Back to the beginnings:
The age of the history of bacterial pathogens

Mark Achtman FRS
Warwick Medical School, Coventry
UK

My prior interests (1965-1977:
Bacterial genetics in general
Plasmids in general
Bacterial conjugation
F factor
Gene mapping
Protein synthesis *in vitro*
Outer membrane proteins

Falkow, Ann Rev Microbio. 2008: 'I was especially lucky that first year to have Staffan Normark as a sabbatical visitor. He, like Gordon Dougan and **Mark Achtman**, who were visitors in Seattle, had decided to change his focus from … to the study of pathogenicity'

My new goals (1977-1978):
Investigating the bacterial properties that were responsible for disease
Focus on *E. coli* K1 (Bureau of Biologics, Bethesda; John Robbins and Richie Silver)
Learning how to perform an animal model (infant rats) for *E. coli* K1 meningitis (Univ. of Maryland, Richard Moxon)

Memories from sabbatical (Seattle 1978-79)

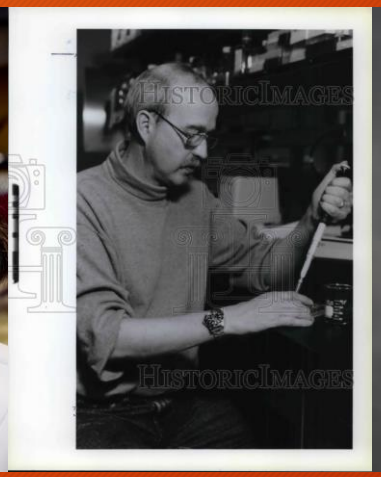
Daily meetings with Stan for 8 months. General discussions about medical microbiology and pathogens

To students: ‘Give me a data fix!’

To me: ‘Mark, U Wash is a non-smoking area! I have found a cubicle in the library where you can smoke behind closed doors, but you can’t tell anybody’

To me: ‘Mark, why don’t you do your sonication of *Vibrio cholerae* in a safety cabinet. Seattle is a port and if we had a case of cholera in Seattle, they would quarantine all of Seattle. You don’t want that on your academic record’

Never published the exciting finding: SDS-PAGE of *E. coli* membrane preps from urine showed different OMPs than after single colony purification



Subsequent work

- *E. coli* K1 animal model (1979-1984)
  - Demonstrated that anti-LPS antibodies were protective against newborn meningitis in infant rats.
  - Developed an interest in population structures of pathogenic bacteria
    - Robert Selander and Howard Ochman used MLEE to investigate the population structure of *E. coli* and defined the representative EcoR collection
  - Stopped using animal models and working on pathogenesis when I demonstrated that human babies with newborn meningitis had more anti-LPS antibodies than healthy babies (unpublished)

- Epidemic meningitis due to serogroup A *N. meningitidis* (1984-2001)
  - Molecular epidemiology
  - Serological analyses of epitopes in outer membrane proteins
  - Population genetic studies, which shifted in the 1990s from MLEE to DNA sequence analyses, including phylogenetic trees.

Important encounters 1978-2014

Me: I'm having extreme stomach pains.
Stan: Stop drinking espresso coffee.

Me: I'm feeling weak and feverish every afternoon.
Stan: You've got atypical pneumonia due to *Mycoplasma*. Take erythromycin.

I suspect strongly that Stan was one of the external reviewers whose recommendations were crucial to me obtaining my current job.

I feel very privileged to have met Stan.

Multilocus sequence typing (MLST)

(Maiden et al., PNAS **95**: 3140-45, 1998) Cited 3,506x

•*fumC* Sequence (500 bp) •Allele

Sequences

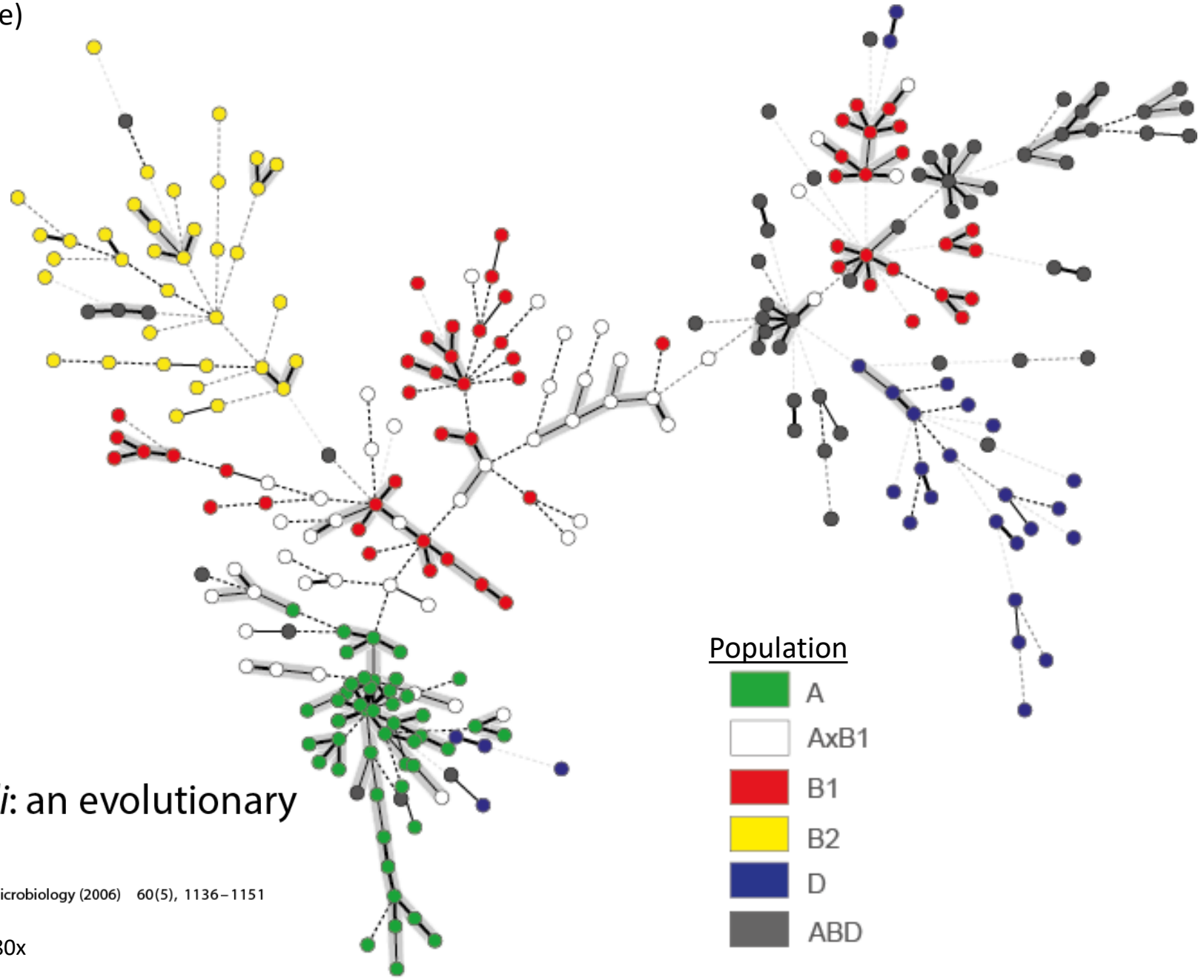
	abcZ	*fumC*	*gyrB*	*recA*	*mdh*	*ureI*	*icd*
Alleles:	1	2	1	5	22	39	7

Sequence Types (STs):

Sequence type (ST)	Alleles
1	1 2 1 5 22 39 7
2	1 2 1 **7** 22 39 7
3	1 **5** 1 5 22 39 7
4	1 2 1 5 **25** 39 7

Minimal spanning tree of 460 *E. coli* according to MLST.
Recombination is extremely frequent (grey/white)

MLST schemes for *Escherichia*
MLST: 7 housekeeping genes
rMLST: 51 ribosomal genes
cgMLST: 2,512 core genome genes
wgMLST: 25,002 pan-genome genes



Sex and virulence in *Escherichia coli*: an evolutionary perspective

Thierry Wirth, Daniel Falush, Ruiting Lan, Frances Colles, Patience Mensa, Lothar H. Wieler, Helge Karch, Peter R. Reeves, Martin C. J. Maiden, Howard Ochman and Mark Achtman

