Evolutionary Dynamics of the *Yersinia enterocolitica* Complex

2

Sandra Reuter, Nicholas R. Thomson, and Alan McNally

2.1 Introduction

Yersinia enterocolitica is one of the three species within the *Yersinia* genus pathogenic to humans, and is the species most commonly associated with human disease episodes (Bottone 1999). It is a zoonotic pathogen causing self-limiting gastro-intestinal infection, with disease heavily associated with the consumption of undercooked or contaminated food products of porcine origin (Bottone 1999; Fredriksson-Ahomaa and Korkeala 2003; McNally et al. 2004; Milnes et al. 2008). The species is subclassified into biotypes based on the utilisation of carbon sources, with biotype classification also corresponding to levels of pathogenesis in a mouse infection model (Wauters et al. 1987). Biotype 1B isolates are highly pathogenic in a mouse infection model resulting in mortality. Biotypes 2–5 are low pathogenic and result in mild gastro-intestinal disease in the model, and Biotype 1A isolates are completely non-pathogenic. Biotype 1B strains are most commonly described in North America and are termed new world strains, with isolation rare in the Eurasian land mass (Bottone 1999). Biotypes 2–5 are the most commonly isolated strain types in Eurasia in human disease cases and are very common in veterinary livestock, particularly pigs (Fredriksson-Ahomaa and Korkeala 2003; Milnes et al. 2008). Biotype 1A are ubiquitous in the environment and studies have reported them as the most common strain types in both humans and animals (McNally et al. 2004).

Decades of work have elucidated the complex pathogenesis mechanisms of the pathogenic *Yersinia*, including *Y. enterocolitica*, with most work on the species conducted on a small number of widely disseminated BT 1B isolates, notably the human clinical isolate 8081. Numerous classical virulence factors have been defined for the genus and species, including the well-described pYV virulence plasmid containing the archetypal Ysc type III secretion system responsible for delivery of toxic Yop effector proteins directly into eukaryotic cells (Cornelis 2002), the invasin protein responsible for internalisation into intestinal epithelial cells (Finlay and Falkow 1988), and the adhesin factors Ail, YadA, and Myf, as well as a wide range of accessory factors (Cornelis 1994). Additionally all these described virulence factors are absent in the non-pathogenic BT 1A isolates.

Similarly the past decade has seen seminal work conducted on the evolution of the pathogenic *Yersinia*. However almost all research has

S. Reuter • N.R. Thomson
Pathogen Sequencing Group, Wellcome Trust Sanger Institute, Cambridge, UK

A. McNally (✉)
Pathogen Research Group, Nottingham Trent University, Nottingham, UK
e-mail: alan.mcnally@ntu.ac.uk

A.M.P. de Almeida and N.C. Leal (eds.), *Advances in Yersinia Research*,
Advances in Experimental Medicine and Biology 954, DOI 10.1007/978-1-4614-3561-7_2,

Table 2.1 General properties of the sequenced reference genomes of *Y. enterocolitica*

Property	BT 1B O:8 (8081)*	BT 1A O:5 (5303)	BT 2 O:9 (21202)	BT 3 O:9 (5603)	BT 3 O:5,27 (14902)	BT 4 O:3 (1203)	BT 5 O:2a,2b,3 (309496)
Genome size (bp)	4,615,899	4,940,199	4,565,009	4,564,085	4,596,413	4,527,945	4,665,902
Number of genome contigs	1	1	8	8	6	4	5
Plasmid name and size (contiguated)	pYV 67 kb	pSR2-1A 8 kb pSR3-1A 4 kb	pYV 75 kb pSR1-2 42 kb	pYV 75 kb	pYV 75 kb	pYV 75 kb	pYV 75 kb pSR4-5 100 kb
G+C content (%)	47.27	48.42	48.09	48.10	48.16	48.22	48.24
Number of CDS	4,053	4,365	3,922	3,917	3,978	3,893	4,051
Coding density (%)	84.7	86.7	85.9	82.7	83.3	80.7	79.8
Average gene size (bp)	963	965	981	982	975	976	964
IS elements	60	31	87	85	79	101	110

