EVOLUTIONARY DYNAMICS OF THE *YERSINIA ENTEROCOLITICA* COMPLEX

SANDRA REUTER

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

Collaborating Establishment: Wellcome Trust Sanger Institute, Cambridge

September 2011

Abstract

The genus Yersinia consists of a heterogeneous collection of organisms, comprising the highly pathogenic species *Yersinia pestis*, enteropathogens *Yersinia pseudotuberculosis* and *Yersinia enterocolitica* as well as environmental species. The evolutionary history of *Y. pestis* has been well documented, but information on the evolutionary relationship between the other Yersiniae is less characterized. *Y. enterocolitica* is a diverse species classed into six different biotypes (BT), but only a single genome sequence for high-pathogenic BT 1B was available at the start of the project.

This project looked into the dynamics shaping the pathogenic lineages in Yersinia. For each of the *Y. enterocolitica* BTs, a reference genome was sequenced, annotated and analysed in detail. Analysis of biochemical characteristics was carried out using a phenotypic microarray. A further 98 *Y. enterocolitica*, 36 *Y. pseudotuberculosis*/*Y. pestis* isolates and 81 environmental species were sequenced and analysed to investigate the evolutionary dynamics of the genus *Yersinia*.

The *Y. enterocolitica* BTs form three distinct groups according to non-, low-, and high-pathogenic BT, each with a different set of accessory genes. The *Y. enterocolitica* core genome comprises ~3,300 CDSs. The non-pathogenic BT contains genes involved in environmental survival, whereas the high-pathogenic BT has a high-pathogenicity island and the unique Ysa type III secretion system. The low-pathogenic BTs share a second flagella cluster and an insect pathogenicity island. Apart from the virulence plasmid pYV, only 40 CDSs are shared in all pathogenic BTs including the adhesion invasion locus *ail*. The phenotypic microarray confirmed higher metabolic flexibility of the non-pathogenic BT, and decreased biochemical abilities in the low-pathogenic BTs, mirroring an increased number of pseudogenes.

Looking at the whole genus, pathogenic lineages are visible at opposite ends of an evolutionary tree. *Y. pestis* has evolved out of *Y. pseudotuberculosis*, and both are distinct from *Y. enterocolitica* which developed out of the environmental Yersiniae. Fish pathogen *Y. ruckeri* forms a third independent pathogenic lineage. Despite the large evolutionary distance, comparing dynamics in *Y. enterocolitica* to *Y. pestis* it is clear that independent acquisition of the virulence plasmid and metabolic streamlining to specific hosts drove pathogen evolution. Whilst this has concurrently happened in *Y. pestis*, this development has been weighted differently in lineages of *Y. enterocolitica* leading to a high-pathogenic, human-adapted biotype and low-pathogenic biotypes, which can be adapted to human or animal hosts.

Contents

Abstract i
Contents ii
Acknowledments v
Declaration vi
Abbreviations vii
List of tables ix
List of figures x
1 Introduction 1
1.1 The Genus *Yersinia* 1
1.1.1 *Yersinia pestis* and *Yersinia pseudotuberculosis* 2
1.1.2 *Yersinia enterocolitica* 5
1.1.3 The special case of *Yersinia ruckeri* 8
1.1.4 Environmental *Yersiniae* 9
1.2 Pathogenicity in the genus *Yersinia* 11
1.2.1 Relationship of pathogenic *Yersiniae* 11
1.2.2 Plasmid-encoded basis for pathogenicity 12
1.2.2.1 Shared virulence plasmid pYV 12
1.2.2.2 Plasmids specific to *Y. pestis*: pPCP and pMT 13
1.2.2.3 Plasmids specific to *Y. pseudotuberculosis*: pVM82 and pIB 15
1.2.3 Chromosomal markers of pathogenicity 15
1.2.3.1 Invasion and adhesion proteins: Inv and Ail 15
1.2.3.2 Yersinia adhesion pathogenicity island (YAPI) 16
1.2.3.3 High-pathogenicity island (HPI) - Pigmentation / hemin storage locus 17
1.2.3.4 *Yersinia pseudotuberculosis* superantigen YPM 18
1.2.3.5 Ysa type 3 secretion system 19
1.2.3.6 Further chromosomally encoded virulence factors 19
1.3 Evolution of the genus *Yersinia* 20
1.3.1 Differentiation and speciation in *Yersinia* 20
1.3.2 16S rDNA analysis and MLST scheme 21
1.3.3 Current model of evolution 23
1.4 Next-generation sequencing techniques 26
1.4.1 Capillary sequencing 27
1.4.2 454 Pyrosequencing 29
1.4.3 Illumina sequencing 31
1.4.4 Comparison of capillary, pryo- and Illumina sequencing and applications 34
1.5 Aim of this project 37
1.5.1 The *Yersinia* Paradox 37
1.5.2 Aims of the study 37
2 Materials and Methods 39
2.1 Bacterial Strains 39
2.1.1 Reference Genomes 39
2.1.2 Worldwide Collections 39
2.1.3 Additional strains 40
2.2 Sequencing, Assembly and Annotation 40
2.2.1 Sequencing Technologies 40
2.2.2 Annotation and Sequence Analysis 42
2.3 Plasmid profiling 46
2.3.1 Reagents 46
2.3.2 Protocol 46
2.4 Phenotypic Microarray 47
2.4.1 Experimental setup 47
2.4.2 Analysis 48
2.5 Phylogenetic Analyses 51
2.5.1 Phylogenetic reconstruction of *Y. enterocolitica*: whole genome alignment 51
2.5.2 Phylogenetic reconstruction of the genus *Yersinia*: extended MLST scheme 51

2.5.3 Phylogenetic reconstruction of the virulence plasmid pYV 53
2.5.4 Serotyping from sequence data 53
2.5.5 Phylogenetic reconstruction using 16S rDNA sequence 53
2.5.6 Phylogenetic reconstruction including *Y. entomophaga* and *Y. nurmii* 53

3 Sequencing and Analysis of the Reference Genomes for *Y. enterocolitica* 55
3.1 Introduction 55
3.2 Results 56
3.2.1 Sequencing, assembly and sequence improvement of the *Y. enterocolitica* biotypes genomes 56
3.2.2 Genome annotation 64
3.2.3 Genome comparison 66
3.2.4 General genomic features of the reference genomes 67
3.2.5 General characteristics of reference genomes 70
3.3 Discussion 72
3.3.1 Generation of Improved High-Quality Draft genome sequences 72
3.3.2 General genomic composition of a *Y. enterocolitica* genome 74

4 Comparative analysis of the *Y. enterocolitica* biotypes 77
4.1 Introduction 77
4.2 Results 78
4.2.1 The core- and accessory genome of *Y. enterocolitica* 78
4.2.2 Regions of difference in *Y. enterocolitica* biotypes 87
4.2.2.1 *Yersinia* Genus Type 3 Secretion System 87
4.2.2.2 Variation of the Plasticity Zone 89
4.2.2.3 Lipopolysaccharide operons 91
4.2.3 Pseudogene distribution amongst the *Y. enterocolitica* biotypes 95
4.2.4 The role of IS elements and recombination in *Y. enterocolitica* biotypes 103
4.3 Discussion 108
4.3.1 High-, low-, and non-pathogenic biotypes have distinct genomic features 108
4.3.2 Metabolism, pseudogenization and IS elements are important evolutionary forces in *Y. enterocolitica* 110
4.3.3 Differences in niche adaptation 112

5 Plasmid Variety and Variability in *Y. enterocolitica* Biotypes 115
5.1 Introduction 115
5.2 Results 116
5.2.1 Genetic Organization of the virulence plasmid pYV in *Y. enterocolitica* BT 116
5.2.2 Generic Map of pYV 120
5.2.3 Variability in pYV core gene sequence 121
5.2.3.1 *yopM* 121
5.2.3.2 *lcrV* 121
5.2.3.3 *yscP* 123
5.2.4 Plasmid Variety: Unique plasmids 127
5.2.4.1 Unique plasmids in the reference strain for BT1A: pSR2-1A and pSR3-1A 127
5.2.4.2 Plasmids in atypical BT 1A/1B strains 129
5.2.4.3 Unique plasmid in the reference strain for BT2: pSR1-2 130
5.2.4.4 Unique plasmid in the reference strain for BT5: pSR4-5 134
5.3 Discussion 136
5.3.1 The virulence plasmid differs in high- and low-pathogenic BTs 136
5.3.2 Sequence variation in single genes: *lcrV*, *yopM* and *yscP* 138
5.3.3 Plasmid Repertoire in Yersinia 140

6 Metabolic profiling of the *Y. enterocolitica* biotypes 141
6.1 Introduction 141
6.1.1 Biotyping and speciation based on biochemical properties 141
6.1.2 Phenotypic microarray technology 142
6.2 Results 145
6.2.1 Metabolic activity in *Y. enterocolitica* biotypes at optimum temperature 145
6.2.1.1 Metabolism of carbon sources 148
6.2.1.2 Utilization of essential nutrients nitrogen, phosphorus and sulphur 149
6.2.1.3 Biotyping reactions 150
6.2.2 Differences relating to genetic differences 152
6.2.2.1 L-arabinose 153
6.2.2.2 D-trehalose and D-sorbose 156
6.2.2.3 N-acetyl-D-galactosamine 157

6.2.2.4 D-arabitol....158
6.2.2.5 m-inositol....159
6.2.3 Metabolic activity of *Y. enterocolitica* biotypes at human body temperature....162
6.3 Discussion....164
6.3.1 Metabolic profiling of *Y. enterocolitica* confirms genomic analysis....164
6.3.2 Linking genomic and phenotypic differences....166
6.3.3 The influence of temperature on metabolism in *Y. enterocolitica*....168

7 Phylogeny of the Genus *Yersinia* and the species *Y. enterocolitica*.... 171
7.1 Introduction....171
7.2 Results....172
7.2.1 Phylogeny of the species *Y. enterocolitica*....172
7.2.2 Phylogeny of the genus *Yersinia*....180
7.2.3 Phylogeny of the genus *Yersinia* based on 16S rDNA sequences....186
7.2.4 Acquisition of the virulence plasmid pYV in pathogenic *Yersiniae*....188
7.3 Discussion....191
7.3.1 The evolution of pathogenic *Y. enterocolitica* lineages....191
7.3.2 The genus *Yersinia* and the virulence plasmid pYV – independent evolution of pathogenicity....193
7.3.3 Differentiation of species in the genus *Yersinia*....194

8 Overall Discussion and Future Work.... 196
8.1 Analysis of reference genomes and plasmids....196
8.2 Analysis of phenotypic microarray....197
8.3 Revised model for the evolution of pathogenicity in the genus *Yersinia*....198
8.4 Future Work....201

9 Appendix.... 204

References.... 259

Acknowledments

I would like to thank my supervisors Dr Alan McNally and Dr Georgina Manning, Nottingham Trent University, and Dr Nicholas R Thomson, Wellcome Trust Sanger Institute, for the opportunity to carry out this PhD project and for their excellent support and encouragement throughout this project. Their enthusiasm for the project was infective and I very much enjoyed working with them.

Many thanks go to Nico Petty, Matt Holden, Simon Harris, Maria Fookes, Nick Croucher and Tom Connor at the Wellcome Trust Sanger Institute for answering all my questions and for a lot of bioinformatics support. Thanks to Theresa Feltwell for help with the phenotypic microarray and for providing a British experience including Sunday roast dinners and apple crumbles; Danielle Walker for all the Ts shared, and the Gs, Cs, and As as well; and Lars Barquist for implementing the phenotypic microarray data into a visually nice heatmap.

I would also like to thank Prof Brendan Wren, Dr Mike Prentice, and Dr Thilo M Fuchs for stimulating and thoughtful discussions.

Thanks also go to past and present members of the office ED104 at Nottingham Trent University, the Microbiology research team and the prep room for an enjoyable working environment.

Special thanks go to my parents for their never-ending support and encouragement throughout all my studies.

This research was funded by a Nottingham Trent University Vice Chancellor scholarship and core funding to the Wellcome Trust Sanger Institute.

Declaration

The conclusions reached in this thesis are my own based on the experiments reported herein. Experiments were performed in the Pathogen Research Group at Nottingham Trent University and in the Pathogen Genomics group at the Wellcome Trust Sanger Institute.

Dr Timothy Hetherington at Nottingham Trent University designed a six-way Venn diagram. Implementation of the phenotypic microarray data in Limma and generation of a heatmap was carried out by Lars Barquist at the Wellcome Trust Sanger Institute. Phylogenetic studies were carried out in collaboration with Drs Simon Harris and Thomas Connor at the Wellcome Trust Sanger Institute. Bayesian Analysis of Population Structure was performed by Dr Thomas Connor.

Abbreviations

ACT	Artemis comparison tool
b,bp	bases, base pairs
BT	biotype
CDS	coding sequence
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
ddNTP	di-deoxynucleoside triphosphate
g	gram
HPI	high-pathogenicity island
IS element	insertion sequence element
kb	kilo bases = 1,000 bases
kDa	kilodalton
l	litre
LB agar/broth	Luria-Bertani medium
LPS	lipopolysaccharide
M	Molar = mol/l
mg	milligram = 10^{-3} g
mM	millimolar = 10^{-3} mol/l
ml	millilitre = 10^{-3} litre
MLST	multi-locus sequence typing
Mb	mega bases = 1,000,000 bases
NTU	Nottingham Trent University
O	outer antigen
PCR	polymerase chain reaction
pl	picolitre = 10^{-12} litre
PM	phenotypic microarray
pMT / pTox / pFra	murine toxin plasmid in *Yersinia pestis*
pPCP / pPst / pPla	plasminogen activator plasmid in *Yersinia pestis*
PTS	phosphotransferase system
pYV / pCD1	*Yersinia* virulence plasmid
PZ	plasticity zone
RE	restriction enzyme
rDNA	ribosomal DNA

RNA	ribonucleic acid
rpm	rounds per minute
SC	species cluster
SNP	single nucleotide polymorphism
SPI	*Salmonella* pathogenicity island
ST	serotype
T2SS	type II secretion system
T3SS	type III secretion system
tc PAI	toxin complex pathogenicity island
V	volts
WTSI	Wellcome Trust Sanger Institute
YAPI	*Yersinia* adhesion pathogenicity island
YGI	*Yersinia* genomic island
YGT	*Yersinia* genus T3SS
Yop	*Yersinia* outer protein
YPM	*Yersinia pseudotuberculosis* mitogen
µl	microlitre = 10^{-6} litre

List of tables

Table 1.1: Selection of biochemical characteristics of members of the genus *Yersinia*.....1
Table 1.2: Biotyping scheme according to Wauters, Kandolo and Janssens (1987) and Bottone (1997). 6
Table 1.3: Comparison of sequencing technologies. 35
Table 2.1: List of reference strains sequenced in this project. 39
Table 2.2: Additional strains. 40
Table 2.3: Function and basis of the algorithms used for gene prediction and characterization. 42
Table 2.4: Housekeeping genes used for construction of the phylogenetic relationship within the genus *Yersinia*. 52
Table 3.1: Contig ordering in the *Y. enterocolitica* reference genomes for BT 1A, 2 - 4. 60
Table 3.2: Contig ordering in the *Y. enterocolitica* reference genome for BT 5. 61
Table 3.3: General properties of the sequenced reference genomes of *Y. enterocolitica* 68
Table 4.1: Regions of interest identified in the pangenome of *Y. enterocolitica* (Figure 4.7). 86
Table 4.2: IS elements in the *Y. enterocolitica* BT 104
Table 4.3: Syntenic blocks identified in the reference genomes of *Y. enterocolitica*. 106
Table 4.4: Distribution of important regions and islands of *Y. enterocolitica* in reference genomes...109
Table 5.1: General properties of the sequenced pYV. 116
Table 6.1: Utilization of N, P, and S sources absent from BT 1A. 149
Table 6.2: Comparison of biotyping and phenotypic microarray reactions. 150
Table 7.1: Numbers of SNPs between different lineages of *Y. enterocolitica*. 174
Table 7.2: Pseudogenes in the core genome of *Y. enterocolitica* lineages. 176
Table 9.1: *Y. enterocolitica* strain collection. 204
Table 9.2: Worldwide strain collection of *Yersinia spp.* 206
Table 9.3: Class key 209
Table 9.4: Colour-coding key of CDSs. 213
Appendix 9.5: perl script for identification of sets of orthologous genes in seven reference genomes for *Y. enterocolitica* 214
Table 9.6: Matrix of orthologous gene sets. 225
Table 9.7: Pseudogenes and their orthologs in the *Y. enterocolitica* reference genomes. 227
Table 9.8: Number of total CDSs, pseudogenes, and partial genes per group for the *Y. enterocolitica* reference genomes. 240
Table 9.9: Signal values for *Y. enterocolitica* biotypes. 242
Table 9.10: Positive signal values and significant changes compared to BT 1B strain 8081 for *Y. enterocolitica* biotypes. 254

List of figures

Figure 1.1: Life cycle of *Y. pestis* and worldwide distribution. 3
Figure 1.2: Evolution of the genus *Yersinia*. 24
Figure 1.3: Workflow of capillary, pyro- and Illumina sequencing. 28
Figure 1.4: Preparation of paired-end fragments for pyrosequencing. 31
Figure 1.5: Paired-end fragment preparation in Illumina sequencing. 33
Figure 1.6: Multiplex preparation of DNA fragments for Illumina sequencing. 34
Figure 2.1: Example of coverage of *Y. enterocolitica* strains sequenced as part of the worldwide strain collection. 41
Figure 2.2: Identification of orthologs and paralogs. 44
Figure 2.3: Example of a 3-way comparison for identification of orthologous gene sets. 45
Figure 2.4: Preparation of fluids for phenotypic microarray. 47
Figure 2.5: Example for utilization of carbon sources in BT 1A and BT 1B. 49
Figure 3.1: Sequencing timeline for the *Y. enterocolitica* reference genomes. 57
Figure 3.2: Initial assembly based on single-end 454 reads and comparison of the six *Y. enterocolitica* reference genomes. 58
Figure 3.3: Pairwise comparison of final assemblies after gap closure of *Y. enterocolitica* reference genomes. 62
Figure 3.4: Flow chart of the preparation of a genome sequence for annotation and analysis. 64
Figure 3.5: Identifying sets of orthologous CDSs. 67
Figure 3.6: Distribution of classes amongst CDSs of all reference genomes. 71
Figure 4.1: Connections between individual reference genomes based on analysis of orthologous genes 79
Figure 4.2: Orthologous genes shared in the low-pathogenic BTs 80
Figure 4.3: Shared genes in non- and low-pathogenic BTs 81
Figure 4.4: Orthologous genes clusters abesent from BT 4 and BT 5 81
Figure 4.5: Orthologous genes shared in the pathogenic BTs 82
Figure 4.6: Combined comparison of orthologous accessory gene sets in the reference genomes for the *Y. enterocolitica* biotypes. 83
Figure 4.7: The pangenome of *Y. enterocolitica*. 85
Figure 4.8: Comparison of SPI-2 of *Salmonella* Typhi with the novel YGT. 87
Figure 4.9: Comparison of YGT within several *Yersinia spp.* 88
Figure 4.10: Arrangement of the plasticity zone in the different biotypes. 90
Figure 4.11: Lipopolysaccharide structure of *Y. enterocolitica* serotypes 91
Figure 4.12: Genomic context of LPS O:8 operon and outer core. 92
Figure 4.13: Genomic context of O:9 and O:3 operons. 93
Figure 4.14: Putative O:5 LPS operon. 95
Figure 4.15: Pseudogene and partial gene distribution in the *Y. enterocolitica* reference genomes 98
Figure 4.16: Top 3 pseudogene categories for *Y. enterocolitica* BTs. 103
Figure 4.17: Position of the predominant IS element IS1667 in the reference genome assemblies 105
Figure 4.18: Exemplar position of six syntenic blocks. 107

Figure 5.1: Alignment and comparison of the sequenced pYV.. 117
Figure 5.2: Backbone of pYV and variable regions in different *Y. enterocolitica* biotypes. 120
Figure 5.3: Protein sequence alignment for LcrV. 122
Figure 5.4: Protein sequence alignment of YscP protein sequences of different *Yersinia* spp. 125
Figure 5.5: Modular nature of YscP in different *Yersinia* spp.. 126
Figure 5.6: Linear plasmid map of pSR2-1A found in *Y. enterocolitica* BT 1A O:5 strain 5303. 128
Figure 5.7: Linear plasmid map of pSR3-1A found in *Y. enterocolitica* BT 1A O:5 strain 5303. 128
Figure 5.8: Plasmid profile of atypical BT 1A/1B isolates. 129
Figure 5.9: Circular plasmid map of pSR1-2 found in *Y. enterocolitica* BT 2 O:9 strain 21202. 131
Figure 5.10: Plasmid profile of BT 2 isolates. 132
Figure 5.11: Comparison between *Salmonella* Dublin pMAK-2 and *Y. enterocolitica* pSR1-2. 133
Figure 5.12: Plasmid map of pSR4-5 as found in *Y. enterocolitica* BT 5 O:2a,2b,3 strain 3094. 134
Figure 5.13: Comparison between *Y. enterocolitica* pYE854 and pSR4-5. 135
Figure 6.1: Respiration is coupled to conversion of tetrazolium violet dye 142
Figure 6.2: Workflow of phenotypic microarray. 143
Figure 6.3: Signal values (A) and significant changes towards BT 1B (B) depicted as heatmaps. 146
Figure 6.4: Number of sources that are expressed significantly higher than in BT 1B strain 8081. 147
Figure 6.5: Number of sources that are expressed significantly lower than in BT 1B 8081. 147
Figure 6.6: Kinetic plots for raffinose in *Y. enterocolitica* biotypes 151
Figure 6.7: Operon for L-arabinose utilization contrasted with signal values and significant changes compared to BT 1B. 154
Figure 6.8: L-arabinose utilization kinetic plots over time for *Y. enterocolitica* biotypes 155
Figure 6.9: Mutations specific for biotyping of BT 5 contrasted with signal values and significant changes compared to BT 1B. 156
Figure 6.10: Operon structure for utilization of N-acetyl-D-galactosamine in representative BTs contrasted with signal values and significant changes compared to BT 1B. 157
Figure 6.11: Unique operon for D-arabitol utilization in BT 1A, contrasted with signal values and significant changes compared to BT 1B. 159
Figure 6.12: Absorbance curves over time for m-inositol utilization 160
Figure 6.13: Point mutation in the *idh* gene is the basis for BT 1A negative m-inositol phenotype. 161
Figure 6.14: Signal values depicted as heatmap. 163
Figure 6.15: Comparison of biotyping reactions and phenotypic microarray. 164
Figure 7.1: The maximum likelihood phylogenetic tree of *Y. enterocolitica* based on the SNP differences across the whole genome excluding laterally acquired phages. 173
Figure 7.2: Distribution of the 85 housekeeping genes used for construction of the phylogenetic relationship within the genus *Yersinia*. 181
Figure 7.3: The phylogeny of the genus *Yersinia* and the virulence plasmid pYV. 183
Figure 7.4: The phylogeny of the genus *Yersinia* based on *recA*, *glnA*, *gyrB*, and *hsp60*. 185
Figure 7.5: The phylogeny of the genus *Yersinia* based on 16S rDNA sequences. 187
Figure 7.6: Phylogeny of the plasmid pYV compared to the phylogeny of the chromosome in *Y. enterocolitica*.. 189
Figure 8.1: Revised model of the evolution of pathogenicity in the genus *Yersinia*. 200

1 Introduction

1.1 The Genus *Yersinia*

The heterogeneous genus *Yersinia* comprises three human-pathogenic species, *Y. pestis*, *Y. pseudotuberculosis*, and *Y. enterocolitica*, the fish-pathogen *Y. ruckeri*, and a number of environmental, seemingly non-pathogenic species (Carniel and Mollaret, 1990; Bottone, 1997; Bottone, 1999; Sulakvelidze, 2000; Wren, 2003). The *Yersiniae* belong to the family of *Enterobacteriaceae*, and are rod-shaped, facultative anaerobic bacteria (Carniel and Mollaret, 1990; Bottone, 1999; Wren, 2003).
Historically, speciation in *Yersinia* is based on biochemical differences (Sulakvelidze, 2000). A selection of biochemical tests for species designation is listed in Table 1.1.

Table 1.1: Selection of biochemical characteristics of members of the genus *Yersinia* [adapted from Bottone, 1997; Sulakvelidze, 2000; Hurst, et al., 2011; Murros-Kontiainen, et al., 2010a,b].

	Rhamnose	Sorbose	Cellobiose	Raffinose	Melibiose	Sucrose	Indole reaction	Voges-Proskauer test	Inositol	D-xylose	Urease
Y. aldovae	+	-	-	-	-	-	-	+	+	+	+
Y. aleksiciae	-	+	+	-	-	-	+	-	-	+	+
Y. bercovieri	-	-	+	-	-	+	-	-	-	+	+
Y. enterocolitica	-	v	+	-	-	+	v	-	+	v	+
Y. entomophaga	-	-	+	+	+	+	-	?	-	-	+
Y. frederiksenii	+	+	+	-	-	+	+	v	+	+	+
Y. intermedia	+	+	+	+	+	+	+	+	-	+	+
Y. kristensenii	-	+	+	-	-	-	v	-	-	+	+
Y. massiliensis	-	+	+	v	-	+	+	-	+	+	+
Y. mollaretii	-	+	+	-	-	+	-	-	+	+	+
Y. nurmii	-	-	(+)	-	-	+	-	+	-	-	-
Y. pekkanenii	-	-	+	-	-	-	-	-	-	+	+
Y. pestis	v	-	-	-	v	-	-	-	-	-	-
Y. pseudotuberculosis	+	-	-	-	+	-	-	-	-	+	+
Y. rohdei	-	+	+	+	v	+	-	-	-	+	+
Y. ruckeri	-	-	-	+	-	-	-	-	-	-	-
Y. similis	+	-	-	-	-	-	-	-	-	+	+

+ – positive; (+) – delayed positive; v – variable; - – negative; ? – information unavailable

There is no one typing scheme used for *Yersinia*, and the most commonly used key tests are presented (Bottone, 1997; Stock, Henrichfreise and Wiedemann, 2002). The description of further species generally relies on new tests being added, resulting in data that is then not available for all species. Differences in DNA-DNA hybridization values are used to furthermore classify the environmental *Yersiniae* (Bottone, 1997). Most recently differences in the 16S rDNA sequence have been employed alongside biochemical signatures and DNA-relatedness to define new species (Sprague and Neubauer, 2005; Hurst, et al., 2011).

1.1.1 *Yersinia pestis* and *Yersinia pseudotuberculosis*

Y. pestis is the causative agent of plague, the notorious Black Death of the middle ages, responsible for the death of millions of Europeans (Brubaker, 1991; Perry and Fetherston, 1997; Wren, 2003). The life cycle of *Y. pestis* is depicted in Figure 1.1A. *Y. pestis* is a host-dependent organism, and has limited abilities to survive outside either the flea vector or a warm-blooded host (Perry and Fetherston, 1997; Parkhill, et al., 2001a; Wren, 2003). In the flea vector, the bacteria block the foregut. The flea cannot feed on a new host, but in trying will regurgitate the bacteria directly under the skin and into the blood stream of the mammalian host (Wren, 2003). This mammalian host can be wild living rodents, such as rabbits or gerbils (Carniel and Mollaret, 1990). This is called the sylvatic cycle, and the ability to hibernate and circulate in wild rodent populations precludes eradication. *Y. pestis* may cause disease in wild rodents as well, and its persistence might be explained either by hibernation in the ground of abandoned rodent burrows, which may subsequently become re-inhabited, or reduced virulence in rodents due to mutations (Carniel and Mollaret, 1990).

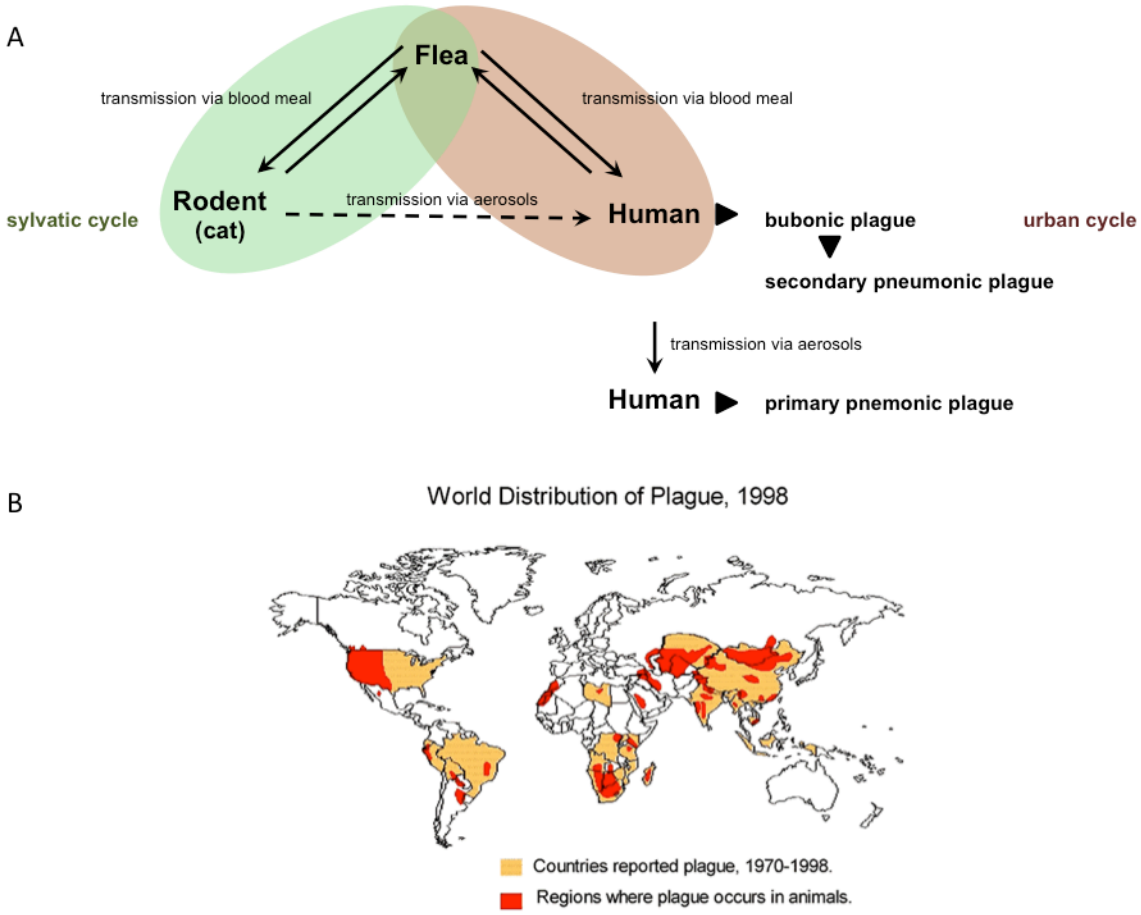


Figure 1.1: Life cycle of *Y. pestis* and worldwide distribution. A. Sylvatic and urban transmission routes of *Y. pestis*. B. Current endemic foci of plague [according to CDC, 2009].

