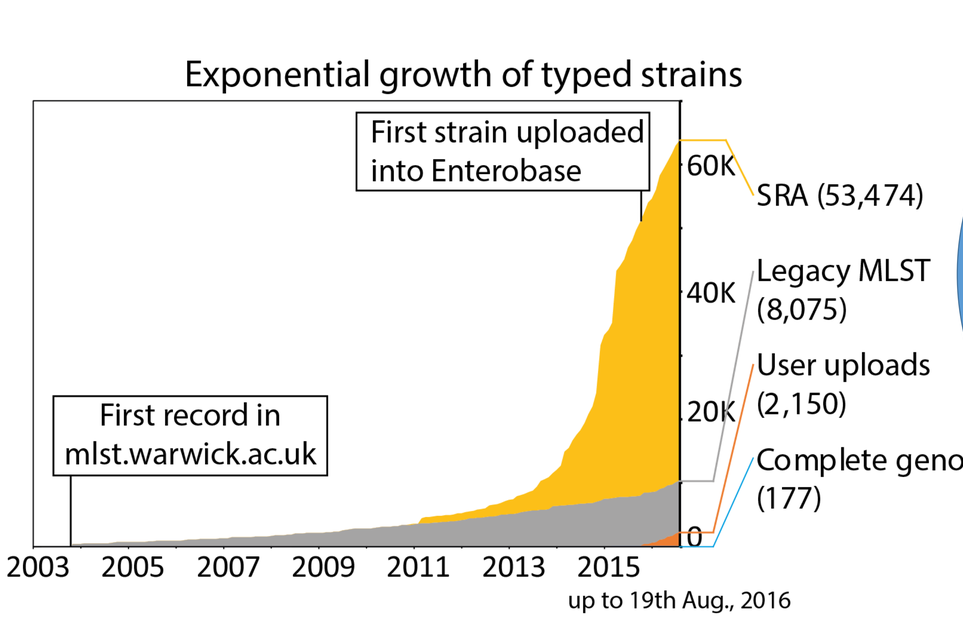
Using MLST to decipher the population structure of *Salmonella enterica*

Zhemin Zhou, Nabil-Fareed Alikhan, Martin Sergeant, Mark Achtman
Warwick Medical School, University of Warwick, Coventry, UK, CV4 7AL

Abstract

Currently, there are already over 50K sets short reads from *Salmonella spp.* in public sequence repositories, with tens of thousands of new *Salmonella* reads being added every year. However, there is a lack of standardised typing approaches to analyse this amount of data. We have developed automatic pipelines within EnteroBase to type all the currently available genomes with different multi-locus sequence typing (MLST) schemes. Over 98% of genomes in *S. enterica* subsp. I were assigned in one of the 296 sub-populations (reBGs). We have also calculated genus trees for core genes in *Salmonella*, in order to reconstruct its evolutionary history.



Big Data

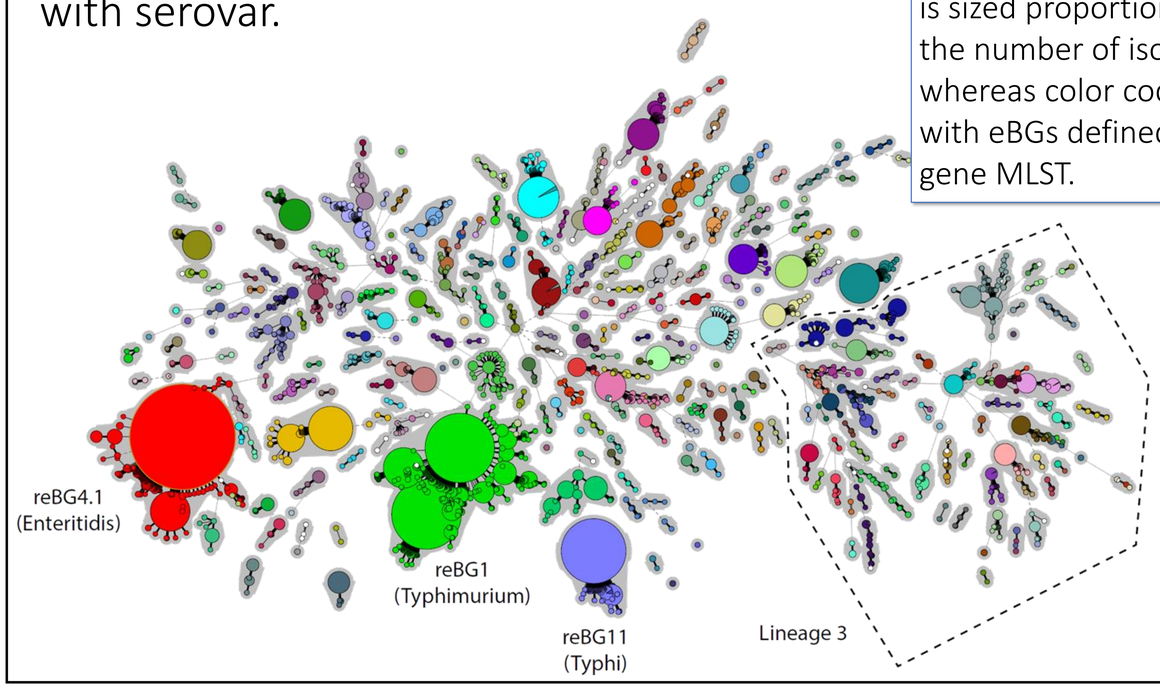
Now we have over 55K *Salmonella* genomes, how can we best manage them?