focussed on the evolution of *Yersinia pestis* from *Yersinia pseudotuberculosis* employing a combination of multilocus sequence typing (Achtman et al. 1999; 2004) and whole genome sequence comparison approaches (Chain et al. 2004; Chen et al. 2010; Parkhill et al. 2001; Thomson et al. 2006). To date all molecular epidemiology on *Y. enterocolitica* has been performed using techniques such as amplified fragment length polymorphism (AFLP) and Pulsed field gel electrophoresis (PFGE), and despite the obvious variation across the species suggested by the range in biotypes, only one isolate has been sequenced, the human clinical BT 1B isolate 8081 (Thomson et al. 2006). The only thorough attempt to investigate genetic variation across the *Y. enterocolitica* species, and the evolution of this species within the context of the genus, has been performed using microarray comparative genomic hybridisation analysis (Howard et al. 2006; Thomson et al. 2006), and comparison of the available genome sequenced isolates across the genus (Chen et al. 2010). The work suggested that *Y. enterocolitica* was genetically split into distinct biotypes, with BT 1A displaying increased heterogeneity compared to the other biotypes. In addition there was no demarcation between animal and human clinical isolates. Whole genome comparisons showed that *Y. pestis* and *Y. pseudotuberculosis* were much more closely related to each other than to *Y. enterocolitica*, and that there were a large number of genes unique to *Y. enterocolitica* compared to the other two human pathogenic species (Thomson et al. 2006).

However the phylogenetic relationship between high, low, and non-pathogenic *Y. enterocolitica* is still not fully understood, nor is the genetic variation across the species. To address this a representative isolate from each biotype which had been previously extensively characterised both phenotypically and genotypically (Howard et al. 2006; McNally et al. 2006; Thomson et al. 2006) was sequenced using a combination of 454, Illumina and Sanger sequencing to create a reference genome for each *Y. enterocolitica* biotype (Table 2.1). A total of 100 *Y. enterocolitica* isolates representing a temporal and geographical cross section of all biotypes were sequenced using Illumina technology. This data was used to study the phylogenetic relationship between the different *Y. enterocolitica* biotypes, as well as the phylogeny of the *Yersinia* genus.

2.2 Differences in Genomic Content Across *Y. enterocolitica* Biotypes

The low-pathogenic biotypes (BT 2–5) all have a very similar genome size of about 4.5 Mb. The genome of the high-pathogenic strain BT 1B is slightly larger at 4.6 Mb. The non-pathogenic BT

Table 2.2 IS elements in the *Y. enterocolitica* BT

IS name	BT 1B – 8081	BT 1A – 5303	BT 2 – 21202	BT 3 – 5603	BT 3 – 14902	BT 4 – 1203	BT 5 – 3094
IS3	2	0	0	0	0	0	0
IS285	1	0	0	0	0	0	0
IS1222	1	1	0	0	0	0	0
IS1328	5	0	8	8	5	4	8
IS1329	6	0	0	0	0	0	0
IS1330	8	0	0	0	0	1	4
IS1400	4	0	2	2	3	1	2
IS1541	3	0	0	0	0	0	0
IS1660	6	0	0	0	0	0	0
IS1664	3	0	0	0	0	0	0
IS1665	5	0	0	0	0	0	0
IS1666	1	0	1	1	2	0	0
IS1667	8	0	37	35	30	53	50
IS1668	2	0	10	10	9	10	9
IS1669	5	0	2	3	2	6	3
Unclassified IS	–	30	27	26	28	26	34
Total	60	31	87	85	79	101	110

1A has the largest genome with 4.9 Mb, which is also the genome with the highest number of CDSs. The GC content and coding density as well as average gene size are very similar in all of the sequenced biotypes (Table 2.1).