Molecular Microbiology (2006) 60(5), 1136–1151

Cited 1380x

Bacterial genomics and epidemiology

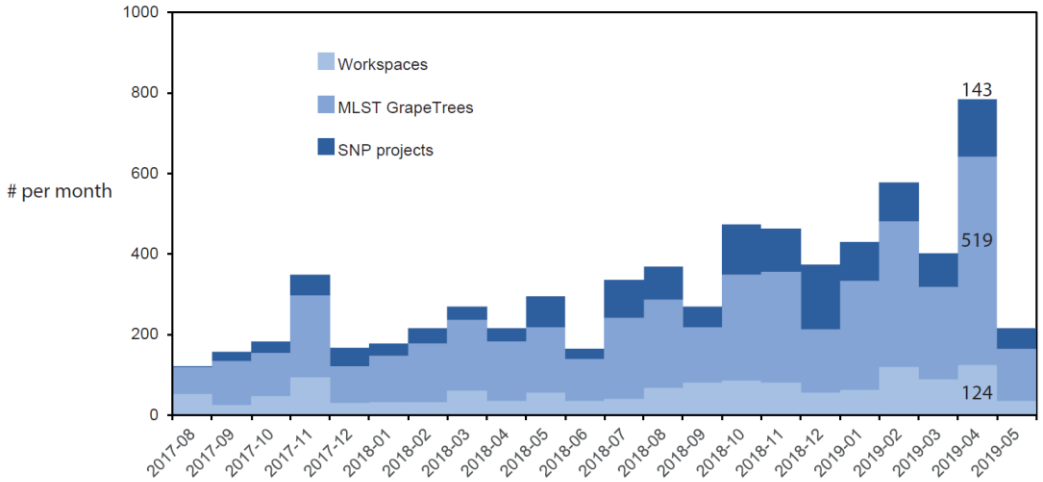
- High throughput sequencing has become economical and common since 2013 (Illumina HiSeq & MiSeq).
- Large numbers of genomes have been sequenced and their short reads deposited in short read archives. Until very recently, the genomes for these short reads were not available for public access.
- Many microbiologists struggle to acquire the necessary bioinformatic skills to evaluate genomic sequences, especially if they are in the form of short reads.
- The most common solution is to hire one or more bioinformaticians.
- We developed EnteroBase, which enables microbiologists to work directly with genomes without assistance by IT specialists.

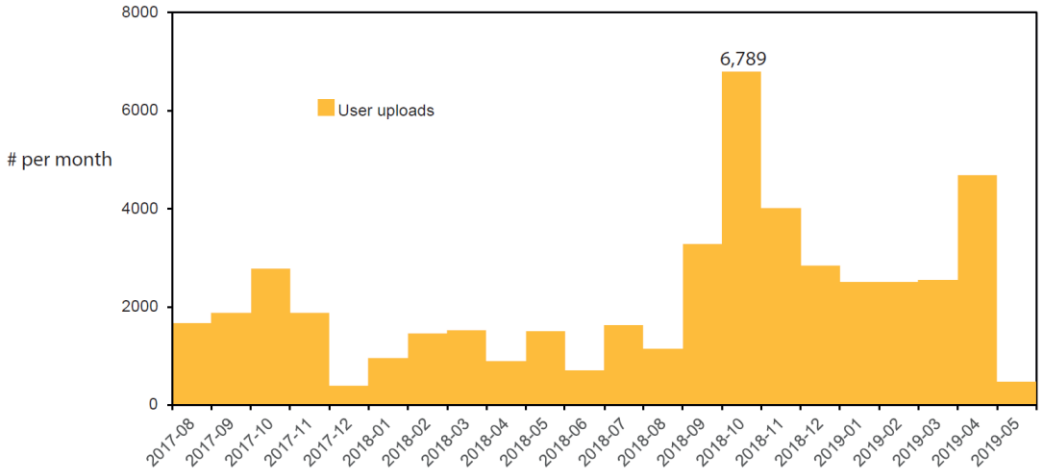
EnteroBase (http://enterobase.warwick.ac.uk/)

- Database of genotypes from *Salmonella*, *Escherichia*, *Yersinia*, *Clostridioides*, *Vibrio*, *Helicobacter*
- Can help microbiologists go from Illumina sequence reads to looking at closest relatives in 150 min
- Predicts serovar
- Calls MLST, rMLST, cgMLST, wgMLST genotypes
- GrapeTree presentation of MLST and SNP trees
- Can help in epidemiological investigations

The user's guide to comparative genomics with EnteroBase.
Three case studies: micro-clades within *Salmonella enterica* serovar Agama, ancient and modern populations of *Yersinia pestis*, and core genomic diversity of all *Escherichia*
Zhemin Zhou, Nabil-Fareed Alikhan, Khaled Mohamed, the Agama Study Group, Mark Achtman
bioRXiv April 18 2019 https://doi.org/10.1101/613554

User usage of EnteroBase tools and short read uploads





Salmonella

Strains:207214

Assembled
- Legacy:5555
- From NGS:201659
- In Progress:10

Schemes
- Achtman 7 Gene MLST:207172
- cgMLST V2 + HierCC:201582
- CRISPR:50416
- rMLST:201530
- wgMLST:201021

Database Home

Escherichia/Shigella

Strains:105633

Assembled
- Legacy:9614
- From NGS:96019
- In Progress:378

Schemes
- Achtman 7 Gene MLST:105584
- cgMLST V1 + HierCC:96004
- rMLST:95969
- wgMLST:95721

Database Home

Clostridioides

Strains:13963

Assembled
- From NGS:13963
- In Progress:0

Schemes
- cgMLST V1 + HierCC:13951
- Griffiths 7 Gene:13722
- rMLST:13933
- wgMLST:13948

Database Home

Vibrio

Strains:6928

Assembled
- From NGS:6928
- In Progress:6

Schemes
- rMLST:6928

Database Home

Yersinia

Strains:3556

Assembled
- Legacy:881
- From NGS:2675
- In Progress:1

Schemes
- Achtman 7 Gene:3326
- cgMLST V1 + HierCC:2677
- McNally 7 Gene:2899
- rMLST:2676
- wgMLST:2671