In the urban cycle, humans are infected as accidental hosts of the flea. Once inside the human body, the bacteria will spread to the regional lymph nodes, either by surviving inside macrophages or through tissue infection (Wren, 2003). Once there, *Y. pestis* multiplies and gives rise to the classic symptoms of inflamed lymph nodes of primary bubonic plague (Carniel and Mollaret, 1990; Perry and Fetherston, 1997). The incubation time of infection is 2 to 6 days after the bite (CDC, 2005). Once the bacteria have entered the lymph nodes of the human host, they then progress to the lungs, liver and spleen (Carniel and Mollaret, 1990; Wren, 2003). The bacteria enter the blood stream causing septicaemia (Carniel and Mollaret, 1990). The flea feeding on an infected, bacteraemic individual closes the cycle from mammalian host back to the flea. A transmission from cats to humans via aerosols has been reported, but is very rare (CDC, 2009). In crossing sylvatic and urban cycles, rats play an important role as hosts to infected fleas and their proximity to towns and settlements (Carniel and Mollaret, 1990; Perry and Fetherston, 1997).

The terminal phase of *Y. pestis* is associated with lodging and multiplication of the bacteria in the lung (Carniel and Mollaret, 1990). Upon reaching this site of infection, secondary pneumonic plague ensues with a fatality rate of over 50% (Wren, 2003; CDC, 2005). Additionally at this stage, an arthropod host is redundant as transmission via aerosols to other humans is possible and highly effective (Carniel and Mollaret, 1990; Wren, 2003). Following this infection route, the bacteria can rapidly spread to the bloodstream, circumventing the need for prior establishment in lymph nodes, and death occurs within a few hours (Carniel and Mollaret, 1990).

Today, *Y. pestis* is still present in sylvatic foci found in North and South America, sub-Saharan Africa, and Asia with a sufficient rodent population (Figure 1.1B; CDC, 2005). It is still a relevant human disease with 1,000 to 2,000 cases a year worldwide (CDC, 2005), and multidrug resistant strains have been identified (Wren, 2003). It also poses a bioterrorism threat due to extremely efficient person-to-person spread, and delayed onset of symptoms enabling a geographic spread (Carniel and Mollaret, 1990; CDC, 2005). The infective dose is also very low, at 10 cells or less (Brubaker, 1991).

Three classic biovars of *Y. pestis* have been associated with the three main pandemics (Achtman, et al., 1999; Wren, 2003). Bacteria that gave rise to the Justinian plague are related to Biovar Antiqua, and those that gave rise to the Black Death in the Middle Age relate to biovar Mediaevalis. Modern plague is associated with biovar Orientalis (Achtman, et al., 1999). These are distinguished on the basis of glycerol fermentation and nitrate reduction.

Y. pseudotuberculosis is closely related to *Y. pestis* (Chapter 1.3), but has very different ecology. It is found widely in animals and the environment. Humans are infected via the fecal-oral route following the ingestion of contaminated food sources (Carniel and Mollaret, 1990; Carniel, 2003; Wren, 2003). In animals, the bacteria invade the small intestine and spread from lymphatic vessels to distal lymph nodes and subsequently to liver and spleen (Carniel and Mollaret, 1990). At these sites, abscesses form, and the bacteria can enter the bloodstream with fatal outcome (Carniel and Mollaret, 1990). In human infections, the bacteria enter the small intestine as well, but disease is usually milder (Carniel and Mollaret, 1990). *Y. pseudotuberculosis* targets microfold cells (M cells) for invasion of the gut epithelium (Carniel and Mollaret, 1990; Wren, 2003). It then disseminates widely, and causes mesenteric lymphadenitis (Wren, 2003). Mild diarrhoea can be observed, but

abdominal pain is the most frequent and constant symptom (Carniel and Mollaret, 1990). No further colonization of deeper tissues or secondary organs ensues though, and usually the disease is self-limiting (Carniel and Mollaret, 1990; Carniel, 2003). Seasonality of the disease is suspected, as the susceptible animal hosts appear to shift from being healthy carriers to infected hosts after exposure to stresses such as cold, humid weather and starvation periods (Carniel and Mollaret, 1990).

In countries of the Far East, infection with *Y. pseudotuberculosis* occurs with additional systemic symptoms such as skin rashes and toxic shock syndrome, which is also known as Far East scarlet-like fever (Fukushima, et al., 2001; Eppinger, et al., 2007). It has been proposed that the presentation of *Y. pseudotuberculosis* as Far East scarlet-like fever arises predominantly in isolates that carry the superantigen *Yersinia pseudotuberculosis* mitogen (YPM), which is absent in strains causing enteric disease (Fukushima, et al., 2001; Eppinger, et al., 2007).

In *Y. pseudotuberculosis*, 21 different serogroups are known which are combined in six serotypes (Carniel and Mollaret, 1990). There is no correlation between serotype and disease severity, in contrast to *Y. enterocolitica* (Carniel and Mollaret, 1990). *Y. pseudotuberculosis* infections are less common than *Y. enterocolitica* infections (Carniel and Mollaret, 1990).

1.1.2 *Yersinia enterocolitica*

Similar to *Y. pseudotuberculosis*, *Y. enterocolitica* causes enteric disease. *Y. enterocolitica* infections are most commonly due to ingestion of contaminated food products although cases of infections through contaminated blood transfusion have been reported (Bottone, 1997). Survival and proliferation at refrigeration temperatures make it a significant food-borne pathogen (Bottone, 1997; Bottone, 1999). Following ingestion, the bacteria colonize the gastrointestinal tract, specifically the terminal ileum and proximal colon, by attaching and invading M cells in the Peyer's patches (Carniel and Mollaret, 1990; Bottone, 1997; Bottone, 1999). They then multiply and spread horizontally through the mucosal epithelium (Bottone, 1997). This manifests in inflammatory symptoms such as enteritis, enterocolitis, acute mesenteric lymphadenitis, and terminal ileitis, and is also associated with fever and diarrhoea, which can occasionally be bloody (Carniel and Mollaret, 1990; Bottone, 1997; Bottone, 1999). It is thought that serotype O:8 strains present with more aggressive disease than O:3 infections (Bottone, 1999). Often, the infection is

self-limiting, as these symptoms subside on their own account after 2 weeks in adults and longer periods in infants (Pepe and Miller, 1993; Bottone, 1997; Carniel, 2003). The bacteria are able to invade deeper tissue layers as well, and may enter the bloodstream with subsequent spread to liver, spleen, and mesenteric lymph nodes (Bottone, 1997; Bottone, 1999). Arthritis and erythema nodosum, an inflammation under the skin, have been described as the most common sequelae of *Y. enterocolitica* infections (Bottone, 1997; Bottone, 1999).

The species *Y. enterocolitica* is very heterogeneous with respect to both biochemical differences and causing disease. Some strains are able to cause severe enteritis in humans, other strains are responsible for milder symptoms, whilst a third group of strains is not associated with any disease in humans (Carniel and Mollaret, 1990; Bottone, 1999; Sulakvelidze, 2000; Virdi and Sachdeva, 2005). Therefore, the species is differentiated into six different biotypes (BT) based on biochemical differences (Table 1.2; Wauters, Kandolo, and Janssen, 1987).

Table 1.2: Biotyping scheme according to Wauters, Kandolo and Janssens (1987) and Bottone (1997).

	Biotype	**BT 1A**	**BT 1B**	**BT 2**	**BT 3**	**BT 4**	**BT 5**
	Pathogenicity potential	**No**	**High**	**Low**	**Low**	**Low**	**Low**
1	Salicin AP	+	-	-	-	-	-
2	Pyrazinamidase activity	+	-	-	-	-	-
3	Esculin hydrolysis	+/-	-	-	-	-	-
4	Lipase activity	+	+	-	-	-	-
5	Indole production	+	+	v	-	-	-
6	Xylose AP	+	+	+	+	-	v
7	Trehalose AP	+	+	+	+	+	-
8	Sorbose AP	+	+	+	+	+	-
9	Nitrate reduction	+	+	+	+	+	-
10	Ornithine decarboxylase	+	+	+	+	+	+ (+)
11	Voges-Proskauer test	+	+	+	+	+	+ (+)
12	Inositol AP	+	+	+	+	+	+

AP – acid production, + – positive, - – negative, v – variable, (+) – delayed positive.

There are several reactions, which are unique to BT 1A, whereas there is only one reaction to distinguish BT 4 from BT 3, and one variable reaction to separate BT 2 and BT 3. Generally, biotyping identification is based on the absence of utilization of a nutrient source compared to BT 1A rather than unique properties specific for each biotype. This results in a “step-wise” appearance of positive reactions.

In line with its heterogeneity there are also approximately 60 serogroups in *Y. enterocolitica*, but only certain bio-serotypes are frequently associated with human infections, which are BT 1B O:8, and O:21, BT 2 O:9, and O:5,27, BT 3 O:1,2,3, O:9, and O:5,27, BT 4 O:3, and BT 5 O:2,3 (Carniel and Mollaret, 1990; Bottone, 1997; Bottone, 1999; Sulakvelidze, 2000; Virdi and Sachdeva, 2005). In early studies, these bio-serotypes have been associated with invasion of epithelial cells, autoagglutination and conjunctivitis in guinea pigs (Pedersen, Winblad and Bitsch, 1979; Schiemann and Devenish, 1982). BT 5 potentially presents an exception and does not invade epithelial cells (Pedersen, Winblad and Bitsch, 1979). It was also noted specifically that BT 1B O:8 and O:21 serotypes generate more severe disease in the guinea pig model (Schiemann and Devenish, 1982).

In a mouse infection model, the biotypes also exhibit differences in disease severity. BT 1B is considered the high-pathogenic BT, as it kills mice in infection models, and BTs 2 – 5 are the low-pathogenic BTs that cause infection in mice but are not lethal. Because it does not cause disease in mice, BT 1A is designated the non-pathogenic BT (Wren, 2003). Recently, two subspecies have been proposed (Neubauer, et al., 2000). Based on differences in the 16S rDNA sequence, it has been suggested to refer to the high-pathogenic BT 1B as *Y. enterocolitica* subsp. *enterocolitica*, and the non- and low-pathogenic BTs 1A, 2 – 5 as *Y. enterocolitica* subsp. *palearctica* (Neubauer, Hensel, Aleksic and Meyer, 2000; Neubauer, et al., 2000; Neubauer and Sprague, 2003).

Y. enterocolitica is a ubiquitous organism, found in the environment as well as the gastrointestinal tract of animals (Bottone, 1999). Pigs are a common reservoir for nearly all *Y. enterocolitica*, and this is considered a major route of infections for humans although a clear link has not yet been discovered (Carniel and Mollaret, 1990; Bottone, 1999; Neubauer and Sprague, 2003). The non-pathogenic BT 1A is also found in a range of environmental sources such as animals, food, and water, and is also isolated from faeces of healthy and sick humans (Bottone, 1999; Tennant, Grant and Robins-Browne, 2003; Tennant, Skinner, Joe and Robins-Browne, 2005). BT 5 is a rarely isolated biotype, and its main ecologic distribution is in hare and goats. It is therefore commonly referred to as the “hare-biotype”, and isolation from humans has never been reported (Swaminathan, Harmon and Mehlman, 1982; Carniel and Mollaret, 1990; Wuthe and Aleksic, 1997; Bartling, et al.,

2004). In hares, infection can cause acute septicaemic disease leading to epidemic outbreaks (Bartling, et al., 2004).

1.1.3 The special case of *Yersinia ruckeri*

Amongst the *Yersiniae*, *Y. ruckeri* is a special case. Compared to other members of the genus, it is the most distantly related species, and it has been questioned whether it is indeed a member of the genus (Bottone, 1997; Sulakvelidze, 2000). Studies on DNA relatedness indicated distant relationships to both *Serratia* and *Yersinia*, and a designation into the genus *Yersinia* was based on similarity in biochemical reactions as well as the GC content (Sulakvelidze, 2000; Fernández, Méndez and Guijarro, 2007).

Y. ruckeri causes redmouth disease in salmonid fish, especially in rainbow trout, which presents as acute or chronic septicaemia with internal and external haemorrhages (Tobback, et al., 2009). *Y. ruckeri* has also been implicated in fatal septicaemia in carp (Sulakvelidze, 2000; Fernández, Méndez and Guijarro, 2007). As such, it is considered an important economic factor in the fish industry (Tobback, et al., 2009; Ryckaert, et al., 2010). An effective vaccine is available, yet outbreaks under stress conditions can spread quickly from faeces of ill and carrier fish through water (Fernández, Méndez and Guijarro, 2007). Recently, new strains have been reported, which are unaffected by vaccination, possibly due to changes in surface antigens (Fernández, Méndez and Guijarro, 2007).

Little is known about the molecular mechanisms of pathogenesis. Both abilities to form biofilms and to survive for prolonged time in aquatic, nutrient-restricted environments are thought to play a role, as well as an iron-acquisition system (Fernández, Méndez and Guijarro, 2007). A large plamid that is in common to American and French isolates has also been reported (Toranzo, Barja, Colwell and Hetrick, 1983; Guilvout, et al., 1988). This plasmid could be correlated with pathogenicity as the French isolates were from a redmouth disease outbreak, and because it is found in the most pathogenic serovars of *Y. ruckeri*, but a link with disease has not been established (Lesel, Lesel, Gavini and Vuillaume, 1983; Garcia, Dominguez, Larsen and Pedersen, 1988). Expression of virulence genes is potentially regulated by the amount of available iron and temperature changes, and a decrease in temperature from 28°C to 18°C might be crucial (Fernández, Méndez and Guijarro, 2007). It was shown that gills could be an important entry point for *Y. ruckeri* into the host (Tobback, et al., 2009). The bacteria attach to the gills, which carry a large number of blood vessels for respiration. This opens up a direct route for

dissemination through the host via the blood stream (Tobback, et al., 2009). *Y. ruckeri* can also persist in macrophages, similar to the human pathogenic *Yersinia*, as means of protected environment and immune evasion (Ryckaert, et al., 2010).

Y. ruckeri is a very homogeneous, clonal species, in contrast to the human pathogens *Y. enterocolitica* and *Y. pseudotuberculosis*, which are both considered heterogeneous (Sulakvelidze, 2000). It can be differentiated into different bio-serotypes, with the most virulent strains found in biotype 1 O:1 (Fernández, Méndez and Guijarro, 2007).

1.1.4 Environmental *Yersiniae*

Environmental species of the genus *Yersinia* are found in various environments, such as water and aquatic ecosystems, sewage, wild and domestic animals as well as foods (Sulakvelidze, 2000). They have not been associated with human infections, although some species have been isolated from sick and healthy individuals (Sulakvelidze, 2000). The environmental *Yersiniae* are also collectively referred to as '*Y. enterocolitica*-like' species, as they were speciated from *Y. enterocolitica* (Sulakvelidze, 2000). They do present individual species although the extent of their distinction may be underestimated (Sulakvelidze, 2000). 14 species have been described. Of these, eight species could be considered classical, established lineages speciated from *Y. enterocolitica* between 1978 and 1995 (Sulakvelidze, 2000). The other six species have been portrayed from 2005 onwards and are therefore less well characterized (Sprague and Neubauer, 2005; Murros-Kontiainen, et al., 2010a,b; Hurst, et al., 2011; Merhej, et al., 2008; Sprague, Scholz, Amann and Busse, 2008).

The eight species classically recognized as *Y. enterocolitica*-like species are *Y. aldovae*, *Y. bercovieri, Y. frederiksenii*, *Y. intermedia*, *Y. kristensenii*, *Y. mollaretii*, *Y. rohdei*, and *Y. ruckeri* (Chapter 1.1.3). In contrast to other environmental species, little is known about *Y. aldovae*.

Y. frederiksenii is a heterogeneous species consisting of four genomospecies. These show differences in DNA-hybridization, 16S rDNA sequence comparison and when comparing the electrophoretic properties of specific enzymes (Sulakvelidze, 2000; Virdi and Sachdeva, 2005). As they do not differ on the biochemical level, these genomospecies can currently not be distinguished (Sulakvelidze, 2000).

The naming of *Y. intermedia* arises from the fact that it shows common biochemical properties with both *Y. enterocolitica* and *Y. pseudotuberculosis*, and therefore was

thought to be ‘midway’ between the two (Sulakvelidze, 2000). Several different biotypes have been described based on differential utilization of Simmons citrate and fermentation of different sugars (Martin, Leclercq, Savin, and Carniel, 2009).

Y. kristensenii was originally designated for trehalose-positive, sucrose-negative strains. This comprised a heterogeneous collection, with some isolates showing a closer genotypic similarity to *Y. mollaretii* (Sulakvelidze, 2000). These were later speciated into *Y. aleksiciae* (Sprague and Neubauer, 2005).

Y. mollaretii and *Y. bercovieri* were originally grouped as biotype 3A and 3B of *Y. enterocolitica*. They were then separated from that species based on differences in ecology, DNA hybridization, and antigen patterns (Sulakvelidze, 2000). *Y. bercovieri* is known to be genetically heterogeneous based on experiments distinguishing several ribotyping and pulse-field gel electrophoresis patterns (Sulakvelidze, 2000).

Y. rohdei is a homogeneous group of organisms thought to be most closely related to *Y. frederiksenii* (Sulakvelidze, 2000).

Recently described species are *Y. aleksiciae*, *Y. entomophaga*, *Y. massiliensis*, *Y. nurmii*, *Y. pekkanenii*, and *Y. similis*. Similarly to the classical environmental species, these have been isolated from various sources including water, symptomatic and asymptomatic humans, food products and animals such as reindeer and pigs.

Y. aleksiciae was speciated as atypical isolates from *Y. kristensenii* (Sprague and Neubauer, 2005). They differ from *Y. kristensenii* in 16S rDNA sequence and lysine-decarboxylase phenotype.

Both *Y. nurmii* (Murros-Kontiainen, et al., 2010a) and *Y. entomophaga* (Hurst, et al., 2011) are newly described species in which the closest relative appears to be *Y. ruckeri*. This is interesting as *Y. ruckeri* is thought to be the most distant member of the genus (Bottone, 1997; Sulakvelidze, 2000). *Y. entomophaga* was isolated from a diseased insect larva (Hurst, et al., 2011), and *Y. nurmii* was isolated from meat (Murros-Kontiainen, et al., 2010a). It is unknown how these two species relate to each other. Out of 19 biochemical tests with results available for both species, they differ in raffinose and melibiose utilization. Cellobiose and lactose give weak positive reactions for *Y. nurmii*, whereas *Y. entomophaga* is clearly positive (Hurst, et al., 2011; Murros-Kontiainen, et al., 2010a).

Y. massiliensis was initially isolated from water (Merhej, et al., 2008), but can be found in food sources as well (Souza, Falcão, and Falcão, 2011). The species is

thought to be closely related to *Y. bercovieri*, *Y. mollaretii*, and *Y. frederiksenii* and can be distinguished from these in indole, inositol, rhamnose, and citrate utilization patterns (Merhej, et al., 2008; Souza, Falcão, and Falcão, 2011).

Based on commercially available biochemical tests, *Y. pekkanenii* cannot be distinguished from *Y. pseudotuberculosis* (Murros-Kontiainen, et al., 2010b). It was shown though that classical virulence determinants are absent, and that these environmental isolates form a unique cluster related to *Y. mollaretii* and *Y. aldovae* (Murros-Kontiainen, et al., 2010b).

Y. similis isolates were also originally attributed to *Y. pseudotuberculosis* but phenotypic characterization as well as 16S rDNA analysis and DNA-DNA hybridization suggested a separate species (Sprague, Scholz, Amann and Busse, 2008).

1.2 Pathogenicity in the genus *Yersinia*

1.2.1 Relationship of pathogenic *Yersiniae*

The pathogenic *Yersiniae* have a number of characteristics that are thought to be involved in pathogenicity and virulence. Despite the difference in the diseases caused by *Y. pestis*, *Y. pseudotuberculosis*, and *Y. enterocolitica* some of these genes and islands are shared in all pathogenic *Yersiniae* (Revell and Miller 2001; Carniel, 2002). There is a common theme in the infections caused by pathogenic *Yersiniae*, in that iron overload of the host can increase pathogenicity, and there is marked tropism in all three species for lymphatic tissues: *Y. pestis* targets the lymph nodes, *Y. pseudotuberculosis* the lymphatic chain, and *Y. enterocolitica* infects the Peyer's patches and lymphoid follicles (Carniel and Mollaret, 1990; Carniel, 2003).

It is thought that *Y. pestis* expresses antiphagocytic, invasive, and toxic activities to cause an acute infection and effective transmission because it is less adapted to environmental survival than the enteropathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*, which favour a chronic infection by production of adhesins and invasins (Brubaker, 1991).

A considerable number of genes and regions have been implicated with a role in pathogenesis in *Yersinia*, yet not all of them present essential genes in the infection process. It is also becoming clear that some pathogenicity markers or islands that were thought to provide survival advantages under host infection conditions are in fact also present in environmental isolates and non-pathogenic BT 1A strains (Grant,

Bennet-Wood and Robins-Browne, 1998; Chen, et al., 2010). Some genes identified on the basis of essentiality in infection are housekeeping genes rather than specific virulence factors, therefore only genes involved in specific steps of the infection process are reviewed here.

1.2.2 Plasmid-encoded basis for pathogenicity

1.2.2.1 Shared virulence plasmid pYV

The *Yersinia* virulence plasmid, pYV, is common to all pathogenic *Yersinia*, and varies in size between 40 to 48 kDa (Portnoy, Moseley and Falkow, 1981), equivalent to 60 to 70 kb. Hybridization experiments have indicated gene content variations suggesting the pYV plasmids are a "closely related family" of plasmids (Portnoy, Moseley and Falkow, 1981). In *Y. enterocolitica*, pYV is only present in high- and low-pathogenic strains, never in BT 1A (Carniel, 2002; Wren, 2003). It is also found in *Y. pseudotuberculosis* and *Y. pestis*; in the latter, it is called pCD1 for calcium dependency as expression of proteins encoded for on the plasmid is triggered under low-calcium conditions (Carniel, 2002; Chain, et al., 2004).

Portnoy, Moseley and Falkow (1981) were among the first to link pathogenicity to the presence of pYV and showed that strains cured of the plasmid had reduced lethality in gerbils. The lethal dose of plasmid-cured strains was higher than 10^6 organisms compared to plasmid-bearing strains with a lethal dose of 10^3 organisms. Using the Sereny test it has been shown that infection of guinea pig conjunctiva is dependent on the presence of pYV, as cured strains did not cause an infection (Heesemann, et al., 1983). The lack of pYV in BT 1A is the main reason why it is considered to be non-pathogenic (Carniel, 2002).

The pYV encodes a type III secretion system (T3SS) secreting Yersinia outer proteins (Yops) at 37°C in low calcium conditions (Cornelis, 2002; Carniel, 2002). The functionality of this replicon is conserved in all pYVs, although the overall organization may vary (Carniel, 2002). This system as been extensively studied as a model system for type 3 secretion in the past (Portnoy and Falkow, 1981; Portnoy, et al., 1984; Cornelis, et al., 1987; Cornelis, et al., 1998).

The concerted action of Yops leads to prevention of phagocytosis and resistance to the immune response through three steps of adhesion to host cells, injection of

effectors using the T3SS, and modulation of target cell function (Cornelis, 2002; Carniel, 2002). The chromosomally encoded invasin Inv (1.2.3.1) and plasmid encoded YadA both mediate adhesion (Carniel, 2002). YadA is an outer membrane adhesin only expressed in enteropathogenic *Yersiniae* due to a frameshift mutation in *Y. pestis* (Brubaker, 1991; Cornelis, 1994). It is known to be involved in autoagglutination and erythrocyte agglutination as well as collagen and fibronectin binding (Brubaker, 1991; Cornelis, 1994). Contact with the host cell activates expression of structural and effector proteins of the T3SS, which is known as contact-dependent secretion (Carniel, 2002; Cornelis, 2002). The apparatus genes are encoded by the *ysc* (*Yersinia* secretion) injectisome, whose 27 proteins make up the channel structure and protruding needle (Cornelis, 2002). The Yop effector proteins (YopEHMPT and YpkA) are then transported from the bacterial cell through the needle and injected straight into the target host cell where they have an effect on cellular pathways (Carniel, 2002). YopEHT and YpkA act on the cytoskeleton of the host cell and thus prevent phagocytosis (Cornelis, 1994; Carniel, 2002; Cornelis, 2002). YopP and YopH down-regulate the immune response by inhibition of transcription activators resulting in decreased secretion of signal molecules (Carniel, 2002; Cornelis, 2002). This also promotes intracellular survival of *Yersinia* inside macrophages (Cornelis, 2002). The exact nature of action for YopM is not known, but putatively involves influence on gene transcription (Cornelis, 2002).

1.2.2.2 Plasmids specific to *Y. pestis*: pPCP and pMT

Apart from the virulence plasmid pYV, *Y. pestis* has two additional unique plasmids that are implicated in pathogenicity; pPCP and pMT (Brubaker, 1991; Wren, 2003).

The plasmid pPCP has a size of ~10 kb and can be referred to as pPla for plasminogen activator, pPst for pesticin, or pPCP1 for pesticin-coagulase-plasminogen activator (Brubaker, 1991; Perry and Fetherston, 1997; Wren, 2003). Pesticin has a limited range of activity, and is restricted to acting on atypical *Escherichia coli*, *Y. pseudotuberculosis* serogroup 1A and 1B, and *Y. enterocolitica* serogroup O:8 (Brubaker, 1991). An essential role in disease has not been established and pesticin might be more important in competing with other bacteria in the same niche (Brubaker, 1991; Carniel, 2002). The plasminogen activator has fibrinolytic activity, and can coagulate rabbit plasma (Brubaker, 1991). This protein is involved in dissemination of *Y. pestis* in the infected host (Brubaker, 1991; Achtman, et al., 1999; Carniel, 2002).

The pMT plasmid encodes the murine toxin Ymt and has a size of 96 kb. It is also known as pTox or pFra, because it encodes the fraction 1 capsule-like antigen (Brubaker, 1991; Carniel, 2002; Wren, 2003).

The fraction 1 capsule protein is an outer membrane protein that has been postulated to be involved in resistance to macrophage uptake during the initial stage of infection (Brubaker, 1991). As it presents an outwardly displayed antigen, it might be important in the interaction between the bacterium and host cells, especially of the immune response, and has been the target for vaccine development against plague (Carniel, 2002; Eppinger, et al., 2010). Capsule-deficient *Y. pestis* biovar Angola was shown to lack lethality in the guinea pig model after subcutaneous infection, and was also less virulent in the subcutaneous mouse model but not in the aerosol model (Eppinger, et al., 2010). The exotoxin Ymt is responsible for the death of infected mice possibly due to blocking host cell receptors in the terminal, septicaemic stage of disease resulting in hypotension and vascular collapse (Brubaker, 1991; Hinnebusch, et al., 2002). Ymt might be more essential for colonization of the flea (Parkhill, et al., 2001a; Hinnebusch, et al., 2002). It was shown that Ymt acts as an intracellular phospholipase that is involved in neutralizing agents in the flea midgut, which might otherwise lead to cell lysis (Hinnebusch, et al., 2002; Carniel, 2002). Through being able to persist and proliferate in the flea midgut, *Y. pestis* can then spread and block the proventriculus for effective transmission (Hinnebusch, et al., 2002). Without the protection through Ymt, the bacteria cannot establish colonization of the flea and are cleared (Hinnebusch, et al., 2002).

A chimeric plasmid consisting of one pMT copy and two tandemly repeated copies of pPCP called pMT-PCP has been reported in *Y. pestis* strain Angola (Eppinger, et al., 2010). It is thought that the presence of an insertion sequence (IS) element enabled the co-integration of the plasmids via recombination (Eppinger, et al., 2010). The effect of this is unclear, but there are potentially dosage-related effects, as the chimeric plasmid is present in higher copy numbers than other plasmids (Eppinger, et al., 2010). There are three copies of the genes encoding the plasminogen activator. Conversely, the genes encoding the capsular antigen are absent on the chimeric plasmid, and it has been shown that this affects virulence following subcutaneous infection in the mouse model but not the aerosol route (Eppinger, et al., 2010). A chimeric plasmid might be an advantage in vertical plasmid progression and might ensure maintenance of all plasmid-encoded virulence factors (Eppinger, et al., 2010).

1.2.2.3 Plasmids specific to *Y. pseudotuberculosis*: pVM82 and pIB

Both pVM82 and pIB have been isolated from strains of *Y. pseudotuberculosis* that have been implicated in Far East scarlet-like fever (Carniel, 2002; Eppinger, et al., 2007). A region on pVM82 has been linked to immunosuppression and ability to resist phagocytosis (Eppinger, et al., 2007). Both plasmids carry genes encoding a pilus similar to the pilin gene cluster found on the Yersinia adhesion pathogenicity island (YAPI) although of different evolutionary origin (Eppinger, et al., 2007). This three-fold presence of pilin gene clusters might explain the host immune response associated with Far East scarlet-like fever (Eppinger, et al., 2007).

Further plasmid-encoded genes that might potentially be involved in pathogenicity include a type 4 secretion system and a putative phytotoxin on pVM82, and a plasmid-transfer system on pIB (Eppinger, et al., 2007).

1.2.3 Chromosomal markers of pathogenicity

Whilst the virulence plasmid shared in the pathogenic *Yersiniae* undoubtedly plays a major role in infection, studies have shown that its loss in *Y. enterocolitica* does not completely abolish virulence, indicating the presence of essential chromosomally encoded virulence factors (Revell and Miller, 2001).

1.2.3.1 Invasion and adhesion proteins: Inv and Ail

Both Inv and Ail have been described as important proteins for promoting the invasion of host cells (Miller and Falkow, 1988; Miller, Farmer, Hill and Falkow, 1989). The invasin protein Inv can confer an invasive phenotype on *E. coli* cells if cloned from the high-pathogenic *Y. enterocolitica* BT 1B (Miller and Falkow, 1988). Invasin mutants were also shown to be defective in colonization and dissemination in a mouse infection model (Pepe and Miller, 1993; Revell and Miller, 2001). The attachment and invasion locus *ail* was found to change invasive behaviour, but it appeared to be dependent on the cell type used suggesting a different mode of action of the two invasins (Miller and Falkow, 1988). Ail was shown to be involved not only in invasion, but also in adhesion to cells and in serum resistance (Pierson and Falkow, 1993). It is thought that Inv and Ail act in concert, with Inv important for priming of bacteria at lower temperatures, and Ail being responsible for acting once the infection process has started inside the host (Pepe and Miller, 1993).