The biggest notable difference can be seen in the total number and distribution of specific insertion sequence (IS) elements (Table 2.2). Biotype 1A shows the lowest number of IS elements. When looking at the distribution of specific IS elements, one can see that BT 1A has none of the characterised IS elements that BT 1B has. There are no copies of IS1328, IS1667, IS1668, or IS1669. Instead, it has copies of transposases similar to ISEhe3 and IS621 found in *Escherichia coli*, transposases similar to IS116/IS110/IS902 of *Stenotrophomonas*, and transposases similar to IS2 and IS3. All low- and high-pathogenic biotypes have similar numbers of IS1328 and IS1669. The low-pathogenic biotypes have a slightly higher copy number of IS1668, there are eight to ten copies as compared to only two copies in BT 1B, but the biggest difference is the expansion of IS1667, especially in BT 4 and 5. In BT 1B there are only eight copies of IS1667. BT 2 and 3 shows about 30 copies, and BT 4 and 5 have the highest number with over 50 copies each. This makes it the most prevalent IS element in the low-pathogenic biotypes. The high-pathogenic biotype does not show any prevalence for a specific IS element, there is not a single IS element with more than eight copies.

The *Y. enterocolitica* biotypes share a large common backbone. Obvious regions of difference are composed of genomic islands and prophages. It is also evident that there are regions which are only shared amongst a subset of the BTs. The *fes-fep* operon at base position 700,000 (respective to 8081 genome sequence) for example is present in all of the biotypes except for BT 4. The *Yersinia* Genus Type III secretion system (YGT) is a novel TTSS which has been briefly mentioned in recent genomic studies of the genus but remains to be characterised (Chen et al. 2010; Wang et al. 2011). It is located at position 4,500,000 and is present in all non- and low-pathogenic biotypes, but is truncated in the high-pathogenic BT, with only one remaining CDS.

Using information about orthologous genes sets, a grouping analysis was carried out to more

accurately define those CDS sets which are shared between certain biotypes or all BTs. Considering all possible combinations, more than 60 groups of CDSs shared between various biotypes are possible. From this several major groups were identified. One group relates to both BT 4 and BT 5 which show signs of having undergone independent genome decay. BT 4 has lost the *fes-fep* iron-uptake operon. A putative amino acid transport system and a phospho-transferase system have been lost from BT 5. Both BTs are missing an additional putative sugar transport system. This loss of metabolic properties is also evident in the biotyping scheme used for differentiation of *Y. enterocolitica* biotypes. The second group includes all low- and non-pathogenic biotypes, and excludes the high-pathogenic BT 1B. It consists of 125 CDSs, which include the YGT, a LPS outer core biosynthesis region, phosphate transport systems for *N*-acetyl-D-galactosamine, cytochrome c biogenesis, and dimethylsulfoxide metabolism. The third group of about 35 CDSs includes all pathogenic biotypes, excluding the non-pathogenic BT 1A. The CDSs include genes for threonine/serine metabolism, a repeat toxin protein, the attachment invasion locus gene *ail,* and hypothetical proteins. The threonine/serine operon however is presumably only functional in BT 1B, as it contains additional genes encoding a dehydrogenase/reductase, an aldo/keto reductase, and a transcriptional regulatory protein. All low-pathogenic biotypes are included in the next group and contain approximately 170 CDSs. It is interesting that of these, 32 are transposases. Some of the other CDSs encode for the insect toxin complex pathogenicity island (Bresolin et al. 2006), fimbria-related proteins, and the second flagella cluster Flag-2 (Bresolin et al. 2008). Biotype 1A and 1B share 86 CDSs. These include metabolic pathway genes such as sugar transporters, permeases, kinases, and reductases. The group also includes a nickel–cobalt efflux system, the arsenic resistance cluster of BT 1B located outside the plasticity zone, YGI-2 (a glycolipoprotein) and some parts of YGI-4 (an integrated plasmid).

The overall impression is that the low-pathogenic biotypes are very closely related to each other. The unique regions of BT 3 O:5,27, BT 4, and BT5 consist of phages, whereas BT 2 and BT 3 O:9 share the same phages. IS elements are often conserved with respect to their position in the low-pathogenic BT, possibly indicating ancient and stable insertion events. The high-pathogenic BT 1B and the non-pathogenic BT 1A are distinct from the low-pathogenic BT and each have more than 500 unique CDSs. However they also share a 86 CDSs relating to metabolic functions that are absent from the low-pathogenic BT. It is striking that all of the pathogenic biotypes only have 35 CDSs in common that separate them from the non-pathogenic BT. It is also notable that the non-, low-, and high-pathogenic strains separate without consideration of the virulence plasmid pYV.