Database Home

Helicobacter

Strains:2427

Assembled
- From NGS:2427
- In Progress:1

Schemes
- rMLST:2427

Database Home

Moraxella

Strains:1052

Assembled
- Legacy:416
- From NGS:636
- In Progress:0

Schemes
- Achtman 7 Gene:1053
- rMLST:636

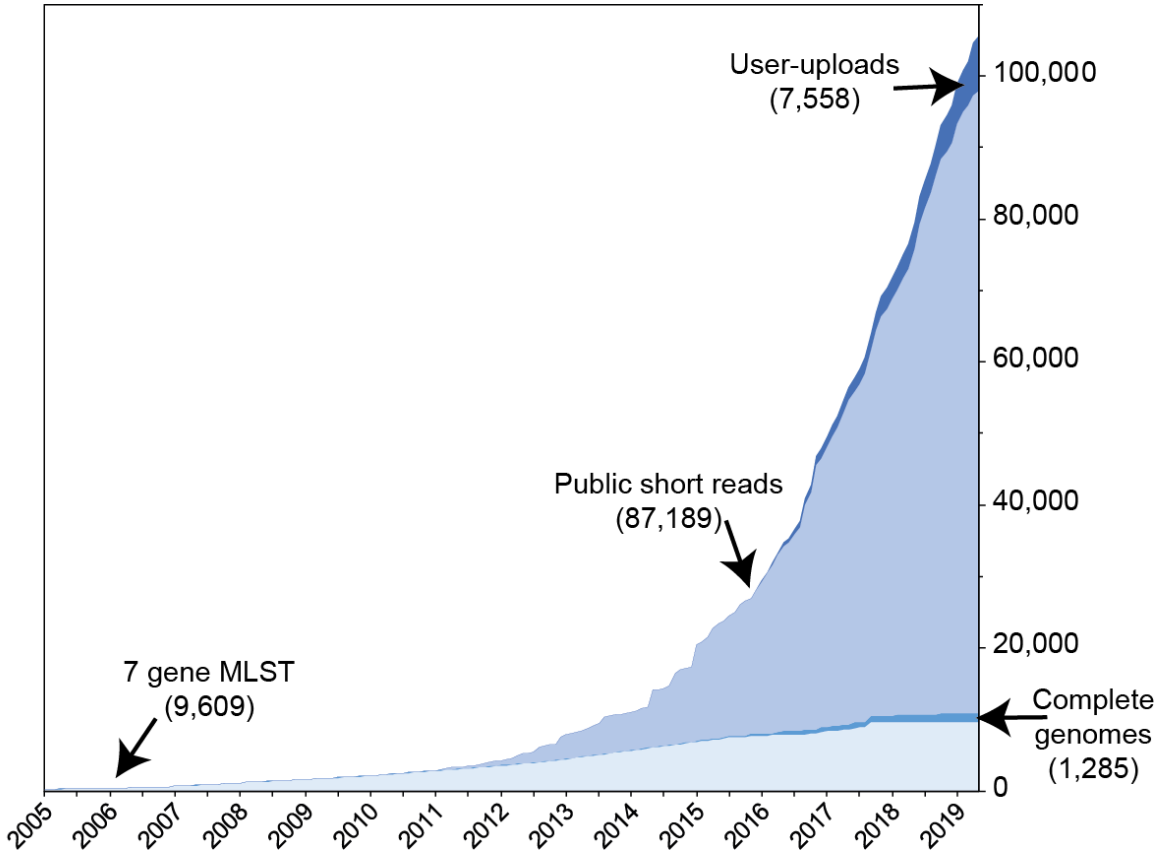
Database Home

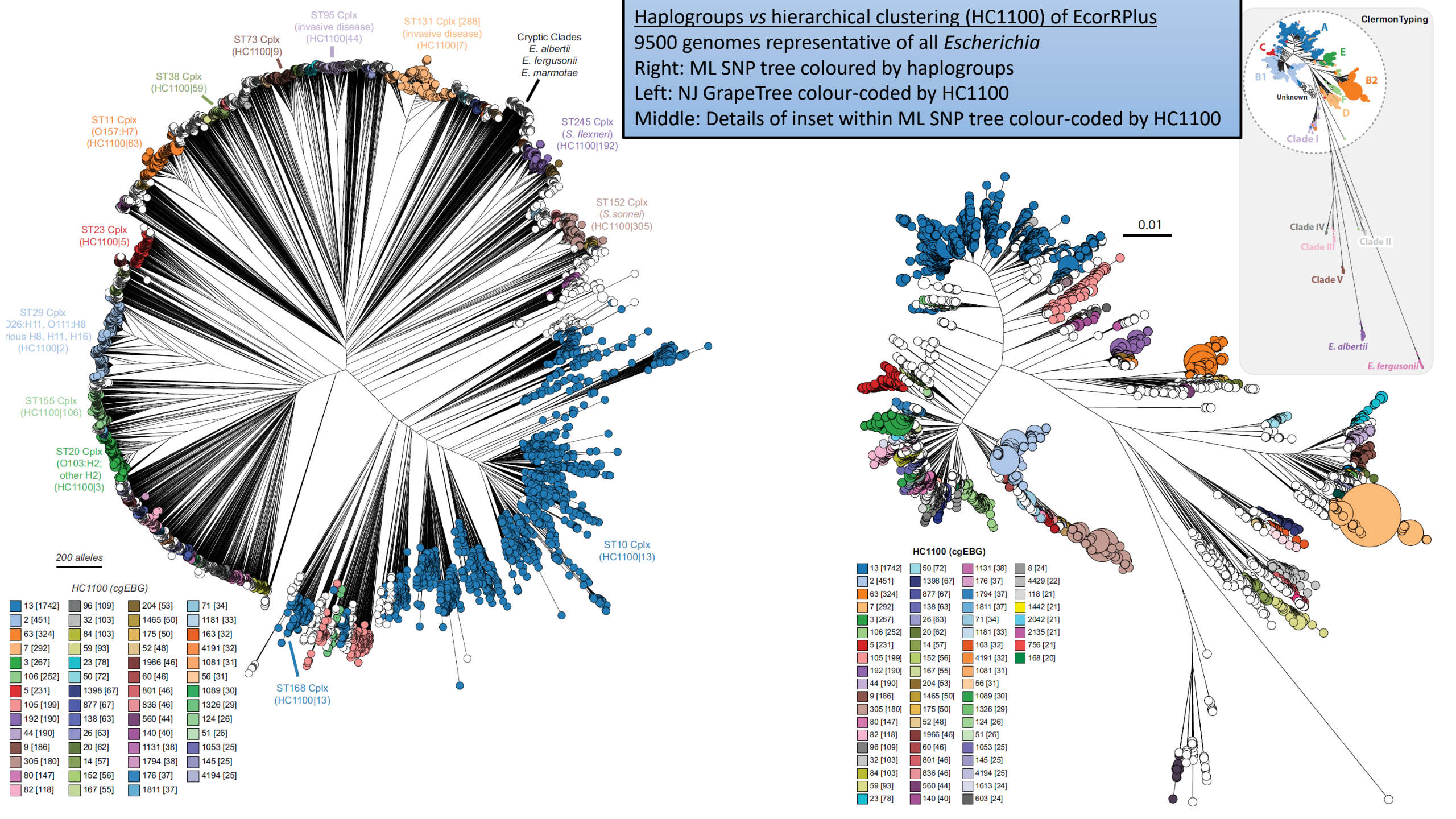
Genotypes and genomes publicly available in EnteroBase (May, 2019)

EnteroBase provides a catalogue of ever increasing genomic data

Genus	Assembled genomes (05/2019)	Monthly increase
Salmonella	200,000	5,000
Escherichia	96,000	2,800
Clostridioides	14,000	600
Yersinia	2,675	50
Helicobacter	2,400	
Vibrio	6,900	

