ROSLIN

Setting the scene: applications of whole genome sequencing in a ‘one health’ context

Prof. David Gally, Division of Infection and Immunity, The Roslin Institute, University of Edinburgh

THE UNIVERSITY *of* EDINBURGH

BBSRC
20 Years of Pioneering
Great British Bioscience

One health and WGS

Sequencing **captures diversity** and is re-defining bacterial taxonomy

The ‘bottom line’ of a sequence provides **precision** in terms of identification

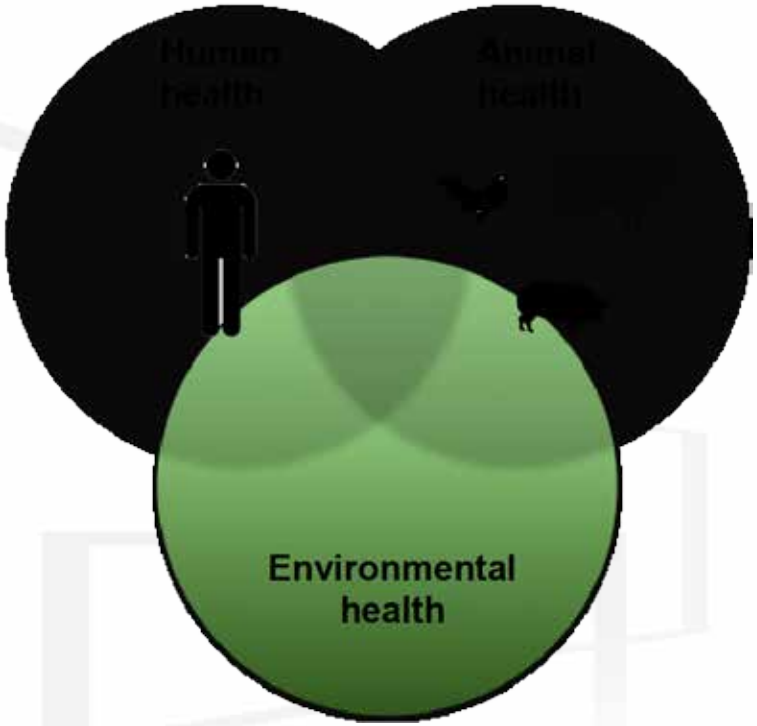
Identify which specific bacterial sub-clusters are associated with disease (or another trait)

Track human-human and animal-to-human (or visa versa`) spread

Identify **origins and vehicles** of transmission

Predict pathogenic isolates and isolates with zoonotic potential

Can focus on particular genes, e.g. those associated with **antibiotic resistance**

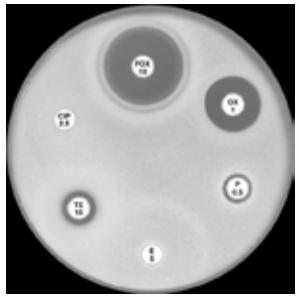


WGS offers precise identification and phenotype prediction

Sample

Isolate

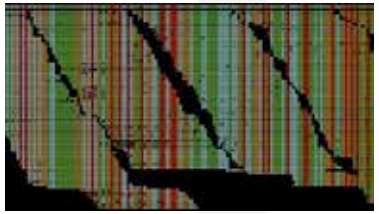
Resistance profile



Good guy or bad guy

DNA

Sequence



Analysis

What bacteria

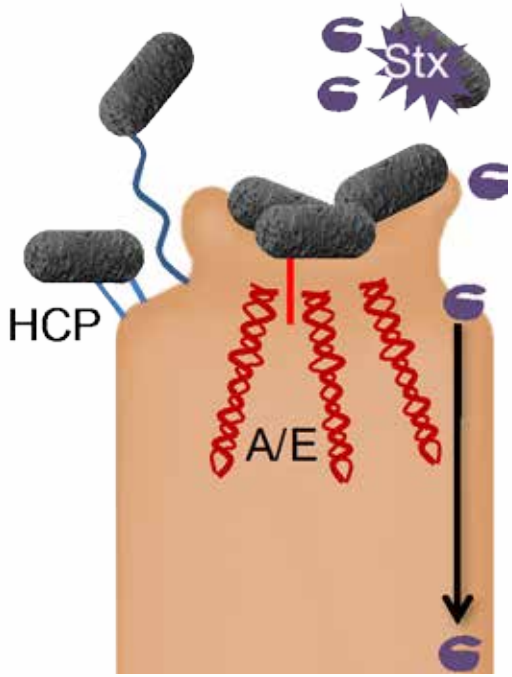
What sub-type

What virulence genes

What resistance genes

Previous detection

Enterohemorrhagic *E. coli* O157:H7 analysis as an example



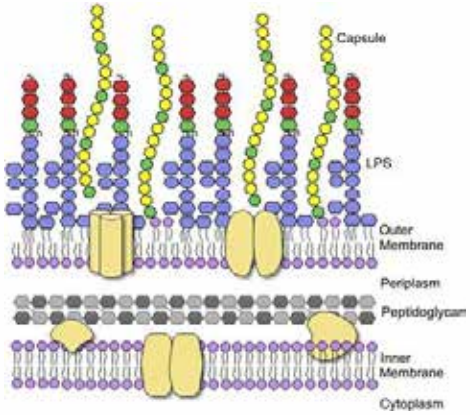
Specific type of *E. coli* considered ‘commensal’ in ruminants, especially cattle, but associated with severe disease in humans.

EHEC defined by possession of **type 3 secretion system** which drives host colonisation and **Shiga-toxin** (Stx) which is associated with severe disease pathology in humans

About 1000 human cases a year in the UK, >15,000 in the USA.

Originally known as ’the burger bug’ but now the majority of infections through contamination of ‘raw’ foodstuffs and direct contact

From the isolate sequence we can extract valuable typing data



From the sequence we can analyse specific genes to define:

1. **Serotype such as O157 LPS and H7 flagella**
2. **Potential virulence by examining toxins and adherence mechanisms. For example certain Shiga toxin subtypes (Stx2a) is associated with more severe pathology**

Can also extract data to examine an isolates relationship with others based on older typing methods such as PFGE & phage typing

Precision of relatedness – SNP address can be related to ZIP code

Reference genome

Isolate sequence reads

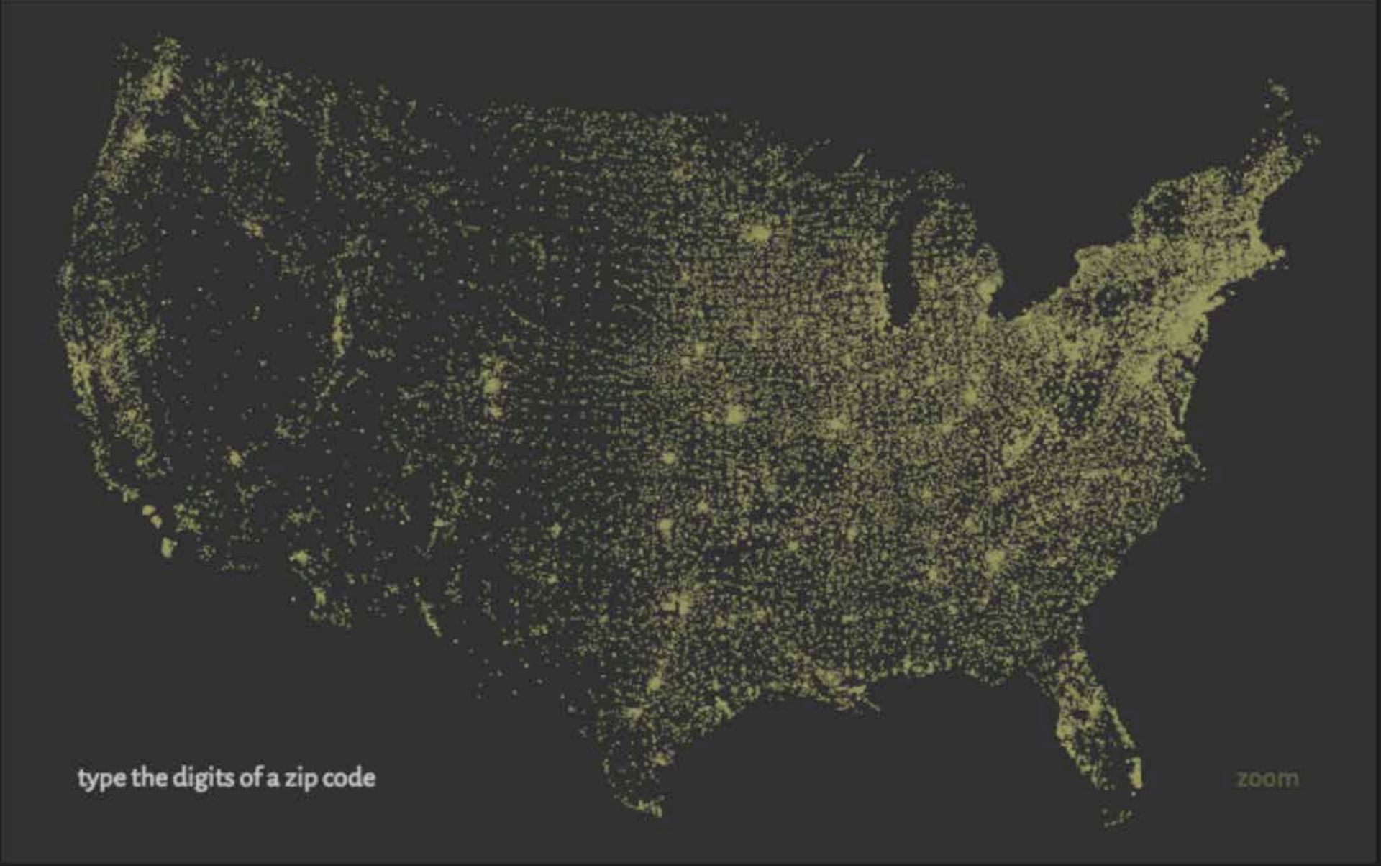
A	**G**	**T**	**C**	**G**	**C**	**G**	**T**	**A**	**T**	**G**	**T**	**C**	**T**	**G**	**A**	**C**	**C**	**C**
A	G	T	C	G	C			A	T	G	T	A	T					
			C	G	C	G	T	A			T	A	T	G	A	C	C	C
A	G	T	C	G	C	G						A	T	G	A	C	C	C
						G	T	A	T	G	T	A						
		T	C	G	C	G	T			G	T	A	T	G	A			
A	G	T	C	G	C	G						A	T	G	A	C		
				G	C	G	T	A	T									
									T	G	T	A	T					
					C	G	T	A	T	G								