Homologs of the invasin gene *inv* have been described in all *Yersinia* sp., but functionality has only been demonstrated in enteropathogenic *Yersiniae* (Miller and Falkow, 1988; Miller, Farmer, Hill and Falkow, 1989; Pierson and Falkow, 1990; Grant, Bennet-Wood and Robins-Browne, 1998). It is thought to be cryptic in *Y. pestis* due to disruption (Brubaker, 1991; Revell and Miller, 2001). Functionality in BT 1A could not be established, as it is neither expressed in the non-pathogenic BT, nor does it confer an invasive phenotype on *E. coli*, although no inactivating mutations were detected (Pierson and Falkow, 1990). The global regulator, *rovA* regulates the expression of invasion, and modulates the ability to infect and colonize deeper tissue layers in the mouse infection model (Revell and Miller, 2001).

Distribution of the attachment and invasion locus protein Ail is very specific to pathogenic isolates, namely *Y. pseudotuberculosis*, *Y. pestis* and the *Y. enterocolitica* biotypes 1B, and 2 – 5 (Miller, Farmer, Hill and Falkow, 1989; Revell and Miller, 2001). Artificial overexpression of Ail in non-pathogenic BTs of *Y. enterocolitica* leads to transcription and localization to the cell surface, but fails to induce invasive behaviour, possibly due to differences in the cell surface structure of high- and non-pathogenic BTs (Pierson and Falkow, 1993).

1.2.3.2 Yersinia Adhesion Pathogenicity Island (YAPI)

The Yersinia adhesion pathogenicity island YAPI shows varied distribution amongst *Y. pseudotuberculosis* and *Y. enterocolitica* BT 1B strains. In *Y. pseudotuberculosis* it has been found in the strain causing Far East scarlet-like fever and a serotype I strain (Eppinger, et al., 2007). Only two BT 1B strains of *Y. enterocolitica* have been described containing YAPI suggesting independent acquisition events of this mobile genetic element (Thomson, et al., 2006).

The conserved part of YAPI encodes a polycistronic pilin cluster that is involved in adhesion of bacteria to the respiratory epithelium in a mouse infection model (Collyn, et al., 2004; Eppinger, et al., 2007). The variable part of YAPI in *Y. pseudotuberculosis* contains metabolic genes, a restriction-modification system, IS elements and phage genes, which do not appear to contribute to pathogenicity (Collyn, et al., 2004). In *Y. enterocolitica*, the island is smaller than in *Y. pseudotuberculosis*, and apart from the shared pilin cluster encodes for a possible hemolysin, toxin/antitoxin system, and an arsenic resistance operon (Thomson, et al., 2006).

It is thought that YAPI might have derived from the *Salmonella* Typhi island SPI-7 due to extensive homology and is also found in other *Enterobacteriaceae*, such as *Photorhabdus luminescens* (Collyn, et al., 2004; Thomson, et al., 2006; Eppinger, et al., 2007).

1.2.3.3 High-pathogenicity island (HPI) - Pigmentation / hemin storage locus

The high-pathogenicity island (HPI) has been described in isolates of *Y. enterocolitica*, *Y. pseudotuberculosis*, and *Y. pestis* (Carniel, Guilvout and Prentice, 1996; Carniel, 2002). In *Y. enterocolitica*, it is only present in the high-pathogenic BT 1B. Its presence is variable in *Y. pseudotuberculosis* isolates (Carniel, Guilvout and Prentice, 1996; Bach, et al., 1999) and was notably absent in the representative *Y. pseudotuberculosis* clone held responsible for Far East scarlet-like fever (Eppinger, et al., 2007). The HPI has been shown to play a major role in systemic dissemination during infection (Eppinger, et al., 2007).

There are variations in gene content, and the HPI could be considered as two joined segments (Bach, et al., 1999; Carniel, 2002). The total island in *Y. pestis* occupies ~100 kb and is referred to as pigmentation or hemin storage locus (*hms / pgm*; Carniel, Guilvout and Prentice, 1996). This locus has been described in *Y. pestis* as the ability to absorb exogenous hemin and form pigmented colonies (Brubaker, 1991; Carniel, Guilvout and Prentice, 1996). *hms* encodes an acquisition and storage system for the essential nutrient iron, which is limited in the host (Brubaker, 1991; Carniel, 2002). Also located on the pigmentation locus are *fyuA-psn*, which encodes the double protein FyuA acting as the siderophore yersiniabactin and bacteriocin pesticin, and *irp1-5*, iron-regulated high-molecular-weight proteins of which protein 2 appears to have special importance (Carniel, Guilvout and Prentice, 1996; Bach, et al., 1999). In experimental studies, pigmentation mutants were shown to have a marked decrease in virulence (Brubaker, 1991; Carniel, Guilvout and Prentice, 1996). The locus is required in flea colonization as well, as it blocks the flea's midgut leading to regurgitation of *Y. pestis* upon feeding on a host and thus ensuring efficient transmission (Parkhill, et al., 2001a).

The HPI is significantly shorter in *Y. enterocolitica* BT 1B than in *Y. pestis*, and only stretches over 45 kb (Carniel, Guilvout and Prentice, 1996). The only genes conserved with the *Y. pestis* locus are the genes involved in iron capture: *fyuA* and *irp1-5* (Carniel, Guilvout and Prentice, 1996; Bach, et al., 1999). Instead of the pigmentation locus, the rest of the HPI is occupied by IS and repeat elements in

Y. enterocolitica BT 1B (Carniel, Guilvout and Prentice, 1996). It is thought that this absence of iron-chelating proteins in the low-pathogenic *Y. enterocolitica* BTs 2 – 5 contributes to their lowered virulence potential and less invasive disease outcome (Carniel, Guilvout and Prentice, 1996).

The HPI in *Y. pseudotuberculosis* has the same genetic content as *Y. pestis*, and differs only in the terminal sites (Bach, et al., 1999). The presence of repeats and functional integrases in *Y. pseudotuberculosis* and *Y. pestis* leads to frequent deletion of the HPI, whereas due to a mutated integrase the region is highly stable in *Y. enterocolitica* (Bach, et al., 1999). Deletions of the HPI in *Y. enterocolitica* are possible but infrequent and comprise a much larger genomic region, with additional loss of phenotypes such as motility and growth defects (Bach, et al., 1999). The HPI has been described in other *Enterobacteriaceae* as well, including *Klebsiella*, *Citrobacter*, and *E. coli*, supporting the mobile nature of this locus (Bach, et al., 1999; Carniel, 2002).

1.2.3.4 *Yersinia pseudotuberculosis* superantigen YPM

The *Y. pseudotuberculosis*-derived mitogen (YPM) is a superantigenic exotoxin that has been linked with *Y. pseudotuberculosis* infections presenting as Far East scarlet-like fever (Eppinger, et al., 2007). Similar to superantigens found in Gram-positive bacteria such as *Staphylococci* or *Streptococci*, the *Y. pseudotuberculosis* superantigen can activate the host immune system in an excessive, uncontrolled manner. The similar clinical manifestation of Far East scarlet-like fever and scarlet fever caused by *Streptococcus pyogenes* is probably due to an analogous action of the superantigens (Eppinger, et al., 2007). YPM can present in three different variants encoded by *ypmA*, *ypmB*, and *ypmC*. Whilst *ypmB* and *ypmC* have been correlated with environmental and animal isolates, *ypmA* is established in clinical isolates (Eppinger, et al., 2007). It has been suggested that the last common ancestor of *Y. pseudotuberculosis* bore an ancestral *ypm* gene, which subsequently diverged and has been lost in individual lineages of *Y. pseudotuberculosis* as it is contained in an unstable locus in the genome (Ch'ng, et al., 2011). It has been suggested that YPM+ HPI- strains cause Far East scarlet fever, whereas YPM- HPI+ strains are linked with enteric disease (Fukushima, et al., 2001; Ch'ng, et al., 2011; Eppinger, et al., 2007).

1.2.3.5 Ysa type 3 secretion system

Y. enterocolitica BT 1B strains possess a second T3SS called the Ysa T3SS for *Yersinia* secretion apparatus that is absent from non- and low-pathogenic BTs and other *Yersinia* species and that is thought to be involved in virulence (Haller, Carlson, Pederson and Pierson, 2000; Foultier, et al., 2002). At least eight *Yersinia* secreted proteins, Ysps, depend on secretion by the Ysa T3SS (Haller, Carlson, Pederson and Pierson, 2000). It was shown that secretion by this system is activated at 28°C and high salt conditions, and that this secretion is independent from the plasmid-encoded Yop secretion that is triggered at 37°C and low calcium conditions (Haller, Carlson, Pederson and Pierson, 2000). This has prompted speculation that the physiological role of this system is early on in the infection process and helps to prime the bacterial cells (Haller, Carlson, Pederson and Pierson, 2000; Foultier, et al., 2002). A study was undertaken to determine the relatedness of this new, chromosomal T3SS to the established plasmid-encoded T3SS (Foultier, et al., 2002). It was shown that the two systems are very distinct, with the Ysa T3SS closely related to the *Shigella* Mix/Spa-type system and the *Salmonella* SPI-1 T3SS, and the Yop T3SS related to the Psc T3SS in *Pseudomonas aeruginosa* and Bsc in *Bordetella* (Foultier, et al., 2002). It appears that there is a third, different T3SS present in *Y. pseudotuberculosis* and *Y. pestis* that is similar to the *Salmonella* SPI-2 T3SS, but so far no in-depth study into this system has been undertaken (Foultier, et al., 2002).

1.2.3.6 Further chromosomally encoded virulence factors

Other genes or operons have been implicated in a role in disease (Revell and Miller, 2001). Amongst these is *yst* encoding the heat-stable enterotoxin. It is thought that its action induces diarrhoea, specifically in infants, similar to the heat-stable toxin of *E. coli* (Cornelis, 1994; Revell and Miller, 2001). Some non-pathogenic BT 1A strain may have homologous sequences to *yst* as indicated by hybridization experiments (Grant, Bennet-Wood and Robins-Browne, 1998), but their expression and functionality need further study.

The *myf* operon encodes fimbriae similar to pili of *E. coli*, one of them being the pH6 antigen (Cornelis, 1994; Revell and Miller, 2001). The operon is present in all of the pathogenic *Yersiniae*, and mediates binding to host cells (Revell and Miller, 2001). Similarly to *yst*, *myf*-homologous sequences might be present in the non-pathogenic BT (Grant, Bennet-Wood and Robins-Browne, 1998).

Using *in vivo* expression technology, IVET (Mahan, Slauch and Mekalanos, 1993), a number of genes were identified that are expressed under stress conditions but not under laboratory conditions. This led to the characterization of host responsive elements, *hre* (Young and Miller, 1997), regulator of systemic colonization RscR (Nelson, Young and Miller, 2001), and systemic invasion factors, *sif*, all of which have been attributed an undefined function in the mechanism of disease (Revell and Miller, 2001).

It is possible that a combination of different virulence factors is required for full infectivity and that the different factors contribute to various stages of disease, as individual mutants usually fail to completely attenuate the strain (Revell and Miller, 2001).

1.3 Evolution of the genus *Yersinia*

1.3.1 Differentiation and speciation in *Yersinia*

Correct identification of pathogenic *Yersiniae* is crucial for disease management. In the past, phenotypic tests and serotyping have been used, with the limitations of strains presenting with atypical biotyping reactions and with serotyping only possible at reference laboratories (Virdi and Sachdeva, 2005). Early on, autoagglutination, calcium dependency, uptake of crystal violet and Congo red, as well as serum resistance have been used as indirect indicators of virulence (Pedersen, Winblad and Bitsch, 1979; Schiemann and Devenisch, 1982; Robins-Browne, et al., 1989). Drawbacks of these techniques are atypical and contradictory results for a number of isolates (Robins-Browne, et al. 1989). Commercially available phenotypic tests were also lacking some of those sugars and tests that more reliably differentiate *Yersiniae*, leading to decreased sensitivity with respect to identification of environmental species (Neubauer, et al., 1998).

These disadvantages prompted investigation into the use of DNA hybridization using probes for potential virulence genes for the detection of pathogenic isolates (Robins-Browne, et al., 1989). Probes for the invasin Inv derived from *Y. pseudotuberculosis* only detected *Y. pseudotuberculosis* and *Y. pestis* isolates, but probing with *Y. enterocolitica* Inv gave positive results in environmental as well as pathogenic species of *Yersinia* (Robins-Browne, et al., 1989). In contrast, probes for Ail proved to be highly specific, and depending on design to include adjacent DNA regions can

either detect *Y. enterocolitica* BT 1B or all pathogenic *Y. enterocolitica* biotypes (Robins-Browne, et al., 1989). Probes designed for detection of the virulence plasmid pYV detected plasmid-bearing strain of *Y. pseudotuberculosis*, *Y. pestis*, and *Y. enterocolitica* (Robins-Browne, et al., 1989), but it has to be considered that the plasmid can be easily lost in laboratory culture and thus could have an influence on virulence assessment of isolates. The detection of virulence-associated genes was further developed into assays using polymerase chain reaction (PCR) to amplify parts of the virulence plasmid or chromosomal fragments including *inv*, *ail*, or the enterotoxin *yst* (e.g. Nakajima, et al., 1992; Ibrahim, Liesack and Stackebrandt, 1992).

New methods such as multilocus enzyme electrophoresis (MLEE), restriction endonuclease analysis of chromosomal and plasmid DNA (REAC, REAP), ribotyping, pulse-field gel electrophoresis (PFGE), and typing based on variable number of tandem repeats (VNTR) have been applied to study heterogeneity of *Y. enterocolitica* and '*Y. enterocolitica*-like' species (reviewed in Virdi and Sachdeva, 2005; Souza, Pitondo-Silva, Falcão and Falcão, 2010). It was found that *Y. enterocolitica* has three consistent lineages of high-, low-, and non-pathogenic biotypes (Virdi and Sachdeva, 2005). For some methods, limited data is available on the environmental species, precluding any deduction of the evolutionary relationship of *Y. enterocolitica* and the environmental species (Virdi and Sachdeva, 2005). Further more, these methods are more suitable for short-term study of isolates during an epidemic outbreak, but are less appropriate to determine the historic evolutionary relationships between the *Yersinia* species (Kotetishvili, et al., 2005). Nucleotide analysis of 16S rDNA gene or housekeeping genes in multi-locus sequence typing (MLST) is considered a more unbiased approach for speciation (Kotetishvili, et al., 2005; Souza, Pitondo-Silva, Falcão and Falcão, 2010).

1.3.2 16S rDNA analysis and MLST scheme

Initial phylogenetic studies in the genus *Yersinia* focussed on the 16S rDNA gene (Ibrahim, et al., 1993). The 16S rDNA gene is present in all bacteria and thus used as a gold standard to establish the relationship between different genera (Janda and Abbott, 2007). For *Yersinia*, five different lineages were identified (Ibrahim, et al., 1993; Ibrahim, et al., 1997). *Y. enterocolitica* forms a heterogeneous cluster with the possibility of splitting them into two distinct subclusters of American and European isolates. *Y. pseudotuberculosis* and *Y. pestis* together with *Y. frederiksenii* form a

separate cluster. *Y. rohdei* and *Y. ruckeri* each form independent clusters, and the environmental species *Y. mollaretii*, *Y. intermedia*, *Y. bercovieri*, *Y. kristensenii*, and *Y. aldovae* form another heterogeneous cluster (Ibrahim, et al., 1993; Ibrahim, et al., 1997). Later, a sixth lineage was discovered which includes atypical *Yersinia* isolates, which have a different 16S signature (Ibrahim, et al., 1997). Biochemically, these belong to *Y. frederiksenii* and have been designated as genomospecies 2 (Ibrahim, et al., 1997). Generally, the sequence similarities vary between 96.9 – 99.8%. This is considered too high to infer evolutionary relationships between different *Yersinia* lineages and a distinction between species is less reliable (Ibrahim, et al., 1993; Janda and Abbott, 2007). To improve speciation, 16S rDNA analysis in conjunction with DNA hybridization and biochemical characteristics was suggested (Ibrahim, et al., 1993; Neubauer, Hensel, Aleksic and Meyer, 2000).

Upon reconstruction of a phylogenetic tree based on 16S rDNA sequences only, it was found that *Y. ruckeri* does not present an outlier as opposed to relationships based on other housekeeping genes (Ibrahim, et al., 1993; Ibrahim, et al., 1997; Souza, Pitondo-Silva, Falcão and Falcão, 2010). This might be due to an exchange of 16S rDNA sequences as has been noted in other bacteria before (Asai, Zaporojets, Squires and Squires, 1999; Schouls, Schot and Jacobs, 2003; Souza, Pitondo-Silva, Falcão and Falcão, 2010), and this recombinatorial potential might preclude this gene from use in establishing evolutionary connections.

Recent exploration of use of the 16S rDNA sequence has lead to the proposition that *Y. enterocolitica* should be grouped into two subspecies (Neubauer, Hensel, Aleksic and Meyer, 2000; Neubauer, et al., 2000). DNA hybridization experiments indicate that there are possibly three groups within *Y. enterocolitica* corresponding to high-pathogenic 'New-World' strain, low-pathogenic 'Old-World' strains, and non-pathogenic 'Old-World' strains (Neubauer, et al., 2000). Studies on the 16S rDNA gene have found regions to distinguish between 'American' BT 1B strains and 'European' BT 1A, 2 – 5 strains (Ibrahim, et al., 1993; Ibrahim, et al., 1997; Neubauer, Hensel, Aleksic and Meyer, 2000; Kotetishvili, et al., 2005). The comparison of 1367 bases of the 16S genes showed a difference in 12 nucleotides between isolates from America and Europe, which forms the grounds for establishing two subspecies: the 'New World' BT 1B designated as *Y. enterocolitica* subsp. *enterocolitica* and the 'Old World' BTs 1A, 2 – 5 belonging to *Y. enterocolitica* subsp.

palearctica (Neubauer, Hensel, Aleksic and Meyer, 2000; Neubauer, et al., 2000; Neubauer and Sprague, 2003).

An MLST scheme has been attempted using the 16S rDNA sequence as well as the four housekeeping genes *glnA*, *gyrB*, *recA*, and *hsp60* (Kotetishvili, et al., 2005). Only the 11 classically described *Yersinia* species were included, as the new species were proposed after publishing of the scheme. *Y. pseudotuberculosis* and *Y. pestis* proved inseparable by the gene fragments investigated, indicating their close genetic relationship that has been noted before (Achtman, et al., 1999; Kotetishvili, et al., 2005). It was found that *Y. aldovae*, *Y. bercovieri*, *Y. intermedia*, *Y. rohdei*, and *Y. ruckeri* as well as *Y. pestis* and *Y. pseudotuberculosis* are homogeneous species, whereas the other four species *Y. enterocolitica*, *Y. kristensenii*, *Y. frederiksenii*, and *Y. mollaretii* are heterogeneous (Kotetishvili, et al., 2005). The 16S rDNA sequences had less discriminatory power than the proposed MLST scheme, and the biochemical species definition was more in agreement with the MLST scheme than 16S rDNA gene (Kotetishvili, et al., 2005), which could indicate recombination amongst 16S genes and thus masking of the evolutionary signal. Some strains were found to be distinct from their respective species based both on MLST and 16S rDNA analysis, and would warrant further investigation as potentially new species within the genus *Yersinia* (Kotetishvili, et al., 2005). *Y. frederiksenii* is known to be a heterogeneous species and the atypical isolates might be related to the newly proposed *Y. massiliensis* (Merhej, et al., 2008; Souza, Falcão, and Falcão, 2011). Atypical *Y. kristensenii* were later speciated into *Y. aleksiciae* (Sprague and Neubauer, 2005).

1.3.3 Current model of evolution

The current model for evolution of the genus Yersinia is presented in Figure 1.2 (Carniel, 2002; Carniel, 2003; Wren, 2003).

It is thought that the pathogenic *Yersiniae* have evolved from the environmental species following the acquisition of the virulence plasmid pYV (Carniel, 2002; Wren, 2003). *Y. enterocolitica* and *Y. pseudotuberculosis* then developed independently. This split is thought to have occurred 42 – 187 million years ago, similar to the distance of 140 million years for the last common ancestor of *Escherichia coli* and *Salmonella enterica* (Achtman, et al., 1999). *Y. enterocolitica* is signified by the acquisition of the heat-stable toxin Yst. It then splits into three different lineages of non-, high-, and low-pathogenic isolates. The non-pathogenic BT 1A loses pYV to

once again become a predominantly environmental lineage. The New World, high-pathogenic BT acquired several pathogenicity islands, including HPI, as well as unique type 2 and 3 secretion systems, explaining the increase in virulence potential. The low-pathogenic lineage includes BTs 2 – 5.

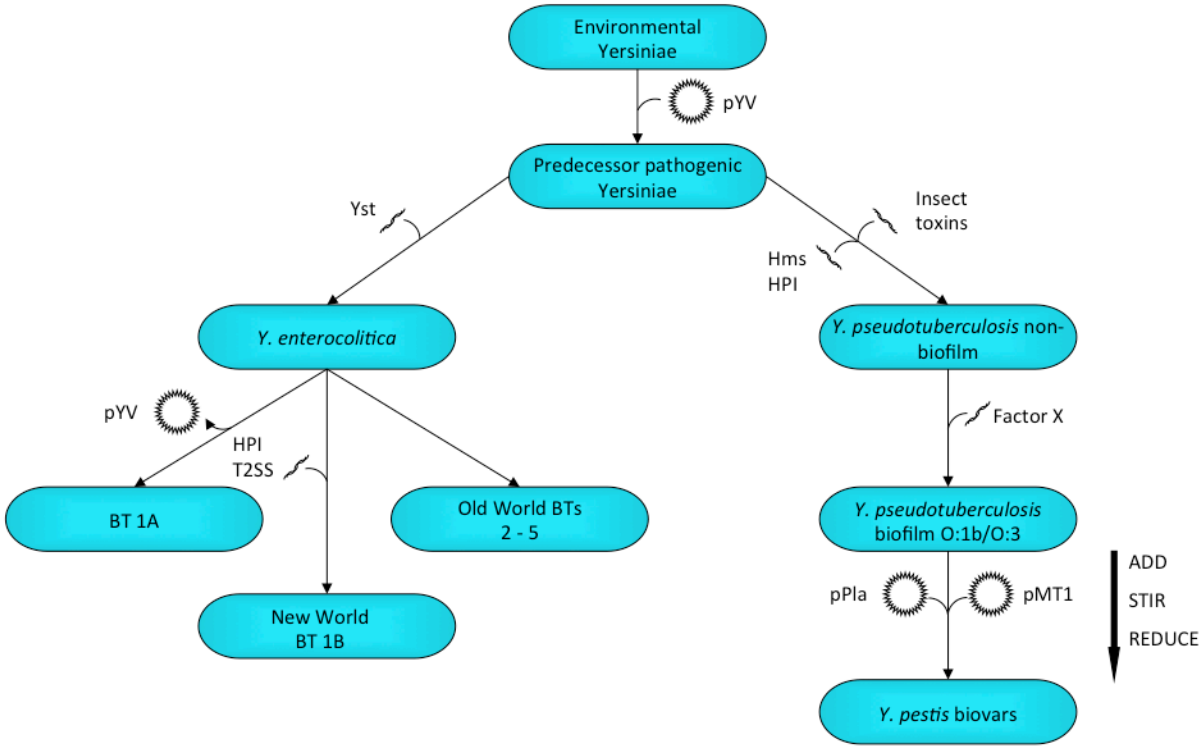


Figure 1.2: Evolution of the genus *Yersinia* [adapted from Carniel, 2002; Wren, 2003].

Y. pseudotuberculosis also acquired pathogenicity islands. Similar to *Y. enterocolitica* BT 1B, it obtained the HPI, but it also has the hemin storage locus, and the insect toxin encoded on the tcPAI. A subset of *Y. pseudotuberculosis* then evolves the ability of biofilm formation, and the acquisition of the plasmids pPCP and pMT gives rise to the *Y. pestis* lineages. The last common ancestor of *Y. pseudotuberculosis* and *Y. pestis* is thought to have existed 0.4 – 1.9 million years ago (Achtman, et al., 1999). The evolution of the *Y. pestis* biovars is thought to have occurred within the last 1,500 to 20,000 years, through a number of bottlenecks (Achtman, et al., 1999).

For the evolution of *Y. pestis* three principles have been identified: add DNA, stir and reduce (Parkhill, et al., 2001a; Wren, 2003; Chain, et al., 2004). The acquisition of large genomic islands, such as the hemin storage locus, and additional plasmids pPCP/pPla and pMT/pFra has enabled the colonization of new niches (Wren, 2003). It is thought that the acquisition of the plasmid pMT occurred first, possibly donated

by *Salmonella*, which was important for colonization of and transmission via fleas (Achtman, et al., 1999; Carniel, 2003). Both chromosomal and plasmid factors are required for full virulence potential, acting together to enable *Y. pestis* to colonize the flea and be transmitted by it, and are also needed to avoid the mammalian immune system, and to disseminate in the mammalian host (Carniel, 2003; Wren, 2003).

Compared to *Y. pseudotuberculosis*, the *Y. pestis* genome shows an increase in the number of insertion sequence (IS) elements that are present (Parkhill, et al., 2001a; Deng, et al., 2002; Wren, 2003; Chain, et al., 2004; Eppinger, et al., 2007; Eppinger, et al., 2010). IS elements are DNA sections encoding a transposase flanked by perfect inverted repeats. The transposase can mediate the transposition of the element to a different genomic location. Recombination between IS elements have led to wide-ranged genome rearrangements in *Y. pestis*, which are still ongoing in culture *in vivo* (Parkhill, et al., 2001a; Deng, et al., 2002; Chain, et al., 2004; Eppinger, et al., 2010). IS elements have inserted into the coding sequences of other genes, thus disrupting the expression of functional proteins and promoting isolate-specific polymorphisms (Parkhill, et al., 2001a; Chain, et al., 2004; Eppinger, et al., 2010).

The loss of function plays a major part in the evolution of *Y. pestis*, and may be even more important than gene acquisition (Chain, et al., 2004). Some of the genomic islands identified in *Y. pestis* have orthologs in *Y. pseudotuberculosis* and thus precede the split of these two lineages, but there is evidence of gene decay in *Y. pestis* (Parkhill, et al., 2001a; Chain, et al., 2004). For instance, some proteins encoded in the toxin complex pathogenicity island (tcPAI) responsible for insect toxicity have been inactivated in *Y. pestis* (Parkhill, et al., 2001a). Instead of killing the flea, it might be more important for *Y. pestis* to establish persistence for transmission (Achtman, et al., 1999; Parkhill, et al., 2001a; Wren, 2003). Later studies however showed that the insecticidal toxins are not mutated in other *Y. pestis* biovars, leaving in question the exact role of functionality or non-functionality of the tcPAI in survival in the flea (Chain, et al., 2004). In comparing *Y. pestis* and *Y. pseudotuberculosis* it was found that up to 13% of the *Y. pestis* genome is inactivated (Chain, et al., 2004). These genes account for a number of functions relating to environmental and enteric survival that have become redundant in a host-adapted life-style and include metabolic pathways, motility and lipopolysaccharides (Brubaker, 1991; Achtman, et al., 1999; Parkhill, et al., 2001a; Chain, et al., 2004). This way, *Y. pestis* has now restricted itself to be reliant on the host for the provision

of nutrients (Brubaker, 1991; Achtman, et al., 1999). It is also evident that *Y. pseudotuberculosis* possessed a number of colonization and virulence factors that primed *Y. pestis* in establishing its pathogenic potential (Parkhill, et al., 2001a).

Recently the grouping scheme for *Y. pestis* has been revised based on genetic relatedness rather than biochemical differences (Achtman, et al., 2004; Morelli, et al., 2010). The high diversity of strains can be a sign of geographical source, indicating *Y. pestis* originated in Asia (Achtman, et al., 2004; Morelli, et al., 2010). From *Y. pseudotuberculosis*, the evolutionary route develops into a branch 0, from which biovar Pestoides (0.PE) branches off about 2,600 years ago (Morelli, et al., 2010), confirming the ancestral status of the Pestoides group (Eppinger, et al., 2010). After further evolution along branch 0, there is a split into branch 1 and branch 2 about 700 years ago (Achtman, et al., 2004; Morelli, et al., 2010). Branch 1 contains isolates of the most recent pandemic sweep of plague, and covers isolates of biovar designation Antiqua and Orientalis, hence 1.ANT and 1.ORI. It could be shown that this clone arose >200 years ago, and it subsequently spread out of China and around the world distributed by plague ships (Morelli, et al., 2010). Single introductory events into North America and Madagascar lead to local clonal evolution (Morelli, et al., 2010). Branch 2 of the phylogenetic tree in *Y. pestis* is evolutionary older than branch 1 (>500 years ago) and is associated with biovars Medievalis and Antiqua (2.MED, 2.ANT; Achtman, et al., 2004; Morelli, et al., 2010). The distribution of clones and their accumulating genomic differences implies that *Y. pestis* was spread along the silk trade route in medieval times (Morelli, et al., 2010). Isolates of biovar Antiqua are found in both branches 1 and 2, indicating that the genomic differences do not relate to phenotypic differences, and that the phenotype possibly arises from independent mutations (Achtman, et al., 2004).

1.4 Next-generation sequencing techniques

Recent advances in our knowledge of *Yersinia* evolution such as tracing plague along the silk trade route (Morelli, et al., 2010) are due to rapid advances in whole genome sequencing capacities. Sequencing of bacterial genomes provides a powerful tool to investigate the complete genetic makeup of a strain and allows for identification of regions unique to particular organisms by comparison of strains against each other. Until recently, sequencing was carried out using automated

Sanger or chain-termination sequencing (Mardis, 2008a; Ansorge, 2009; Metzker, 2010). As this technique is costly, it was only applied to high-interest representative species of model organisms and pathogens, available at larger genome sequencing centres such as the Wellcome Trust Sanger Institute (WTSI) (Hall, 2007; Mardis, 2008a).

Next-generation sequencing techniques have been available since 2004 (Mardis, 2008a), and have found widespread applications. The term next-generation sequencing is generally used to describe techniques that allow for massive parallel sequencing approaches and circumvent the need for cloning in *Escherichia coli*. This gives the opportunity to sequence larger strain collections in shorter periods of time compared to traditional Sanger sequencing (Mardis, 2008a; Ansorge, 2009; Metzker, 2010). The different techniques have different strengths and weaknesses, making them more suitable for some scientific questions than others. In the following, the main sequencing techniques used in this study will be described and compared.