Sources of *Escherichia* MLST genotypes with time

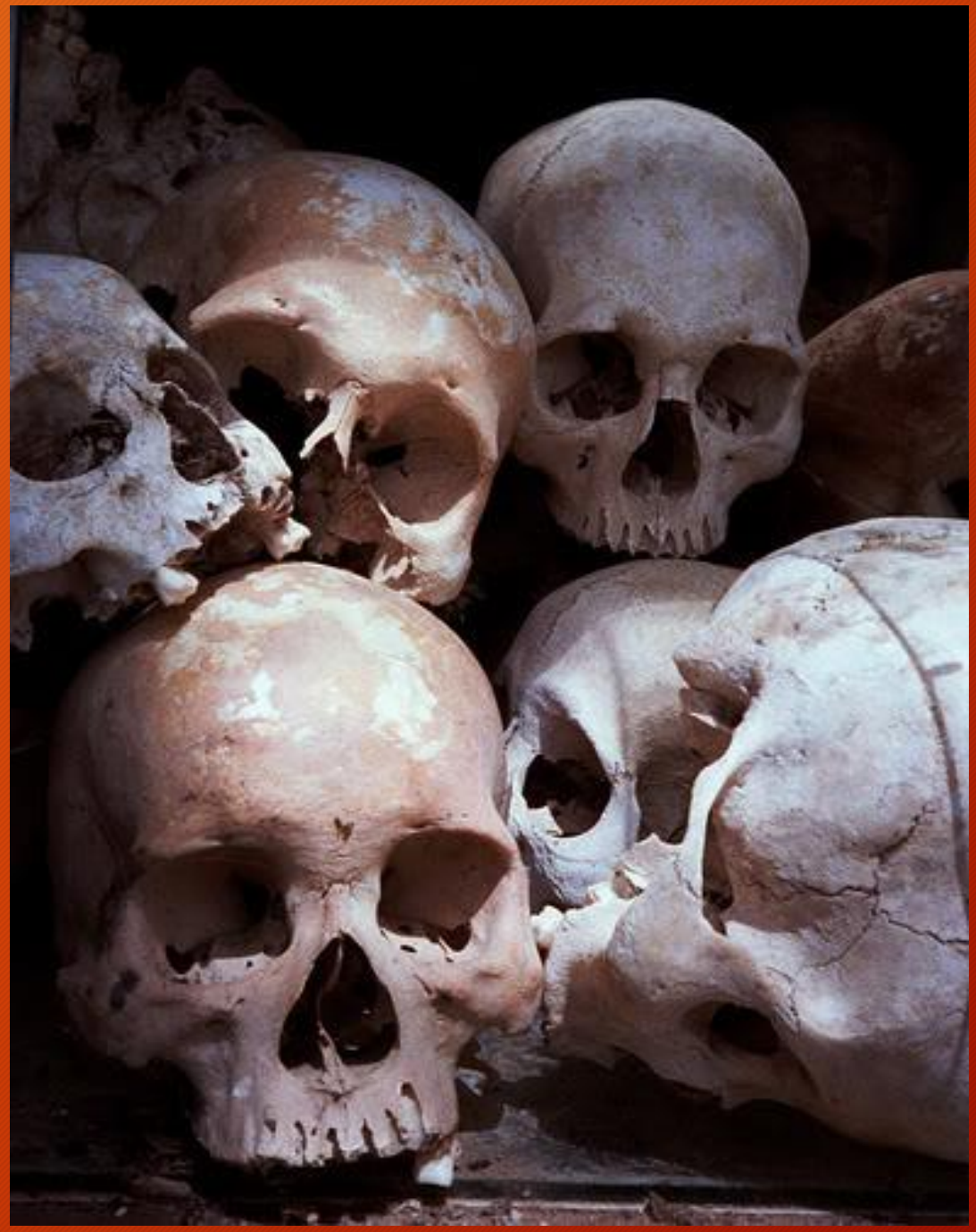



Haplogroups *vs* hierarchical clustering (HC1100) of EcorRPlus
9500 genomes representative of all *Escherichia*
Right: ML SNP tree coloured by haplogroups
Left: NJ GrapeTree colour-coded by HC1100
Middle: Details of inset within ML SNP tree colour-coded by HC1100
ST95 Cplx (invasive disease) (HC1100|44)
ST131 Cplx [288] (invasive disease) (HC1100|7)
ST73 Cplx (HC1100|9)
Cryptic Clades
E. albertii
E. fergusonii
E. marmotae
ST38 Cplx (HC1100|59)
ST11 Cplx (O157:H7) (HC1100|63)
ST245 Cplx (*S. flexneri*) (HC1100|192)
ST152 Cplx (*S.sonnei*) (HC1100|305)
ST23 Cplx (HC1100|5)
ST29 Cplx
O26:H11, O111:H8
rious H8, H11, H16)
(HC1100|2)
ST155 Cplx (HC1100|106)
ST20 Cplx (O103:H2; other H2) (HC1100|3)
ST10 Cplx (HC1100|13)
ST168 Cplx (HC1100|13)
200 alleles
HC1100 (cgEBG)
13 [1742]
2 [451]
63 [324]
7 [292]
3 [267]
106 [252]
5 [231]
105 [199]
192 [190]
44 [190]
9 [186]
305 [180]
80 [147]
82 [118]
96 [109]
32 [103]
84 [103]
59 [93]
23 [78]
50 [72]
1398 [67]
877 [67]
138 [63]
26 [63]
20 [62]
14 [57]
152 [56]
167 [55]
204 [53]
1465 [50]
175 [50]
52 [48]
1966 [46]
60 [46]
801 [46]
836 [46]
560 [44]
140 [40]
1131 [38]
1794 [38]
176 [37]
1811 [37]
71 [34]
1181 [33]
163 [32]
4191 [32]
1081 [31]
56 [31]
1089 [30]
1326 [29]
124 [26]
51 [26]
1053 [25]
145 [25]
4194 [25]
ClermonTyping
A
B1
C
E
B2
F
D
Unknown
Clade I
Clade IV
Clade III
Clade II
Clade V
E. albertii
E. fergusonii
0.01
HC1100 (cgEBG)
13 [1742]
2 [451]
63 [324]
7 [292]
3 [267]
106 [252]
5 [231]
105 [199]
192 [190]
44 [190]
9 [186]
305 [180]
80 [147]
82 [118]
96 [109]
32 [103]
84 [103]
59 [93]
23 [78]
50 [72]
1398 [67]
877 [67]
138 [63]
26 [63]
20 [62]
14 [57]
152 [56]
167 [55]
204 [53]
1465 [50]
175 [50]
52 [48]
1966 [46]
60 [46]
801 [46]
836 [46]
560 [44]
140 [40]
1131 [38]
176 [37]
1794 [37]
1811 [37]
71 [34]
1181 [33]
163 [32]
4191 [32]
1081 [31]
56 [31]
1089 [30]
1326 [29]
124 [26]
51 [26]
1053 [25]
145 [25]
4194 [25]
1613 [24]
603 [24]
8 [24]
4429 [22]
118 [21]
1442 [21]
2042 [21]
2135 [21]
756 [21]
168 [20]

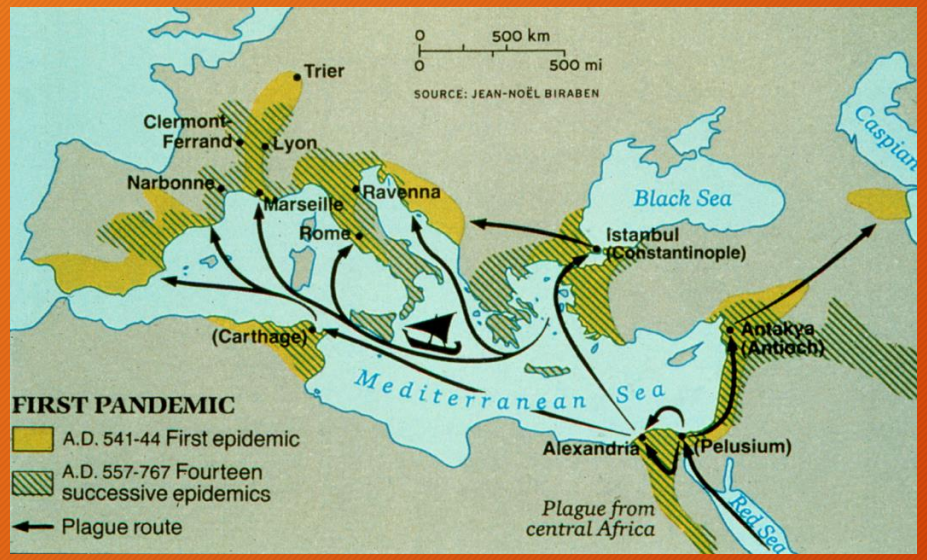
Recent developments for historical reconstructions (financial support by BBSRC & Wellcome Trust)

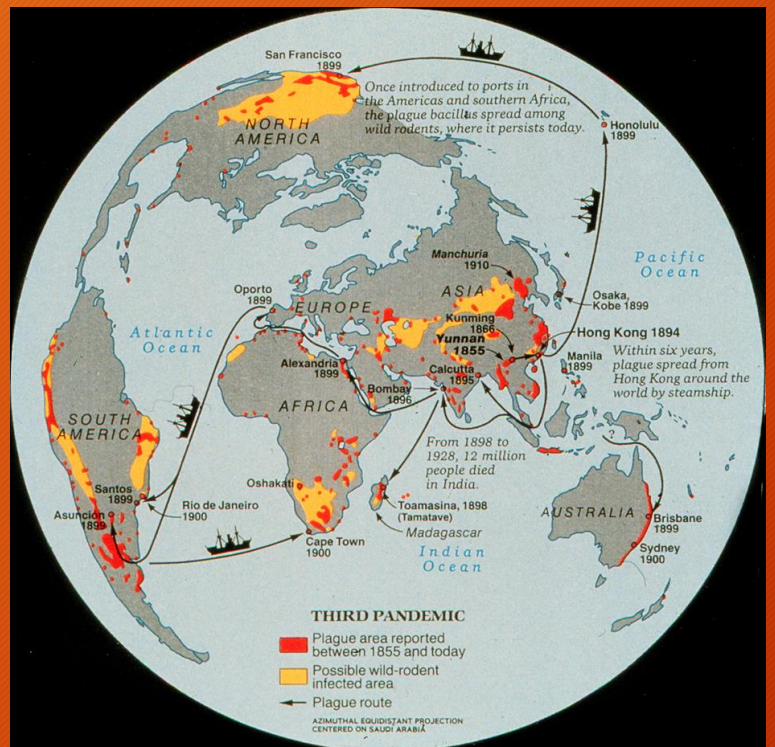
- Use defined SNP matrices in EnteroBase to integrate sparse genomic data from aDNA over 5,000 years with modern genomes
- Plague
- Invasive salmonellosis
- Biofilms in dental calculus