SNP

READ FROM RIGHT TO LEFT

1.2.3.158.199.222.243

- 250 SNP level
- 100 SNP level
- 50 SNP level
- 25 SNP level
- 10 SNP level
- 5 SNP level
- 0 SNP level



http://benfry.com/zipdecode/

BBSRC
20 Years of Pioneering
Great British Bioscience

http://benfry.com/zipdecode/

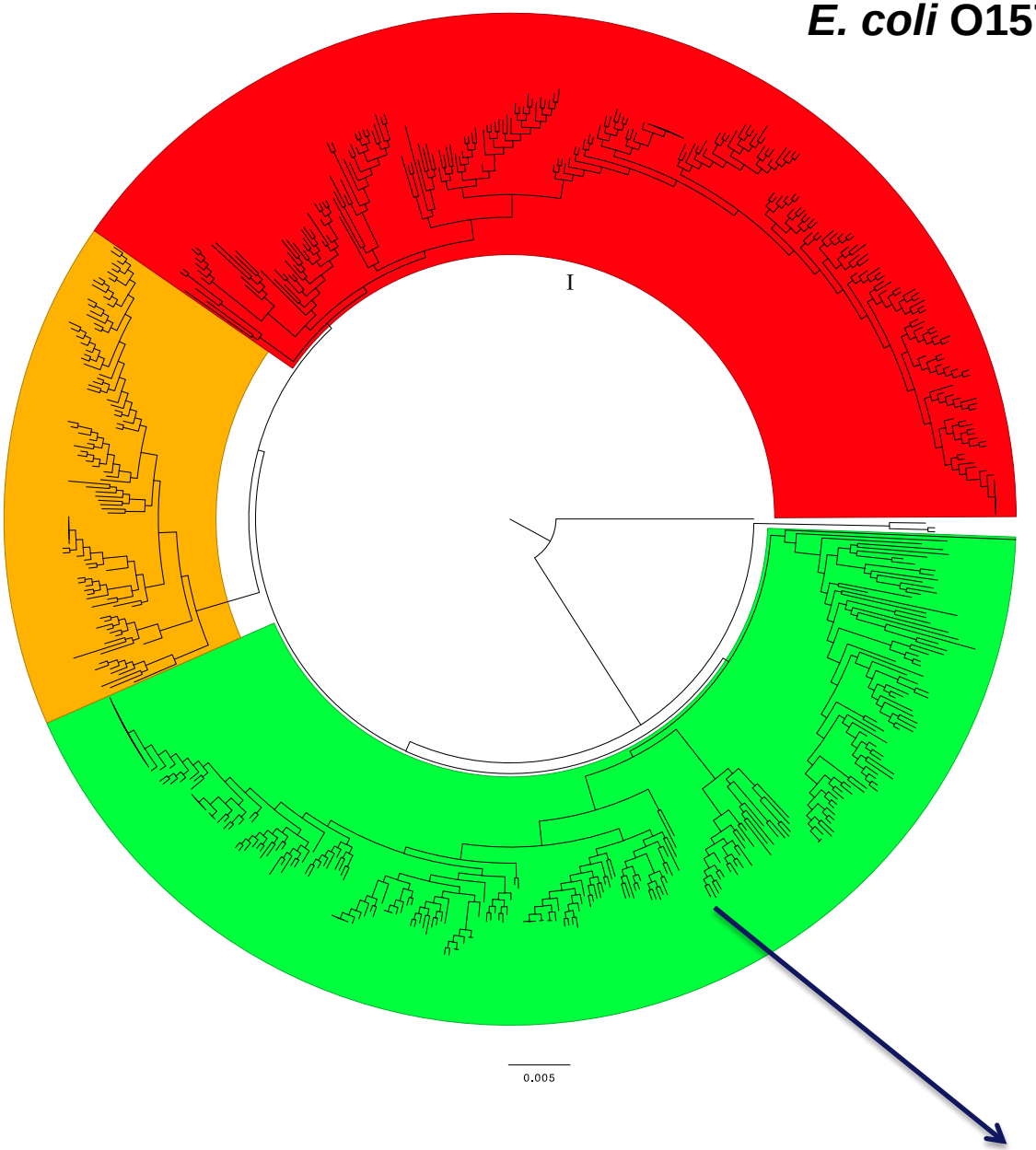
http://benfry.com/zipdecode/

http://benfry.com/zipdecode/

http://benfry.com/zipdecode/

http://benfry.com/zipdecode/

E. coli O157 – 2527 genomes sequenced

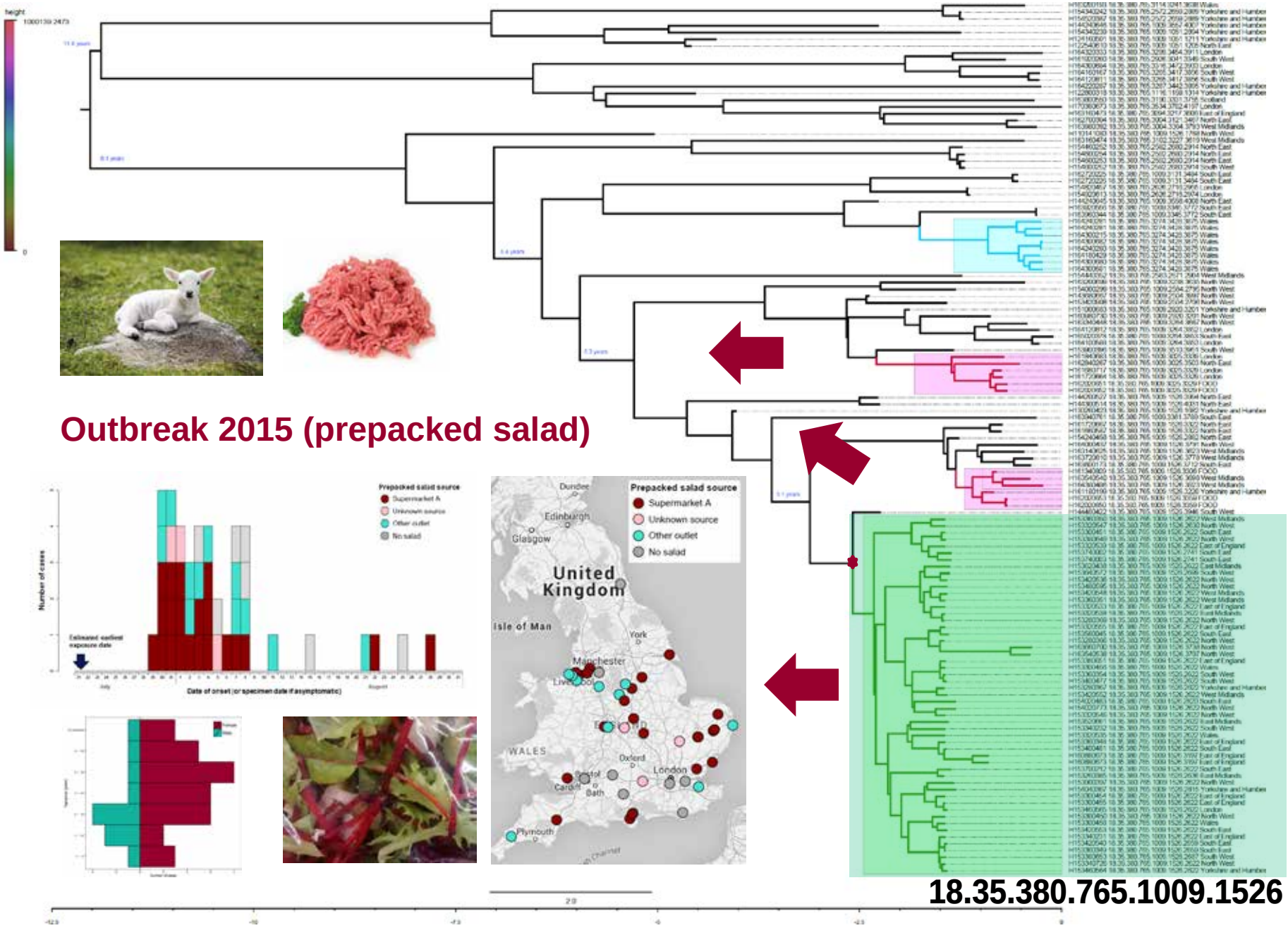


Clustering (relatedness) is at different levels depending on how many SNPs are allowed between groups

79	t250 clusters
240	t100 clusters
519	t50 clusters
1001	t25 clusters
1423	t10 clusters
1519	t5 clusters
1885	t0 clusters

and so a 4 number ZIP code can be generated at the 25 SNP level

18.35.397.765



- 49 cases can be confidently linked despite wide distribution (green cluster) as share SNP address (to 5 SNPs)
- Later isolates (pink) from which outbreak isolates are descended were from lamb products

British *E. coli O157* cattle survey (2014-2015)

Scotland: 110 farms (26 positive)
England & Wales: 160 farms (34 positive)

	Herd level	CI	Pat level	CI
Scotland	0.236	0.166 – 0.325	0.104	0.093 – 0.116
England and Wales	0.213	0.156 – 0.283	0.080	0.071 – 0.091

Spatial distribution of *E. coli* O157 sub-types in the country helps inform local risk management and the tracing of infection source