2.3 Phylogenetic Relationship Amongst *Y. enterocolitica* Biotypes

Given the unexpected nature of the genome content comparisons, a whole genome-based phylogeny of the species was performed. This was conducted to answer 2 key questions. Firstly were the reference genome strains truly representative of the *Y. enterocolitica* population, and secondly what is the exact phylogenetic relationship between the high-, low-, and non-pathogenic strains. Genome resequencing was performed on a global collection of 94 *Y. enteroclitica* isolates using the Illumina sequencing platform and the reference genomes of each biotype as reference scaffolds. A de novo assembly was then performed on each genome to create a FASTA file for each strain, with SSAHA used for mapping of Illumina reads against the reference strain 8081 (Ning et al. 2001). An alignment was created from SNP sites using progressiveMAUVE (Darling et al. 2010) and a maximum likelihood phylogeny created using RaxML (Stamatakis et al. 2005) (Fig. 2.1).

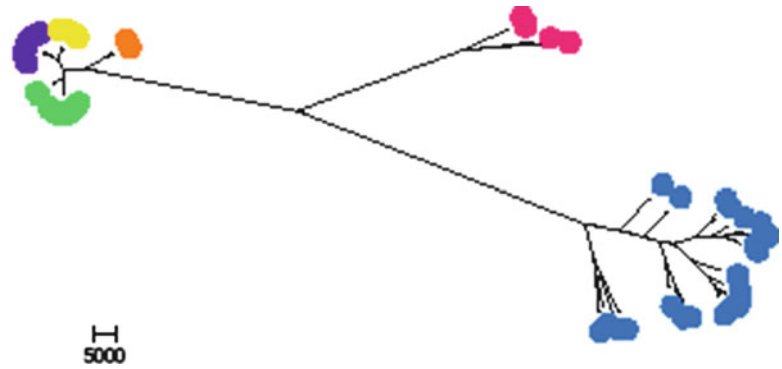

Fig. 2.1 Phylogeny of the *Y. enterocolitica species.* Maximum likelihood phylogenetic inference of the *Y. enterocolitica* species. Biotype 1A strains are indicated by *blue circles*, Biotype 1B strains by *red circles*, Biotype 5 strains by *orange circles*, Biotype 4 O:3 strains by *purple circles*, Biotype 3 O:5,27 strains by *yellow circles*, and Biotype 2 and Biotype 3 O:9 strains by *green circles*

The resulting phylogenetic tree shows clearly that the reference genome created for each biotype is perfectly indicative of the genomic content for that biotype, and that these can be considered as reference genomes for each of the *Y. enterocolitica* biotypes.

From a phylogenetic perspective the most striking observation is that the non-pathogenic BT 1A and high-pathogenic BT 1B are more closely related to each other than to the low-pathogenic biotypes, with phylogeny inferring that non-pathogenic BT 1A is the progenitor for modern *Y. enterocolitica*. Extrapolation of the phylogeny to the genome content data would agree that the emergence of the pathogenic biotypes has occurred as a result of gain of pYV, but more crucially extensive gene decay within operons encoding metabolic function, mirroring the observation well recorded within *Y. pseudotuberculosis* and *Y. pestis* (Chain et al. 2006).