History and phylogeography of *Yersinia pestis* and the origins of plague



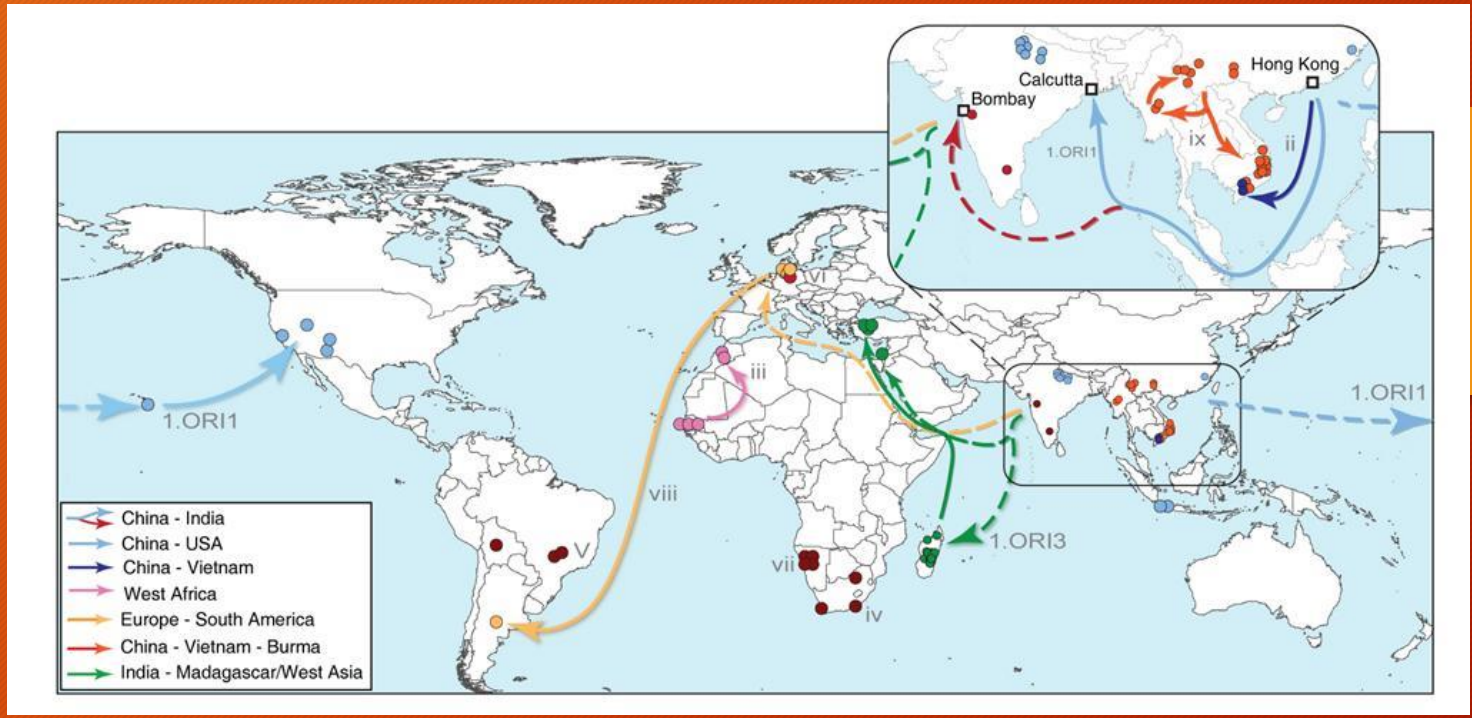
Three plague pandemics

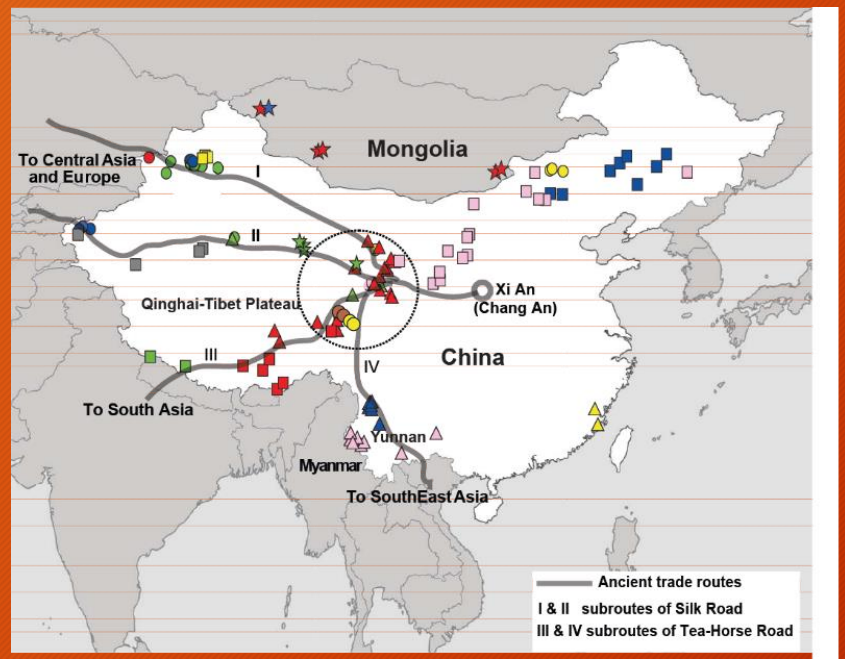




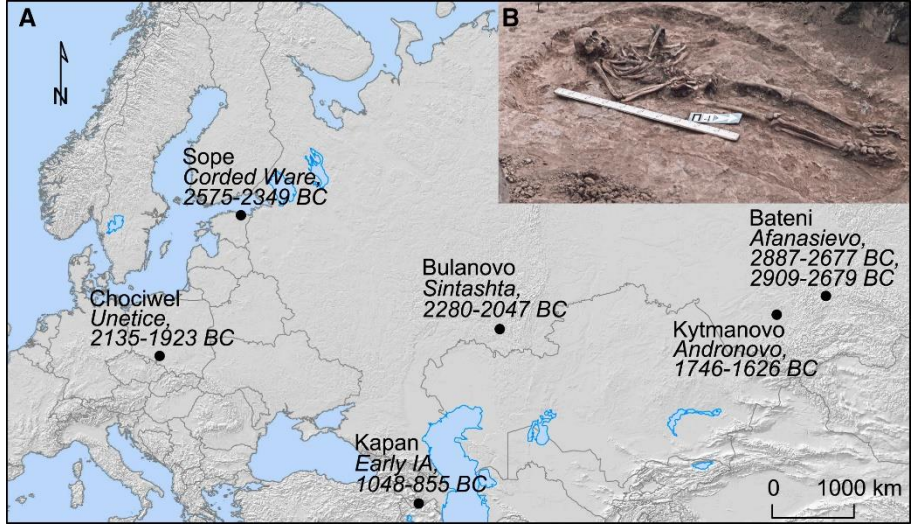
The identity between modern plague and historical plague (Black Death, Justinianic) was hotly discussed for >130 years.
Since 1894, microbiologists believed that historical plague was caused by *Y. pestis* but historians, epidemiologists and zoologists doubted this due to different epidemiological patterns.
These disputes were resolved through a combination of ancient DNA analyses of tooth pulp from skeletons in historical mass graves (Drancourt, Bramanti, Krause, Wagner) and my reconstructions together with Elisabeth Carniel and many others of historical migrations based on a detailed overview of modern genetic diversity.
(Figure by E. Carniel)

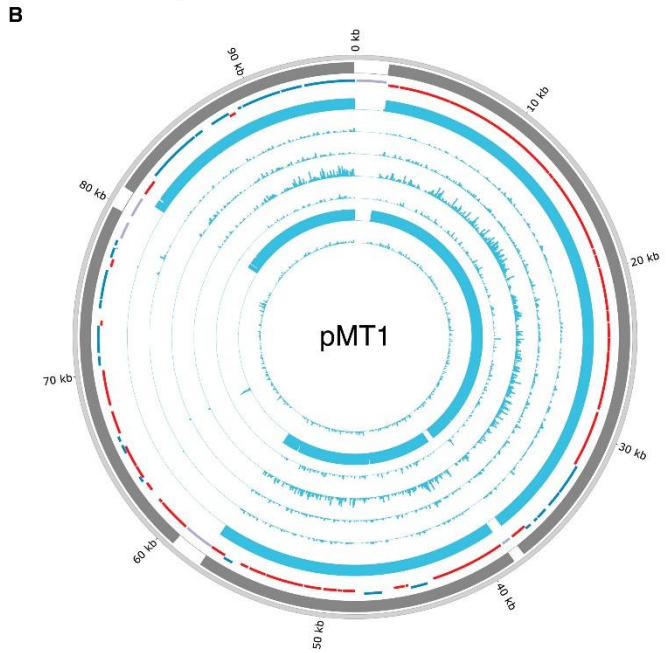
Third pandemic (Morelli *et al.*, 2010)





Spread from China (Cui *et al.*, 2013)

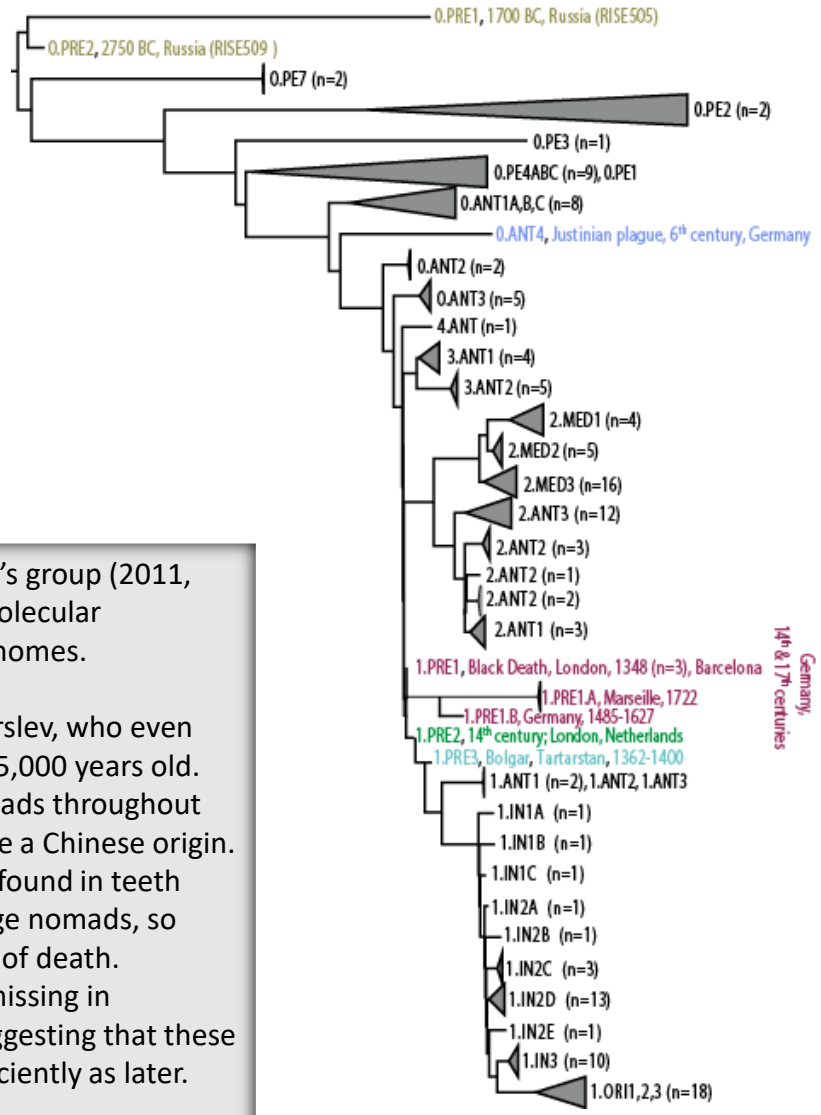




Kirstin Bos in Johannes Krause's group (2011, 2016-2018) has been doing molecular epidemiology on historical genomes.

So have Rasmussen and Willerslev, who even sequenced genomes that are 5,000 years old. Those genomes are from nomads throughout western Eurasia, and invalidate a Chinese origin. *Y pestis* sequence reads were found in teeth from 7/107 random Bronze Age nomads, so plague was a common source of death. *ymt* gene for murine toxin is missing in genomes before 1,000 BC, suggesting that these bacteria did not spread as efficiently as later.