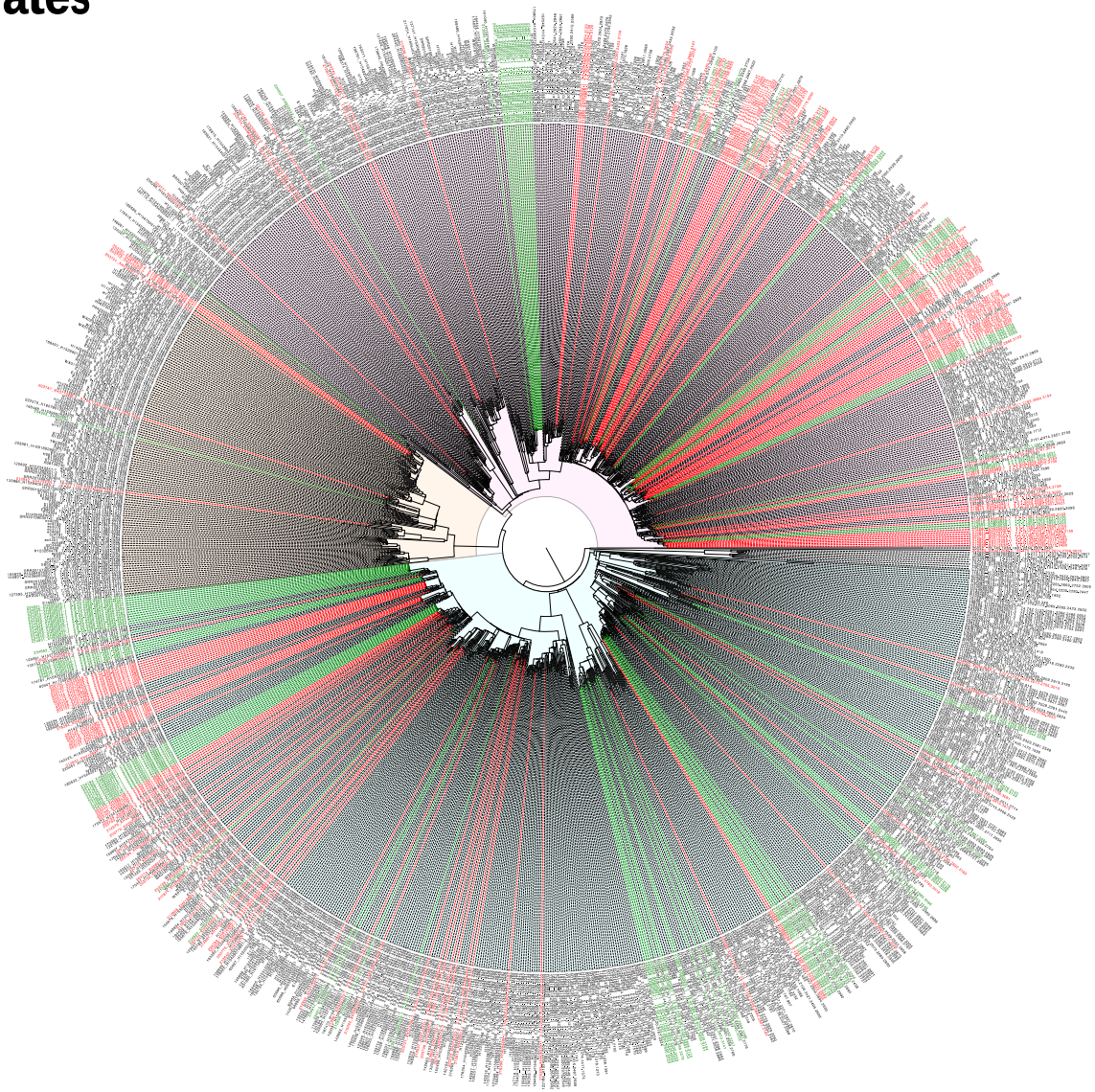
Genome Sequencing of isolates

1. Can determine if the isolate is part of an outbreak.

2. Can determine if it is a local isolate or one that is imported; and to some extent where from.

3. Can we predict which subset of isolates in animal reservoirs are a threat to human health?

Red = human
Green = cattle



Microbial Genomics, 10.1099/mgen.0.000096
Microbial Genomics, 10.1099/mgen.0.000029
Proc Natl Acad Sci USA, 10.1073/pnas.1606567113

To improve prediction capacity it helps to use all the genome information

Illumina sequences

Core

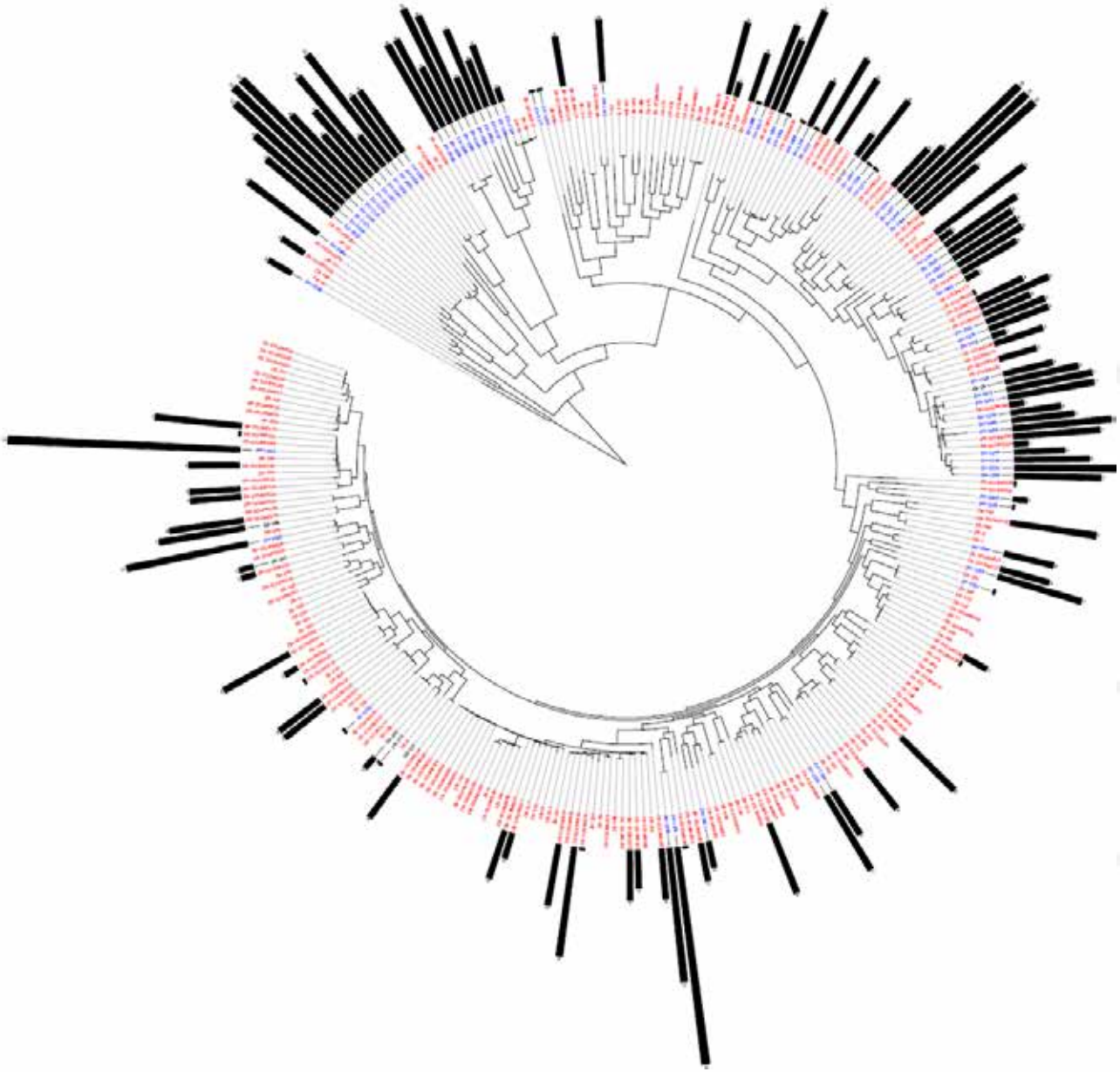
- Map to reference
  Phylogenetic –SNP based
- Type (MLST, phylogroup, serogroup, etc)

Core + Accessory

- De novo assembly
- Annotate/identify
- Cluster proteins (e.g. Roary)
- Predictions (SVM, R e1071)

Including accessory and core genome sequence in analyses to improve evaluation of the risk posed by an isolate

Extracting AMR gene content from *E. coli* sequence data

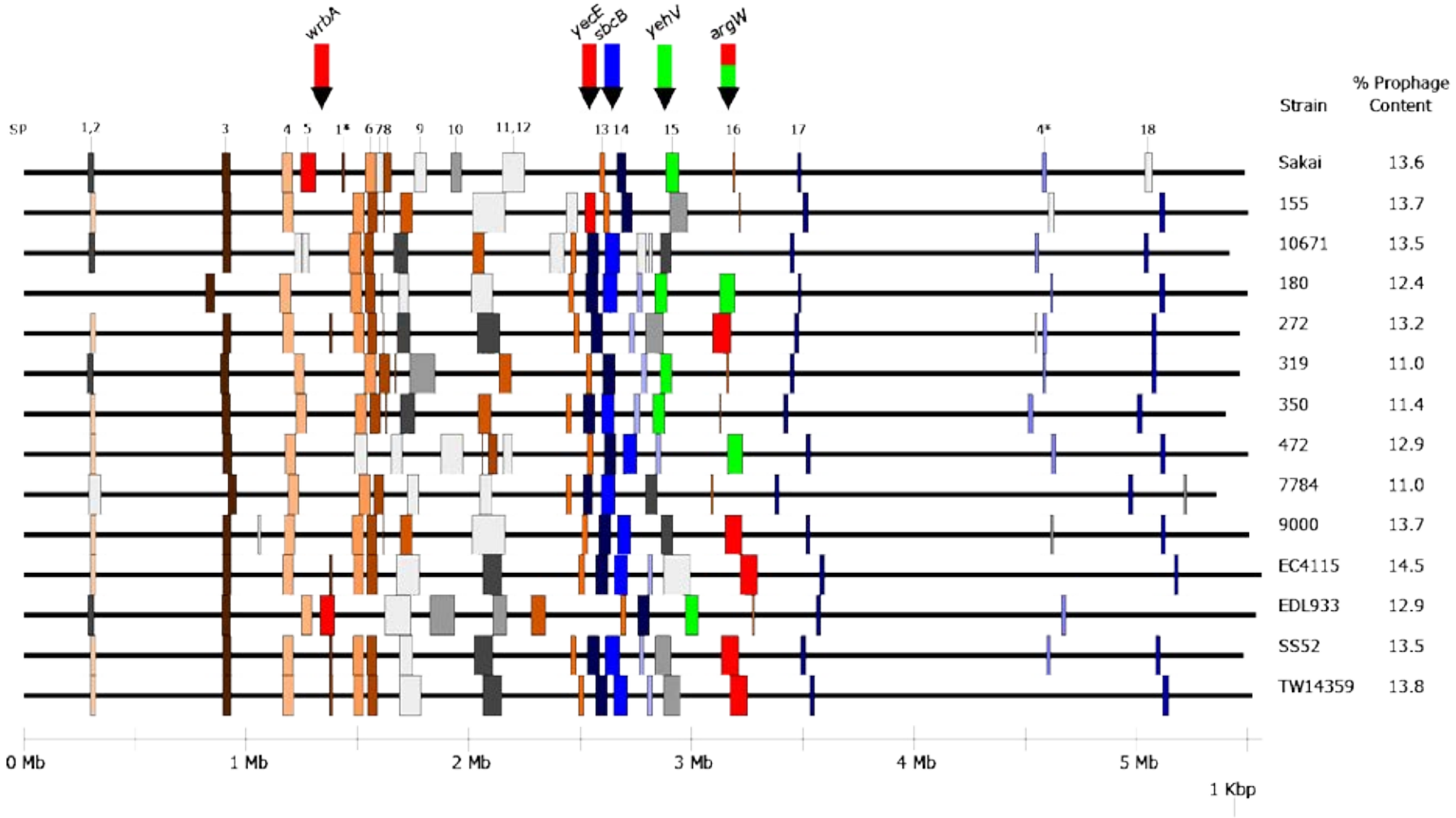


(a) Sequencing *E. coli* isolates from cattle (red) and humans (blue) in around Lusaka, Zambia.

(b) From de novo assemblies extract resistance genes. Number of resistance genes are shown with the bar plots (n= 1 to 17)

(c) Evidence that the majority of human isolates encode multiple resistances; less resistance in bovine isolates.