Closer examination of the whole genome-based species phylogeny highlights some interesting novel observations, the most obvious of which is the clustering of strains according to serotype and not biotype within the low-pathogenic clade. This has curious potential ramifications in strain typing and epidemiology, in that *Y. enterocolitica* are more closely related at the whole genome level based on serotype than on biotype. *Y. enterocolitica* biotypes are often only differentiated by the utilisation of a single metabolite, results of which can often vary between labs based on prior growth conditions of the organism. The clear elucidation that serotyping is far more accurate in determining grouping and relatedness of strains may make strain typing and epidemiology easier for *Y. enterocolitica*. However of equal frustration is that even at a whole genome level of comparison it is impossible to differentiate between human and animal strains, with no clear grouping of human clinical isolates. Rather all human and animal strains of the same serotype seem equally related based on our phylogeny inferred from 100 isolates. This is in agreement with molecular epidemiological studies based on more primitive genotyping methods, and suggests that *Y. enterocolitica* is unlike other enteropathogens such as *E. coli* O157 (Kim et al. 1999) and *Campylobacter jejuni* (Champion et al. 2005) where subsets of animal isolates have been shown to be related to, and therefore more likely to cause disease in, human clinical isolates. There is a possibility that full genome-based phylogeny on a more substantial strain set may elucidate a clear epidemiological relationship between animal subset

and human strains, and this is the focus of current research within our group.

Phenotypic microarray experiments conducted using Biolog technology support this observation, with the reference BT 1A strain exhibiting the greatest degree of metabolic flexibility, followed by the BT 1B high pathogenic, leading down to the low-pathogenic biotypes which show the most restricted level of metabolic activity. In particular the BT 3 O:5,27 strain shows high levels of metabolic restriction, consistent with epidemiological observations that this pathotype of *Y. enterocolitica* is very rarely associated with human disease cases, but is ubiquitous in livestock (McNally et al. 2004). In addition the BT 5 reference strain shows an extremely restricted metabolic range. BT 5 is very rarely isolated from humans or the environment, and the few references to this pathotype in the literature refer to isolation from hares, leading to this pathotype being referred to as the hare biotype (Wuthe and Aleksic 1997). The extremely restricted metabolic profile of BT 5 is consistent with niche restriction of pathogens and reduced genome/metabolism observed both within the genus in *Y. pestis* (Chain et al. 2006), and in other non-related genera such as in the case of *Mycobacterium leprae* (Cole et al. 2001).

2.4 Phylogenetic Inference of the *Yersinia* Genus Based on Illumina Genome Resequencing Data

In an attempt to place the phylogeny of the *Y. enteroclitica* complex in the context of the wider genus, our genome sequence data set was combined with all publicly available *Yersinia* genus genome sequences including the existing *Y. pestis* and *Y. pseudotuberculosis* genome sequences as well as short read archive data for environmental species to create a whole genus phylogeny. Attempts to do this in a similar fashion to the methods employed for the species phylogeny were hampered by the enormous heterogeneity across the genomes across the genus, making it impossible to obtain whole genome alignments of sufficient quality to obtain meaningful phylogeny. To subvert this problem, the sequences of the housekeeping genes previously used for MLST studies in *Y. pestis* and *Y. pseudotuberculosis* (Achtman et al. 1999) were extracted from all of our genome data sets along with all publicly available *Yersinia* genome sequences, and concatenated to produce a sequence profile for each strain. The resulting concatenated sequences were then aligned using Mega, and a maximum likelihood phylogeny obtained using RaxML (Fig. 2.2).

Confidence in the robust nature of the phylogenetic inference of the genus tree is obtained from the observation that the phylogeny of the sequenced *Y. enterocolitica* species is identical in the concatenated MLST tree as it is for the previous whole genome sequence-derived species phylogeny. From this the most striking observation is that the human pathogenic *Y. enterocolitica* and *Y. pestis/Y. pseudotuberculosis* complexes are at opposite ends of the phylogeny, suggesting that they evolved in a manner completely independent of each other, and yet using strikingly similar mechanisms of gene gain and metabolic function encoding gene loss. This is further strengthened by the positioning of all the environmental *Yersinia* species between the two complexes. The elucidation that *Y. pestis/Y. pseudotuberculosis* and *Y. enterocolitica* are only slightly genetically related organisms opens debate on the way in which pathogenesis and fundamental biology of these organisms is studied. Up until now most *Yersinia* biology research conducted on one of the human pathogenic species has been acceptably extrapolated to be applicable to the others. However the phylogenetic inference obtained from genomic data suggests that the genetic distance between the two complexes is such that, despite obvious common strategies employed by both pathogenic complexes, there is a need to study the two in independence of each other to garner a full understanding of the biological and pathogenic spectrum of this complex and fascinating genus.