Tree at right made by me by hand!!!



Rasmussen *et al.* (2015). Cell **163**:571-582.

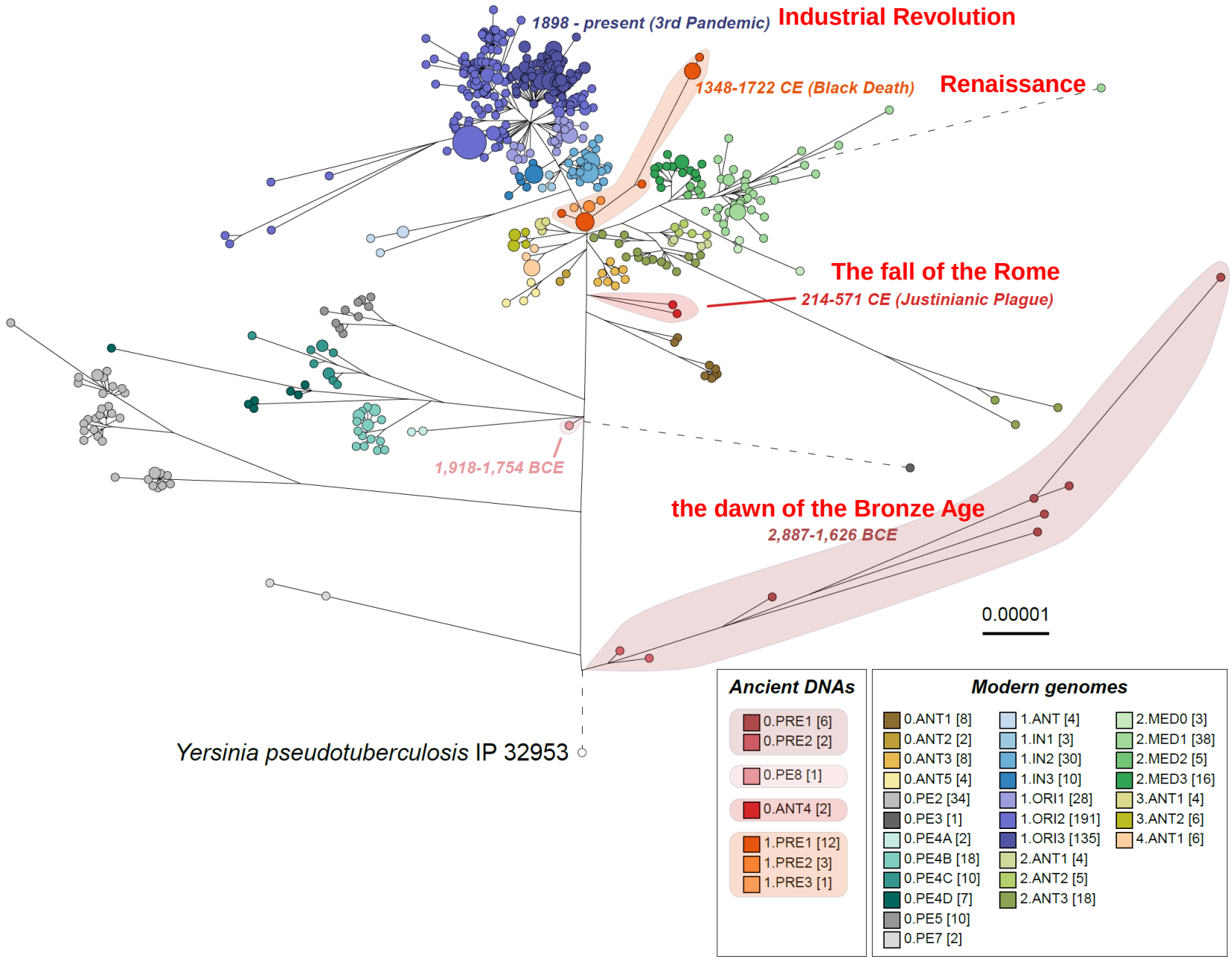
Integrating aDNA with modern genomes of *Y. pestis*

ML tree of 25,795 non-repetitive core SNPs (3.9Mb)

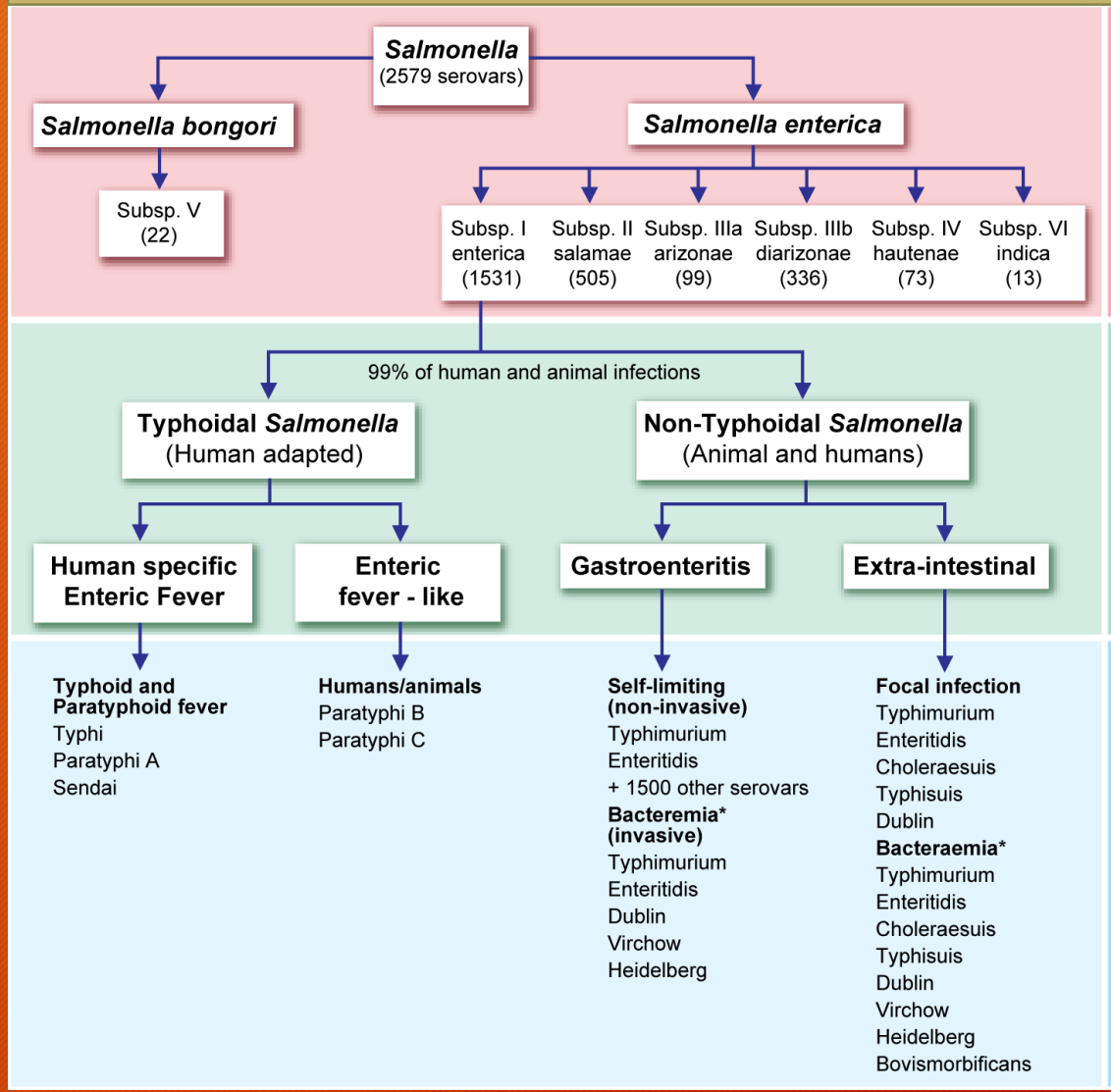
624 representative modern genomes
27 ancient genomes from the last 5000 years
4 ancient lineages
3 pandemics
Colour-coded by population grouping

New genomes from Justinianic on bioRXiv but not yet included because not yet publicly available

Interactive access at
http://tinyurl.com/Ypestis-ancSNP



Classical view of *Salmonella*: Species, subspecies, serovars, phage types, PFGE fingerprints, MLVA



Species and subspecies were originally defined by DNA-DNA hybridisation, confirmed by MLSA and are currently differentiated by biochemistry.

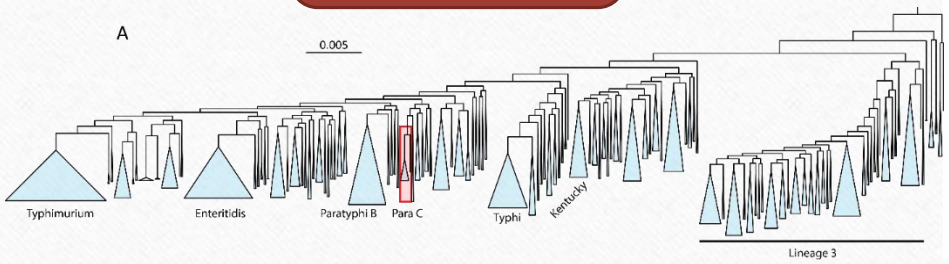
The split between typhoidal and non-typhoidal is based on the disease syndrome. Enteric fever is a prolonged fever, whilst extra-intestinal is usually acute and metastatic. Gastroenteritis is characterised by diarrhea.