1.4.1 Capillary sequencing

The traditional sequencing technique is based on the chain termination method published by Sanger, Nicklen and Coulson (1977). DNA is isolated and fragmented into pieces of known size (Figure 1.3). These fragments are then cloned into a plasmid containing universal sequencing primers, thus creating a library of fragments (Mardis, 2008b; Shendure and Ji, 2008). The library is used to transform the vector into *E. coli*, using the bacterial host for amplification of the vector. Individual clones that have arisen from a single transformation event are then picked and the DNA is isolated for sequencing. The sequencing reaction occurs in several cycles by supplying it with two types of nucleotides: normal deoxynucleotides (dNTPs) that can be incorporated into the growing chain as usual, and di-deoxynucleotides (ddNTPs) which have a modified 3' end (Sanger, 1988). The integration of a ddNTP leads to the termination of the chain, as another base cannot be added due to the blocked 3' end, hence this method is known as chain-termination sequencing. Additionally, the ddNTPs are labelled with different fluorophores corresponding to the different bases. As the incorporation of ddNTPs is competitive with the dNTPs, the resulting chains are of different lengths with the terminal base marked by the fluorescence-labelled ddNTP (Shendure and Ji, 2008).

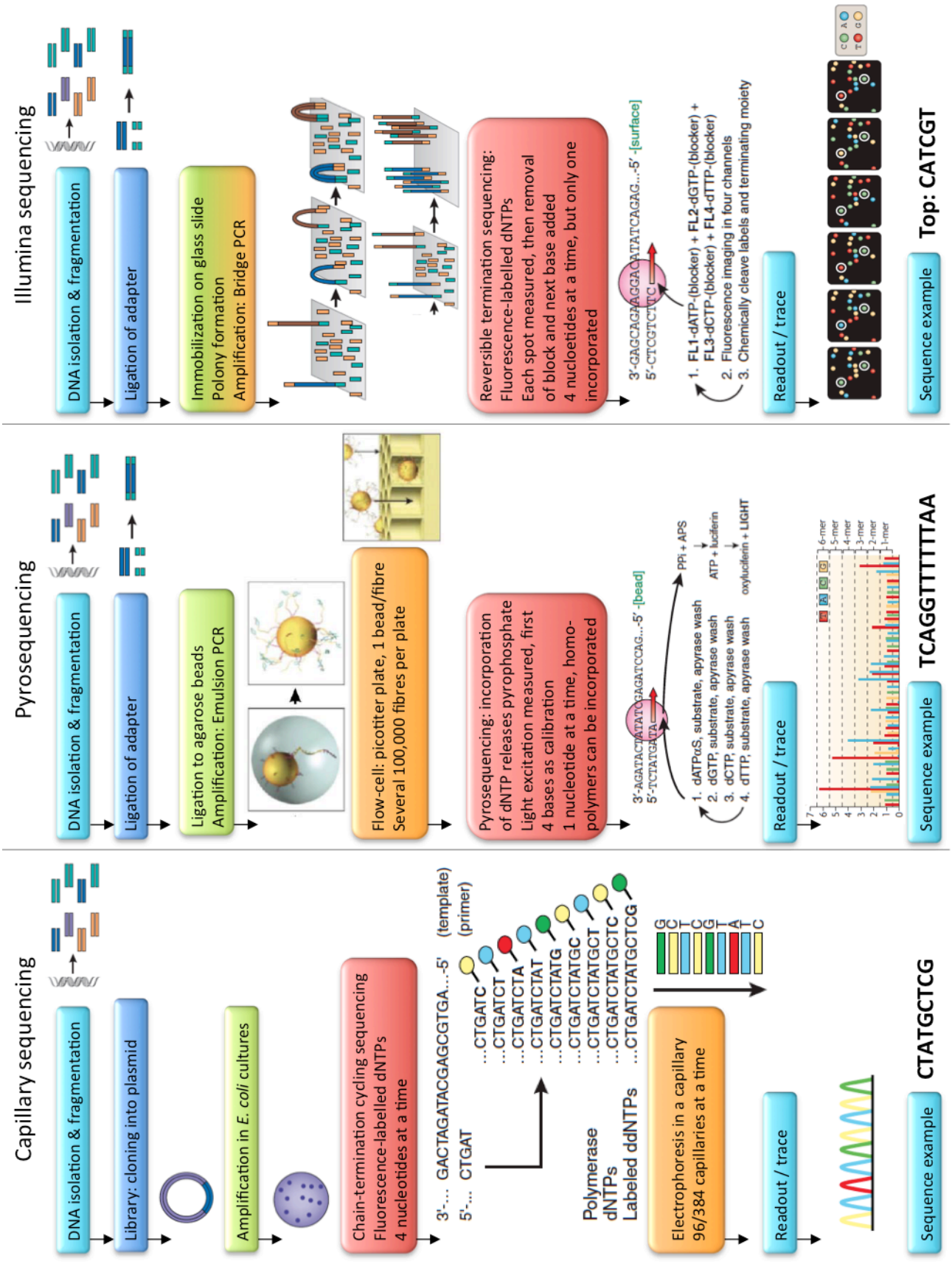


Figure 1.3: Workflow of capillary, pyro- and Illumina sequencing [adapted from Shendure and Ji, 2008; Ansorge, 2009; Meztker, 2010].

For visualization of the different chains, separation is carried out in a gel electrophoresis step. In early sequencing, this meant running long gels and photographing them (Sanger, Nicklen and Coulson, 1977). Smaller fragments run faster than longer fragments, and the progression through the gel is proportional to the length of the molecule. A read-out of the 5' to 3' sequence is possible starting from the bottom upward. In more recent times polymer-coated capillaries are used which allow the elution of the chains according to length with high-resolution (Hall, 2007; Shendure and Ji, 2008). The chains are excited with a laser and the fluorophore that terminated the chain can be read out.

Capillary electrophoresis allows for parallelization of 96 or 384 independent capillaries. This limitation is the main drawback of Sanger sequencing, as only a specific number of sequencing reactions can be observed in parallel (Hall, 2007; Shendure and Ji, 2008; Ansorge, 2009).

1.4.2 454 Pyrosequencing

Pyrosequencing was the first commercially available platform introduced in 2004 (Margulies, et al., 2005; Mardis, 2008a). The associated platform is the 454 FLX (Titanium) produced by Roche.

DNA is isolated and fragmented, similar to the traditional sequencing technique (Figure 1.3). Instead of cloning and transformation into *E. coli* though, the DNA is ligated to specific adapters. These adapters have complimentary sequences located on the surface of agarose beads, so that the adapter-ligated DNA fragments can attach to the beads (Metzker, 2010). A 1:1 ratio of fragments to beads ensures that only one DNA fragment is attached per bead (Margulies, et al., 2005; Mardis, 2008b). The beads are then added to a water:oil mixture containing PCR reagents. Vigorous vortexing results in small micelles, which are beads in an aqueous phase containing PCR reagents and which are surrounded by oil. Universal primer binding sites incorporated in the adapter sequences allow for clonal amplification *in vitro* of the attached fragments resulting in approximately one million copies per bead (Mardis, 2008b; Metzker, 2010). This has been termed emulsion PCR (Margulies, et al., 2005). By this means a single-stranded fragment library is created (Mardis, 2008b).

In the next step, the beads are added to a flow-cell. This is a picotiter plate consisting of several hundred thousand fibres per plate. The dimensions of the fibre only allow for one bead to be associated with a well in each fibre, thus immobilizing the

fragments on the plate in a specific position for monitoring of sequencing progress (Margulies, et al., 2005; Mardis, 2008a; Shendure and Ji, 2008; Ansorge, 2009).
The underlying principle for pyrosequencing is that the incorporation of dNTP releases pyrophosphate. This in turn produces energy, which can be used to break down luciferin, resulting in light emittance (Ansorge, 2009). The primer sequence is located within the adapter, and the first four bases that are sequenced correspond to the end of the adapter. As the adapter sequence is fixed, these four bases are always the same and are used for calibration of the detector for signal strength (Mardis, 2008a). Each nucleotide is added separately, and the number of nucleotides incorporated is proportional to the emitted light (Mardis, 2008a,b; Shendure and Ji, 2008). This poses a problem in regions where there are multiples of the same nucleotide, called homopolymeric runs (Mardis, 2008a; Ansorge, 2009). The detector can be saturated, and a differentiation between signal intensities is complicated. An additional or missing nucleotide within a coding sequence will show up as a frameshift in the sequence, implying that the gene would not be encoding a functional protein. This insertion/deletion error is the main inaccuracy in pyrosequencing (Mardis, 2008a; Metzker, 2010). Substitution errors in which the wrong base is incorporated are rare as nucleotides are added individually (Mardis, 2008a,b).
Pyrosequencing produces 250 bp reads on average, although they can be longer (Metzker, 2010).

A way of improving assemblies and counteracting the assembly of a large number of small contigs is paired-end sequencing. The initial fragmentation into ~3 kb molecules remains unchanged, but before the addition of adapters additional steps are introduced (Figure 1.4).
The ends of the fragments are biotinylated so that circularization is possible. Through this circularization the ends of the 3 kb fragment come in close proximity to each other and are linked by the biotinylated fragment. The circularized molecules are fragmented into a few hundred base pairs and fragments with the linker can be identified using streptavidin. The adapters are then ligated to the ends of the fragments and are subject to emulsion PCR (454 Life Sciences, 2008).

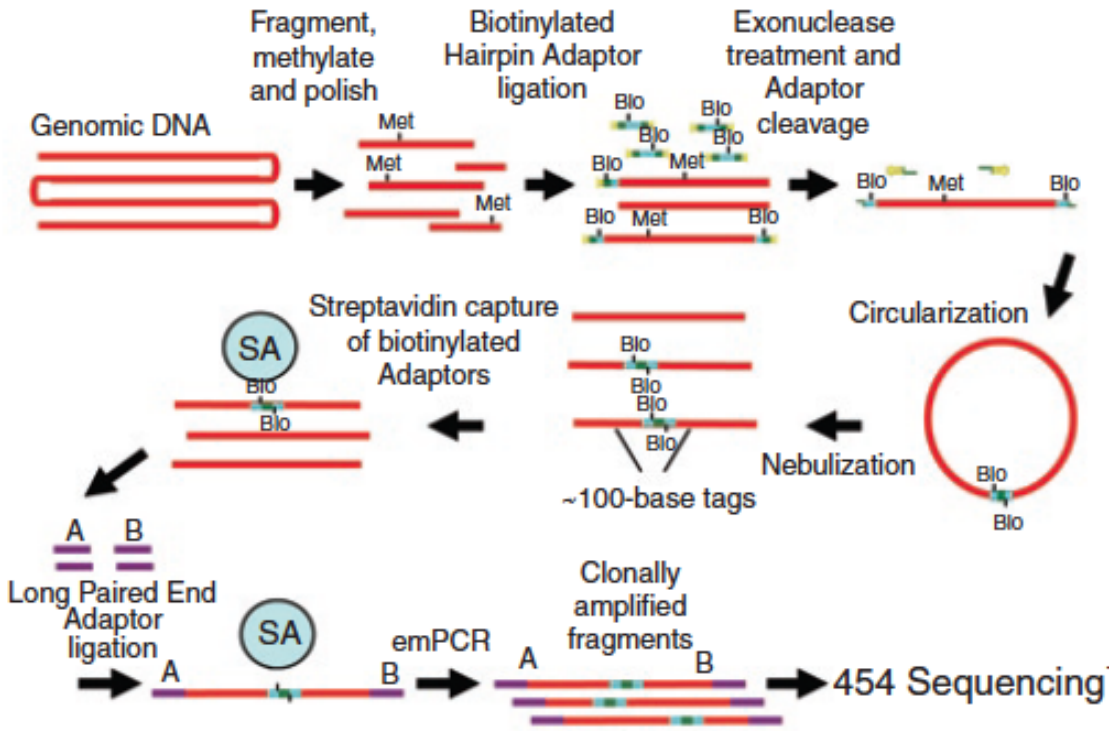


Figure 1.4: Preparation of paired-end fragments for pyrosequencing [taken from 454 Life Sciences, 2008].

This results in paired sequencing reads for which the distance, typically 3 kb, between them is known. This is particularly useful in genomes that have a high number of repeat units such as IS elements (Pop and Salzberg, 2008). A typical IS element including transposase and repeat sequences is about 1,200 bp long. The short reads produced by pyrosequencing are unable to bridge this length with reads with unique end sequence, and the multiple copies of repeat sequences thus make it difficult to find the exact location in the genome. Using the information that mate-pairs are 3 kb apart, contigs can be linked across the IS element simply because the read-pairs can be located on either side of it (Pop and Salzberg, 2008).

1.4.3 Illumina sequencing

Illumina produces the Illumina / Solexa genome analyser.

Illumina or Solexa sequencing also starts with DNA isolation, fragmentation and ligation to adapters (Figure 1.3) (Mardis, 2008a). Instead of attachment to beads, the DNA fragments are immobilized on a glass slide in Illumina sequencing, which at the same time acts as the place where sequencing and detection takes place (Mardis, 2008b). The glass slide is covered to a high density with adapters and adapter-homologous fragments (Mardis, 2008a; Ansorge, 2009). Once the fragments have

been immobilized, PCR reagents are added for amplification. The DNA fragments will bend over to form a bridge, across which PCR can take place (Shendure and Ji, 2008; Ansorge, 2009). The bridge PCR products are then denatured into single-stranded format so that a new round of amplification can take place. Over time a cluster of amplified DNA fragments has grown, which is called a “polony” for a polymerase-triggered, spatially clustered colony (Ansorge, 2009).

The sequencing reaction uses reversible termination technology. ddNTPs are added that contain a reversible block at the 3’end of the base and a fluorophore (Metzker, 2010). All four nucleotides are added at the same time, but only one can be incorporated as the 3’end is blocked from further extension (Mardis, 2008b). Each polony spot is then measured for fluorescence before the fluorophore and the 3’block are removed thus enabling the addition of a new nucleotide (Mardis, 2008a; Shendure and Ji, 2008; Ansorge, 2009; Metzker, 2010). This technology eliminates the chance of wrong multiple calling in homopolymeric runs as each nucleotide is added singularly. Conversely, the most common error observed is substitution when a nucleotide other than the correct one is incorporated by mistake (Shendure and Ji, 2008; Metzker, 2010).

Illumina sequencing produces very short read lengths, currently between 36 and 54 bp are common, but up to 100bp reads or longer are possible (Metzker, 2010).

Paired-end technology has also been adapted for Illumina sequencing, with distances typically 200 – 300 bp. The process is slightly different as the fragments used for inserts are much shorter (Figure 1.5).

Different adapters with individual primers are attached to either side of the fragmented molecule. Polony formation then enriches both sequences, but supply of only one primer in the first sequencing reaction will lead to sequence readout for one end of the molecule. In a separate sequencing reaction with the second primer the paired-end is read out. The spatial clustering allows sequences to be linked as pairs.

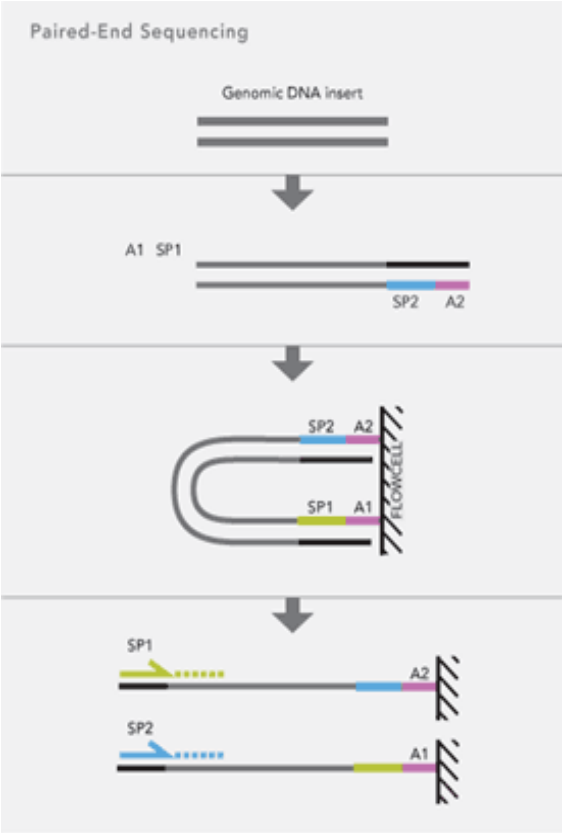


Figure 1.5: Paired-end fragment preparation in Illumina sequencing [taken from Illumina, 2011]. Adapter (A1 and A2) with sequencing primer sites (SP1 and SP2) are ligated onto DNA fragments. Template clusters are formed on the flow cell by bridge amplification and then sequenced by synthesis from the paired primers sequentially.

The paired-end technology as described for pyrosequencing – using larger fragment inserts, biotinylation, and subsequent fragmentation – is also possible with Illumina sequencing, but is termed “mate-pair sequencing”. It uses a preparation protocol similar to that used for pyrosequencing.

Usually the glass slide used for Illumina sequencing contains 8 lanes (Mardis, 2008a). Each lane may contain a single DNA species, but it is also possible to sequence up to 12 different species per lane. This process is called multiplexing and involves adding different sequence tags to the adapters (Figure 1.6).

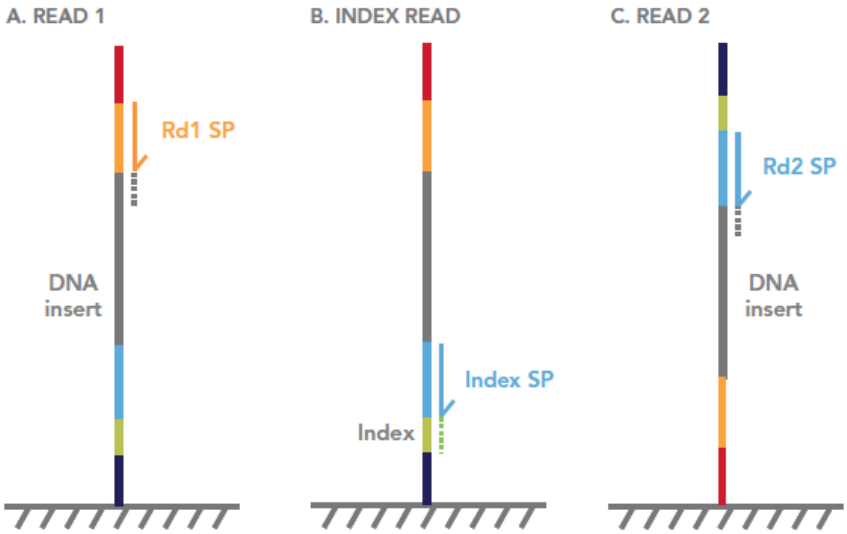


Figure 1.6: Multiplex preparation of DNA fragments for Illumina sequencing [taken from Illumina, 2011].

Each DNA species is treated separately with ligation of adapters for paired-end reads and a specific sequence tag of six bases. The different sequences are then mixed and applied to the lane. Sequencing takes place in three separate steps. The first and third reads supply the paired-end sequences, whilst the second read covers the index sequence. Computational methods then allow for identification of the index sequence and sorting. This increases the number of samples per slide from 8 to 96, but will decrease the coverage for each of the species, as physically less DNA is present. Because fragment density and thus coverage for each genome is very high in Illumina sequencing, this has minor influence on quality of sequence produced.

1.4.4 Comparison of capillary, pryo- and Illumina sequencing and applications

The different technologies described have advantages and disadvantages and are compared in Table 1.3.

The biggest contributor to the need for development of next-generation technology was the time factor. The use of capillary separation presents the most advanced technology that is possible in Sanger sequencing, and does not allow for any more parallelization than the maximum of 384 samples (Shendure and Ji, 2008; Ansorge, 2009).

Table 1.3: Comparison of sequencing technologies [adapted from Mardis, 2008a,b; Margulies, et al., 2005; Shendure and Ji, 2008; Metzker, 2010].

	Sequencing technology		
	Sanger	**454**	**Illumina**
> time factor	96 / 384 reads	Massive parallel sequencing array	Massive parallel sequencing array
		Several 100,000 reads	Tens of millions of reads
> time factor	Vector-based cloning	*In vitro* fragment library &	*In vitro* fragment library &
> cloning bias	libraries	amplification	amplification
	E. coli amplification		
> read length	~1,000 bp	~ 250 bp	36 – 54 bp
> base accurarcy	Very high	Indel error in homopolymeric runs	Substitution errors
> error rate	0.001%	0.6%	0.1%
	1 error in 100,000 bp	6 errors in 1,000 bp	1 error in 1,000 bp
> time factor	Run time 1-2 h	Run time 12 h	Run time: SE*: 2 days
			PE: 4 days
> reagent cost	Reagent volume µl	Reagent volume pl	Reagent volume pl
> cost per Mb^{+}	$ 500	$ 85	$ 6

* SE – single-end read, 36 bp, PE – paired-end read, 36 bp; $^{+}$ data from Shendure and Ji, 2008; Mardis, 2008a.

Library preparation, transformation and amplification in *E. coli* and the number of reads that can be analysed present the rate-limiting steps. Furthermore, there may be cloning bias as not all DNA is readily accepted and amplified in *E. coli* (Hall, 2007; Metzker, 2010). This adds to the comparably high reagent costs and thus the overall cost per megabase.

The next-generation sequencing technologies all share an array approach, allowing massive parallelization of reads (Shendure and Ji, 2008). This reduces the time needed to cover the whole genome of an organism. Library preparation time is reduced as well, as the preparation and amplification takes place *in vitro*. This circumvents the cloning bias given in *E. coli*, and is also quicker (Shendure and Ji, 2008; Metzker, 2010).

The overall time for a sequencing run is comparable with Sanger sequencing, but gives a higher coverage, that is each base is sequenced and recorded 20 – 60X, as opposed to usually 2X in Sanger sequencing. The offset is that the new sequencing technologies give shorter sequence reads and have a higher error rate than Sanger sequencing (Mardis, 2008b; Shendure and Ji, 2008). Sanger sequencing gives reads of up to 1,000 bp, whereas pyrosequencing currently gives on average 250 bp. Ultrashort reads are produced by Illumina sequencing. It should be noted that the read lengths of the next-generation technologies are still under development and are increasing frequently (Shendure and Ji, 2008; Metzker, 2010) so the length limit is not known yet. There have also been studies suggesting that information given by paired-end and mate-pair sequencing might be more important than increasing read

length (Chaisson, Brinza and Pevzner, 2009). Even in very fragmented assemblies most genes will fall within a single contig and not be split in two different ones. That way discovery and annotation should be possible (Chaisson, Brinza and Pevzner, 2009).
The main errors in pyrosequencing are insertion/deletions (indels) in homopolymeric runs, as mentioned before. Substitution errors are rare as each base is supplied individually. For Illumina sequencing, the opposite is true. Substitution errors are the most common error, as all ddNTPs are competing with each other for incorporation into the growing chain. The reversible termination on the other hand precludes mistakes in homopolymeric runs (Shendure and Ji, 2008).
Using an array approach with hundreds of thousands or more parallel reads also allows for a reduced reagent volume, which significantly lowers the cost of sequencing (Shendure and Ji, 2008).

The application for which the sequence information will be used very much determines the sequencing technology that will be most appropriate to use (Hall, 2007; Mardis, 2008b).
Higher costs of Sanger sequencing precludes it from being used in large-scale studies, but for the determination of exact sequence it is still considered gold standard (Hall, 2007; Ansorge, 2009). Given the longer reads, pyrosequencing is used in *de novo* sequencing and assembly, where an appropriate reference genome is not present. Illumina sequencing is used in re-sequencing, when an alignment against a reference is produced and single nucleotide polymorphisms (SNPs) can be determined (Hall, 2007; Mardis, 2008a; Metzker, 2010).

Next generation sequencing technologies have opened up new fields of applications of sequencing data (Mardis, 2008a; Metzker, 2010). A major field is population studies. It is now possible to sequence more strains of single species or even smaller distinctions such as subspecies or multi-locus sequence types. In this way it is possible to determine the normal range of variation within populations and distinguish it from variations that reflect specific adaptations for example in clinical isolates (Fraser-Liggett, 2005; Hall, 2007; Mardis, 2008a). Population studies also aid in mutation discovery, again possibly reflecting adaptational change (Mardis, 2008b; Shendure and Ji, 2008).

The possibility of sequencing of more representatives of species also helps identification of accessory genes and establishes the core genome of a species. Accessory genes can be unique phages as well as regions acquired by horizontal gene transfer, such as the exchange of antibiotic resistance markers and other so-called “virulence genes” (Hall, 2007; Mardis, 2008a,b). The accessory genes can make up a considerable part of the total genome of a strain, and can vary between strains of a species. Considering this variability in possible genome content supports sequencing of more than one representative for each species to establish the species’ core and the overall variety within a species described by the pangenome (Fraser-Liggett, 2005; Hall, 2007). Comparison of the accessory genomes between different eco- and biotypes of a species may also discover mechanisms behind adaptation to specific niches (Fraser-Liggett, 2005).

1.5 Aim of this project

1.5.1 The *Yersinia* Paradox

Yersinia is considered a paradoxical organism, as *Y. pestis* and *Y. pseudotuberculosis* are nearly identical with respect to genomic content, yet cause very different disease. On the other hand, *Y. pseudotuberculosis* and *Y. enterocolitica* are very distantly related, but cause the same disease outcome.

From the current hypothesis for the evolution of the genus, there is another paradox: *Y. pseudotuberculosis*, *Y. pestis*, and *Y. enterocolitica* are thought to share a common pathogenic precursor that evolved out of the environmental *Yersiniae*. Yet, at the same time, *Y. enterocolitica* and the environmental *Yersiniae* are thought to be closely related, since they were originally part of the heterogeneous group of ‘*enterocolitica*’ and are still referred to as ‘*Y. enterocolitica*-like’ species.

1.5.2 Aims of the study

Studies on the pathogenicity of strains of various biotypes of *Y. enterocolitica* have been undertaken in the past, generally using either random mutagenesis, hybridization techniques or insertional inactivation of genes that were shown to be involved in virulence in *Y. pseudotuberculosis* or *Y. pestis.* Although major genetic components were discovered with these methods in a wide range of bacteria, the disadvantages of these approaches are that only a limited number of genes are

observed – either the ones targeted for identification or those that give strong response signals under selective conditions. Hybridization approaches might also be complicated due to sequence divergence thus leading to false negative results. Also in many instances, single strains of that species or biotype are selected and used in these studies to represent all members of that species or biotype, and most often the high-pathogenic members are selected. In the case of *Yersinia*, *Y. enterocolitica* BT 1B has been used extensively as a model for *Yersinia* pathogenesis.

Little information is currently available on the *Y. enterocolitica* biotypes, although it is apparent that there is additional coding capacity within the BTs (Howard, et al., 2006). Only one representative strain for the high-pathogenic BT 1B has been sequenced and analysed in comparison with *Y. pestis* and *Y. pseudotuberculosis* (Thomson, et al., 2006). This comparison showed that about 68% of the genes are conserved between the three pathogenic species. This is much higher than the 20.8% of genes that are shared between all of the *Y. enterocolitica* biotypes as determined by microarray (Howard, et al., 2006), highlighting the sequence diversity inherent in *Y. enterocolitica* and also underlining the limitations of comparative genomic hybridization studies (Thomson, et al., 2006).

This project therefore set out to sequence representative strains for each of the biotypes and comparatively analyse the genomic content. As biochemical traits form an important part of differentiating the *Y. enterocolitica* biotypes, a phenotypic assessment of the representatively sequenced strains was undertaken.

The aims of the projects are:

1) To establish the genetic makeup of each of the biotypes of *Y. enterocolitica* and to define the unique sets of genes for each biotype
2) To characterize what is shared in pathogenic biotypes that is important for mechanism of disease
3) To determine the metabolic capacities of each of the biotype
4) To link phenotypic and genotypic traits for improved annotation and characterization of the *Y. enterocolitica* biotypes

2 Materials and Methods

2.1 Bacterial Strains

2.1.1 Reference Genomes

The reference strains were selected based on the prevalence of the different bio-serotypes of *Y. enterocolitica* isolated from cattle, pigs, and sheep throughout the UK between 1999 and 2000 during a randomised national survey of food pathogens in livestock sent for slaughter (McNally, et al., 2004) and in addition human isolates identified during the Intestinal Infectious Disease study (Tompkins, et al., 1999) carried out over the same time period. BT5 is not prevalent in the UK and poses no health concern. Thus, the reference strain was obtained from Germany from the laboratory of Dr. Thilo M Fuchs. The UK strain collection was extensively characterized genotypically (Howard, et al., 2006; Thomson, et al., 2006) and phenotypically (McNally, et al., 2006) and the strains listed below were chosen as representative strains for each prevalent bio-serotype.

Table 2.1: List of reference strains sequenced in this project.

Strain	Biotype	Serotype	Plasmids	Source
YE53/03	1A	O:5	pYV^-	Human faecal isolate
YE8081*	1B	O:8	pYV^+	Clinical isolate
YE212/02	2	O:9	pYV^+	Pig abattoir isolate
YE56/03	3	O:9	pYV^+	Human faecal isolate
YE149/02	3	O:5,27	pYV^+	Sheep abattoir isolate
YE12/03	4	O:3	pYV^+	Human faecal isolate
YE3094/96	5	O:2a,2b,3	pYV^+	Animal isolate

*: Reference strain, sequenced previously [Thomson, et al., 2006]

2.1.2 Worldwide Collections

The study of the genus Yersinia and a broader characterization of the species *Y. enterocolitica* were carried out on a worldwide collection of different species and strains from collaborating laboratories. The strains are listed in Table 9.1 (p. 204) and 9.2 (p206) in the appendix.

2.1.3 Additional strains

Some strains examined, which are not part of the sequencing collections, are listed in Table 2.2.

Table 2.2: Additional strains.

Strain	Biotype	Serotype	Source	country	Lab
YE2103	1B	O:19	Human control	UK	McNally
NZ6	1B	nd	Human	New Zealand	DuFour
YE21402	3	O:9	Pig	UK	McNally

2.2 Sequencing, Assembly and Annotation

2.2.1 Sequencing Technologies

DNA extraction and genome sequencing was carried out at the Wellcome Trust Sanger Institute (WTSI), Cambridge.

Sequencing for the reference genomes was carried out to improved high-quality draft genome standard (Chain, et al., 2009). The strains were sequenced using three different types of sequencing technologies. Capillary sequencing (ABI), also known as traditional Sanger sequencing, was used for 2X coverage, and produced long high-quality reads. Pyrosequencing (454/Roche GS20/FLX) and Illumina (Illumina GAII Analyser) technologies produce shorter reads, and thus reach higher coverage, in excess of 30- and 60-fold coverage respectively. PCR fragments were designed to span gaps and close them. This process was computerized by using algorithmic based automatic contiguation of assembled shotgun sequence (ABACAS; Assefa, et al., 2009; http://abacas.sourceforge.net). De-novo assemblies were produced using the Velvet assembler (Zerbino and Birney, 2008; http://www.ebi.ac.uk/~zerbino/velvet) and the GS De Novo assembly software (Roche, 2011). The final assembly was corrected with iCORN (Otto, Sanders, Berriman and Newbold, 2010).