MLST schemes implemented in EnteroBase

MLST – Classic *(Achtman et al, 2012)*	Ribosomal MLST *(Jolley et al. 2012)*	Core Genome MLST
7 Loci	51 Loci	3,002 Loci
Conserved Housekeeping genes	Ribosomal proteins	Any conserved coding sequence
Low resolution	Medium resolution	High resolution
Different scheme for each species/genus	Single scheme across tree of life	Different scheme for each species/genus

rMLST defined 296 clusters of genetically closely related strains (reBGs) in *S. enterica* subsp. I.

With minor exceptions, the reBGs were largely consistent with those defined by 7 gene MLST, and were largely concordant with serovar.



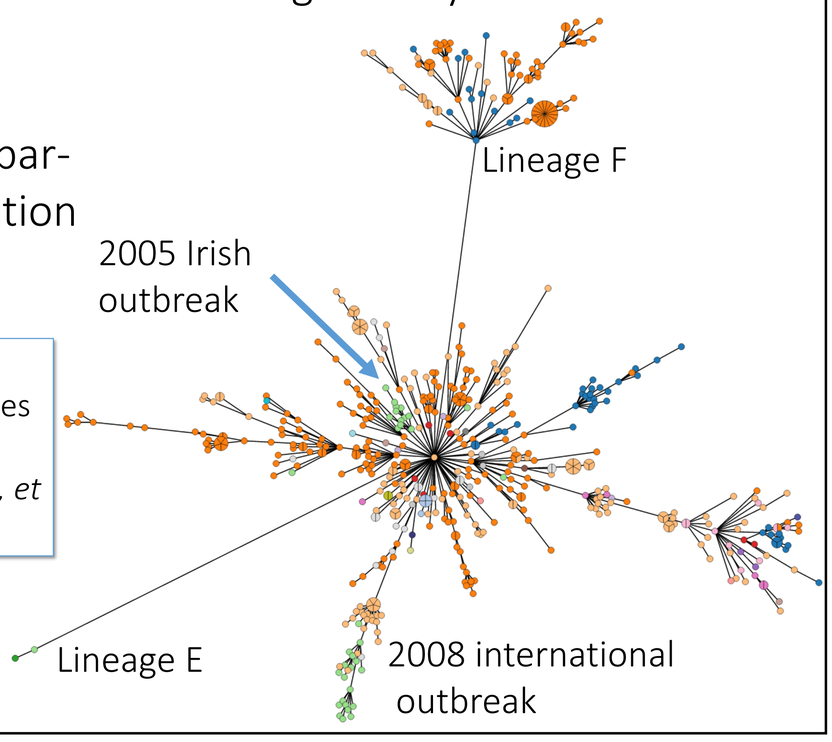
rMLST Minimum spanning tree (MSTree) for all *S. enterica* subsp. I genomes.

Each circle represents one rMLST ST. The circle is sized proportional to the number of isolates whereas color coded with eBGs defined by 7 gene MLST.

cgMLST was constructed from 537 reference genomes. An initial wgMLST scheme with 21,065 orthologs was created from the pan genome of reference genomes.

cgMLST consisted of a subset of 3,002 loci in the wgMLST scheme. These selected loci were generally conserved across the genus.

cgMLST has a comparable level of resolution with SNPs.

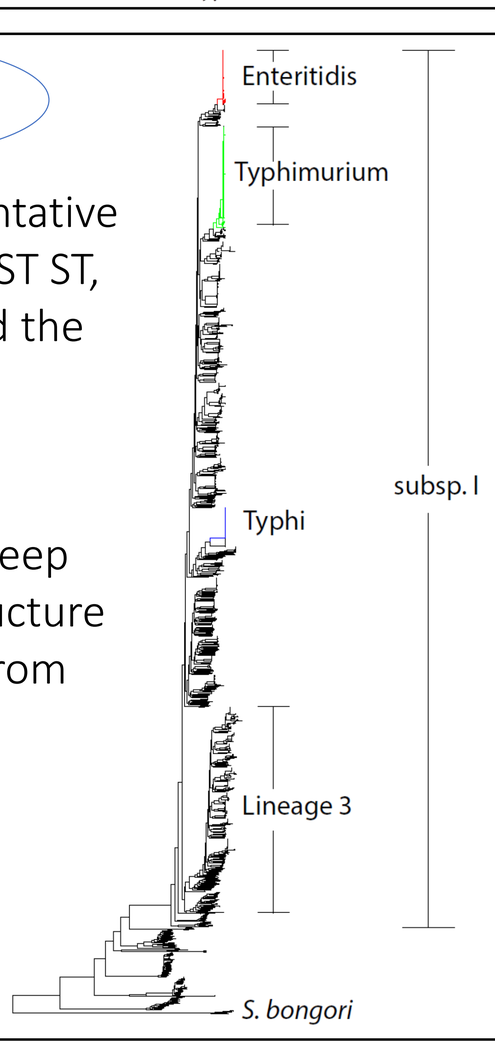


cgMLST MSTree reconstructed deep lineages and clusters of outbreak isolates described in Zhou, *et al.* PLoS Genet 2008

Genus tree

Use one representative genome per rMLST ST, we reconstructed the genus tree for *Salmonella*.

Clear signals of deep phylogenetic structure were identified from the genus tree.



MLST is portable and scalable

	Core SNP	cgMLST
Strains	3,144 representative genomes	All 55,801 strains
Storage	1.05M SNPs in 2.8M core genome	3,002 alleles per strain
	A total of 3.2 GB	A total of 168 MB
Phylogeny	RAxML	MSTree

16 cores
>100 GB RAM

> 2 weeks

1 core
<4 GB RAM

3 hours

WARWICK
Warwick Medical School

http://enterobase.warwick.ac.uk/

BBSRC bioscience for the future