Differentiation of serovars is by agglutination with specific antisera against LPS (O), two phases of flagella (H1 and H2). There are >40 O, 63 H1, 37 H2 and 1 capsule (Vi) antisera which have been described in about 1,500 combinations within subspecies I.

Insights provided by EnteroBase into *Salmonella*

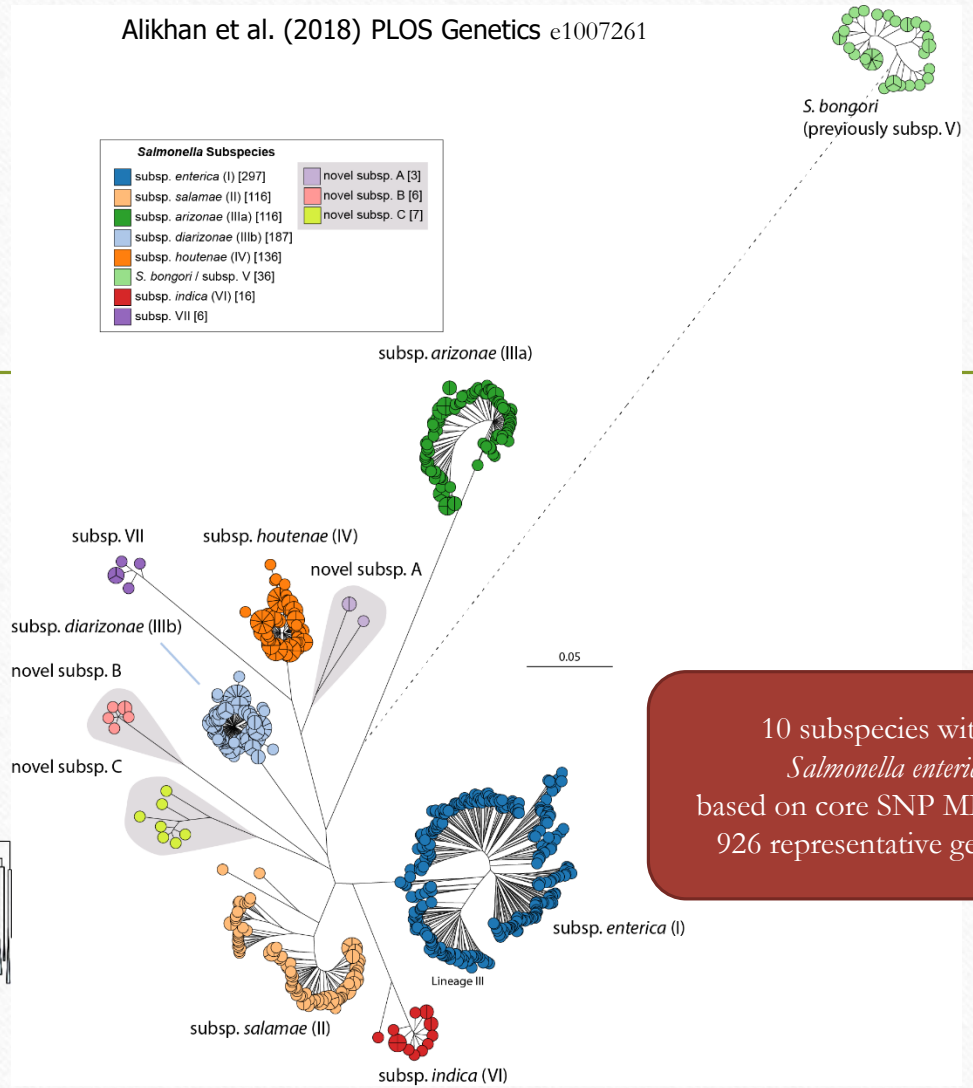
1. Clear distinction between subspecies, and identification of new subspecies
2. Identification of >48 Super-lineages within a super-tree of core gene trees from 2,964 representatives

48 super-lineages within *Salmonella enterica* subsp. *enterica*



Zhou et al. (2018) Current Biology **28:** 2420

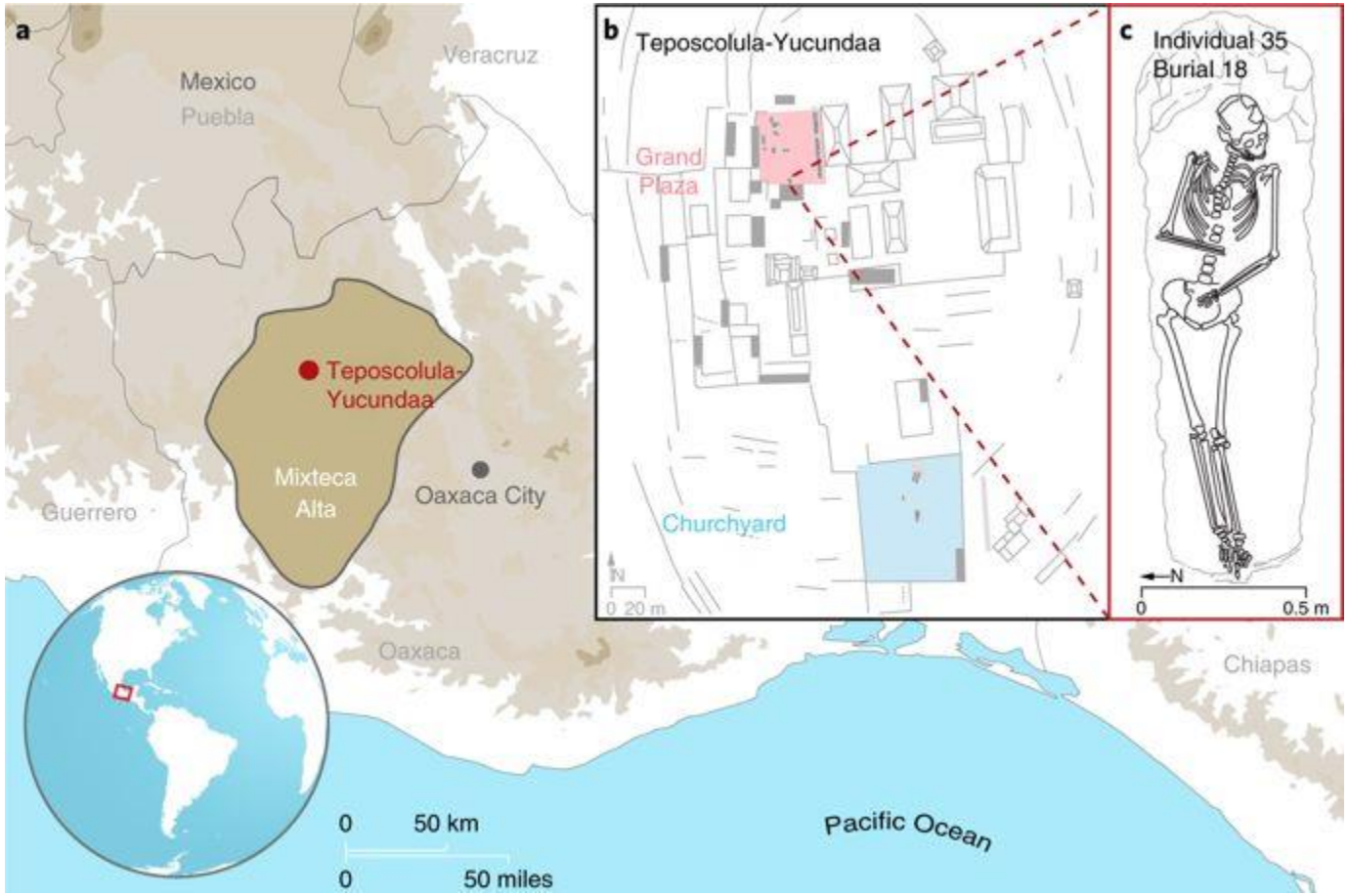
Alikhan et al. (2018) PLOS Genetics e1007261



10 subspecies within *Salmonella enterica* based on core SNP ML tree of 926 representative genomes