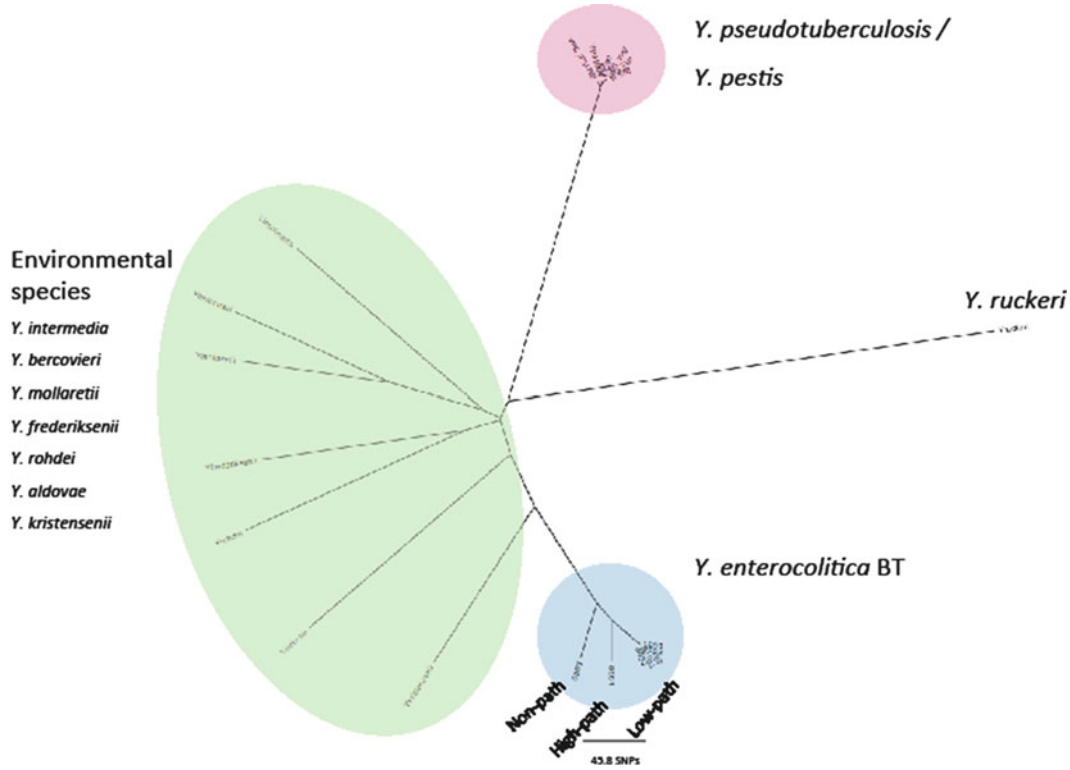

Fig. 2.2 Phylogeny of the *Yersinia genus.* Maximum likelihood phylogeny showing the phylogeny of the *Yersinia* genus

Acknowledgements We would like to thank Mark Achtman for efforts in collating a collection of strains, and all those who contributed to that collection, particularly Thilo Fuchs, Elisabeth Carniel, and Muriel DuFour. This work was funded by a VC studentship awarded to SR and AM, and by the Wellcome Trust Sanger Institute.

References

Achtman M, Zurth K, Morelli G et al (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc Nat Acad Sci USA 96:14043–14048

Achtman M, Morelli G, Zhu P et al (2004) Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc Nat Acad Sci USA 101:17837–17842

Bottone EJ (1999) *Yersinia enterocolitica*: overview and epidemiologic correlates. Microb Infect 1:323–333

Bresolin G, Morgan JAW, Ilgen D et al (2006) Low temperature-induced insecticidal activity of *Yersinia enterocolitica*. Mol Microbiol 59:503–512