Pseudogenes containing stop codons or frameshifts were checked manually for sequence quality, and if necessary further sequencing of that region in question was carried out.

Sequence for the worldwide collection was generated using the Illumina sequencing platform (GAII instruments) with indexed Illumina runs of 12 samples per lane of

sequencing, standard 200-300bp fragments and 76 cycles of sequencing. The 12 IP strains were sequenced separately at the Institute Pasteur using single-end Illumina reads.

The genome coverage is a measure of how many times a base is sequenced on average by different reads, and is therefore an indication of sequence quality in next-generation sequencing technologies. The coverage can be calculated from the average read length (here 76 bp), the number of reads, and the average genome size (5 Mbp) using the following formula:

$$expected\ coverage = \frac{(average\ read\ length * \ number\ of\ reads)}{average\ genome\ size}$$

The resulting coverage for *Y. enterocolitica* strains is plotted in Figure 2.1. The 12 strains obtained from E. Carniel at the Institute Pasteur were not included in this calculation, as they were sequenced separately.

The average coverage is higher than 20X in the strain collection. Variation arises from the number of reads obtained in a sequencing run.

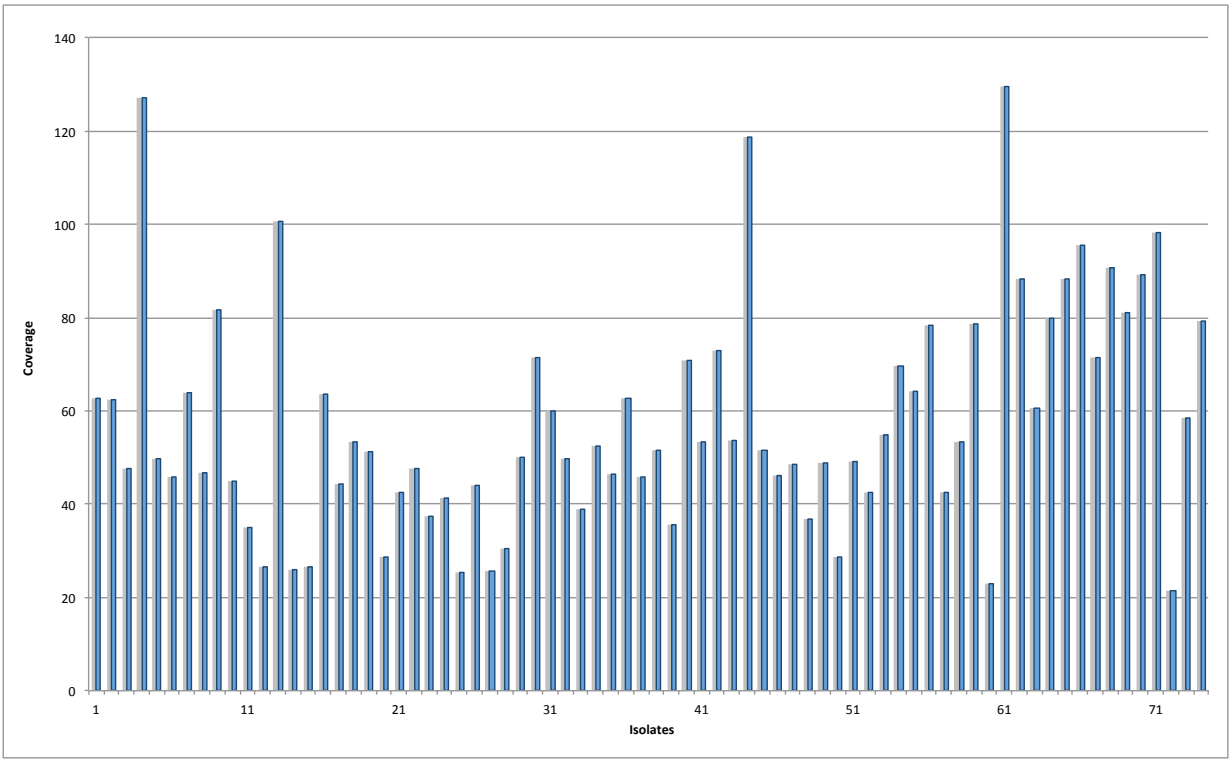


Figure 2.1: Example of coverage of *Y. enterocolitica* strains sequenced as part of the worldwide strain collection. Coverage of strains varies between strains and different runs due to experimental error and DNA quality, but has a minimum of 20X for each strain.

Data for the worldwide collection was assembled using Velvet *de novo* assembly for the environmental samples and assembly against reference BT 1B strain 8081 for *Y. enterocolitica* isolates.

2.2.2 Annotation and Sequence Analysis

The reference genomes for each of the biotypes were annotated and analysed in detail. Algorithms used for gene prediction and characterization of CDSs are given in Table 2.3. Glimmer 3 was used for prediction of coding sequences (CDS; Salzberg, Delcher, Kasif and White, 1998; Delcher, Bratke, Powers and Salzberg, 2007), and searches for motifs (Nielsen, Engelbrecht, Brunak and von Heijne, 1997; Brendtsen, Nielsen, von Heijne and Brunak, 2004; Krogh, Larsson, von Heijne and Sonnhammer, 2001; Finn, et al., 2010; Sigrist, et al., 2002; Hulo, et al., 2006) were employed to characterize potential proteins as detailed in chapter 3.

Table 2.3: Function and basis of the algorithms used for gene prediction and characterization.

Algorithm	Function	Basis
Glimmer 3	Model for prediction of coding sequences (CDSs)	DNA sequence contains 6 reading frames, probability calculation and modelling of most likely open reading frame
TMHMM	Prediction of transmembrane domains	Transmembrane domains will contain hydrophobic amino acids for integration into the lipid bilayer
SignalP	Prediction of signal domains	Proteins transported to the cell membrane contain signals for trafficking and direction
HTH	Prediction of helix-turn-helix motifs (DNA binding proteins)	Probability calculation that a short given stretch of amino acids has the properties to form a tertiary structure in the form of two linked helices
Pfam	Prediction of protein families and domains	Indication for protein functionality by domain comparison with well-studied, functionally assigned examples
Prosite	Prediction of protein families and domains	Indication for protein functionality by domain comparison with well-studied, functionally assigned examples

Comparisons between BT 1B strain 8081 and the newly sequenced BT reference genomes were done using the Artemis Comparison Tool (ACT; Carver, et al., 2005). Regions differing from BT 1B received manual annotation in Artemis (Berriman and Rutherford, 2003). Annotation was based on basic local alignment search tool (BLAST) searches (Altschul, et al., 1990) as well as information gained from motif searches.

Between different sequence versions, the annotation was transferred using an in-house perl script. This script will perform a blast search with the old FASTA file and compare it to the new FASTA file. Matches within CDSs are transferred into a new annotation file, which can be read into the new FASTA file using Artemis. The annotation transfer can be combined with the Glimmer 3 prediction, so that CDSs are highlighted, which have not yet received annotation.

Homologous genes are called orthologous if they have been present in the last common ancestor before speciation. They are propagated vertically via inheritance from mother to daughter cells. The differences observed relate to the evolutionary distance during evolution. As orthologous genes have been present in the ancestor, their genomic context is conserved between different isolates. In contrast, paralogs arise following duplication of genes, and therefore have a different genomic context and differ with respect to distribution in isolate. One common family of paralogs are insertion sequence (IS) elements, and copy number can vary between isolates.
To identify orthologs between the different reference genomes, another in-house perl script was applied. The script performs a homology search of CDSs with 40% homology over 80% of the length of the CDS (Figure 2.2). CDSs are compared pairwise, that is only those are retained as orthologs in which a search of genome 1 against genome 2 yields an ortholog and in which the reciprocal search of genome 2 against genome 1 yields the same ortholog.

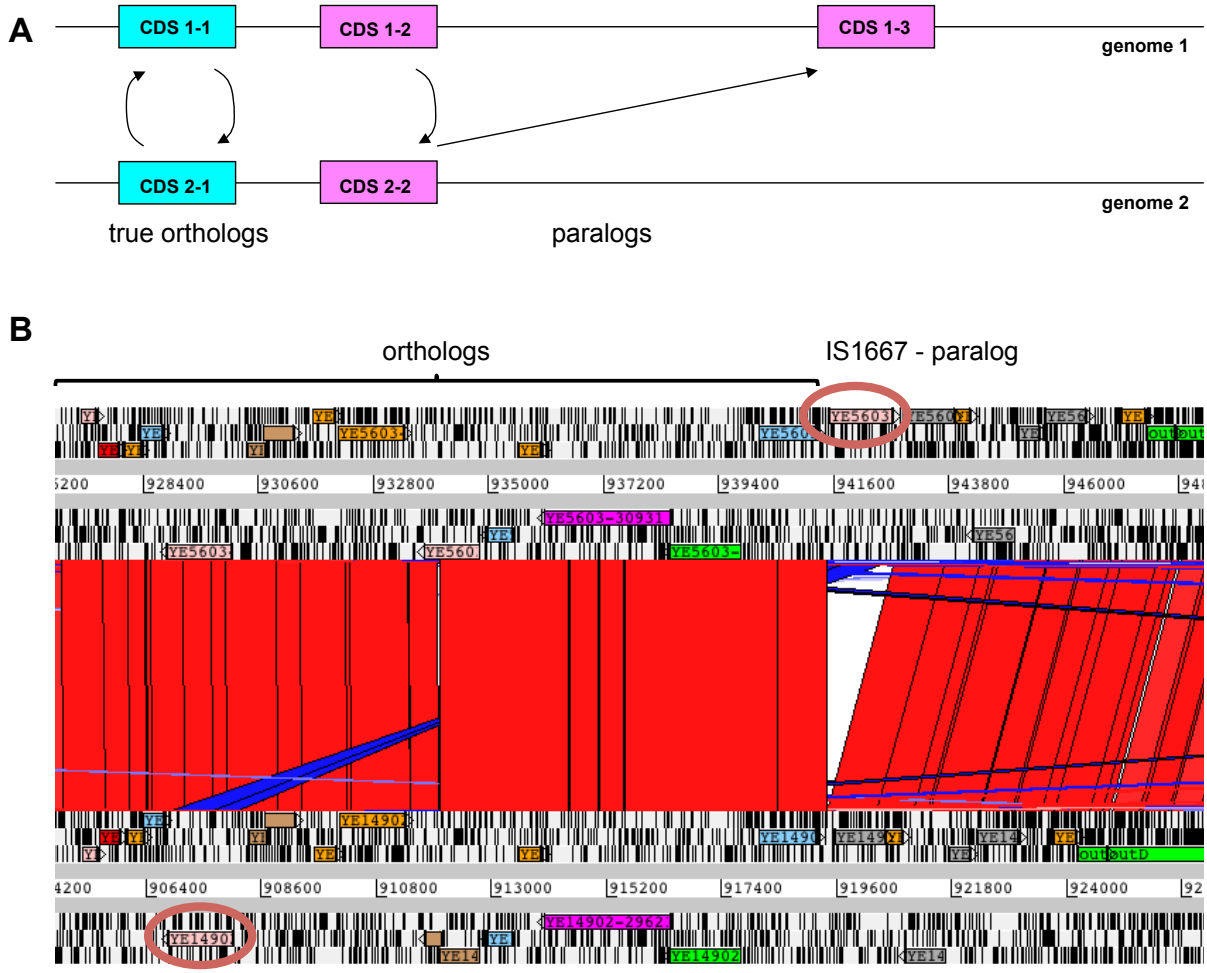


Figure 2.2: Identification of orthologs and paralogs. A. Reciprocal identification of blue CDSs as orthologs. Pink CDSs are not reciprocally identified, CDS 1-2 and CDS 2-2 are orthologs, and CDS 1-3 is a paralog to CDS 2-2. B. Genome example. Most genes are orthologous in both genomes, except an additional copy of IS1667 inserted in the top genome. This copy is paralogous to IS1667 found in the bottom genome, but is potentially misidentified in the reciprocal FASTA search.

The results are comprised in a format readable in Artemis. The identified orthologs need to be checked manually, as pseudogenes might be below the 80% length cut-off, and as IS elements may have been identified wrongly as having orthologs. For this, the orthologs of different combinations of reference genomes were firstly combined in a single annotation file. Then, an in-house perl script was employed to look at genomes in 3-way comparisons (Figure 2.3).

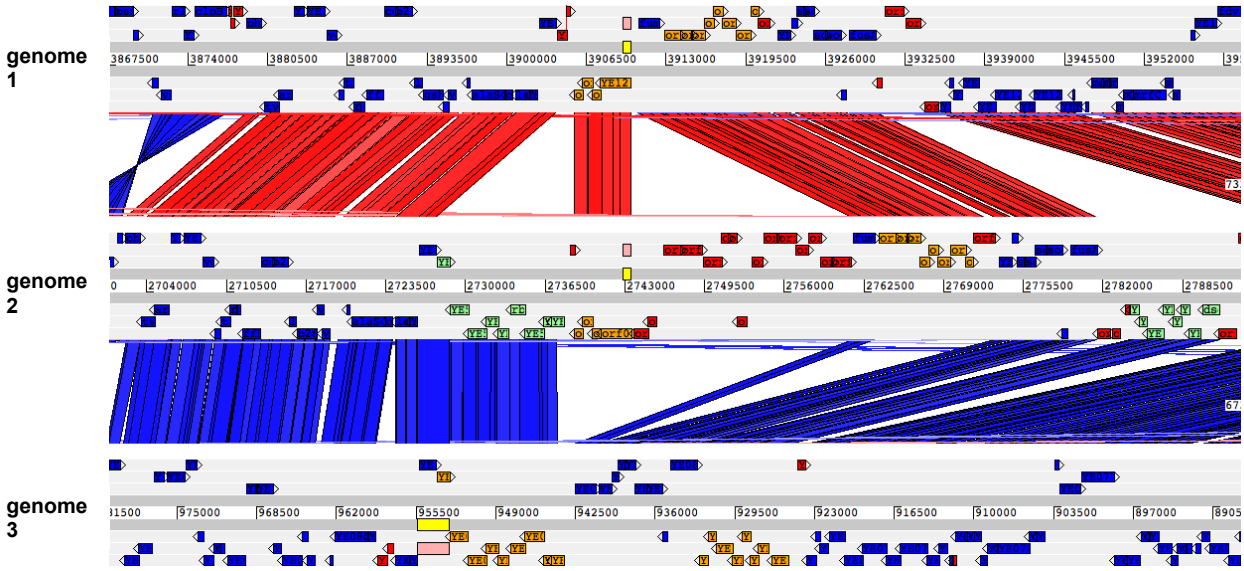


Figure 2.3: Example of a 3-way comparison for identification of orthologous gene sets. Genes present in genomes 1, 2, and 3 are blue. Genes unique to each genome are red. Genes orthologous in genome 1 and 2 are orange in both genomes. Genes orthologous in genome 2 and 3 are green and orange, respectively. Genes shared in genomes 1 and 3 would be green, but are not contained within this comparison.

For each 3-way comparison, CDSs are sorted in four categories and colour-highlighted accordingly: CDSs which are unique to each of the three genomes are highlighted in red, CDSs which are shared in all three genomes are highlighted blue, and CDSs which are shared between only two of the three genomes are highlighted either green or orange. The colouring of CDSs according to the four groups allows for quick identification of those CDSs that have not been coded correctly. For instance, if a red CDS occurs within a region conserved in all three genomes, this can easily be identified and resolved.

The perl script used for identification and colouring of orthologs was then modified to allocate groups instead of colours to specific gene sets (see Appendix 9.5, p. 214). This was done to identify CDSs belonging to the backbone, as well as CDSs that might be shared only within subsets or specific lineages of *Y. enterocolitica* BTs. Sandra Reuter carried out this modification, whereas the Pathogen Genomics group at the Wellcome Trust Sanger Institute provided the other in-house perl scripts.
The information on orthologs was also used to generate circular diagrams, either using the DNA plotter option in Artemis, or an in-house PERL script for high-resolution pictures.

Gene and protein alignments were generated using ClustalW2 (Larkin, et al., 2007) in the Seaview software (Galtier, Gouy, and Gautier, 1996; Gouy, Guindon, and Gascel, 2010).

2.3 Plasmid profiling

To establish a plasmid profile selected strains, the method of Kado and Liu (1981) was used.

2.3.1 Reagents

E buffer

40 mM Tris Base (4.85g in 1 l, 121.14 g/mol; Fisher Scientific)

2 mM Na_2EDTA (0.744 g in 1 l, 372.24 g/mol; Sigma)

Tris base and Na_2EDTA were dissolved in 1 l of distilled water. The pH was then adjusted to 7.9 using glacial acetic acid.

Lysis buffer

0.6055 g Tris Base were dissolved in 40 ml distilled water. 30 ml of 10% SDS (Sigma) were added. Using 8 g in 100 ml distilled water, a 2M solution of NaOH (Fisher Scientific) was made up. To raise the pH, 3.2 ml 2M NaOH were added to the Tris base/SDS solution, and the total volume was adjusted to 100 ml.

2.3.2 Protocol

A 5 ml overnight culture in Luria-Bertani (LB) broth (Fisher Scientific) was set up for strains to be examined. *Y. enterocolitica* was grown at 25 °C, *Salmonella* was grown at 37 °C.

2 ml of the overnight culture was pelleted for 5 minutes at 4,000 rpm at room temperature using a bench-top centrifuge (Hettich Mikro 2000). The pellet was resuspended in 150 µl E buffer. 300 µl lysis buffer were added and gently mixed to observe clearing indicating lysis. The sample was then heated for 60 minutes at 55°C using a heat block. After this incubation, 600 µl of phenol-chloroform (1:1 ratio saturated with Tris; Sigma) were added and gently mixed. To separate the layers, samples were centrifuged for 30 minutes at 14,000 rpm at room temperature in a bench-top centrifuge. From the upper watery layer, 35 to 45 µl of sample was removed and added to 5 µl of loading dye (Promega). For the markers, *Salmonella*

strains SL1344 and PT4 as well as *Y. enterocolitica* strains 5303 and 21202 extractions were mixed in 1:1 ratio.

Visualization was carried out on 0.7%-agarose gels with E buffer containing 10 mg/ml ethidium bromide. Gels were run for one hour at 30V, then a further four hours and 30 minutes at 45 V for best separation.

2.4 Phenotypic Microarray

2.4.1 Experimental setup

The phenotypic microarray experiment was carried out with the help of Theresa Feltwell (WTSI). Bacterial strains were cultured on LB agar plates at 25°C prior to starting the experiment. Inoculation and preparation of the 96-well plates used for phenotypic characterization was carried out as described by the manufacturer (Biolog Inc., Hayward, U.S.A.). PM1, 2A (carbon), PM3B (nitrogen), and PM4A (phosphorus, sulphur), PM9 and PM10 were chosen.

The workflow is depicted in Figure 2.4.

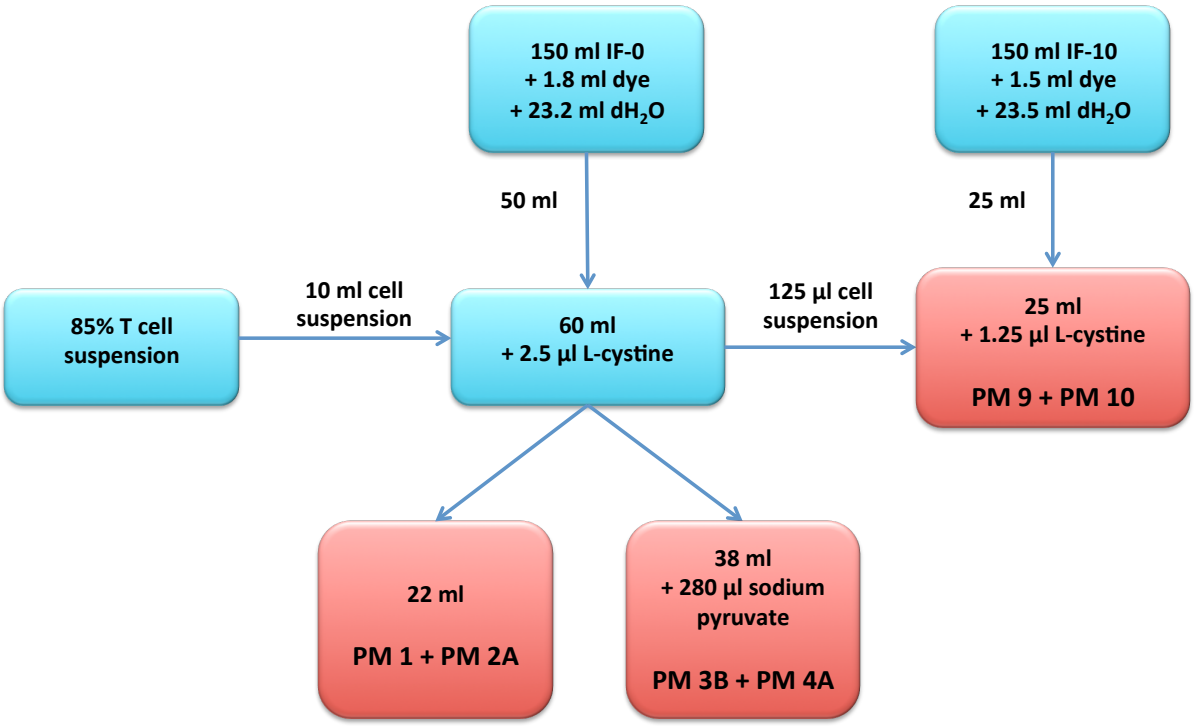


Figure 2.4: Preparation of fluids for phenotypic microarray.

In detail, the inoculation fluids IF-0 for PM 1 – 4 and IF-10 for PM 9 and 10 were split into sterile flasks containing 50 ml each, and sterile water and tetrazolium dye A mix

for Gram negative bacteria was added. Additionally, to each IF-0 flask L-cystine, the homodimer of L-cysteine, (12.5 µM) was added because *Yersinia* needs a reduced from of sulphur for growth. The cell suspension for each strain was prepared by running a sterile swab over the agar plate and transferring the bacteria into a sterile tube containing IF-0. A uniform suspension was created until a turbidity of 42% transmittance (T) ± 2% in the Biolog turbidimeter was reached. 10 ml of this 42% % T cell suspension was then added to 50 ml of IF-0 + dye A to yield a final cell density of 85% T. 125 µl of the 85% T cell suspension was then transferred to a flask containing 25 ml IF-10 + dye A. This suspension was then used to inoculate PM9 and 10, using a volume of 100 µl/well. 22 ml of the 85% T cell suspension in IF-0 + dye was used to inoculate PM 1 and PM 2A with 100 µl/well. To the rest of the 85% T cell suspension 280 µl of 2M sodium pyruvate was added, and PM 3B and PM 4A were inoculated with 100 µl/well. The plates were then placed in the OmniLog incubator/reader, and incubated for 48h, taking readings every 15 min. The experiment was carried out in triplicate for phenotypic characterization at 28°C and 37°C.

2.4.2 Analysis

Evaluation of phenotypic microarray (PM) data has not been standardized. The software that comes with the OmniLog instrument offers a number of variables that can be automatically calculated from the kinetic plots, such as maximum height, average height, area under the curve, and slope. It is also possible to manually select wells that visually stand out.

An example of the current data set is given in Figure 2.5, showing the kinetic plots for the non-pathogenic BT 1A and the high-pathogenic BT 1B. Higher metabolic activity in BT 1A is seen in green, higher activity of BT 1B in red, and the overlapping areas of the curves are shown in yellow.

One challenge with PM data is to determine a valid point to call metabolism versus non-metabolism of nutrient sources. Considering for instance A08, A09, and E01: is there utilization in these wells, or is the activity so low as to be considered background variation?

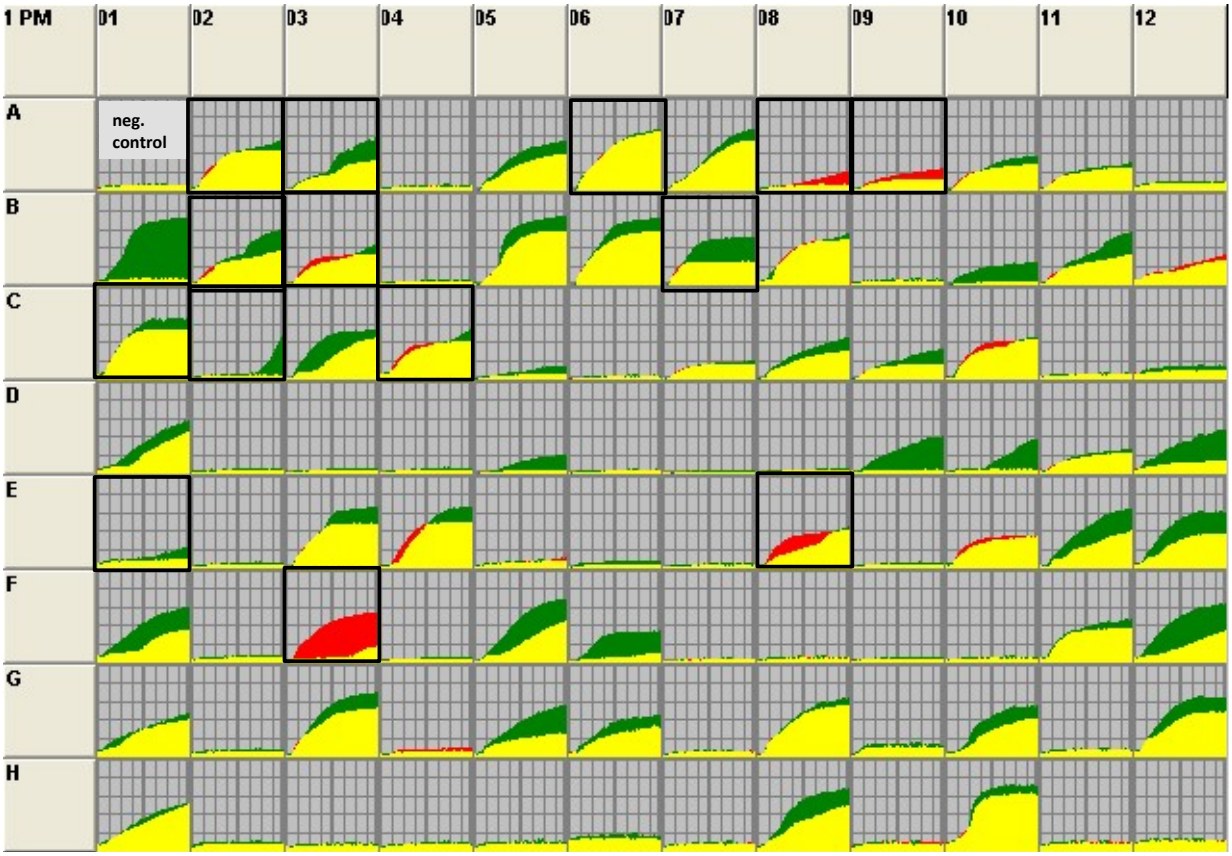


Figure 2.5: Example for utilization of carbon sources in BT 1A and BT 1B. A01 is the negative control. Overlapping curves are shown in yellow, BT 1A is shown in green, BT 1B is shown in red. Highlighted are wells described in the text.

The curves also show very different shapes. Some curves rise steadily right from the start such as A06. Other curves show a sigmoidal progression imitating Michaelis-Menten enzyme kinetics, where the curve rises quickly and then levels off as if there is a maximum turnover of this nutrient with a rate-limiting protein present, such as a transporter or a down-stream enzyme involved with conversion of the nutrient (B07, C01). Other curves resemble diauxic growth curves (A03, B02). How do you compare these curves and are the differences seen relevant or significant? One method would be to use the area under the curve, but there can be curves such as in A02, B03, and C04 with slightly different progression yet same area under the curve. One strain metabolises the nutrient earlier, the other strain utilizes it to a higher degree. Another method would be to look at the endpoint of the measurements. C02 shows a curve that shows no utilization in the strains for most of the measurements and only gives a signal in the last hours. Is this to be treated the same as other curves that show a steady utilization over extended periods of time? Comparison with E08 shows that both curves have the same endpoint, but their progression is entirely different. Also interesting to keep in mind is F03: Is there utilization in BT 1A as the

signal activity is detected very late, or is there no activity as the signal is not very strong?

We chose to adopt an approach by Homann, Cai, Becker and Lindquist (2008) and modified it to be able to apply statistical tests. This was done in collaboration with Lars Barquist (WTSI) who wrote the scripts to process the data and perform the calculations as well as produce heatmaps.
The data was exported from the Biolog File Manager, and further analysis was conducted in R. For each replicate, the average height and maximum height over 48h and the average height over 2h was extracted. The background control was then subtracted for each PM plate. The signal value was calculated using the following formula:

$$signal\ value = \left(\frac{[average\ signal\ over\ 48h + maximum\ signal\ over\ 48h]}{2}\right) - (average\ signal\ over\ first\ 2h)$$

This was done to firstly correct the average height for later data points, to take the progression of the curve into account and to reduce any background signals. A single number can therefore reflect the time course of metabolism.
Each plate showed a clear bimodal distribution of signal values. To call wells in which metabolic activity was taking place, normal distributions were fitted to the two peaks ("off" and "on") of the bimodal distribution. From this the log-odds score for each data point could be calculated, showing how likely each well was to have been produced by the "on" distribution. A conservative cut-off of four times more likely was used for further analysis. To examine differential metabolism between strains, the limma R/Bioconductor package was used (Smyth, 2004). All strains were compared to BT 1B strain 8081, and the Benjamini-Hochberg corrected p-values were used to determine statistical significance of differences controlling for a false discovery rate of 5%.

2.5 Phylogenetic Analyses

2.5.1 Phylogenetic reconstruction of *Y. enterocolitica*: whole genome alignment

The phylogeny of *Y. enterocolitica* was reconstructed using an approach from Harris, et al. (2010). Dr Simon R. Harris (WTSI) provided the necessary scripts. Sandra Reuter and Dr Thomas R. Connor (WTSI) both carried out parts of the workflow.