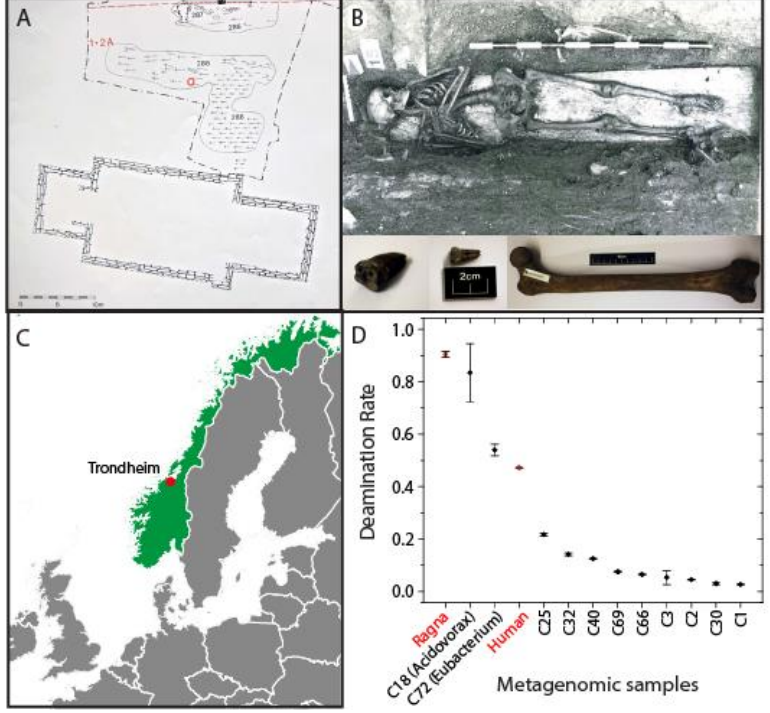
Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico
•Åshild J. Vågene, Alexander Herbig…Kirsten I. Bos & Johannes Krause *Nature Ecology & Evolution* **2**, 520–528 (2018)

Abstract:
Locally, this epidemic was known as '*cocoliztli*', the pathogenic cause of which has been debated for more than a century. Here, we present genome-wide data from ten individuals for *Salmonella enterica* subsp. *enterica* serovar Paratyphi C, a bacterial cause of enteric fever. We propose that *S*. Paratyphi C be considered a strong candidate for the epidemic population decline during the 1545 *cocoliztli* outbreak at Teposcolula-Yucundaa.



What genomes are present in metagenomes from DNA extracted from an 800 year old skeleton (SK152)?

1. Dark matter
2. Recent bacterial and archaeal methanogens that are typical of anoxic, boggy soil.
3. Dental calculus contains an old genome of *Eubacterium*, whose modern relatives cause periodontal disease
4. Low concentrations of DNA from *Salmonella enterica* serovar Paratyphi C in bones and teeth from which a 7fold genome (Ragna) was reconstructed.



Facial reconstruction of SK152

Zhemin Zhou, Inge Lundstrøm, Alicia Tran-Dien, Sebastián Duchêne, Nabil-Fareed Alikhan, Martin J. Sergeant, Gemma Langridge, Anna K. Fotakis, Satheesh Nair, Hans K. Stenøien, Stian S. Hamre, Sherwood Casjens, Axel Christophersen, Christopher Quince, Nicholas R. Thomson, François-Xavier Weill, Simon Y. W. Ho, M. Thomas P. Gilbert, Mark Achtman. Current Biology *2018* **28**: 2420

6,500 year old *Salmonella enterica* genomes link human-host adaptation to animal domestication

- Felix M. Key, Cosimo Posth, Luis R. Esquivel Gomez, Ron Hübler, Maria Spyrou, Gunnar U. Neumann, Anja Furtwängler, Susanna Sabin, Marta Burri, Antje Wissgott, Aditya Kumar Lankapalli, Åshild J. Vågene, Matthias Meyer, Sarah Nagel, Rezeda Tukhbatova, Aleksandr Khokhlov, Andrey Chizhevsky, Svend Hansen, Andrey B. Belinsky, Alexey Kalmykov, Anatoly R. Kantorovich, Vladimir E. Maslov, Stefania Vai, Monica Zavattaro, Alessandro Riga, David Caramelli, Robin Skeates, Jessica Beckett, Maria Giuseppina Gradoli, Noah Steuri, Albert Hafner, Inga Siebke, Sandra Lösch, Yilmaz Selim Erdal, Nabil-Fareed Alikhan, Zhemin Zhou, Mark Achtman, Kirsten Bos, Sabine Reinhold, Wolfgang Haak, Denise Kühnert, Alexander Herbig, Johannes Krause
- Metagenomes sequenced from 5,000 human teeth from skeletons in Europe and western Asia over the last 6500 years. Genomes of six strains were recovered at moderate coverage after oligonucleotide capture enrichment from the aDNA, four of which were in or closely related to the Para C super-lineage.
- No genomes were found from other serovars that might have been expected, such as Typhi, Paratyphi A, Dublin, Virchow, Heidelberg or Infantis.

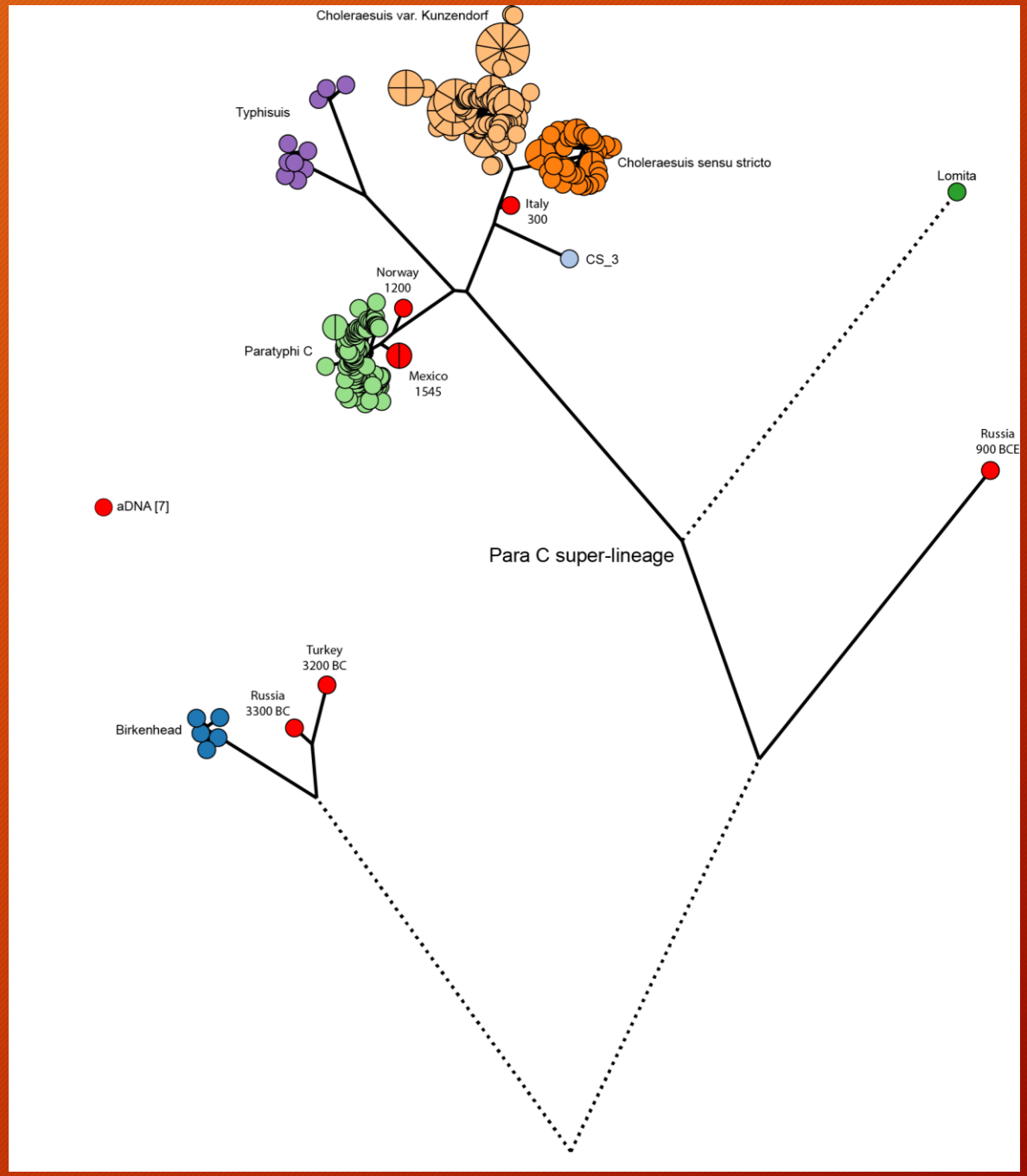
Sources of aDNA *Salmonella* genomes

ML Tree of ancient and modern genomes

The Para C super-lineage includes modern genomes in serovars Paratyphi C, Choleraesuis (variants sensu stricto, Kunzendorf, CS_3), Typhisuis and Lomita. Four aDNA genomes from human teeth mapped within this super-lineage.

The closest relative to the Para C super-lineage is a second super-lineage including Birkenhead. Two aDNAs mapped within this super-lineage.

An 8th aDNA mapped between those two super-lineages, with no known close modern relatives.



Achtman group

- Zhemin Zhou
- Nina Luhmann
- Sandra Bedarida
- Jane Charlesworth
- William Tyne
- Khaled Mohamed
- Yulei Fan

- Previously:
- Nabil-Fareed Alikhan
- Martin Sergeant
- Raffaella Bianucci

aDNA Collaborators

- Gemma Kay
- Felix Key
- Johannes Krause
- Rory Watt
- Camilla Speller
- Laura Weyrich

Funding

BBSRC bioscience for the future

Wellcome