Long read sequencing I: to Identify integrated bacteriophage regions as they can be critical for isolate variation and disease potential: prophage encode Stx genes in EHEC



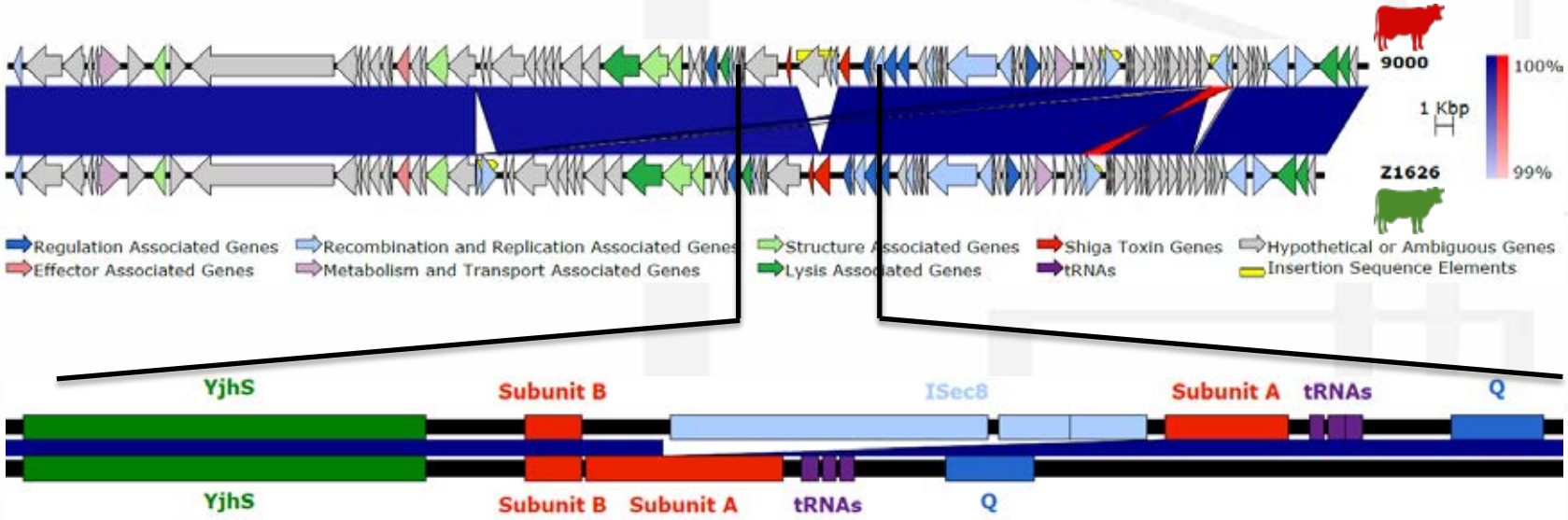
Chromosome alignment of 14 EHEC O157 sequences with phage regions represented as colored blocks

stx2a
stx2c
stx1a
Clusters with 1 Prophage
Clusters with 2 Prophages
Clusters with 3 to 5 Prophages

Microbial Genomics (2016) Vol 2: e000096

Long read sequencing II: resolving the position of repeated sequences

- Short read sequencing finds it very difficult to assess the position of short repeated regions such as insertion sequence (IS) elements
- In the example below the isolate has an IS element that disrupts the Shiga toxin 2a gene, so it now longer expresses this important toxin
- Need longer read sequencing for this level of resolution

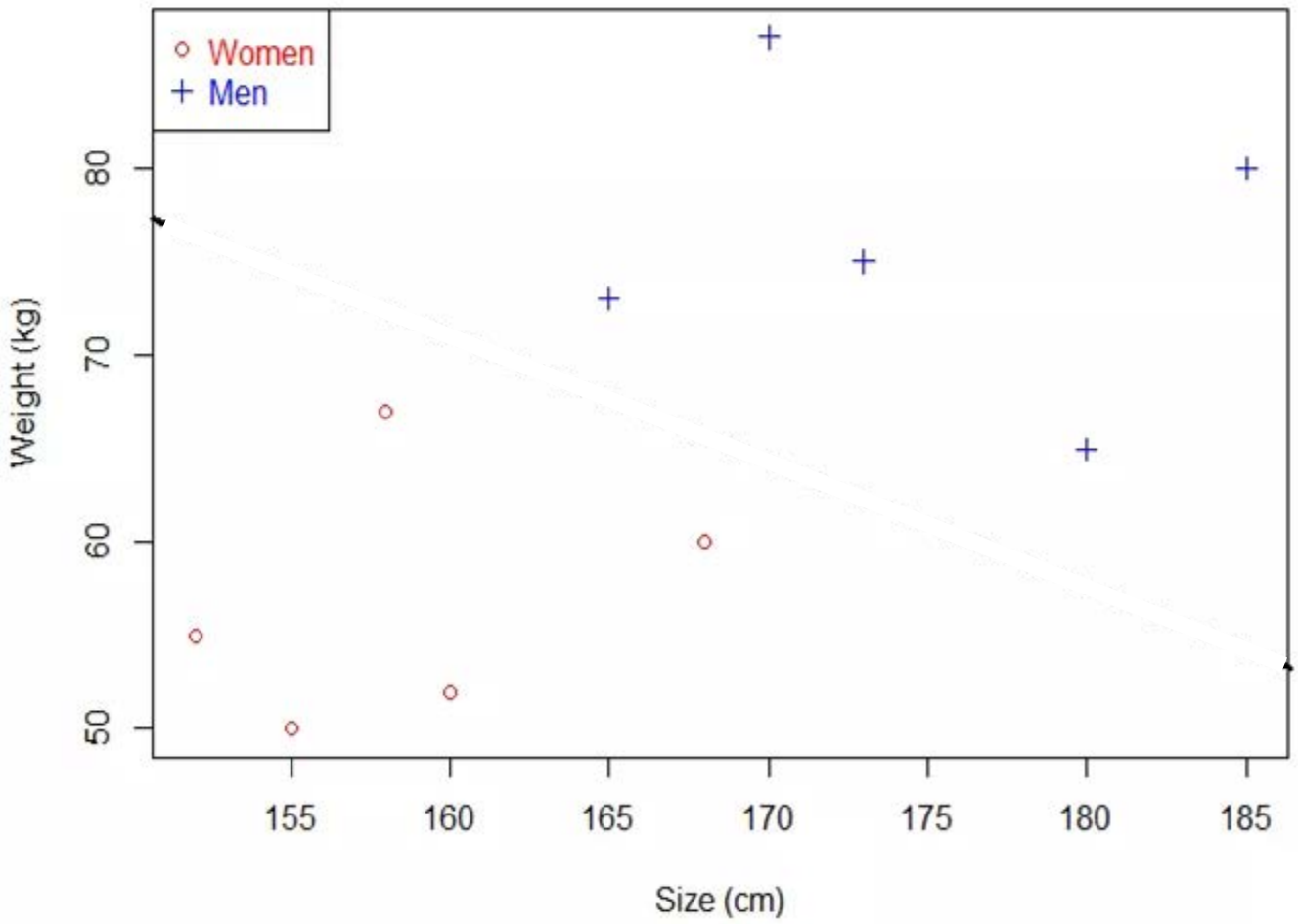


A machine learning approach?

- The capacity of the computer to learn from the experience i.e. to modify its processing on the basis of newly acquired information
  Oxford English dictionary
- 1936, R. A. Fisher suggested the first algorithm for pattern recognition (Fisher 1936)
- Applications: from inbox spam filters, bank fraud, your shopping habits, tumour identification. AlphaGo: DeepMind computer teaches itself to become world's best Go player-18 Oct 2017

Support Vector Machine (SVM) – one type of machine learning

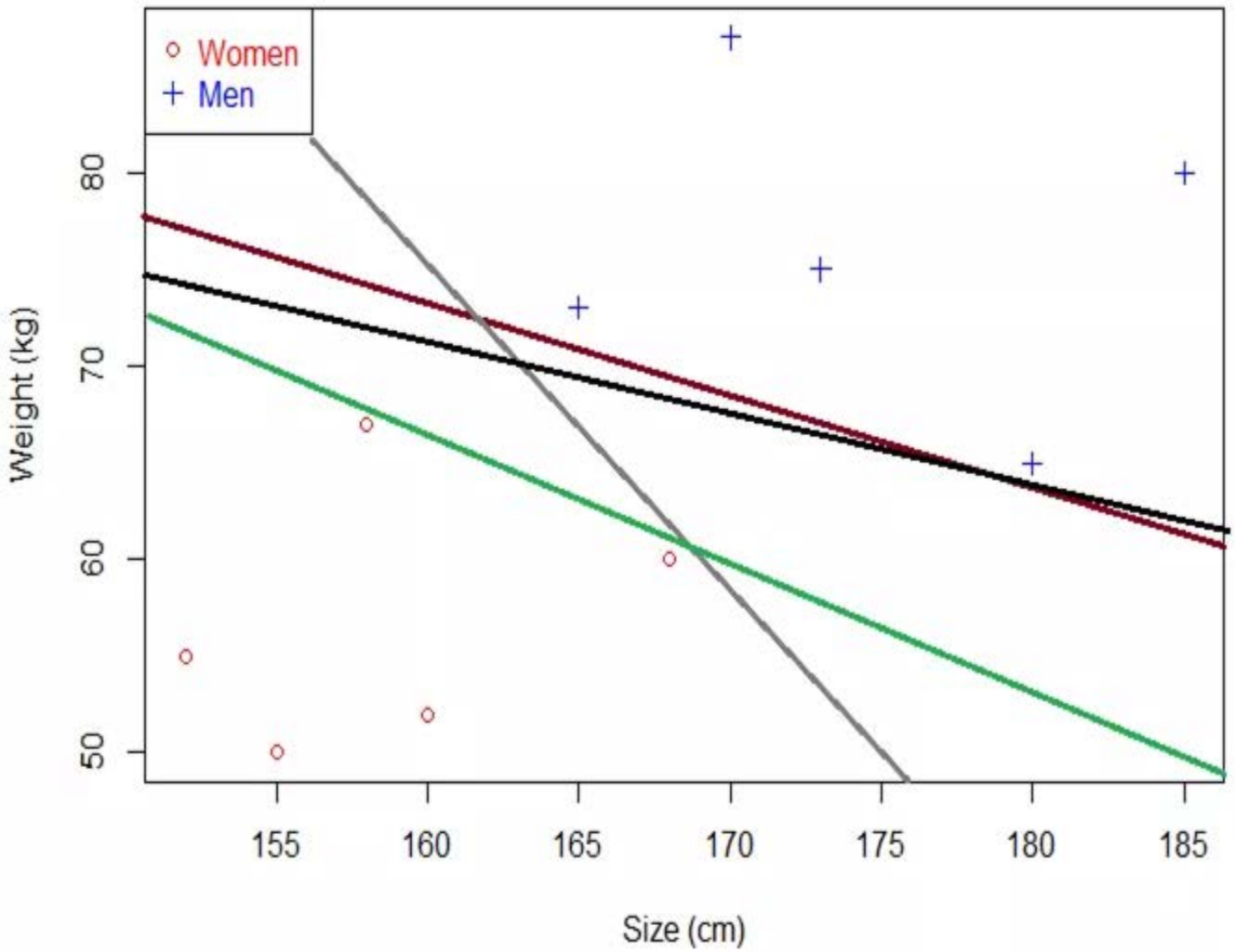
This supervised machine learning approach requires a **training dataset** with defined groups



