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013. doi: 10.1093/bioinformatics/btt086. URL http://bioinformatics.oxfordjournals.org/content/29/8/1072.abstract.

[117] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, 30(14):2068–9, July 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu153. URL http://bioinformatics.oxfordjournals.org/content/30/14/2068.long.

[118] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, 30(9):1312–3, May 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu033. URL http://bioinformatics.oxfordjournals.org/content/early/2014/01/21/bioinformatics.btu033.abstract?keytype=ref&ijkey=VTEqgUJYCDcf0kP.

[119] Andrew J Page, Carla A Cummins, Martin Hunt, Vanessa K Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Jacqueline A Keane, and Julian Parkhill. Roary: Rapid large-scale prokaryote pan genome analysis. Technical report, May 2015. URL http://biorxiv.org/content/early/2015/05/13/019315.abstract.

[120] Olivier Clermont, Julia K Christenson, Erick Denamur, and David M Gordon. The Clermont Escherichia coli phylo-typing method revisited: im-

provement of specificity and detection of new phylo-groups. *Environmental microbiology reports*, 5(1):58–65, February 2013. ISSN 1758-2229. doi: 10.1111/1758-2229.12019. URL http://www.ncbi.nlm.nih.gov/pubmed/23757131.

[121] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC bioinformatics*, 10:421, January 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2803857&tool=pmcentrez&rendertype=abstract.

[122] M. Inouye, H. Dashnow, L. Raven, M. B. Schultz, B. J. Pope, T. Tomita, J. Zobel, and K. E. Holt. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *bioRxiv*, 6(11):006627, January 2014. ISSN 1756-994X. doi: 10.1101/006627. URL http://biorxiv.org/content/early/2014/06/26/006627.abstract.

[123] Katrine G. Joensen, Anna M. M. Tetzschner, Atsushi Iguchi, Frank M. Aarestrup, and Flemming Scheutz. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. *Journal of Clinical Microbiology*, 53(8):2410–2426, aug 2015. ISSN 0095-1137. doi: 10.1128/JCM.00008-15. URL http://jcm.asm.org/lookup/doi/10.1128/JCM.00008-15.

[124] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–7, January 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh340. URL http://nar.oxfordjournals.org/content/32/5/1792.long.

[125] Matthew Kearse, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, Alex Cooper, Sidney Markowitz, Chris Duran, Tobias Thierer, Bruce Ashton, Peter Meintjes, and Alexei Drummond. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics (Oxford, England)*, 28(12):1647–9, June 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts199. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3371832&tool=pmcentrez&rendertype=abstract.

[126] Bruno Contreras-Moreira and Pablo Vinuesa. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and environmental microbiology*, 79(24):7696–701, December 2013. ISSN 1098-5336. doi: 10.1128/AEM.02411-13. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3837814&tool=pmcentrez&rendertype=abstract.

[127] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

[128] David J Edwards and Kathryn E Holt. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial informatics and experimentation*, 3(1):2, jan 2013. ISSN 2042-5783. doi: 10.1186/2042-5783-3-2. URL http://www.microbialinformaticsj.com/content/3/1/2.

[129] Fernanda Morcatti Coura, Soraia de Araújo Diniz, Marcos Xavier Silva, Jamili Maria Suhet Mussi, Silvia Minharro Barbosa, Andrey Pereira Lage, Marcos Bryan Heinemann, and Marcos Bryan Heinemann. Phylogenetic Group Determination of Escherichia coli Isolated from Animals Samples. *The Scientific World Journal*, 2015:1–4, 2015. ISSN 2356-6140. doi: 10.1155/2015/258424. URL http://www.hindawi.com/journals/tswj/2015/258424/.

[130] Mehdy Ratajczak, Emilie Laroche, Thierry Berthe, Olivier Clermont, Barbara Pawlak, Erick Denamur, and Fabienne Petit. Influence of hydrological conditions on the Escherichia coli population structure in the water of a creek on a rural watershed. *BMC microbiology*, 10: 222, aug 2010. ISSN 1471-2180. doi: 10.1186/1471-2180-10-222. URL http://www.ncbi.nlm.nih.gov/pubmed/20723241http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2933670.

[131] Marie-Hélène Nicolas-Chanoine, Xavier Bertrand, and Jean-Yves Madec. Escherichia coli ST131, an intriguing clonal group. *Clinical microbiology reviews*, 27(3):543–74, jul 2014. ISSN 1098-6618. doi: 10.1128/CMR.00125-13. URL http://www.ncbi.nlm.nih.gov/pubmed/24982321http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4135899.