The genome sequence of BT 1B strain 8081 was used as a reference sequence. Regions containing phage were excluded, as they most likely constitute unique acquisition events for the strain and might lead to false positive single nucleotide polymorphisms (SNPs) when compared to related, but independently acquired phages. These SNPs are therefore non-informative in an evolutionary context. The reads obtained from Illumina sequence were mapped against this reference using SSAHA (Sequence Search and Alignment by Hashing Algorithm; Ning, Cox, and Mullikin, 2001). Strains with low mapping percentage (<50%) to the reference were excluded from the analysis. Candidate SNPs were compiled. A conservative approach was chosen to minimize false positive and false negative SNP calls by removing SNPs with low quality scores and by filtering for SNPs that are present in at least 75% of the mapped reads. For each strain, the polymorphisms identified in the previous steps replaced the reference sequence. Thus a SNP alignment was obtained.

For estimation of an evolutionary tree based on maximum likelihood RAxML was used (Stamatakis, 2006). The general time reversal (GTR) model was used to account for differences in rates of nucleotide substitutions, and gamma correction was chosen to model the different positions within a triplet codon. 100 random bootstraps were performed for node support. Again, Dr Simon R. Harris provided scripts for this. In a last step, the SNPs were reconstructed on the tree so that SNPs associated with specific branches can be identified. No outgroup was used for the *Y. enterocolitica* tree, but the tree was rooted in retrospect on the non-pathogenic BT 1A as this is the oldest lineage.

2.5.2 Phylogenetic reconstruction of the genus *Yersinia*: extended MLST scheme

For the phylogenetic relationship of the genus *Yersinia*, a concatenated sequence of 75,651 bases from 85 housekeeping genes (Table 2.4) was used. These genes were

identified in conserved, syntenic blocks in the *Y. enterocolitica* genome. The genes were chosen by Miquette Hall (NTU) on the basis of exhibiting 10-25% SNP divergence between *Y. pestis* and *Y. enterocolitica*.

Table 2.4: Housekeeping genes used for construction of the phylogenetic relationship within the genus *Yersinia*.

ID*	gene	ID	gene	ID	gene	ID	gene	ID	gene
YE0003	asnA	YE0919	gloB	YE2176	ppsA	YE2427	kdsA	YE3331	galR
YE0027	glnA	YE0926	nrdF	YE2209	trpE	YE2431	prfA	YE3374	lysS
YE0062	dfp	YE0927	nrdE	YE2254	xthA	YE2432	hemA	YE3376	prfB
YE0093	tpiA	YE1009	nadB	YE2257	topB	YE2477	chaA	YE3416	tktA
YE0134	sthA	YE1082	guaA	YE2261	ansA	YE2783	udk	YE3419	speA
YE0166	rhlB	YE1281	aroC	YE2281	dadA	YE2794	sfcA	YE3425	endA
YE0168	rho	YE1538	aroA	YE2283	fadR	YE2818	folE	YE3427	gshB
YE0251	udp	YE1640	tmk	YE2284	nhaB	YE2928	nadA	YE3656	parC
YE0258	aarF	YE1643	ptsG	YE2371	minC	YE2977	glnS	YE3673	rfaE
YE0263	hemB	YE1717	phoQ	YE2372	minD	YE3067	rosA	YE3681	gcp
YE0354	groEL	YE1718	phoP	YE2374	rnd	YE3088	hemH	YE3708	uxaC
YE0414	mdh	YE1720	purB	YE2383	zwf	YE3089	adk	YE3766	pyrB
YE0600	thrA	YE1725	icdA	YE2386	msbB	YE3202	proB	YE4132	fdhE
YE0610	dnaJ	YE1784	purT	YE2388	znuA	YE3308	argA	YE4133	fdoI
YE0746	pyrG	YE1793	pip	YE2389	znuC	YE3317	thyA	YE4151	glyQ
YE0773	pcm	YE1887	kduD1	YE2390	znuB	YE3318	lgt	YE4170	gyrB
YE0829	recA	YE2114	tyrR	YE2397	aspS	YE3324	tas	YE4206	atpD

* systematic ID respective to BT 1B strain 8081.

The respective gene sequences for the MLST tree were extracted using a script that performs a tblastn search on an assembled query sequence using a translated nucleotide subject to search for a matching amino acid string. The genes were concatenated into a single super-locus for each strain. Similarly to the phylogeny of the species, the alignment of genes was used to estimate a tree with RAxML based on the GTR model with gamma correction and bootstrapping (Stamatakis, 2006). Dr Thomas R. Connor was responsible for this part of the project. Bayesian analysis of population structure (BAPS; Cheng, et al., 2011) was used to identify clusters of homologous sequences within the 85 housekeeping genes, and formed the basis for the proposition of the new species clusters. Dr Thomas R Connor performed this work as well.

2.5.3 Phylogenetic reconstruction of the virulence plasmid pYV

To determine the phylogenetic relationship of the virulence plasmids, the same approach as for *Y. enterocolitica* was used. Both pYV8081 and pYV5604 were used as references to map Illumina reads using SSAHA in separate approaches. Alignments were created and trees were estimated using RAxML based on the GTR gamma model with 100 bootstraps. Both approaches gave the same tree topology. Dr Thomas R. Connor and Sandra Reuter both performed parts of this experiment.

2.5.4 Serotyping from sequence data

As some of the operons for the serotypes of *Y. enterocolitica* were either known or were identified during the annotation of the reference genomes, this information was used to perform *in silico* serotyping on the larger strain collection. This was done using an in-house perl script by Nicholas Croucher. This script is based on BWA mapping (Burrows-Wheeler Aligner; Li and Durbin, 2009), and will count the number of Illumina reads that map to a reference sequence – the operon of interest – and will return a percentage of mapping. This gives a good indication whether a specific operon is present or absent given the divergence of the different operons.

2.5.5 Phylogenetic reconstruction using 16S rDNA sequence

The 16S sequences were extracted from the Illumina reads by mapping against the BT 1B reference strain 8081. The sequences were then corrected using iterative correction of reference nucleotides (iCORN) to improve sequence quality (Otto, Sanders, Berriman and Newbold, 2010). Sequences for the previously published isolates of the environmental species as well as *Y. pestis* and *Y. pseudotuberculosis* were obtained from the National Center for Biotechnology Information webpage (NCBI; www.ncbi.nlm.nih.gov) and are based on high-quality Sanger sequencing.
Sequences were aligned using Seaview (Galtier, Guoy, and Gautier, 1996) and trimmed for same start and stop of the molecules. A phylogenetic tree was then calculated using the PhyML option in Seaview for a maximum likelihood tree, employing the GTR gamma model with 100 bootstraps.

2.5.6 Phylogenetic reconstruction including *Y. entomophaga* and *Y. nurmii*

The newly established species *Y. entomophaga* and *Y. nurmii* were not included in the large-scale sequencing approach. To determine their relationship with the other

Yersiniae, the deposited sequences for four housekeeping genes *recA*, *hsp60*, *gyrB*, and *glnA* (Hurst, et al., 2011; Murros-Kontiainen, et al., 2010b) were obtained from the NCBI website. These four genes were included in the 85 housekeeping genes, so this sequence information was extracted for the worldwide strain collection. The genes were concatenated into a superlocus and aligned using Seaview. Again, a phylogenetic reconstruction using PhyML in Seaview was carried out.

3 Sequencing and Analysis of the Reference Genomes for *Y. enterocolitica*

3.1 Introduction

At the start of the project, only one genome sequence for the *Y. enterocolitica* high-pathogenic biotype 1B had been published (Thomson, et al., 2006). This strain 8081 was isolated from a fatal human septicaemia case in America in 1980 (Portnoy, Moseley and Falkow, 1981). More recently individual genomes of other BTs have been published, and compared to the reference genome of BT 1B (Batzilla, Heesemann and Rakin, 2011; Fuchs, Brandt, Starke and Rattei, 2011; Wang, et al., 2011; Batzilla, et al., 2011a,b). New regions described are a novel T3SS, potential membrane transporters and associated metabolic pathways, a *toxin complex* pathogenicity island and other potential toxin clusters, and a second flagella cluster Flag-2 (Wang, et al., 2011; Fuchs, Brandt, Starke and Rattei, 2011).

Currently absent from the analyses are BT 5 strains. A comprehensive comparison of "all against all" is also still outstanding but necessary to define the "*enterocolitica*" core and the accessory genes that define the biotypes. A comparison between the biotypes is also required to determine whether the regions identified as unique for individual strains compared to BT 1B are the same as those described in the other studies and whether the regions are indeed functional or show pseudogene acquisition when compared with each other. The genotypic and phenotypic heterogeneity of *Y. enterocolitica* justifies further in-depth comparison non-, low- and high-pathogenic strains with each other. Strains in this study were chosen as representatives based on the prevelance of bioserotypes in the UK in humans and animals (McNally, et al., 2004; Tompkins, et al., 1999) and have been characterized genotypically and phenotypically (Howard, et al., 2006; Thomson, et al., 2006; McNally, et al., 2006).

Sequencing of representative strains for common *Y. enterocolitica* bio-serotypes of non- and low-pathogenic lineages should allow us to identify general characteristics of a *Y. enterocolitica* genome. The advantage of whole genome sequencing is that it gives information about genome organization and topology. It also not only identifies new regions but gives information on whether shared regions are orthologous, that is of the same evolutionary origin and thus conserved in terms of genomic context, or are paralogous, that is if convergent evolution took place. A comparison of all isolates should give a full representation of the *Y. enterocolitica* species.

3.2 Results

3.2.1 Sequencing, assembly and sequence improvement of the *Y. enterocolitica* biotypes genomes

The genomes of six *Yersinia enterocolitica* strains (1203, 5603, 5303, 21202, 14902, 3094; Table 2.1) were sequenced using multiple sequencing technologies. The work was carried out at the core genome sequencing facilities at the Wellcome Trust Sanger Institute (WTSI), Hinxton, UK. The sequencing process, assembly and improvement are outlined in Figure 3.1.

Initially, sequencing started on five representative strains (5303, 21202, 5603, 14902, 1203) of the predominant bio-serotypes 1A and 2 – 4 in the UK. The first round of sequencing was carried out using 454/Roche GS20/FLX shotgun pyrosequencing using 3 kb DNA fragment sizes. This produced in excess of 30-fold coverage of each genome. Polymerase chain reaction (PCR) products were designed to span gaps in contigs and sequenced to close gaps. This was automated by using algorithmic based automatic contiguation of assembled shotgun sequence (ABACAS; Assefa, et al., 2009), which was under development at the time. This process was supervised and carried out by Danielle Walker at the WTSI. The *de novo* assemblies for strains 5303 and 1203 were produced using the Newbler software (454 Life Sciences, Roche). Initial analysis and annotation was started at this stage.

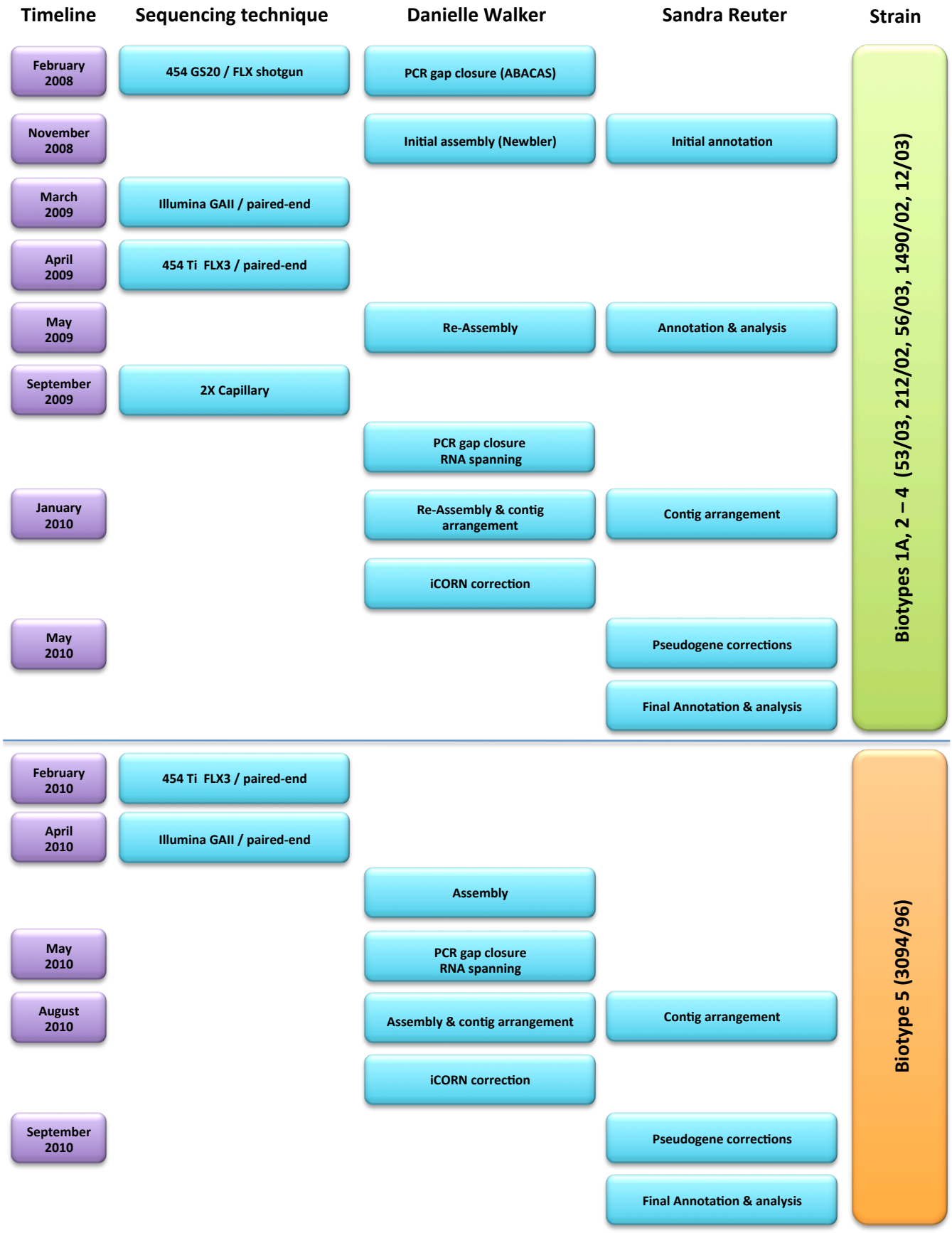


Figure 3.1: Sequencing timeline for the *Y. enterocolitica* reference genomes. The timeline is presented on the left and contrasted with the sequencing progress. The work carried out by Danielle Walker and Sandra Reuter is highlighted. The first five reference genomes are split from the last genome due to difference in sequencing progress.

Analyses were hampered by the fact that some regions of difference compared to the reference genome 8081 were at the end of contigs, thus making it difficult to assess the genomic location as well as the completeness of that region. Also, each of the genomes was present in more than 300 contigs. This was due to a large number of IS elements and repeats as well as several copies of the rRNA operon that are not spanned by a single-ended sequencing technique. The newly sequenced genomes appear o have the same genome architecture as the reference genome 8081, as the large number of contigs can be arranged in a co-linear fashion against the reference (Figure 3.2). The non-pathogenic BT 1A strain 5303 shows the largest genome size. Some gaps are visible which might be due to lower homology hits being filtered out.

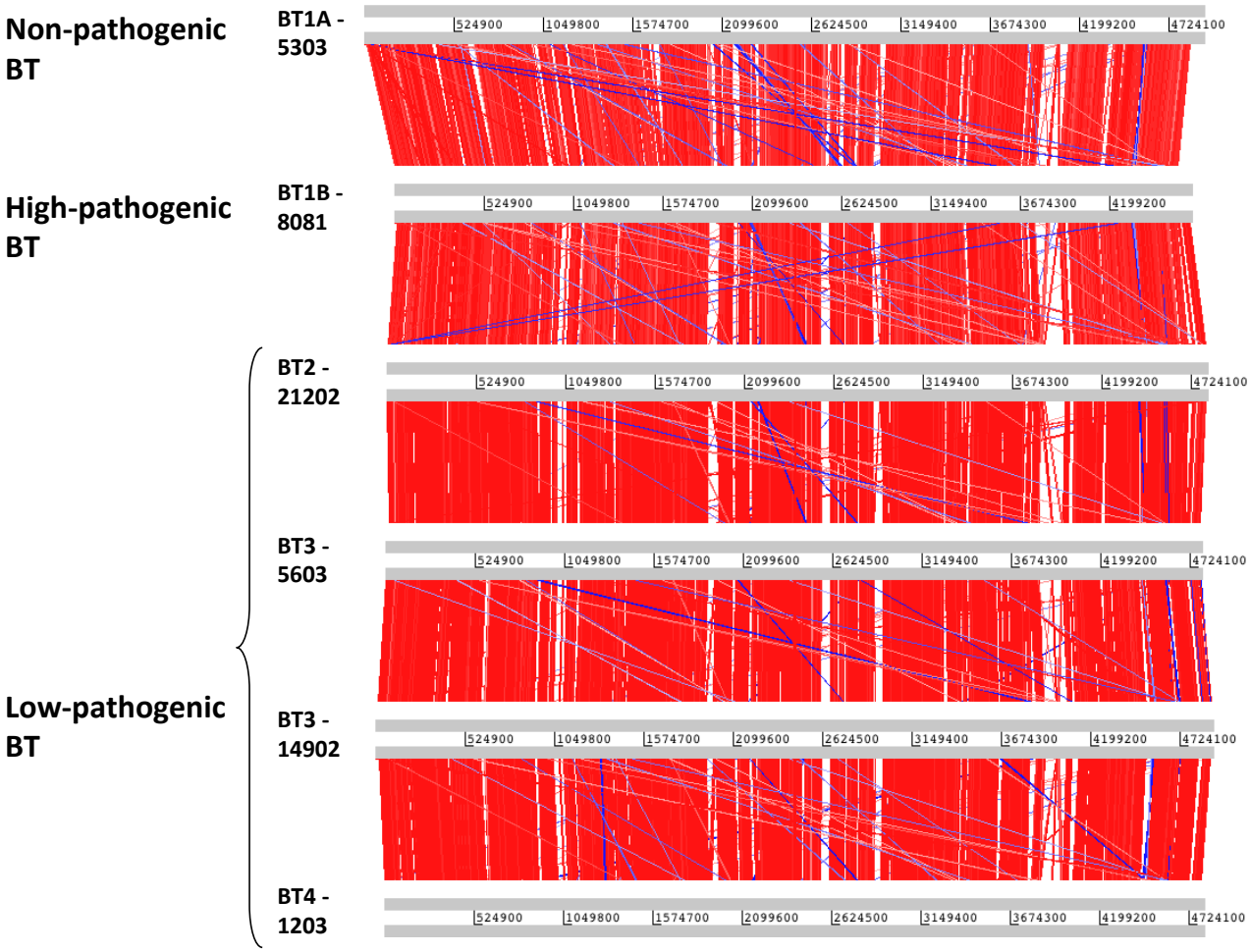


Figure 3.2: Initial assembly based on single-end 454 reads and comparison of the six *Y. enterocolitica* reference genomes. Genomes are compared in a pair-wise fashion. Red bars – co-linear homology >999 bp, blue bars – inverted homology >999 bp. Due to the large number, contigs are not highlighted.

It was then decided to undertake further sequencing. 454/Roche GSFLX Titanium (paired-end) was used to generate sequence in excess of 30-times coverage. It also

used 3 kb DNA fragment sizes, but employed long-read chemistry and paired-end technology (454 Life Sciences, 2008), which had not yet been introduced into the WTSI production pipeline when the first round of pyrosequencing had been carried out. To further improve the assemblies, the five strains were sequenced by more conventional Sanger sequencing using big-dye terminator chemistry on ABI3730 automated sequencers to approximately 2-fold coverage from pOTW12 (insert size 5-6 kb) and pMAQ1Sac_BstxI (insert size 9-12 kb) genomic shotgun libraries. These subsequent rounds of sequencing with the additional information on the distance of both paired-end reads and large insert size libraries were used to help link contigs. Thus the total number of contigs was reduced, as was the chance of mis-assemblies, because unambiguous placement of physically linked fragments was possible. Finally to improve accuracy of the sequence and reduce the possibility of generic errors of 454/pyrosequencing in homopolymeric tracts further sequence was generated using paired-end Illumina sequencing platform on the Illumina GAII Analyser with standard 200-300 bp fragments and 37 cycles of sequencing to give a coverage with just this technology in excess of 60-fold coverage. Danielle Walker again supervised this process.

The Illumina data was assembled, using the *de novo* Velvet assembler (Zerbino and Birney, 2008) with a kmer length of 31 for the 37 cycle runs. The consensus was then extracted and artificially broken down into 500 bp fragments. The GS De Novo assembly software (Roche, 2011) was used to assemble the 454 reads, capillary reads and the artificially fragmented velvet consensus as well as a round of automated PCR reactions designed on the 454 scaffolds using ABACAS. This was then converted for viewing and analysis in the STADEN – Gap4 software package (Bonfield, Smith and Staden, 1995).

The assembled consensus sequence for strains 5303 and 1203 was annotated and this analysis was also used to inform the gap closure process. With additional PCR-based gap closure and sequencing using a primer walking approach especially over large repeat regions such as rDNA operons, Danielle Walker was able to close gaps and scaffold large regions. However, it was still not possible to close all gaps and so for the 16S operons the consensus sequence of BT 1B strain 8081 was used. Finishing this part of the reference genomes was not a priority, and would have been too resource-extensive.

In addition to the gold standard, finished BT 1B genome, only the genome for BT 1A could be contiguated that is assembled into only one large contig. For all other

reference genomes the sequencing was finalized at a point were all the genomes were present in less than eight contigs (Table 3.1). The IS elements present had a great impact on the ability to assemble the genomes because IS elements can be present as multiple perfect repeats that are usually around 1,000 bp long and thus not spanned from unique sequence to unique sequence by paired-end sequencing reads. The number of IS elements varied between the biotypes with BT1A having the smalles number of IS elements and the low-pathogenic BT possessing a larger number. In addition there are 3-8 rRNA operons in the sequenced reference genomes. This also impacted on the assemblies, as the operons present a large repeat region that has recombinatorial potential.

The final assemblies of the low-pathogenic BT genomes were based both on information on rRNA operons linking contigs as well as on comparisons towards both BT 1B and BT 1A. As comparison towards BT 1A and BT 1B yields different contig orders, a preference was given towards BT 1A. This biotype contained the least number of IS elements and repeats, and thus gave the impression of having undergone less genome rearrangement and presenting an ancient lineage. Table 3.1 gives the contig order and the basis for that particular order.

Table 3.1: Contig ordering in the *Y. enterocolitica* reference genomes for BT 1A, 2 - 4.

Strain	Contig #	Size [bp]	Information
5303	1	4,940,199	Contiguated, cut for same start as 8081
21202	1	331,132	Cut for same start as 8081
	2	34,537	Flipped to fit with 5303
	3	106,274	Joins with #4 by read pair, both flipped to fit with 5303
	4	485,939	Joins with #3 by read pair, both flipped to fit with 5303
	5	2,189,202	
	6	1,011,144	Joins with #7 by read pair
	7	6,554	Joins with #6 by read pair
	8	400,227	First half from contig #1
5603	1	19,724	Cut for same start as 8081, flipped to fit 5303
	2	143,745	
	3	197,661	
	4	760,938	
	5	776,795	Joins with #6 by read pair
	6	1,405,148	Joins #5 and #7 by read pair
	7	860,219	Joins with #6 by read pair
	8	399,855	Joins with #1
14902	1	18,110	Cut for same start as 8081
	2	313,931	
	3	701,464	Joins with #4 by read pair
	4	360,673	Joins with #3 by read pair
	5	344,361	
	6	2,857,874	Joins with #1
1203	1	402,778	Cut for same start as 8081
	2	4,317	Based on read pairs
	3	1,725,059	Based on read pairs
	4	2,395,791	Based on read pairs

The 6th strain sequenced in this study, BT 5 strain SZ3094/96 (short: 3094), was sequenced using the 454/Roche GSFLX with long-read GS FLX Titanium chemistry to approximately 25-fold coverage using 8 kb DNA fragment sizes. Sequence was also generated using the Illumina sequencing platform on the Illumina GAII Analyser to approximately 200-fold coverage with standard 200-300 bp fragments and 54 cycles of sequencing. A *de novo* assembly was done for the Illumina reads using Velvet (Zerbino and Birney, 2008) with a kmer length of 45 for the 54 cycle run. Again the consensus was extracted, broken down into 500 bp fragments, and combined with the 454 reads using the GS De Novo assembly software (Roche, 2011). End-sequenced PCR products were used to close gaps and scaffold large repeat regions. This strain was not analyzed until it was in less than 10 contigs, and assembly was based both on information on rRNA operons linking contigs as well as on comparison to both BT 1B and BT 1A. Table 3.2 gives the contig order.

Table 3.2: Contig ordering in the *Y. enterocolitica* reference genome for BT 5.

Strain	Contig #	Size [bp]	Information
3094	1	1,118,114	Cut for same start as 8081
	2	36.977	
	3	262,567	
	4	27,479	
	5	3,220,765	

The final assemblies of the reference genomes after gap closure and corrections can be seen in Figure 3.3. There are striking differences to the initial assemblies of the 454/single-end run (Figure 3.2). Each strain shows several extensive inversions and even translocations compared to other strains. The assemblies have been trimmed for a common start, and thus share beginning and end, but apart from that no common theme of genome organization is visible. The genome rearrangements are visible within contigs, not only at the end of contigs, thus can be considered genuine. The differences to the initial assemblies can be explained partially by the reduced number of contigs in the final assembly. The initial assemblies were missing information on read-pairs. Not only did this result in a more fragmented assembly, the short contigs were then aligned and orientated according to the reference genome as information on linking contigs was missing. These mis-assemblies thus give the impression of co-linear organization of the reference genomes visible in the original comparison (Figure 3.2).

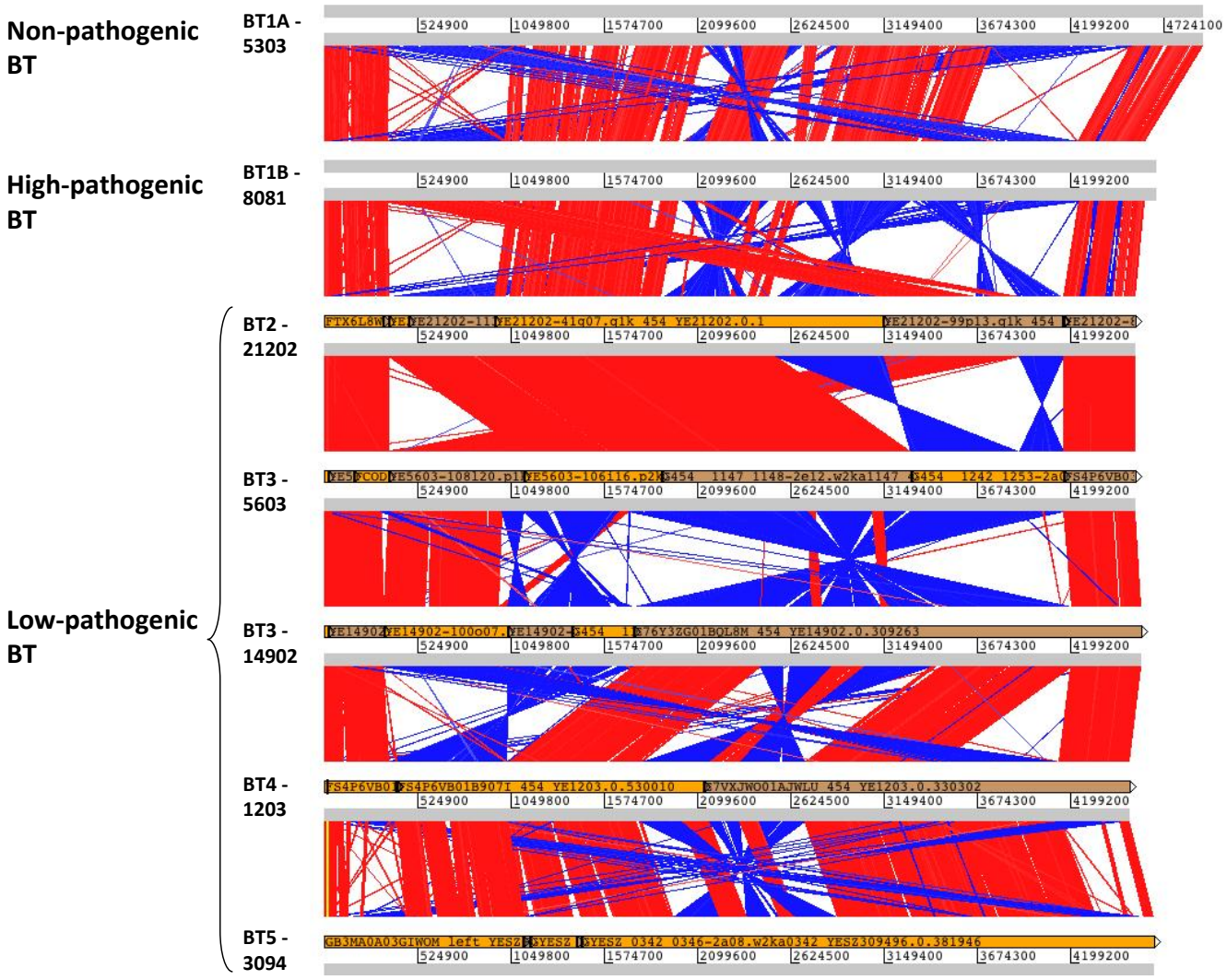


Figure 3.3: Pairwise comparison of final assemblies after gap closure of *Y. enterocolitica* reference genomes. Red bars – homologous, co-linear matches >1,200 bp, blue bars – homologous, inverted matches >1,200 bp, alternating orange/brown – contig borders. The non- and high-pathogenic BTs both consist of only one single contig.

Sequencing of the virulence plasmid pYV contained in BTs 2 – 5 as well as sequencing of unique plasmids of individual strains was carried out as for the genome of the respective strain. Danielle Walker and Sandra Reuter both manually designed primers for closing gaps in the plasmids.

Finally all assemblies were corrected using iterative correction of reference nucleotides (iCORN; Otto, Sanders, Berriman and Newbold, 2010). All strains were completed to an Improved High-Quality Draft Standard (Chain, et al., 2009), and consist of a circular genome with several contigs and varying plasmid numbers.

For further sequence improvement, all CDSs with homopolymeric runs resulting in frameshifts and stop codons introduced in the coding sequence (pseudogenes) were

checked against the original sequencing data for coverage and sequencing technique. If the sequence was questionable in terms of number of reads, or contained data from only one sequencing technique, PCR reactions were designed for improved quality and the sequence corrected accordingly. This was done for all functional classes except those of foreign origin (phage, IS elements) and hypothetical, unclassified proteins. This should present an accurate total number of the pseudogenes as well as give confidence in the nature of the mutations.
During this process, between two and 22 bases were changed in the reference genomes, yielding an error rate beween 0.00004% and 0.0005% equivalent to up to 5 wrong base calls in 1,000,000 bases. This is very much dependent on the sequencing technologies employed. For instance, reference genome YE3094 was sequenced only with pyrosequencing and Illumina technology, and thus has a larger number of errors in the sequence that were identified during subsequent analysis. There is an undetermined error rate with respect to base changes that do not result in immediately visible effects on open reading frames, but judging from the number of changes carried out upon pseudogene inspection it can be assumed to be low.