http://www.svm-tutorial.com/2014/11/svm-understanding-math-part-1/

SVM assigns a decision boundary with the training set

The rule is the decision boundary that will be applied to unseen data

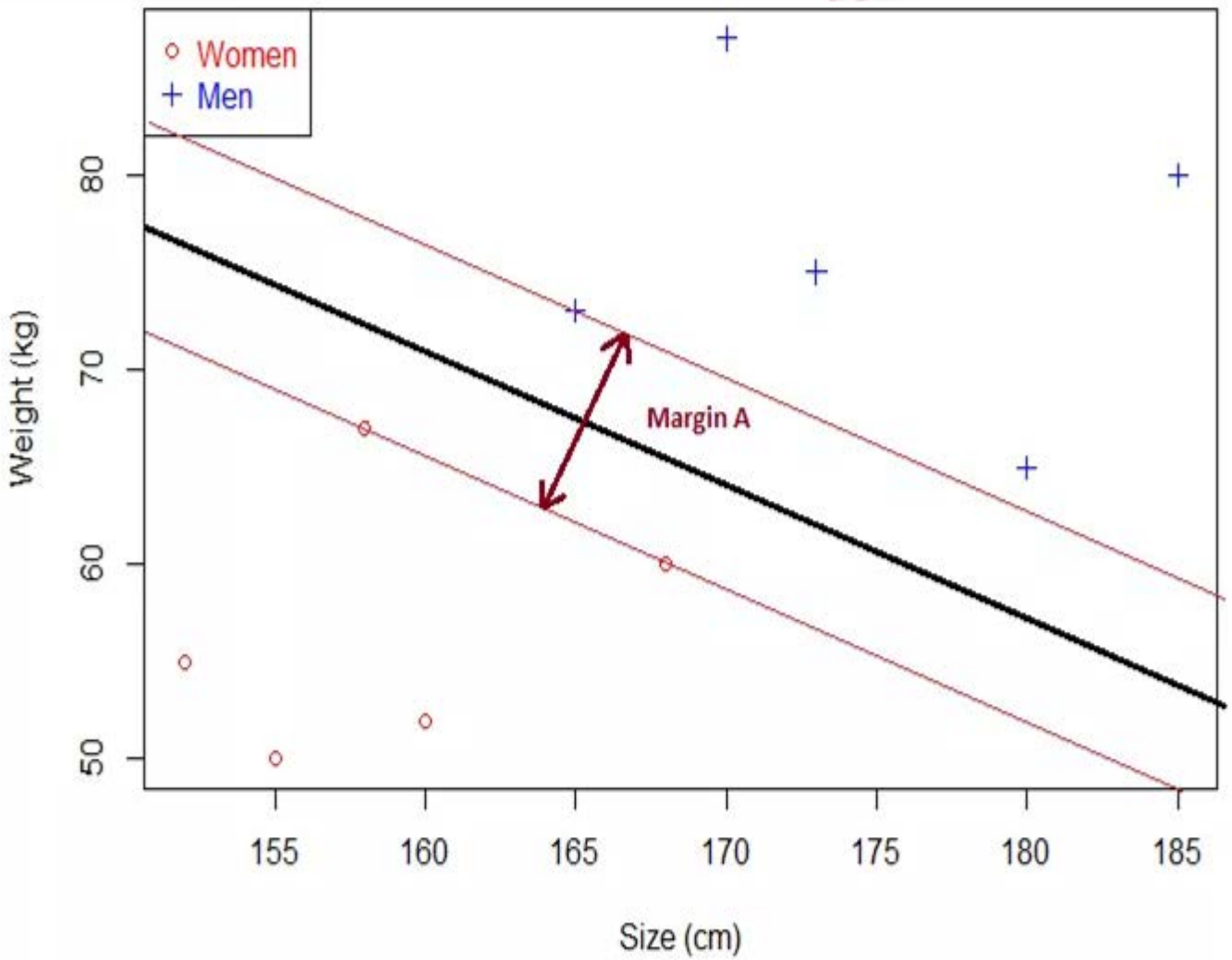


http://www.svm-tutorial.com/2014/11/svm-understanding-math-part-1/

SVM - training

Maximising the margin allows for:

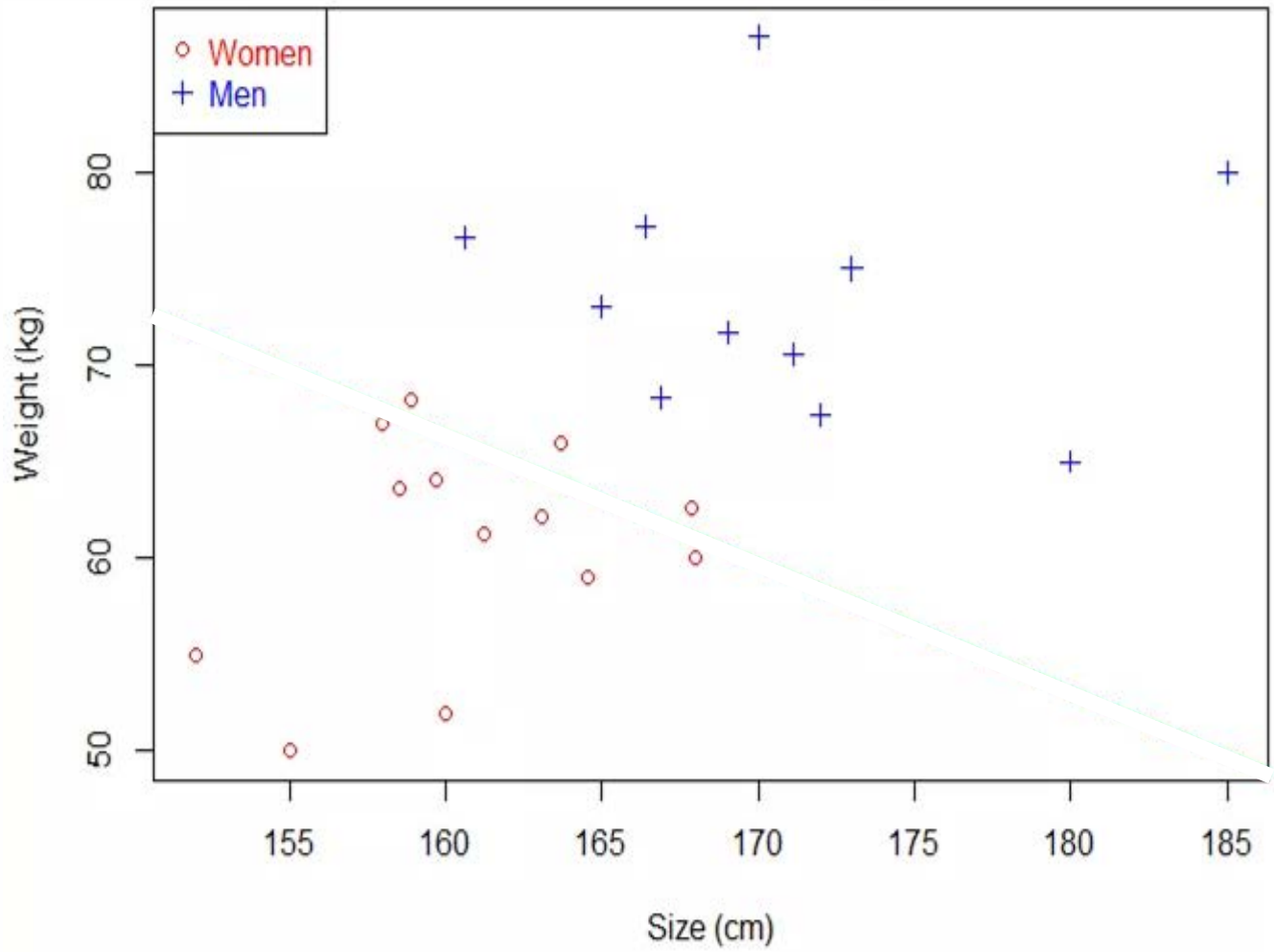
- ➢ Better generalisation
- ➢ More accurate predictions