Bresolin G, Trcek J, Scherer S et al (2008) Presence of a functional flagellar cluster Flag-2 and low-temperature expression of flagellar genes in *Yersinia enterocolitica* W22703. Microbiology 154:196–206

Chain PSG, Carniel E, Larimer FW et al (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. Proc Nat Acad Sci USA 101:13826–13831

Chain PS, Hu P, Malfatti SA et al (2006) Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. J Bacteriol 188:4453–4463

Champion OL, Gaunt MW, Gundogdu O et al (2005) Comparative phylogenomics of the food-borne pathogen *Campylobacter jejuni* reveals genetic markers predictive of infection source. Proc Natl Acad Sci USA 102:16043–16048

Chen P, Cook C, Stewart A et al (2010) Genomic characterization of the *Yersinia* genus. Genome Biol 11:R1

Cole ST, Eiglmeier K, Parkhill J et al (2001) Massive gene decay in the leprosy bacillus. Nature 409:1007–1011

Cornelis GR (1994) *Yersinia* pathogenicity factors. Contrib Microbiol Immunol 192:243–263

Cornelis GR (2002) *Yersinia* type III secretion: send in the effectors. J Cell Biol 158:401–408

Darling AE, Mau B, Perna NT (2010) ProgressiveMAUVE: multiple genome alignment with gene gain, loss and rearrangement. PLOS One 5:e11147

Finlay BB, Falkow S (1988) Comparison of the invasion strategies used by *Salmonella cholerae suis*, *Shigella flexneri* and *Yersinia enterocolitica* to enter cultured animal cells: endosome acidification is not required for bacterial invasion or intracellular replication. Biochimie 70:1089–1099

Fredriksson-Ahomaa M, Korkeala H (2003) Molecular epidemiology of *Yersinia enterocolitica* 4/O:3. Adv Exp Med Biol 539:295–302

Howard SL, Gaunt MW, Hinds J et al (2006) Application of comparative phylogenomics to study the evolution of *Yersinia enterocolitica* and to identify genetic differences relating to pathogenicity. J Bacteriol 188:3645–3653

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 96:13288–13293

McNally A, Cheasty T, Fearnley C et al (2004) Comparison of the biotypes of *Yersinia enterocolitica* isolated from pigs, cattle and sheep at slaughter and from humans with yersiniosis in Great Britain during 1999–2000. J Appl Microbiol 39:103–108

McNally A, Dalton T, Ragione RML et al (2006) *Yersinia enterocolitica* isolates of differing biotypes from humans and animals are adherent, invasive and persist in macrophages, but differ in cytokine secretion profiles *in vitro*. J Med Microbiol 55:1725–1734

Milnes A, Stewart I, Clifton-Hadley F et al (2008) Intestinal carriage of verocytotoxigenic *Escherichia coli* O157, *Salmonella*, thermophilic *Campylobacter* and *Yersinia enterocolitica*, in cattle, sheep and pigs at slaughter in Great Britain during 2003. Epidemiol Infect 136:739–751

Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11:1725–1729

Parkhill J, Wren BW, Thomson NR et al (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 413:523–527

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456

Thomson NR, Howard S, Wren BW et al (2006) The complete genome sequence and comparative genome analysis of the high pathogenicity *Yersinia enterocolitica* strain 8081. PLOS Path 2:e206

Wang X, Li Y, Jing H et al (2011) Complete genome sequence of a *Yersinia enterocolitica* "Old World" (3/O:9) strain and comparison with the "New World" (1B/O:8) strain. J Clin Microbiol 49:1251

Wauters G, Kandolo K, Janssens M (1987) Revised biogrouping scheme of *Yersinia enterocolitica*. Contrib Microbiol Immunol 9:14–21

Wuthe HH, Aleksic S (1997) *Yersinia enterocolitica* serovar 2a, wb, 3:b, c biovar 5 in hares and sheep. Berl Munch Tierarztl Wochenschr 110:176

http://www.springer.com/978-1-4614-3560-0

Advances in Yersinia Research
de Almeida, A.M.P.; Leal, N.C. (Eds.)
2012, XXIV, 367 p., Hardcover
ISBN: 978-1-4614-3560-0