Other single genomes for *Y. enterocolitica* BT have been made available but vary in sequence quality compared to the reference genomes presented here. Two BT 1A strains have been sequenced to draft quality, resulting in 74 and 97 contigs (Batzilla, Heesemann and Rakin, 2011). The published BT 2 O:9 strain W22703 (Fuchs, Brandt, Starke and Rattei, 2011) has been classified as Annotation-Directed Improvement, but gap closure has not been carried out and only one sequencing technique has been employed. Thus the genome is presently in more than 300 contigs, making it difficult to do a full characterization of the genome. The published BT 3 O:9 strain 105.5R(r) (Wang, et al., 2011) might be classified as High-Quality Draft, since two sequencing techniques have been employed and some gap closure has been carried out so that the genome is present in 81 contigs. A detailed analysis of BT 4 O:3 strain Y11 has been published (Batzilla, et al., 2011a,b). Due to gap closure and quality control, this genome may be classified as Improved High-Quality Draft, and so has the same genome standard as the reference genomes generated for this study.

3.2.2 Genome annotation

The individual steps for preparation of a reference sequence for annotation are shown in Figure 3.4.

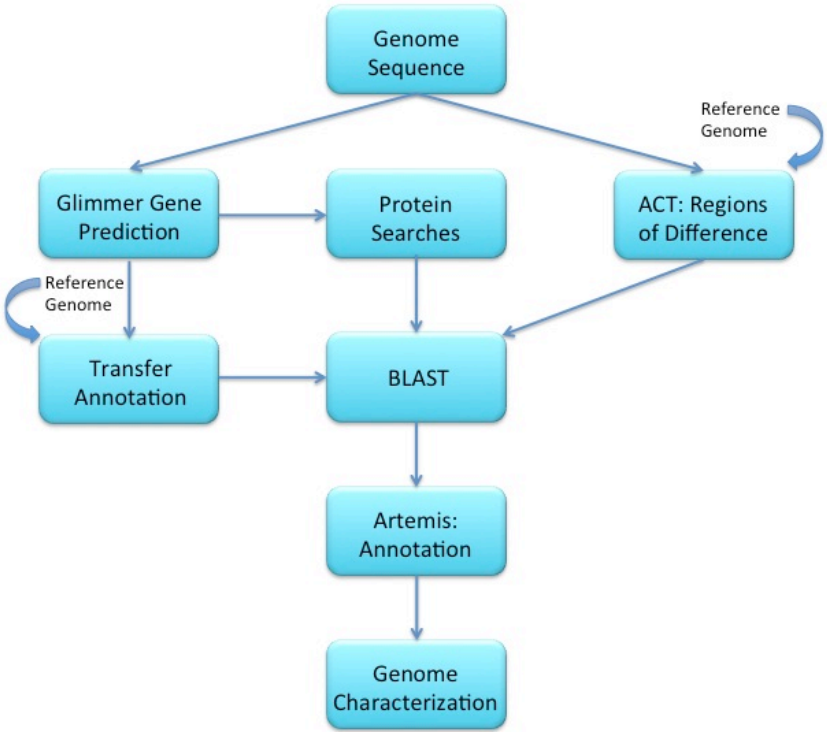


Figure 3.4: Flow chart of the preparation of a genome sequence for annotation and analysis. Gene prediction software is used for finding open reading frames. These are then characterized by protein searches, such as transmembrane domains, signal domains, etc. Annotation is transferred using a reference genome. Both the gene prediction and the glimmer prediction are combined and open reading frames without annotation are blasted and annotated using Artemis. The genome sequence is also visually compared to the reference genome using ACT. Following annotation, the genome can be characterized further with respect to overall composition and orthologous gene sets with other genomes.

First of all, Glimmer 3 was used as a model for prediction of potential open reading frames (Salzberg, Delcher, Kasif and White, 1998; Delcher, Bratke, Powers and Salzberg, 2007). To help characterization of potential new CDSs, searches for transmembrane domains (TMHMM; Krogh, Larsson, von Heijne and Sonnhammer, 2001), signal domains (SignalP; Nielsen, Engelbrecht, Brunak and von Heijne, 1997; Brendtsen, Nielsen, von Heijne and Brunak, 2004), helix-turn-helix motifs (HTH), protein family domains (Pfam; Finn, et al., 2010) and prosite (Sigrist, et al., 2002; Hulo, et al., 2006) were carried out. Using the Artemis Comparison Tool (ACT; Carver, et al., 2005), genomes were compared to the reference BT 1B strain 8081 and to each other to define large-scale regions of difference. The files needed for

comparing two genomes with each other were generated using an in-house bigger blast script, but can also be obtained from WebACT (www.webact.org).

The reference genomes were annotated by transferring annotation from the reference genome BT1B strain 8081. This script compares the reference genome to the subject genome and transfers annotation to CDSs that have more than 90% BLAST identity. The Glimmer prediction and the annotation transfer were combined, and unique regions received new manual annotation using Artemis (Berriman and Rutherford, 2003). All CDSs were blasted against prokaryotes using the UniProt knowledgebase (UniProt, 2011). Annotation was equally based on sequence similarity and indications for protein function.
Annotation was also transferred from one annotated genome to the other, since regions were identified that were absent in BT 1B strain 8081, but shared in the other genomes. The annotation transfer was also employed when transferring annotation from old sequence versions to updated assemblies.
For annotation, each CDS is given a set of qualifiers for identification. Apart from a unique locus tag or a systematic ID that is allocated automatically to each CDS, each CDS is characterized with a product, potentially a gene name if there are well-characterized homologs, a class and a colour. Class and colour qualifiers both reflect the product, but allow easier classification of CDSs into groups (see Table 9.3, p. 209, and 9.4, p. 213, in appendix). The 14 colours used reflect the function and sub-cellular location of a protein, e.g. green proteins are associated with the cell wall and membrane, such as permease or lipoproteins, whereas black proteins are involved in carbon energy metabolism. The class system fulfils a similar purpose, but breaks down the functions in more detail. Thus an exported protein may be classified as being in the inner membrane, or a protein involved in carbon metabolism can be allocated to anaerobic respiration. The product will very specifically give the protein name and function, but the class / colour system allows it to bring proteins of different names together under a common denominator.
Notes were added for further explanations if necessary, and homologs from other bacterial species were added.

Pseudogenes are defined as carrying mutations that would prevent the expression of the gene or would alter the nature of the resultant protein. This includes truncations of N and/or C termini, deletions of other parts of the gene, stop codons introduced

within the coding sequence and frameshifts. All pseudogenes except IS elements and hypothetical proteins were checked for sequence quality and confirmed as described above. IS elements were excluded as they are repetitive. Their function also might not only be in the expression of the transposase itself, but also lies in recombinatorial potential with the same class of IS element in other parts of the genome, which is independent of a functional transposase. Due to the large number of IS elements, their exclusion from confirming pseudogenes can be justified. Similarly, hypothetical proteins were excluded, as their transcription and function is unknown. Pseudogenes were annotated in a different way with labelling the product as pseudogene, partial, or remnant, stating in a note the nature of the mutation, setting the colour to 11 (brown) and adding the qualifiers “/pseudo” and “/partial” where applicable. This gives a set of filters that allows for easy identification of pseudogenes within a given set of CDSs. Pseudogenes are also made to end on the codon before the stop codon (“taken off the stop codon”). These modifications aid in downstream analysis of the genomes and simple tracking of pseudogenes.

RNA searches using Rfam (Griffiths-Jones, et al., 2003; Gardner, et al., 2011) were performed to establish the location of the RNA components. The location of the 16S and 23S rDNA operons was established in comparison with the reference genome of BT 1B strain 8081.
The resultant set of CDSs in the reference genomes was then used for detailed gene-by-gene annotation and analysis.

3.2.3 Genome comparison

For the six biotypes, it was not only interesting to compare strains in a pairwise fashion, but it was more important to see which sets of CDSs might characterize more than two biotypes. For this, orthologous gene sets were identified by reciprocal FASTA searches. This approach is shown in Figure 3.5.

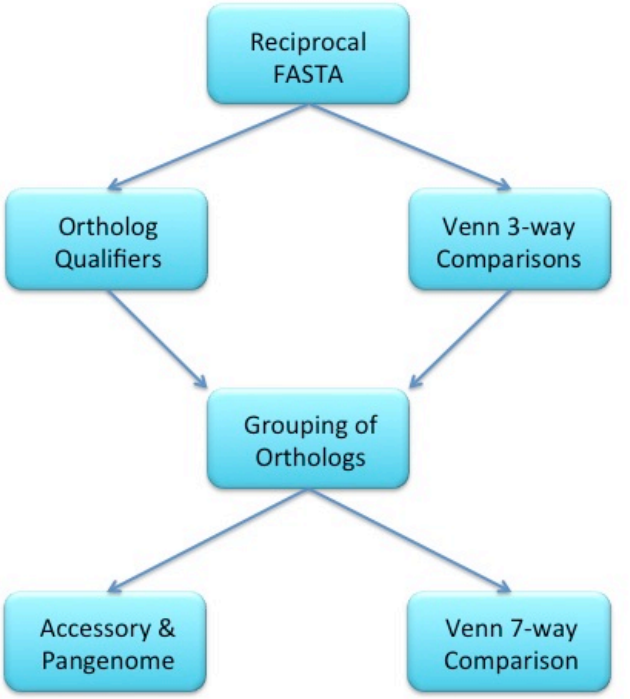


Figure 3.5: Identifying sets of orthologous CDSs. Reciprocal FASTA searches are carried out, the orthologs are combined in a file and manually curated using 3-way Venn comparisons. The identified orthologs can then be grouped. This establishes pan- and accessory genome and allows for a 7-way Venn comparison.

Only those pairs of homologous CDSs were retained for further analysis where the predicted amino acid identity was ≥ 40% over 80% of the protein length. These genes were then subjected to manual curation using gene synteny to increase the accuracy of this analysis. 3-way Venn diagrams were employed to facilitate manual curation. This curation is specifically needed for paralogous gene families such as IS elements, which can be numerous in the genome. Sequence divergence can inhibit direct identification of CDSs as orthologs but looking at genome synteny can help identify these. Truncated genes might also not be identified in the automated analysis as their length and identity can fall below the cutoff. This strategy was applied to pairwise comparisons of the genomes of all sequenced biotypes. These identified orthologs were then combined together with the annotation file.

3.2.4 General genomic features of the reference genomes

The Table 3.3 gives a general overview of the properties of the sequenced reference genomes for each of the biotypes of *Y. enterocolitica*. These basic figures can be used to characterize genome composition.

Table 3.3: 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 (3094)
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 67kb	pSR2-1A 8kb pSR3-1A 4kb	pYV 75kb pSR1-2 42kb	pYV 75kb	pYV 75kb	pYV 75kb	pYV 75kb pSR4-5 100kb
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
Total number pseudogenes	84	81	157	141	146	235	322
Number of partial genes	57	39	96	86	97	126	129
Prophage regions	4	6	2	2	3	2	3
rRNA operons	7	6	3	3	4	8	7
tRNA loci	81	77	72	72	71	72	67
IS elements	60	31	87	85	79	101	110

* numbers updated from Thomson et al., 2006

The genome sizes of the reference genomes vary between 4.53 and 4.94 Mb, with BT1A showing the biggest genome. Relating to the genome size is the number of CDSs, which is higher in larger genomes. The G+C content of all strains is similar and is about 48%. The coding density lies between 80 – 85%, with BT 5 having the lowest and BT 1A having the highest coding density. The average gene size across all biotypes was 970bp.

Comparing the reference genomes to the other five published *Y. enterocolitica* genomes, one can see that they are similar to the results obtained here. Limited information is available on the two BT 1A strain (Batzilla, Heesemann and Rakin, 2011), but their genome size estimated at 4.7 Mb and GC content of 47% are similar to BT 1A strain 5303. Due to their status as draft genomes, a detailed comparison of CDS numbers, pseudogenes, phages, etc. is not possible.

Genome size for BT 2 O:9 strain W22703 (Fuchs, Brandt, Starke and Rattei, 2011) is given as < 4.75 Mb, which is in the same range as the 4.56 Mb obtained for the BT 2 O:9 reference strain 21202. The GC content and coding density of W22703 (46.9%, 84.4%) are slightly lower than 21202 (48.1%, 85.9%). The number of CDSs in

W22703 exceeds the number of CDSs in 21202 (> 4,003, and 3,922, respectively), but given the high number of contigs it could be that some CDSs are in fact split onto different contigs. It is more likely though that the higher number of CDSs reflects the larger genome size. It is not possible to compare the number of rRNA operons nor the number of IS elements as these are likely to present the gaps between the contigs. Both BT 2 O:9 strains carry two prophages, but it cannot be assessed whether these present the same prophage acquired by the BT 2 O:9 lineage or independent phage infections.

The genome size, GC content, coding density, and number of CDSs of BT 3 O:9 strain 105.5R(r) (Wang, et al., 2011) are all comparable to the BT 3 O:9 strain 5603. The number of tRNAs are exactly the same (71). The number of rRNA operons in 105.5R(r) is higher than in 5603 (7, and 4 respectively), whereas the number of IS elements is lower than in 5603 (67, and 79 respectively). 85 pseudogenes are reported for 105.5R(r), which is considerably lower than the 146 pseudogenes found in the reference strain 5603. It could be the definition of a pseudogene that causes this discrepancy between the values. The number of partial genes and gene remnants in 5603 is 97, which is closer to the number stated for 105.5R(r). It could be that Wang et al. (2011), did not consider genes with frameshifts and stop codons as pseudogenes as the complete gene sequence is present.

The genome size and GC content of BT 4 O:3 strain Y11 (Batzilla, et al., 2011a,b) are similar to the reference genome for BT 4 O:3 strain 1203. The number of CDSs (4,355) is higher than for 1203 (3,893). The coding density thus is higher for Y11 than for 1203 (85.0%, and 80.7% respectively). The number of tRNA and rRNA operons in Y11 (70, and 7 respectively) is close to the numbers present in 1203 (72, and 8 respectively). The number of pseudogenes as obtained from the available NCBI entry (accession number FR729477) is considerably lower for Y11 (92) than for 1203 (235), similar to the discrepancy observed between the two BT 3 strains. Again, it could be due to the definition of a pseudogene, but even only considering the 126 partial genes and gene remnants in 1203, this number is higher than that for Y11.

Several plasmids were identified in the reference strains with the virulence plasmid being common to all pathogenic BT. The variations in the virulence plasmid and the unique plasmids found are described in more detail in chapter 5.

In summary, the reference genomes are similar in general composition, with an average genome size of 4.6 Mb ± 0.13 Mb, a GC content of 48% ± 0.3%, and ~4,000 CDSs. Significant variations arise in the number and distribution of IS elements and pseudogenes.

3.2.5 General characteristics of reference genomes

Using the annotation information contained in the assigned class qualifier (see appendix classification scheme table 9.1) for each gene the overall distribution of classes for the whole gene set within the genome could be determined for each of the biotypes (Figure 3.6). The classes of CDSs were combined in 20 different functional groups (see Table 9.1 in appendix).

All bio-serotypes show a similar distribution of CDSs according to class. About a quarter (23.6 – 26.0%) are proteins that have no assigned function or that are hypothetical, unconfirmed open reading frames. The predominant assigned functions are related to transport proteins (11.7 – 12.3%), cell envelope (14.8 – 16.6%), foreign DNA such as phages and IS elements (6.6 – 9.4%), regulation (6.5 – 6.9%), and macromolecule synthesis (5.7 – 6.0%).

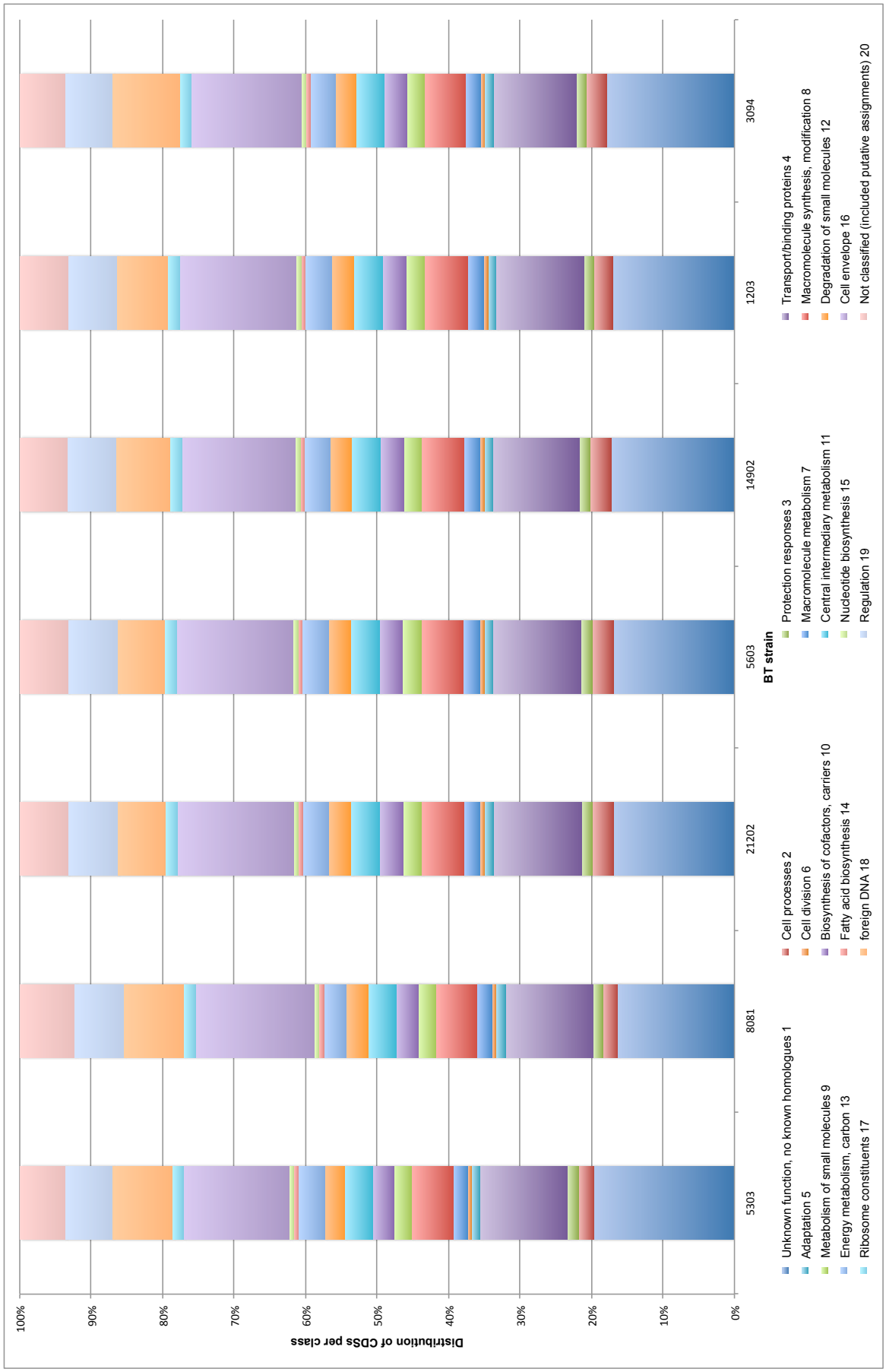


Figure 3.6: Distribution of classes amongst CDSs of all reference genomes. The information on class of a CDS was extracted and individual classes were combined in 20 different functional groups. The functional groups represent the spectrum of activity that can be performed by a bacterial cell. The CDSs in each group were combined to give a percentage of the total number of CDSs.

3.3 Discussion

3.3.1 Generation of Improved High-Quality Draft genome sequences

The generation of the reference genomes in this study tracked the technological advances in next-generation sequencing technologies.

The initial sequencing and analysis was based on a single-end run of pyrosequencing. Although the genome could be assembled into contigs, and gave an approximate size for each genome, the assembly was incomplete and often the interesting regions of difference were at the end of contigs making it difficult to determine the true nature of for example pathogenicity islands or to make statements on the completeness of an operon.

One factor that made assembly of the *Y. enterocolitica* genomes difficult was the presence of large numbers of perfectly repeated IS elements, especially IS1667, in the biotype genomes. At the start of sequencing it was not known that this insertion sequence element was present in over 30 copies in each of the low-pathogenic BT. Other IS elements are also present, plus several copies of the 16S and 23S rRNA operons, which also represent perfect extended repeat regions. Similar numbers of IS elements are also present in the previously sequenced BT 1B strain 8081 as well as in *Y. pseudotuberculosis* and *Y. pestis* (Thomson, et al., 2006). Each *Yersinia* genome contains 6-7 copies of the rRNA operon, and *Y. pestis* has on average more than 100 IS elements of varying classes. Using *Y. pestis* as an example, Pop and Salzberg (2008) hypothesize that sequencing the plague bacterium with short reads would result in obtaining data for only 93% of the genome, with a very fragmented assembly. The length of the read will not be able to bridge the repetitive elements, thus the importance of linking information on either side of the repeat reads is stressed (Pop and Salzberg, 2008). In particular *de novo* assembly is impossible without read-pair information (Chaisson and Pevzner, 2007; Chaisson, Brinza and Pevzner, 2009). Chaisson, Brinza and Pevzner (2007) could even show that the information on read-pairs is more important than an increase in read length.

As a result of this it was then decided to initiate further sequencing. Both Illumina and 454 technologies by that time had the possibility to perform paired-end runs, thus giving additional information on contig linking and scaffolding of sequence reads.

A low coverage Sanger sequencing run was also performed as this was the technology with longest reads and established paired-end technology, thus it was

hoped to improve sequence quality and assembly. As was later seen in the sequencing of BT 5, this would have been redundant, as a combination of paired-end pyrosequencing and paired-end Illumina sequencing produced the same result. The knowledge on read-pairs was sufficient enough in this case to assemble BT 5 to a high quality. Gap closure was still needed as with the other genomes, but a Sanger sequencing run was not necessary.

Similar findings on the assembly of genomes from next-generation sequencing data have been shown before. In one early study, pyrosequencing was combined with Sanger sequencing (Goldberg, et al., 2006). At this point only single-end reads were available for 454 sequencing as the paired-end protocol was to be released in early 2008 (454 Life Sciences, 2008). They found that pyrosequencing data would supplement 5X coverage Sanger sequencing in terms of increasing coverage, and gap reduction. The pyrosequencing reads also reduced regions with poor sequencing due to avoiding amplification bias that is giving in cloning into a vector (Goldberg, et al., 2006). Thus an even coverage of the genome is given. The Sanger sequence, on the other hand, would have balanced the problems of pyrosequencing in homopolymeric runs.
A later study (Aury, et al., 2008) investigated the combination of Sanger, pyro- and Illumina sequencing. They found that with paired-end 454 technology and Illumina sequencing a Sanger sequencing run is dispensible. They suggest using pyrosequencing with paired-ends as the first strategy for assembly due to it having the longer reads. The Illumina reads are then projected or mapped onto the 454 scaffold and can be used to correct the errors in homopolymeric runs. In this way it is possible to circumvent the use of expensive Sanger sequencing (Aury, et al., 2008). The implication of using several sequencing technologies to balance the individual faults of each is also approached by Harismendy, et al. (2009). They compared error rates, coverage, and variant discovery of 454, Illumina, SOLiD, and Sanger technology in the human genome. They found that although the next-generation sequencing platforms produce large amounts of data, not all of it passes the quality filter, and not all of the data passing the quality control will actually map back to the target genome. They suggest taking this into account for coverage calculation. The coverage, although high, can also vary depending repeat sequences in the target sequence. The next-generation technologies showed variation in coverage, but the non-uniform coverage was different in the three technologies compared

(Harismendy, et al., 2009). Thus a combination of technologies might counteract the coverage variation of each other. They also found that the sequencing accuracy is high in all next-generation platforms. Pyrosequencing had the lowest false positive rate for variant discovery, whereas the lowest false negative rate was for Illumina technology.

The reference genomes created can be classified as Improved High-Quality Draft Standard according to Chain, et al. (2009). They confirm with manual and automated gap closure, reduction of number of contigs and lowered possibility of misassemblies. In case of BT 1A strain 5303, the genome could even be contiguated into a single molecule. Misassemblies for the low-pathogenic BT genomes are possible, but in comparing them to both contiguated genomes attempts were made to find the most likely assembly.

All strains have an unresolved tandem repeat in the Orthologs of *YE2684*. The repeat consists of six amino acids PTDTPE corresponding to CCG ACC GAT ACG CCA GAG base sequence at the C-terminal end. This gene encodes a conserved hypothetical protein, so that at the current stage further sequencing was deemed unnecessary.

In contrast to a finished, gold standard genome the reference genomes generated have an unknown error rate, and bases were only manually re-checked for sequence quality in apparent pseudogenes.

3.3.2 General genomic composition of a *Y. enterocolitica* genome

Detailed manual annotation of the reference genomes gives confidence in further downstream analysis. The general features of the biotypes are very similar with respect to GC content, genome size and number of CDSs, as well as coding density and gene size. The main differences can be seen in the number of pseudogenes which is highest in BTs 4 and 5, and in the number of IS elements which is generally higher in the low-pathogenic BT but again highest in BTs 4 and 5.

Comparison to other published genomes also showed that general composition is similar and that the main differences lie in the number of pseudogenes. As mentioned before, the differences in pseudogenes observed between the BT 3 O:9 and the BT 4 O:3 strain to the reference strains generated in this study could be due to the definition of a pseudogene. It could be that the other studies did not include

genes with a frameshift or an introduced stop codon in the category of pseudogene as technically the whole reading frame is still present.

The differences in the number of pseudogenes might also arise due to the difference in pseudogene analysis. Once the phylogenetic relationship between the *Y. enterocolitica* biotypes was established (Chapter 7), the pseudogenes were in fact re-examined and compared. The evolutionary relationship gives putative information and a trajectory on loss versus acquisition events of regions, aiding in pseudogene identification. In this way, even small gene fragments were identified such as a 26 aa remnant of *focB* in the low-pathogenic BTs. Automatic gene prediction software miss these small fragments as they fall below the length cut-off. Their correlation score which relates the amino acid composition to the translation of the three reading frames is probably also lowered. Additionally, the comparative analysis of all BT helped to establish whether genes in BT 1B strain 8081 might possibly be new pseudogenes as compared to the other BTs. These detailed analyses contributed to an increase in the number of pseudogenes. The newly created genomes would need to be compared in depth to the reference genomes of this study to compare which genes exactly are affected and how, if the genes are orthologous, as well as which pathways might be deactivated, in order to establish a true set of pseudogenes for each lineage. A larger strain set will also help to distinguish between strain-specific, niche-specific, and lineage-specific pseudogene distribution.

The variability detected with respect to number of pseudogenes might also reflect differences in niche adaptation. In case of the BT 4 strains it can be said that Y11 is a representative of continental European strains, which is the predominant bio-serotype isolated of *Y. enterocolitica*. On the other hand, strain 1203 is a representative of UK strains. In the UK, the predominant strains isolated are BT 3 O:9 and BT 4 O:3 (McNally, et al., 2004). There is no increased prevalence of a single bio-serotype. The higher number of pseudogenes in the UK isolate might thus reflect decreased survival fitness. A quick comparison of the two strains did not reveal any large difference in genomic content, thus extensive acquisition of virulence factors in the European isolate can be excluded.

The overall comparison of gene functions of the reference genomes is very similar. The predominant functional classes are hypothetical, unknown proteins, transport proteins, proteins involved with the cell envelope, regulation, and macromolecule synthesis as well as DNA of foreign origin.

One could have expected to see an increase in protection responses, adaptation, or DNA of foreign origin in the high-pathogenic BT. These functional classes include proteins such as cell killing, detoxification, drug sensitivity, adaptation to atypical conditions, iron storage, and pathogenicity island-related functions. Such functions could be considered to signify “virulence genes” which could explain the increased virulence potential of high-pathogenic strains. In the overall comparison, this does not seem to be the case, so a closer look into absent, present, and shared regions between the biotypes is necessary and will be done in Chapter 4.

4 Comparative analysis of the *Y. enterocolitica* biotypes

4.1 Introduction

As seen in the previous chapter, the general composition of the *Y. enterocolitica* genomes are different. Certain genomic islands might only be shared in some lineages, such as the HPI in high-pathogenic BT 1B.

The aim of this chapter is to define sets of orthologous gene clusters in the biotypes and to identify biotype-specific marker regions, thus determining the genetic make-up of each bio-serotype and see how this relates to the pathogenic potential.

The main differences observed thus far relate to the number of pseudogenes and partial genes as well as the number of IS elements. Hence, the aim is also to investigate the nature of the pseudogenes to determine in which genes they occur, and to establish the distribution of IS elements.

4.2 Results

4.2.1 The core- and accessory genome of *Y. enterocolitica*

In order to determine which CDSs or regions are shared among the biotypes, the ortholog qualifiers were taken to group each CDS on the basis of number and combination of orthologs in the other strains. The resulting matrix is listed in Table 9.3 in the appendix and was used as the basis for representing the accessory genes shared in different BTs in a 7-way Venn diagram, which was designed in collaboration with Dr Timothy Hetherington, NTU. The information obtained was also used to characterize the core- and accessory genome of the species *Y. enterocolitica*.

For the sets of orthologous gene clusters, an in-house script was used (see Appendix 9.5, p. 214). This script takes one genome at a time and determines which genes have orthologs in other genomes, and the respective combination. Starting with the first genome strain 5303, the other genomes are allocated a genome number, for example strain 8081 is genome1, strain 21202 is genome2, etc (Table. 9.6, p. 225). The different genome combinations are then sorted into different groups. The first group corresponds to the *Y. enterocolitica* backbone, as it contains CDSs, which have orthologs in all of the other genomes. The second group contains CDSs that have orthologs in all genomes except genome6. In the case of strains 5303, 8081, 21202, 5603, 14902, and 1203, this corresponds to strain 3094. In strain 3094, on the other hand, genome6 is strain 1203. This shift in the group numbers has to be taken into account when evaluating table 9.6, and is due to the fact that only orthologs of one genome can be considered at a time.