http://www.svm-tutorial.com/2014/11/svm-understanding-math-part-1/

SVM prediction on a test dataset after training

Test new data according to the rule and make predictions

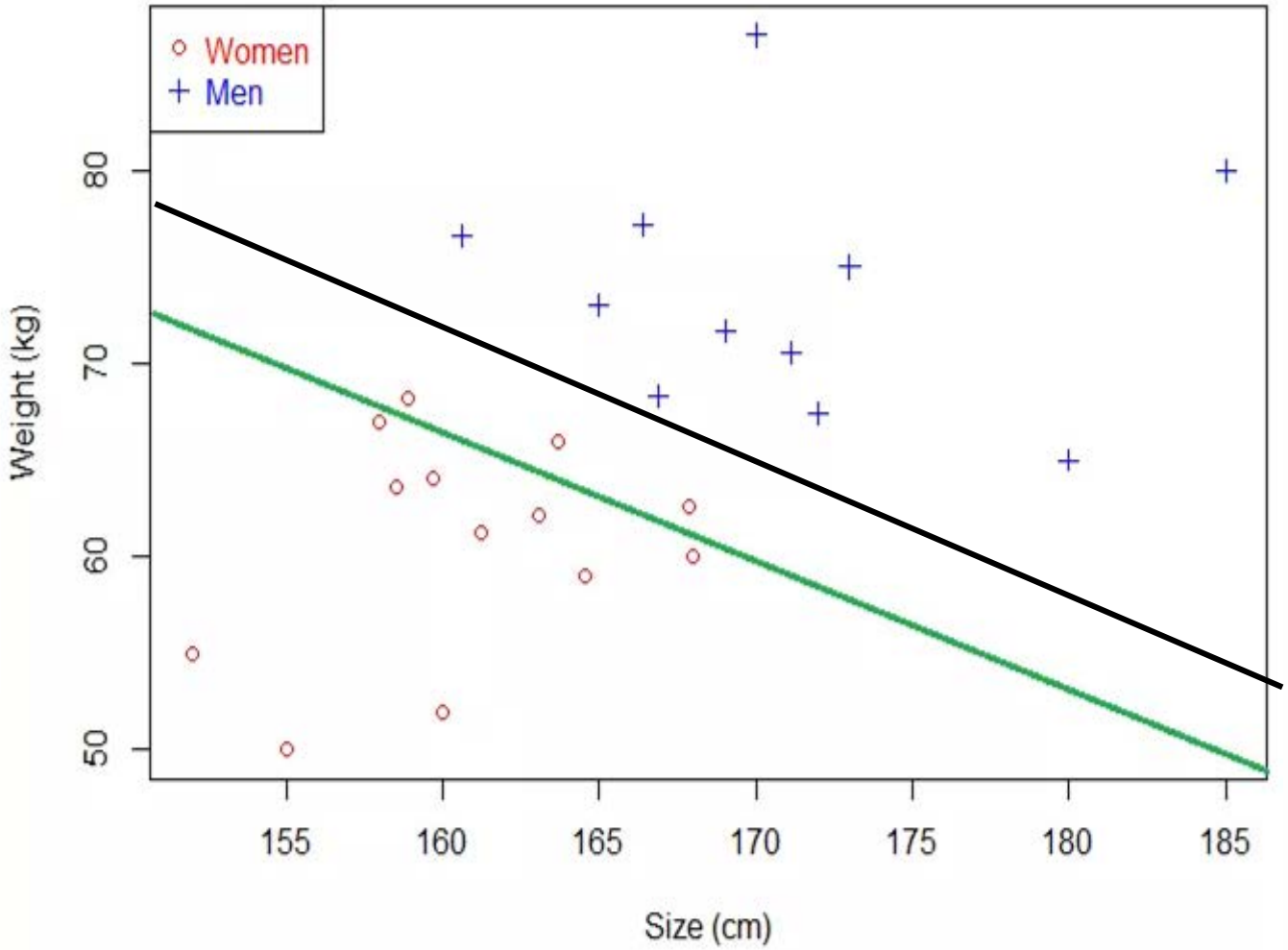


http://www.svm-tutorial.com/2014/11/svm-understanding-math-part-1/

SVM prediction on a test dataset

The black line calculated from the training dataset represents the optimal margin and in this case correctly separates the new test data

The green line represents a suboptimal solution and does not work as well with the test data

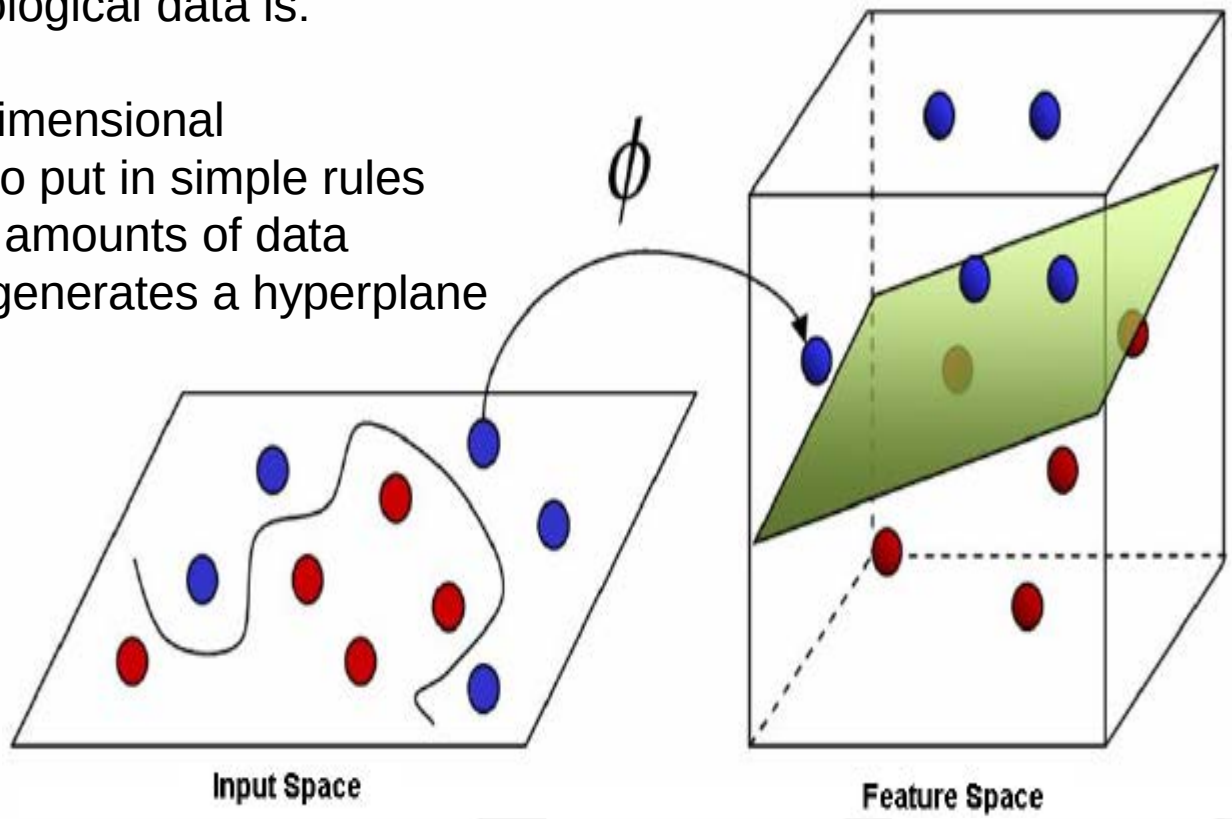


http://www.svm-tutorial.com/2014/11/svm-understanding-math-part-1/

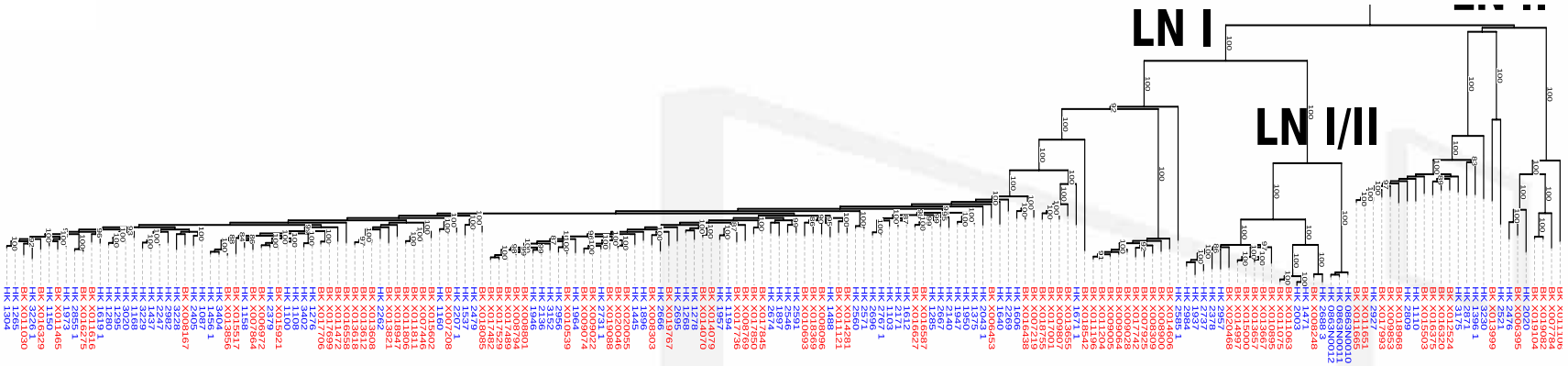
SVM developed for complex data sets

Usually biological data is:

- Multidimensional
- Hard to put in simple rules
- Large amounts of data
- SVM generates a hyperplane