[132] Sam Wagner, Nadejda Lupolova, David L. Gally, and Sally A. Argyle. Convergence of plasmid architectures drives emergence of multi-drug resistance in a clonally diverse Escherichia coli population from a veterinary clinical care setting. *Veterinary Microbiology*, 211:6–14, nov 2017. ISSN 0378-1135. doi: 10.1016/J.VETMIC.2017.09.016. URL https://www.sciencedirect.com/science/article/pii/S037811351730665X.

[133] Fernanda Morcatti Coura, Soraia de Araújo Diniz, Marcos Xavier Silva, Jamili Maria Suhet Mussi, Silvia Minharro Barbosa, Andrey Pereira Lage, and Marcos Bryan Heinemann. Phylogenetic Group Determination of Escherichia coli Isolated from Animals Samples.

TheScientificWorldJournal, 2015:258424, 2015. ISSN 1537-744X. doi: 10.1155/2015/258424. URL http://www.ncbi.nlm.nih.gov/pubmed/26421310http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4572460.

[134] Camila Carlos, Mathias M Pires, Nancy C Stoppe, Elayse M Hachich, Maria Iz Sato, Tânia At Gomes, Luiz A Amaral, and Laura Mm Ottoboni. Escherichia coli phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. *BMC Microbiology*, 10, 2010. URL http://www.biomedcentral.com/1471-2180/10/161.

[135] Lili Wang, Mitsuko Wakushima, Tetsu Aota, Yuka Yoshida, Toshimasa Kita, Tomofumi Maehara, Jun Ogasawara, Changsun Choi, Yoichi Kamata, Yukiko Hara-Kudo, and Yoshikazu Nishikawa. Specific properties of enteropathogenic Escherichia coli isolates from diarrheal patients and comparison to strains from foods and fecal specimens from cattle, swine, and healthy carriers in Osaka City, Japan. *Applied and environmental microbiology*, 79(4):1232–40, feb 2013. ISSN 1098-5336. doi: 10.1128/AEM.03380-12. URL http://www.ncbi.nlm.nih.gov/pubmed/23220963http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3568616.

[136] Marie Touchon, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, Stéphane Bonacorsi, Christiane Bouchier, Odile Bouvet, Alexandra Calteau, Hélène Chiapello, Olivier Clermont, Stéphane Cruveiller, Antoine Danchin, Médéric Diard, Carole Dossat, Meriem El Karoui, Eric Frapy, Louis Garry, Jean Marc Ghigo, Anne Marie Gilles, James Johnson, Chantal Le Bouguénec, Mathilde Lescat, Sophie Mangenot, Vanessa Martinez-Jéhanne, Ivan Matic, Xavier Nassif, Sophie Oztas, Marie Agnès Petit, Christophe Pichon, Zoé Rouy, Claude Saint Ruf, Dominique Schneider, Jérôme Tourret, Benoit Vacherie, David Vallenet, Claudine Médigue, Eduardo P C Rocha, and Erick Denamur. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS genetics*, 5(1):e1000344, January 2009. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000344. URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000344#s3.

[137] David M. Gordon, Olivier Clermont, Heather Tolley, and Erick Denamur. Assigning Escherichia coli strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environmental Microbiology*, 10(10):2484–2496, oct 2008. ISSN 14622912. doi: 10.1111/j.1462-2920.2008.01669.x. URL

http://www.ncbi.nlm.nih.gov/pubmed/18518895http://doi.wiley.com/10.1111/j.1462-2920.2008.01669.x.

[138] Lapo Mughini-Gras, Eelco Franz, and Wilfrid van Pelt. New paradigms for Salmonella source attribution based on microbial subtyping. *Food Microbiology*, 71:60–67, may 2018. ISSN 0740-0020. doi: 10.1016/J.FM.2017.03.002. URL https://www.sciencedirect.com/science/article/pii/S0740002016308917?via{%}3Dihub.

[139] Leonardo V. de Knegt, Sara M. Pires, Charlotta Löfström, Gitte Sørensen, Karl Pedersen, Mia Torpdahl, Eva M. Nielsen, and Tine Hald. Application of Molecular Typing Results in Source Attribution Models: The Case of Multiple Locus Variable Number Tandem Repeat Analysis (MLVA) of Salmonella Isolates Obtained from Integrated Surveillance in Denmark. *Risk Analysis*, 36(3):571–588, mar 2016. ISSN 02724332. doi: 10.1111/risa.12483. URL http://doi.wiley.com/10.1111/risa.12483.

[140] Lapo Mughini-Gras, Eelco Franz, and Wilfrid van Pelt. New paradigms for Salmonella source attribution based on microbial subtyping. *Food Microbiology*, 71:60–67, may 2018. ISSN 0740-0020. doi: 10.1016/J.FM.2017.03.002. URL https://www.sciencedirect.com/science/article/pii/S0740002016308917?via{%}3Dihub.