Applying SVM to predicting zoonotic potential

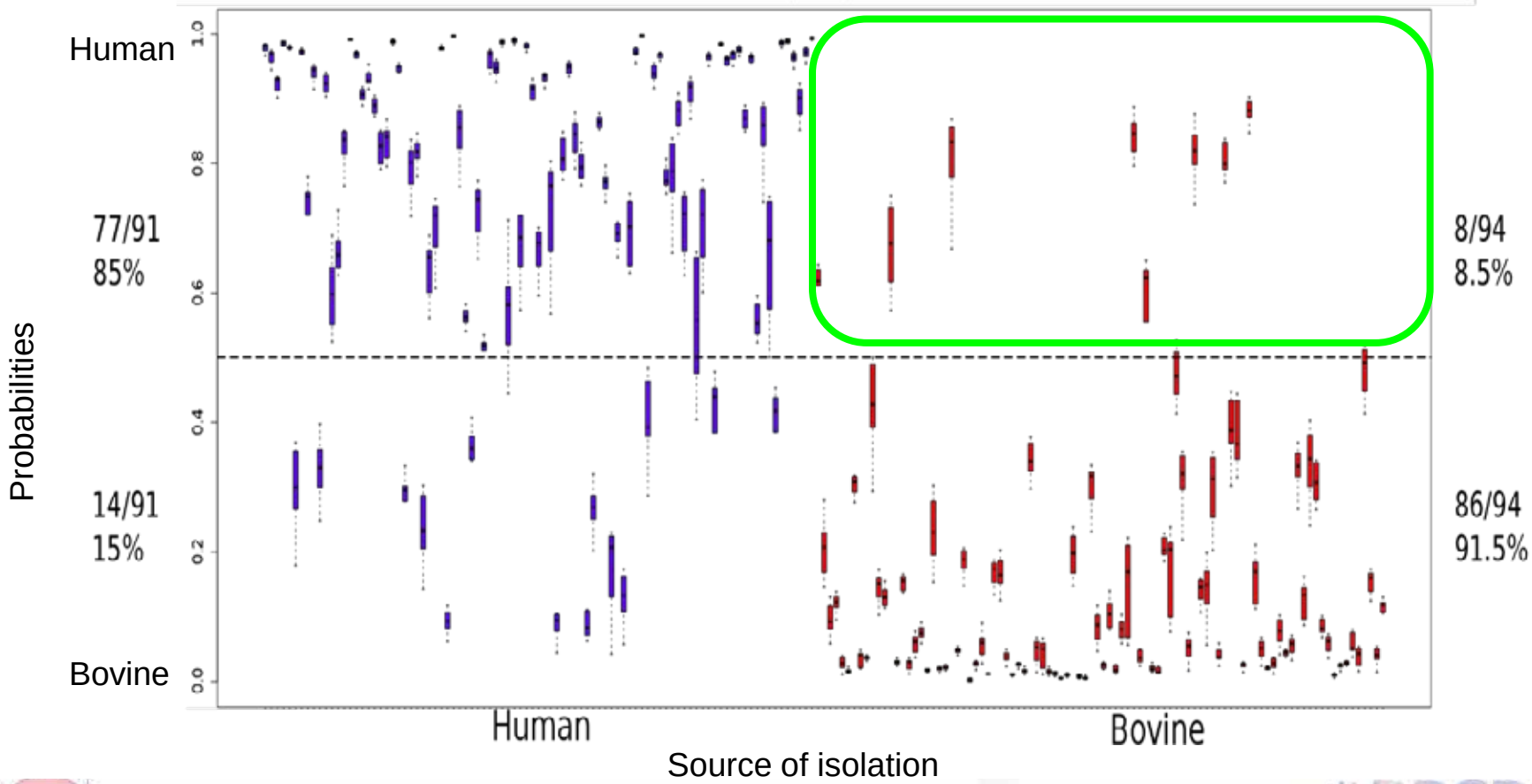


185 *E. coli* O157 strains

- 91 Human
- 94 Bovine

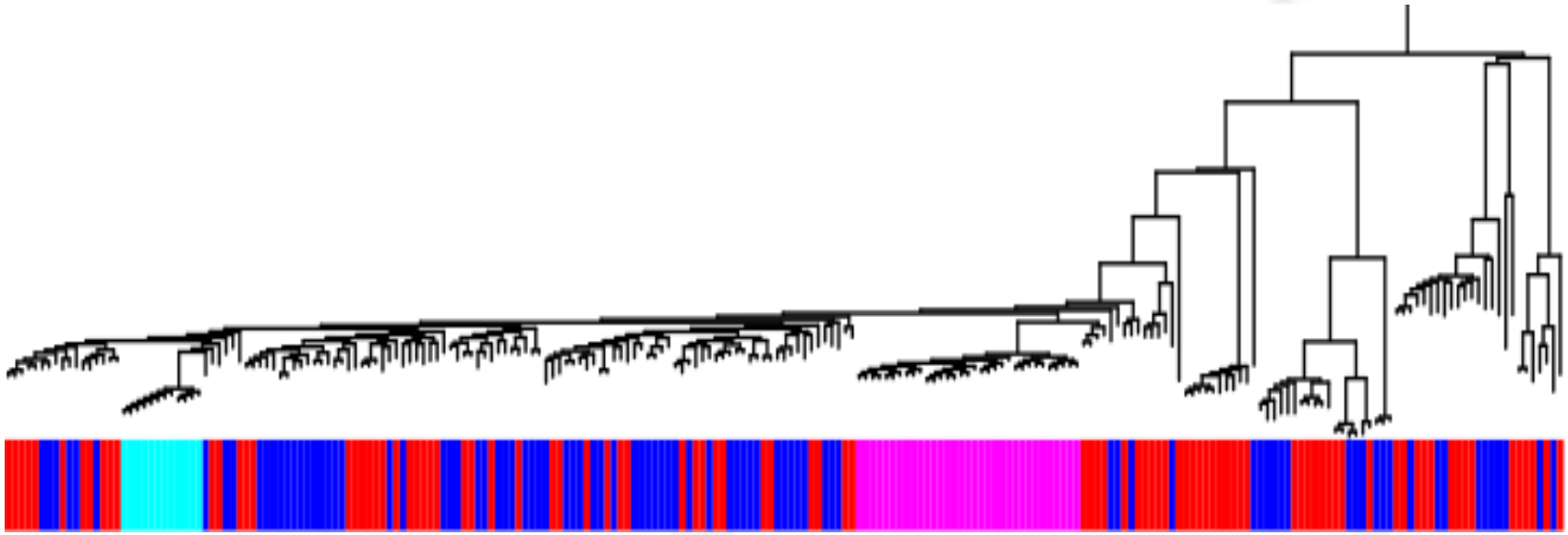
SVM prediction of *E. coli* O157 zoonotic potential

- Train on a subset and test on the remainder. Repeat the process and obtain statistics for prediction scores of isolates being ‘bovine’ or ‘human’
- <10% of bovine isolates were predicted to have high ’human’ scores



We propose that the subset of bovine isolates in the green box represent the more serious risk to human health

What are the predicted for outbreak strains isolated from food and milk?



Milk: 12 Sequences

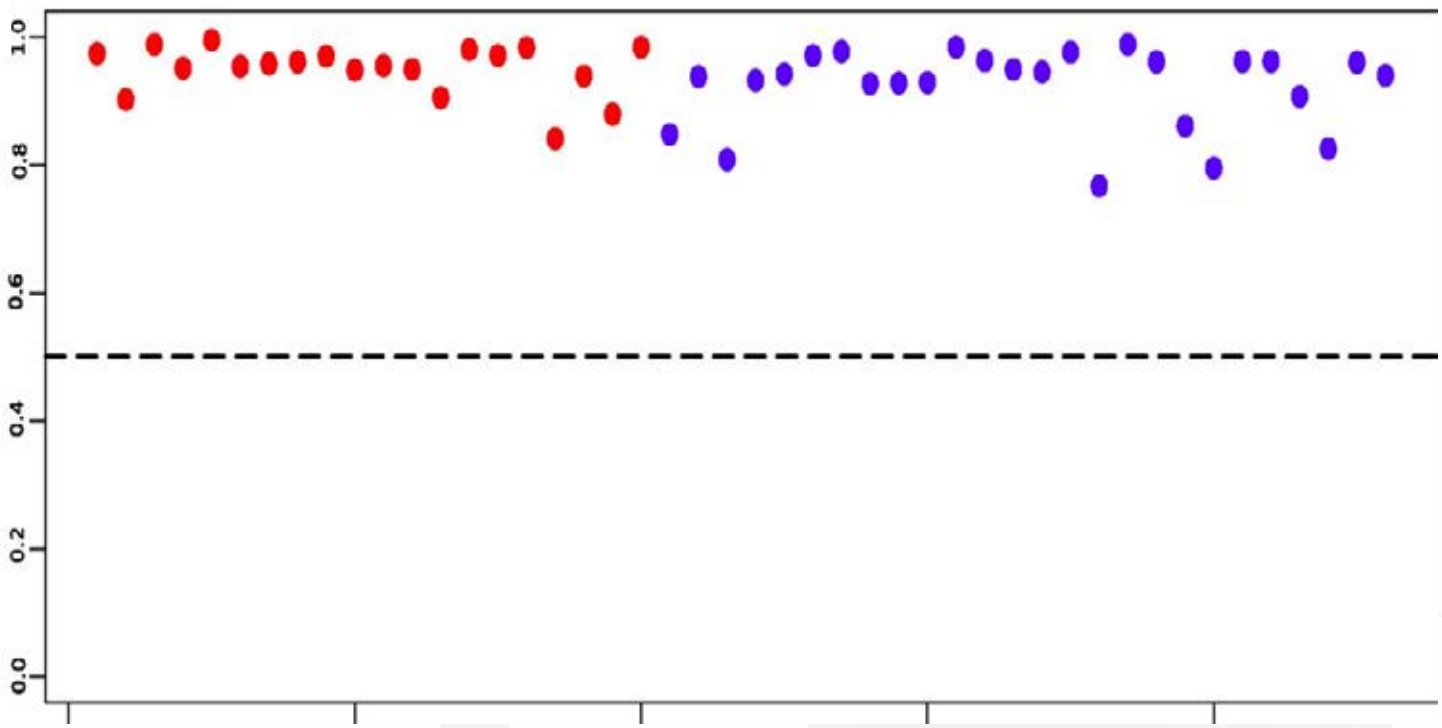
- 9 Human
- 3 Milk

Food: 34 Sequences

- 5 Bovine
- 17 Human
- 12 Hamburger

Predictions for outbreak isolates

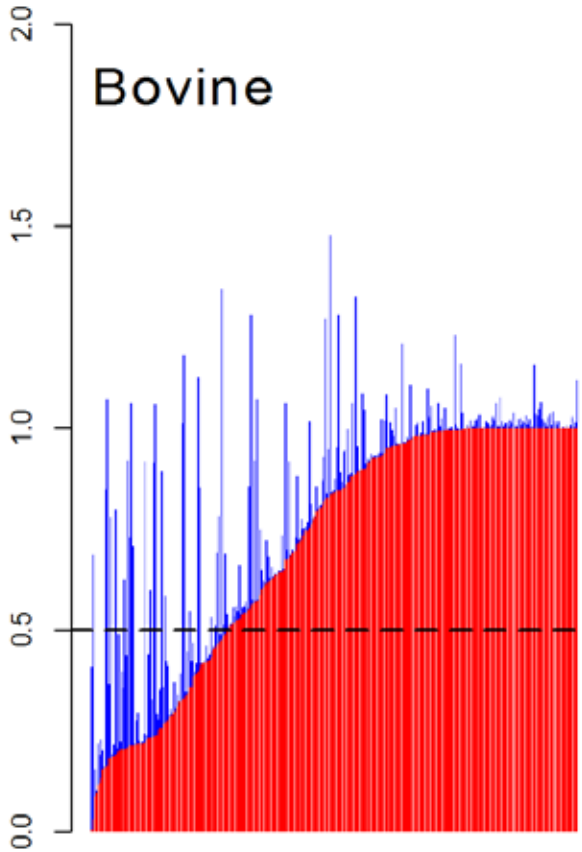
Hamburger
Cattle
Milk
Human



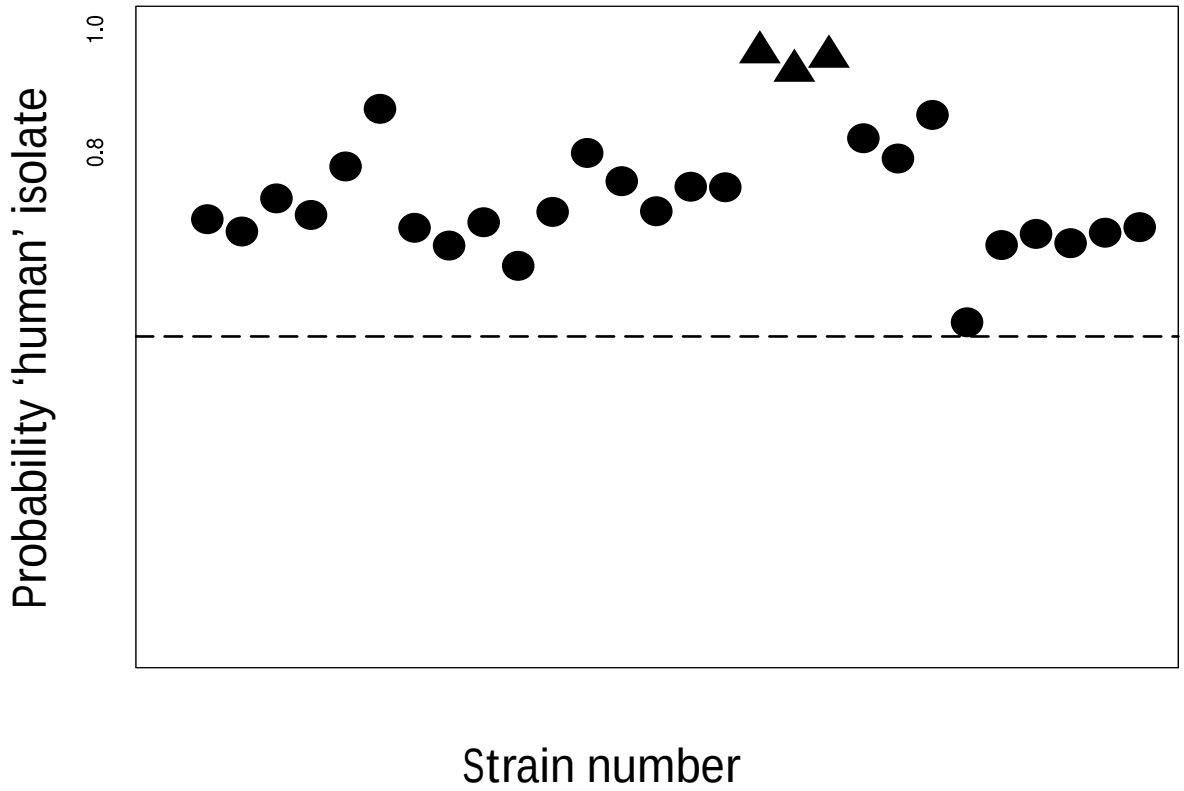
- **Regardless of isolation host all outbreak isolates were assigned "human" scores**

Lupolova *et al* 2016 *PNAS* 113:11312 +

Machine learning to predict isolates that may be more of a zoonotic threat based on SVM analysis



Predicted scores for general bovine *E. coli* isolates: 'Bovine' (red) and 'human' (blue)

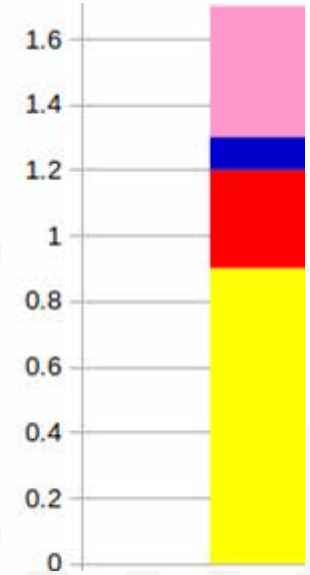


Predicted scores for bovine *E. coli* isolates (red) and human isolates (blue). *E. coli* O157 isolates (dark circles) were tested and shown to have scores more associated with human isolates

Lupolova et al (2017): 10.1099/mgen.0.000135

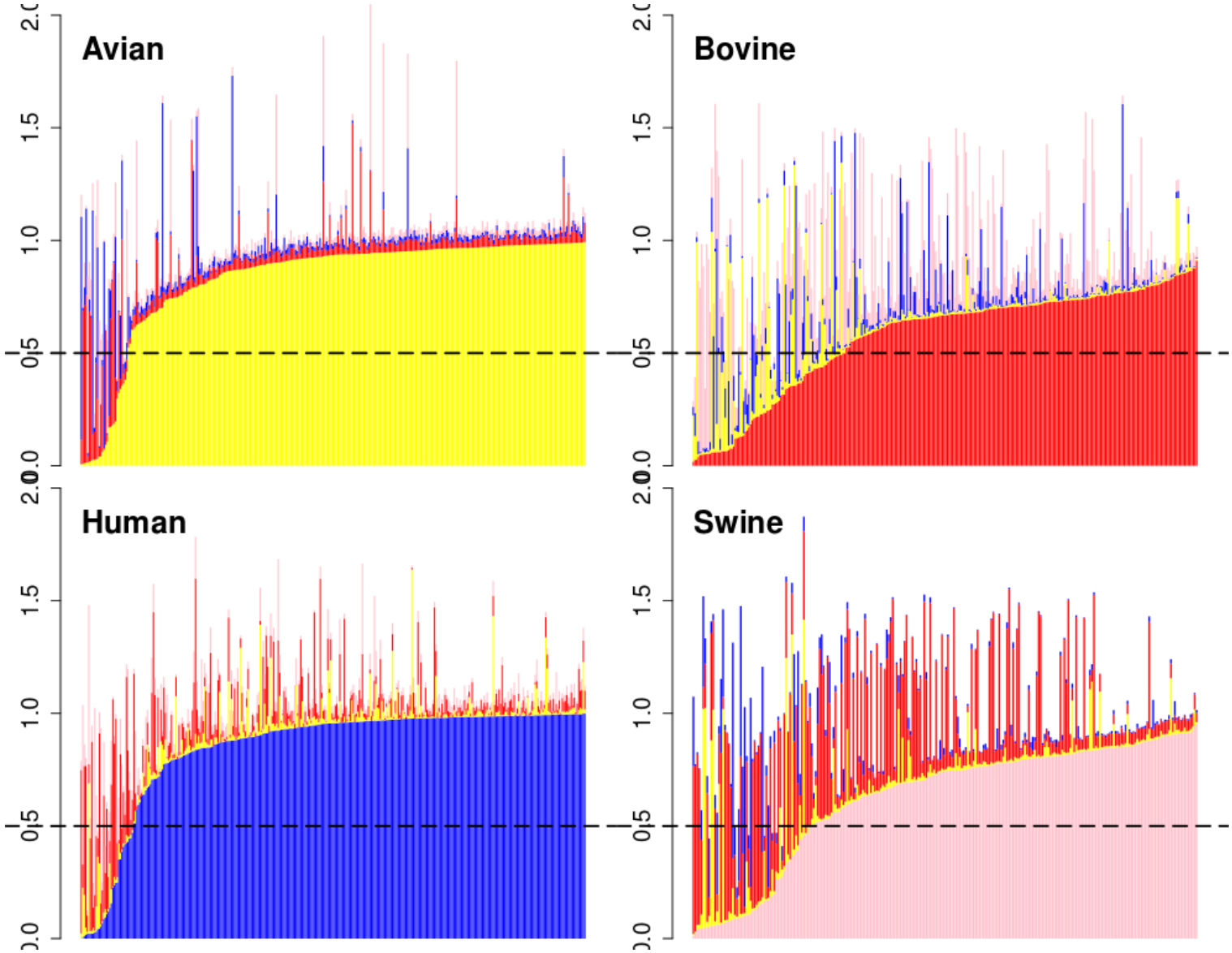
A scoring system for source attribution of an isolate – *Salmonella enterica* isolates

Isolate 1	⇨	is this avian?	⇨	0.9
Isolate 1	⇨	is this bovine?	⇨	0.3
Isolate 1	⇨	is this human?	⇨	0.1
Isolate 1	⇨	is this swine?	⇨	0.4



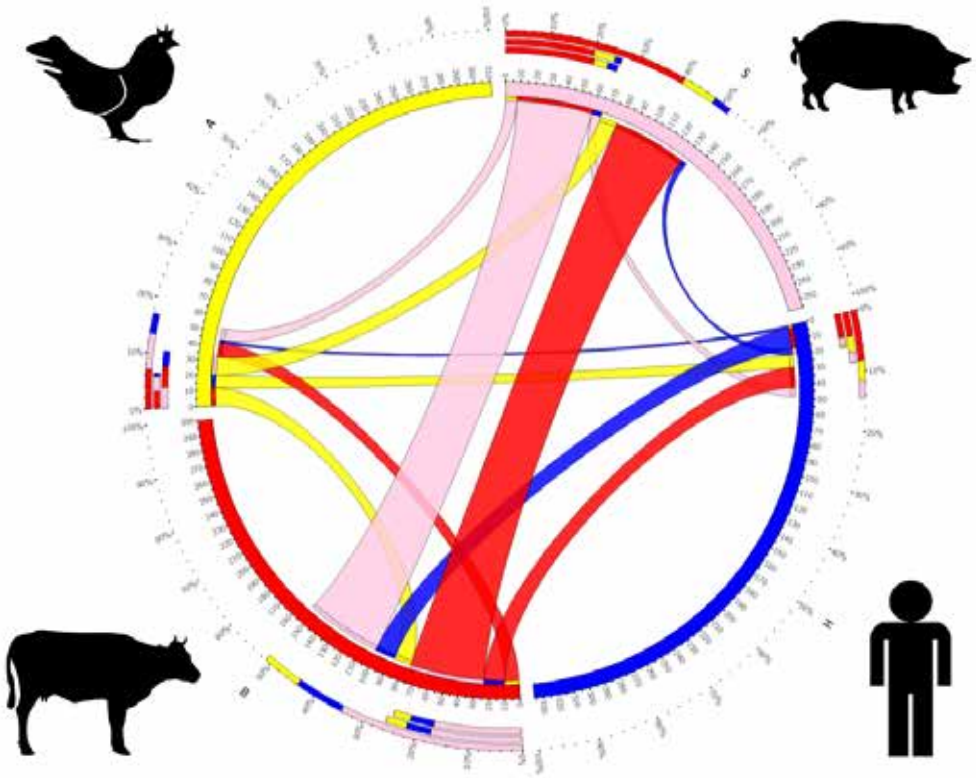
SVM (machine learning) applied in a step wise manner to assign probabilities of an isolate being derived from a particular host. Then plotted as a 'bar' (right)

Machine learning to predict the host source and zoonotic potential *Salmonella enterica*



Lupolova et al (2017): 10.1099/mgen.0.000135

Machine learning to predict the host source and zoonotic potential *S.* Typhimurium



The analysis indicates relative host restriction for the majority of S. Typhimurium isolates from the different hosts - potentially only specific subsets are a risk for human health?

With larger datasets this type of approach can inform us of where human infections originate from to inform risk assessments.

Lupolova et al (2017): 10.1099/mgen.0.000135

Conclusions

- Powerful precision of WGS is transforming our capacity to track infections and trace the sources of bacteria
- Considerable information can be extracted from WGS including basic taxonomy and identification of virulence and AMR genes. This helps determine the threat posed by an isolate
- Short read sequencing can be bolstered by long read methods to more accurately assemble the genome and predict phenotypes including virulence
- Excellent metadata on isolates collected from a broad selection of hosts and environments when coupled with their sequence data will allow us to predict the source off a sequenced isolate and evaluate the risk associated with exposure to that isolate
- The better the ‘database’ of individual isolate sequences the more value this has for identifying what bacteria are present in complex samples (by metagenomic sequencing)

Acknowledgements

University of Edinburgh

Nadejda Lupolova

Sharif Shabaan

Sean McAteer

Stephen Fitzgerald

Amy Beckett

Alex Corbishley

Nur Indah Ahmad

Jason Morgan

Xuefang Xu

Mark Woolhouse

Margo Chase-Topping

SRUC

Sue Tongue

Madeleine Henry

Jude Evans

George Gunn

Public Heath England and SERL

Tim Dallman (supplied many slides)

Claire Jenkins

Lesley Allison

Mary Hanson

Anne Holmes

USDA

Jim Bono

Moredun Research Institute

Tom McNeilly

David Frew

University of Glasgow

Louise Matthews

Jamie Prentice

Andrew Roe