[141] Nigel Goldenfeld and Carl Woese. Biology's next revolution. *Nature*, 445 (7126):369, January 2007. ISSN 1476-4687. doi: 10.1038/445369a. URL http://www.ncbi.nlm.nih.gov/pubmed/17251963.

[142] Massimo Andreatta, Morten Nielsen, Frank Møller Aarestrup, and Ole Lund. In silico prediction of human pathogenicity in the γ-proteobacteria. *PloS one*, 5(10):e13680, oct 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0013680. URL http://www.ncbi.nlm.nih.gov/pubmed/21048922http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2965111.

[143] NIAID Emerging Infectious Diseases/Pathogens | NIH: National Institute of Allergy and Infectious Diseases. URL https://www.niaid.nih.gov/research/emerging-infectious-diseases-pathogens.

[144] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[145] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[146] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[147] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[148] Lior Rokach and Oded Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.

[149] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155 (2):945–959, 2000.

[150] Ola Brynildsrud, Jon Bohlin, Lonneke Scheffer, and Vegard Eldholm. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology*, 17(1):238, dec 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-1108-8. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1108-8.

[151] Jari Oksanen, F Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan Mcglinn, Peter R Minchin, R B O 'hara, Gavin L Simpson, Peter Solymos, M Henry, H Stevens, Eduard Szoecs, Helene Wagner, and Maintainer Jari Oksanen. Title Community Ecology Package. 2018. URL https://github.com/vegandevs/vegan/issueshttps://github.com/vegandevs/vegan.

[152] M. O. Hill. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, 54(2):427–432, mar 1973. ISSN 00129658. doi: 10.2307/1934352. URL http://doi.wiley.com/10.2307/1934352.

[153] R. H. Whittaker. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, 30(3):279–338, feb 1960. ISSN 00129615. doi: 10.2307/1943563. URL http://doi.wiley.com/10.2307/1943563.

[154] Marti J. Anderson, Kari E. Ellingsen, and Brian H. McArdle. Multivariate dispersion as a measure of beta diversity. *Ecology Letters*, 9(6):683–693, jun 2006. ISSN 1461-023X. doi: 10.1111/j.1461-0248.2006.00926.x. URL http://doi.wiley.com/10.1111/j.1461-0248.2006.00926.x.

[155] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. NbClust : An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61 (6):1–36, nov 2014. ISSN 1548-7660. doi: 10.18637/jss.v061.i06. URL http://www.jstatsoft.org/v61/i06/.

[156] Leonard Kaufman and Peter J. Rousseeuw, editors. *Finding Groups in Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, mar 1990. ISBN 9780470316801. doi: 10.1002/9780470316801. URL http://doi.wiley.com/10.1002/9780470316801.

[157] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesú M Pé rez, and Iñ igo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46:243–256, 2012. doi: 10.1016/j.patcog.2012.07.021. URL https://pdfs.semanticscholar.org/a522/fb4646ad19fe88893b90e6fbc1faa1470976.pdf.

[158] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[159] David A Lacher and Edward D O 'donnell. Comparison of Multidimensional Scaling and Principal Component Analysis of Interspecific Variation in Bacteria*. *ANNALS OF CLINICAL AND LABORATORY SCIENCE*, 18(6). URL http://www.annclinlabsci.org/content/18/6/455.full.pdf.

[160] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–25, jul 1987. ISSN 1537-1719. doi: 10.1093/oxfordjournals.molbev.a040454. URL http://www.ncbi.nlm.nih.gov/pubmed/3447015https://academic.oup.com/mbe/article/4/4/406/1029664/The-neighborjoining-method-a-new-method-for.

[161] Satya Dash, Hiroyuki Sano, Justin J Rochford, Robert K Semple, Giles Yeo, Caroline S S Hyden, Maria A Soos, James Clark, Andrew Rodin, Claudia Langenberg, Celine Druet, Katherine A Fawcett, Y C Loraine Tung, Nicolas J Wareham, Inês Barroso, Gustav E Lienhard, Stephen O'Rahilly, and David B Savage. A truncation mutation in TBC1D4 in a family with acanthosis nigricans and postprandial hyperinsulinemia. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9350–5, jun 2009. ISSN 1091-6490. doi: 10.1073/pnas.0900909106. URL http://www.ncbi.nlm.nih.gov/pubmed/19470471http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2695078.

[162] Lars Barquist, Matthew Mayho, Carla Cummins, Amy K Cain, Christine J Boinett, Andrew J Page, Gemma C Langridge, Michael A Quail, Jacqueline A Keane, and Julian Parkhill. The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. *Bioinformatics (Oxford, England)*, 32(7):1109–11, 2016. ISSN 1367-4811. doi: 10.1093/bioinformatics/btw022. URL http://www.ncbi.nlm.nih.gov/pubmed/26794317http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4896371.