Theoretically, certain groups might be of special interest. Apart from the backbone, group 2 and group 8 in table 9.3 could be of interest as they exclude BT 5 and BT 4+5, respectively, and thus could help in identifying the genomic region encoding the biotype reaction. Group 6 excludes BT 1B. Group 7 is of particular interest as this excludes BT 1A, and thus will contain the sets of genes common in all pathogenic BT. Group 22 excludes BT 1A and 1B, giving the shared CDSs of the low-pathogenic BT. The human, low-pathogenic isolates are in group 53, whereas the animal isolates are in group 56. These two groups could indicate sets of genes involved in niche adaptation either to a human or an animal host. Groups 58 to 63 indicate links

between individual genomes, and group 64 will contain CDSs that are unique for each of the biotypes.

From looking at the results (Table 9.6, p.225), only a limited number of sets of orthologous genes are relevant. Some groups, such as group 11, 12, 13, and others, do not contain any genes. For better visualization of these groups, collaboration was set up with Dr Timothy Hetherington, NTU, who did the initial design of a 6-way Venn diagram. Later, this was modified by Sandra Reuter to include the 7th strain 3094. The complete diagram is shown in Figure 4.6, but will be considered in several steps for clarity.

The individual unique connections between strains are observed first (Figure 4.1). BT 1B is only directly connected to BT 1A. Surprisingly, the non- and the high-pathogenic BT share 85 CDSs. The arsenic resistance operon that is located outside the plasticity zone in BT 1B is shared with BT 1A. Two genomic islands are common to BT 1A and 1B: YGI-2 and YGI-4. YGI-2 encodes a glycolipoprotein that is also found in other *Enterobacteriaceae* (Thomson, et al., 2006). YGI-4 represents an integrated plasmid. The original YGI-4 appears to be retained in BT 1A, as the island is 20kb larger than in BT 1B. A common theme of the other CDSs shared between BT 1A and 1B include metabolic regions such as ABC transporters predicted to transport amino acids and sugars, as well as sugar kinases and permeases, exported proteins, and membrane efflux proteins. BT 1A and 1B also share operons encoding a nickel-cobalt efflux system. One of the putative sugar transporters, *YE2603-YE2609*, appears to have been deleted from the low-pathogenic BT, as *YE2604-YE2608* are absent and *YE2603* and *YE2609* are both truncated.

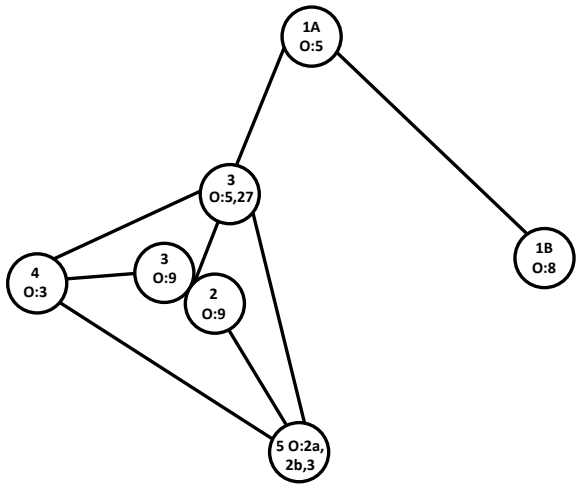


Figure 4.1: Connections between individual reference genomes based on analysis of orthologous genes.

BT 1A and BT 5 share ~30 CDSs, which correspond to a prophage locus. Although the phage has integrated at the same chromosomal location and thus could be considered orthologous, there are differences within the prophage. As phages are mobile elements, the genome locus might also simply present an integrational

hotspot, and the phage would be considered paralogous in BT 1A and BT 5. The connection is thus not highlighted in Figure 4.1.

Connections between some isolates are given through shared parts of the operons for O-specific chain synthesis of the outer membrane lipopolysaccharide (LPS). BT1A and BT 3 O:5,27 share 17 CDSs. This region contains genes encoding for sugar conversions and modifications and a sugar export system and will be described in more detail later on (see 4.2.2.3).

BT 2 O:9 and BT 3 O:9 share a considerable number of CDSs relating to two prophage loci, clustered and unclustered hypothetical proteins as well as the O:9 components of the O:antigen. These two bio-serotypes also share transposases at the exact same genomic positions. BT 3 O:5,27 and BT 4 share transposases and hypothetical proteins, mostly unclustered. BT 4 O:3 and BT 5 O:2a,2b,3 share an operon encoding the O:3 antigen synthesis and transposases.

Moving on to three-way comparisons, the BT 2 and BT 3 isolates have two common prophage loci. The O:9 isolates and BT 4 also share part of a prophage locus.

It appears in general that BTs 2 – 5, the low-pathogenic isolates, are very closely related with respect to their shared accessory gene content (Figure 4.2). They share ~170 CDSs with functions that might relate to environmental survival. These are a combination of groups 22 and 40 (Table 9.3) as BT 4 has lost some CDSs independently but did contain the complete set initially. The low-pathogenic BTs have unique fimbrial proteins that are absent from BT 1A and 1B and there are also two genomic islands, which have been described in the past as specific for low-pathogenic BT.

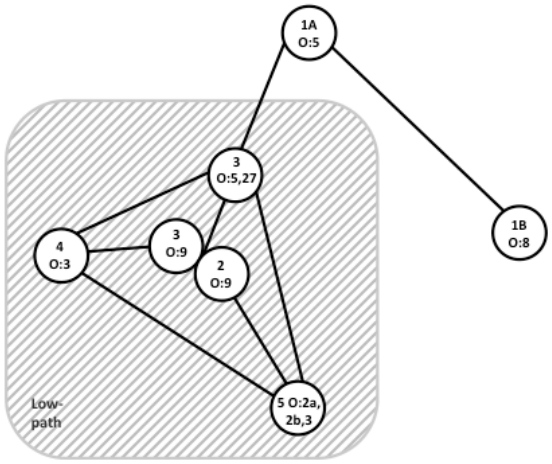


Figure 4.2: Orthologous genes shared in the low-pathogenic BTs.

The first is the *tc* pathogenicity island (PAI) which confers toxicity against insects to *Y. enterocolitica* (Bresolin, et al., 2006a; Fuchs, et al., 2008). Except for one strain, these have not been found in the non- or high-pathogenic BTs. The other island corresponds to the second flagella cluster Flag-2 (Bresolin, Trček, Scherer and Fuchs, 2008). Part of this cluster has been deleted in BT 4 (group 40) due to the

action of transposases. Additionally the low-pathogenic BTs share further hypothetical and phage-related proteins. The low-pathogenic BTs are signified by the expansion of IS element IS1667. Some of the IS1667 insertions appear to be ancient, as the position of the copies is conserved in all low-pathogenic lineages.

The non- and low-pathogenic BTs share ~120 accessory CDSs associated with different metabolic traits such as cytochrome c-type biogenesis, utilization of N-acetyl-D-galactosamine, amino acid metabolism, and anaerobic dimethyl sulfoxide reduction (Figure 4.3). These BTs also possess the same LPS outer core region, which has been replaced by the O:8 operon in the high-pathogenic BT 1B. There are also some functions that might relate to pathogenesis such as a novel hemolysin and its activator protein, A and B subunits of a cytolethal distending toxin, a putative repeat toxin RtxA, and an alternative T3SS. A similar T3SS has been identified in the recently published environmental *Yersinia* (Chen, et al., 2010) and it is also found in *Y. pestis* and *Y. pseudotuberculosis*. We thus propose to call it the *Yersinia* Genus T3SS (YGT) and it will be described in more detail later (4.2.2.1).

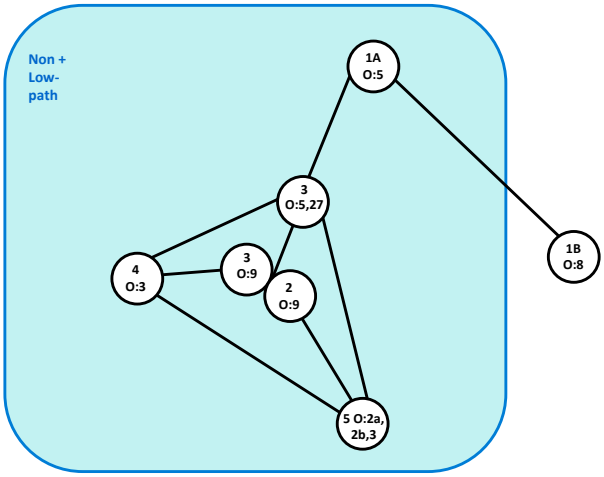


Figure 4.3: Shared genes in non- and low-pathogenic BTs.

Both BT 4 and BT 5 (Figure 4.4) 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. Further unclustered CDSs that have been lost from BT 5 are also putatively associated with metabolic pathways and membrane transporters. Both BTs are missing a putative sugar

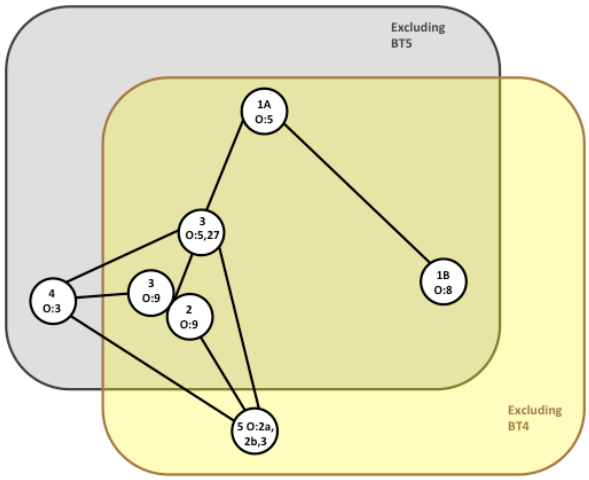


Figure 4.4: Orthologous gene clusters absent from BT 4 and BT 5.

transport system that is present in all other biotypes. This loss of metabolic properties is also evident in the biotyping scheme used for differentiation of *Y. enterocolitica* biotypes, with BT 5 testing negative or variable for 11 of the 12 reactions. BT 4 has five of those negative reactions in common with BT 5 and tests positive for the rest. As the absent sugar transporters are missing a definite pathway assignment, a direct link between the biotyping reaction and genome locus is not possible. The bases of the sorbose and trehalose negative phenotype of BT 5 are mutations in *sorB* and *treB*, respectively.

One further group is relevant: the "pathogenic core", shared in low- and high-pathogenic BTs (Figure 4.5). There are only ~40 chromosomal CDSs in addition to the virulence plasmid pYV (~90 CDSs) that are shared exclusively by all pathogenic BTs. These functions include the adhesion gene *ail*, some clustered exported proteins, non-clustered hypothetical and conserved hypothetical proteins, a putative repeat toxin family protein, and one locus encoding genes involved in threonine and serine metabolism. This last 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. As the hypothetical proteins are not organized in clusters, but are rather scattered throughout the genome, a function is questionable.

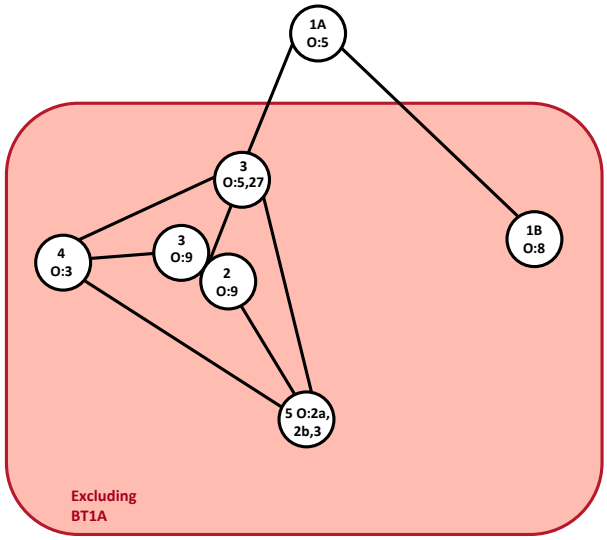


Figure 4.5: Orthologous genes shared in the pathogenic BTs.

The attachment invasion locus gene *ail* is shared in pathogenic BTs, but it is located in a different genomic context compared to the low-pathogenic BTs. This is likely to be due to IS-mediated transposition, as fragments of *IS1328, IS1400* and *IS1668* are visible in both high- and low-pathogenic BTs. The virulence plasmid pYV and the chromosomally located *ail* consequently constitute the only shared, well-characterized "virulence factors". None of the other identified CDSs present exclusively in the pathogenic BTs have known functions.

The individual groups described above are combined in Figure 4.6. It is important to note the number of connections between the different low-pathogenic bio-serotypes based on shared accessory gene content. It is also striking how both the non- and the high-pathogenic BTs stand singled out, with only few, selected connections to the other bio-serotypes.

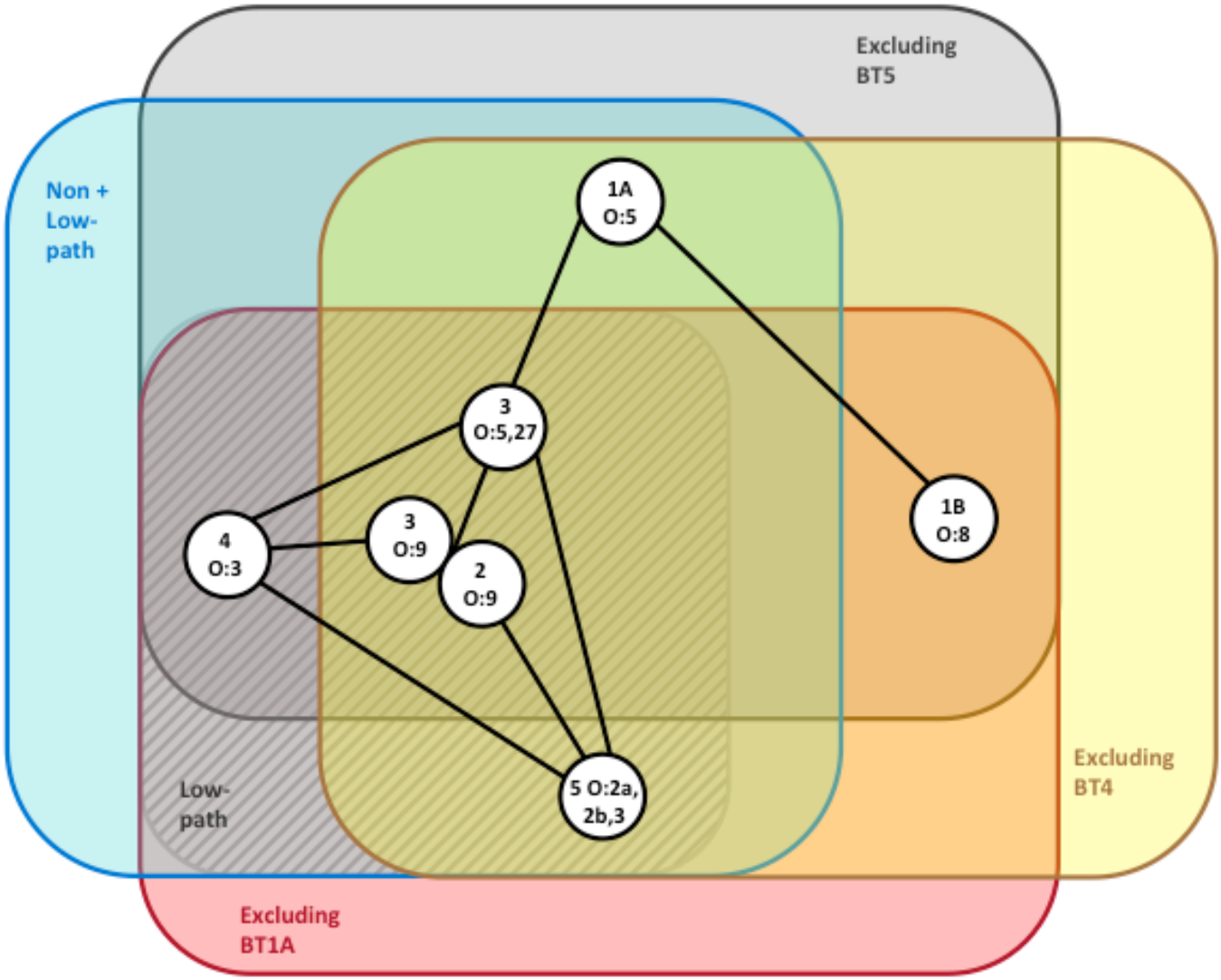


Figure 4.6: Combined comparison of orthologous accessory gene sets in the reference genomes for the *Y. enterocolitica* biotypes. As described above, the single lines represent connections between isolates. The low-pathogenic BTs are highlighted with a shaded box. Further orthologous gene sets are highlighted with boxes for non- and low-pathogenic BTs (blue), all BTs except BT 4 (yellow), all BTs except BT 5 (grey), and all BTs except BT 1A (red; the pathogenic BTs).

After consideration of sets of orthologous genes, it was also possible to determine the number and type of CDSs that are unique to each bio-serotype.

BT 1A and BT 1B both show a considerable number of unique CDSs, 728 (16.7% of total CDSs) and 540 (13.3% of total CDSs) respectively. For BT 1B, these have been

previously described as being specific for high-pathogenic strains and include the acquisition of the high-pathogenicity island (HPI), the Ysa T3SS, and the general secretion pathway Yst-1. Also unique to the high-pathogenic BT is the Yersinia adhesion pathogenicity island, YAPI. Many of the CDSs unique to BT 1A appear to extend its metabolic repertoire compared to the other biotypes, including proteins involved in transport and metabolism of sugars. Additionally, the non-pathogenic BT 1A isolate carries several gene clusters involved in environmental survival such as metal resistance and an arsenic resistance operon. The low-pathogenic BT isolates generally have a lower number of unique CDSs compared to BT 1A and 1B. The reference BT 2 O:9 genome carries five genes that are unique to this strain, and reference BT 3 O:9 strain carries no unique genes (Table 9.6 p. 225). Both serotype O:9 isolates of BT 2 and 3 are genetically very close, sharing phage and transposases, thus it is not surprising that they do not have many unique genes. BT 3 O:5,27, BT 4 O:3, and BT 5 have 171, 134, and 274 unique CDSs, respectively. The majority of these genes are related to several prophage loci, hypothetical and conserved hypothetical proteins and transposases. Prophages thus are the largest contributors to the accessory genome.

All *Y. enterocolitica* BT have 3,306 CDSs in common which represent the core functions of the species (Figure 4.7, Table 4.1). It has been previously speculated that the *Yersinia* genomic island 1 (YGI-1) encoding the tight adherence (*tad*) operon was associated with enteropathogenicity and specific for *Y. pseudotuberculosis* and the pathogenic *Y. enterocolitica* biotypes (Howard, et al., 2006; Thomson, et al., 2006). However the *tad* operon is actually part of the core genome being found at the same locus in all isolates including the non-pathogenic BT 1A strain. The presence of the *tad* operon was also confirmed for the environmental species (Chen, et al., 2010), therefore this cluster is probably ancestral. The same is true of other functions known to be important for virulence including *inv* encoding invasion, its regulator RovA and the *myf* fibrillar operon (Badger and Miller, 1998; Ellison, Lawrence and Miller, 2004; Iriarte and Cornelis, 1995).

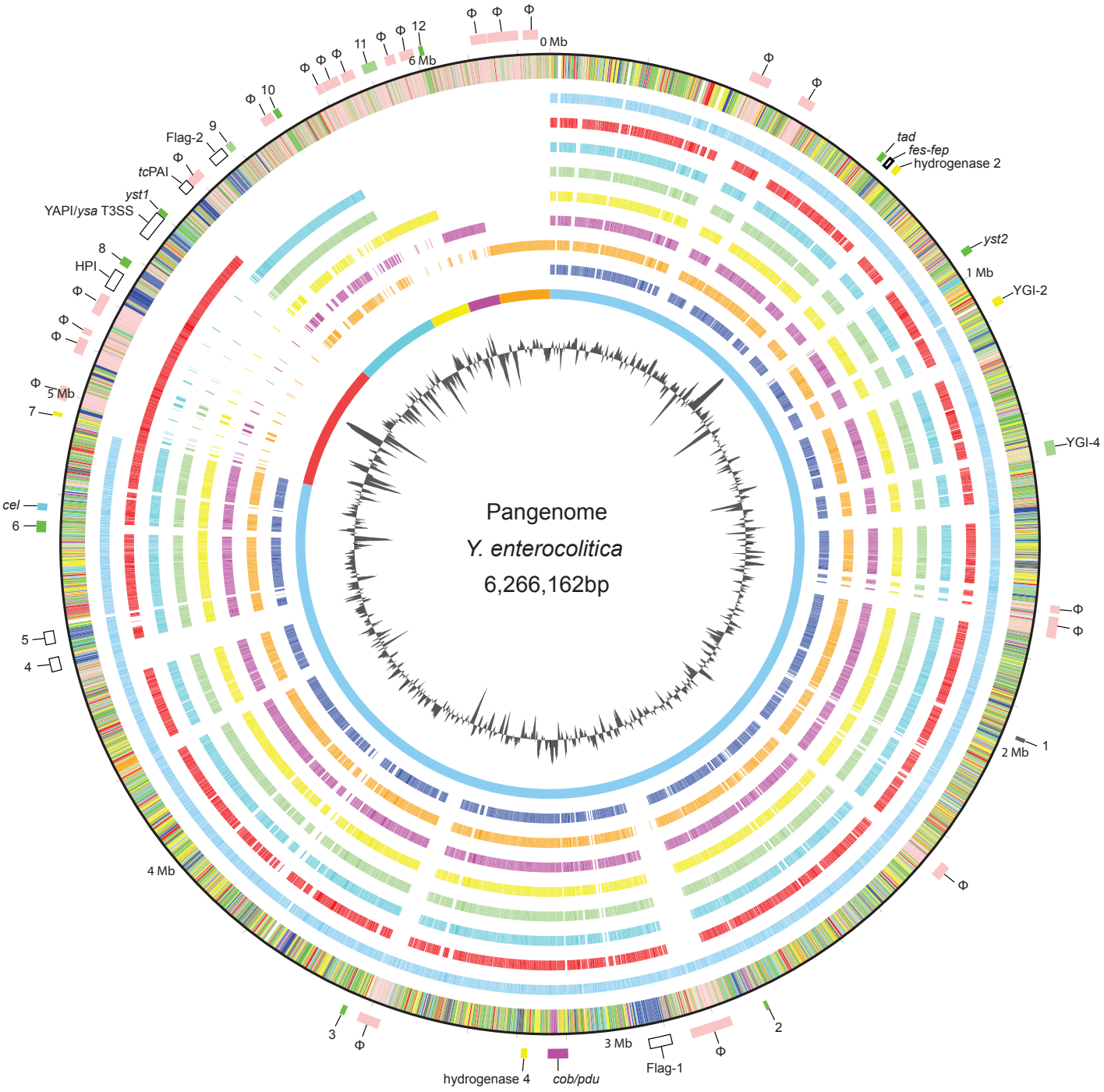


Figure 4.7: The pangenome of *Y. enterocolitica*. The outer scale shows the size in Mbps. From the outside in, circle 1 shows the position of CDSs (for colour codes see below). Circle 2 (CDSs coloured blue) marks the position of *Y. enterocolitica* strain 5303 / BT1A O:5 genes. Circle 3 (red) shows the position of strain 8081 / BT1B O:8 genes. Circle 4 (light blue) 21202 / BT2 O:9. Circle 5 (light green) 5603 / BT3 O:9. Circle 6 (yellow) 14902 / BT3 O:5,27. Circle 7 (pink) 1203 / BT4 O:3. Circle 8 (orange) 3094 / BT5 O:2a,2b,3. Circle 9 (purple) highlights genes shared in all *Y. enterocolitica* biotypes. Circle 10 represents the fasta record from which strain the CDSs were written out. Circle 11 shows a plot of G+C content (in a 10-kb window). Genes in circles 1 are colour-coded according to the function of their gene products: dark green, membrane/surface structures; yellow, central or intermediary metabolism; dark blue, pathogenicity island-related functions; cyan, degradation of macromolecules; red, information transfer/cell division; cerise, degradation of small molecules; pale blue, regulators; salmon pink, pathogenicity/adaptation; black, energy metabolism; orange, conserved hypothetical; pale green, unknown; and brown, pseudogenes. The position of important regions (mentioned in the text) are marked and labeled in the outermost ring. See Table 4.1 for a description.

Table 4.1: Regions of interest identified in the pangenome of *Y. enterocolitica* (Figure 4.7).

Region of interest #	Base position	Content
1	1955006..1962863	cytochrome c-type biogenesis proteins
2	2691400..2698800	Phosphotransferase system for N-acetyl-D-galactosamine
3	3544600..3554299	outer core LPS biosynthesis operon
4	4448957..4475254	silver and copper resistance proteins
5	4502803..4527755	*Yersinia* Genus T3SS (YGT)
6	4722584..4745103	putative O:5 LPS biosynthesis cluster
7	4952491..4960977	serine/threonine metabolism
8	5274712..5291988	O:8 LPS biosynthesis proteins
9	5583332..5597425	replaces YGI-3
10	5697260..5709091	O:9 LPS biosynthesis proteins
11	5890466..5918695	putative genomic island, similar to Escherichia blattae proteins and phage proteins
12	6006539..6015652	O:3 LPS biosynthesis proteins

The two hydrogenase complexes, hydrogenase 2 (*hyb*) and hyrdogenase 4 (*hyf*), were thought to be associated with colonisation and survival of the gut. Both clusters are present in all *Y. enterocolitica* biotypes. The hydrogenase 2 operon shows signs of truncation in the low-pathogenic BTs as the formate transporter *focB* (*YE2812*) is fragmented. The hydrogenase 4 complex is truncated in *hypB* (*YE3603*) in BT 3 O:5,27 strain 14902 and BT 4 O:3 strain 1203. In contrast these clusters are probably functional in the non-pathogenic isolate.

The cellulose biosynthesis operon is also present in all BTs. In *Salmonella* an extracellular matrix is formed through the production of cellulose, which could increase the time the bacteria are retained in the gut. A similar idea has been proposed for *Y. enterocolitica* (Thomson, et al., 2006). Again, this cluster shows signs of mutations. In BT 4 O:3 strain 1203 a stop codon has been introduced in *bscC* (*YE4072*). An opal stop codon (UGA) has been introduced in *bscA* (*YE4074*) in BT 1A, which could potentially be decoded by selenocysteine.

4.2.2 Regions of difference in *Y. enterocolitica* biotypes

Some of the regions identified as regions of difference between the *Y. enterocolitica* bio-serotypes have not been extensively characterized previously. These regions will be discussed in turn below.

4.2.2.1 *Yersinia* Genus Type 3 Secretion System

The first interesting region of difference is an alternative T3SS we propose to call the *Yersinia* Genus T3SS (YGT) since orthologs have been described in the environmental species as well as *Y. pseudotuberculosis* and *Y. pestis*. The system is located in in the same genomic context in all species, indicating an ancient acquisition before the split into separate species, justifying the denomination as a common T3SS of the genus. By sequence, this system is most similar to the *Salmonella* pathogenicity island 2 (SPI-2, Figure 4.8).

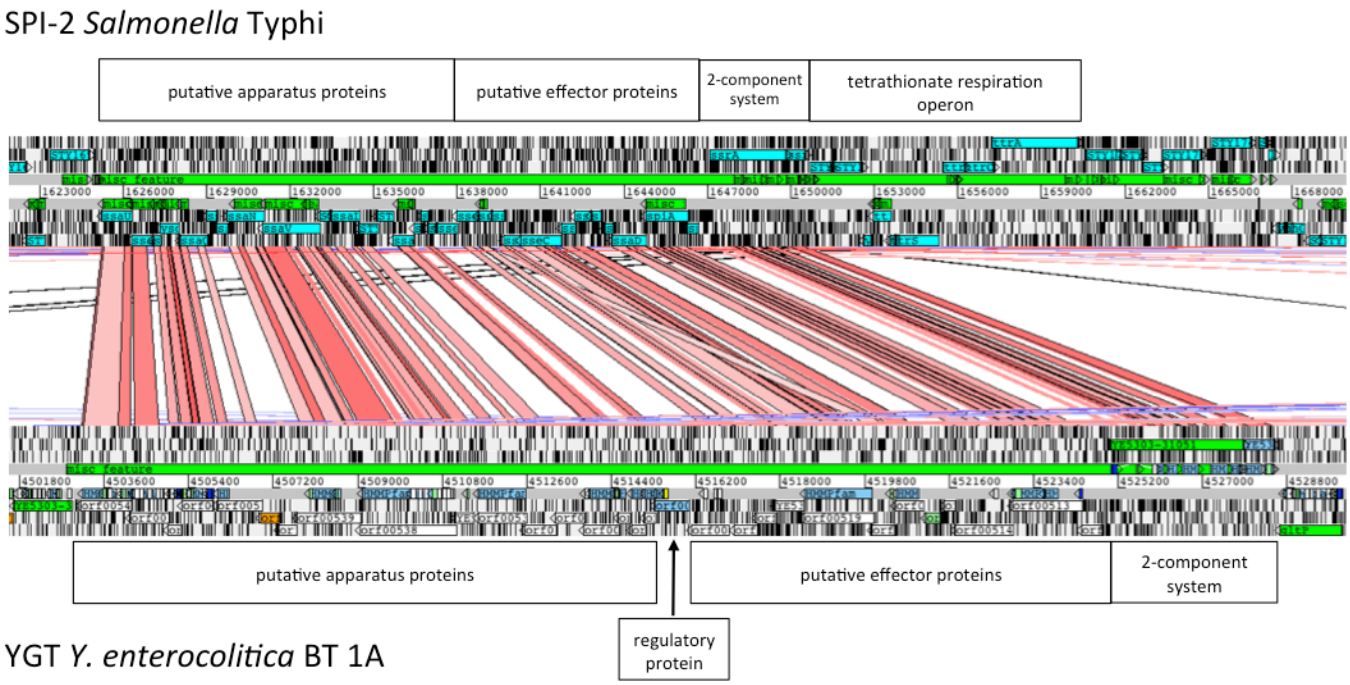


Figure 4.8: Comparison of SPI-2 of *Salmonella* Typhi with the novel YGT. The apparatus and effector regions are highlighted, as well as the 2-component system located upstream of the T3SS in both cases. In SPI-2, the YGT is preceeded by the tetrathionate respiration operon. In *Y. enterocolitica*, the tetrathionate respiration operon is located in a different genomic context.

The sequence similarity shared between SPI-2 and YGT is low but is consistent across the whole YGT region, and ranges on average around ~40% similarity. Upstream of YGT and SPI-2 T3SSs is a 2-component system encoding a sensor kinase and a response regulator that might be involved in governing the expression

of the T3SS. In SPI-2, the tetrationate respiration operon is loacted within the pathogenicity island, whereas for *Y. enterocolitica*, a highly related operon is located elsewhere in the chromosome.

Comparison of the YGT within different assembled *Yersinia spp.* shows differences in the gene content (Figure 4.9).

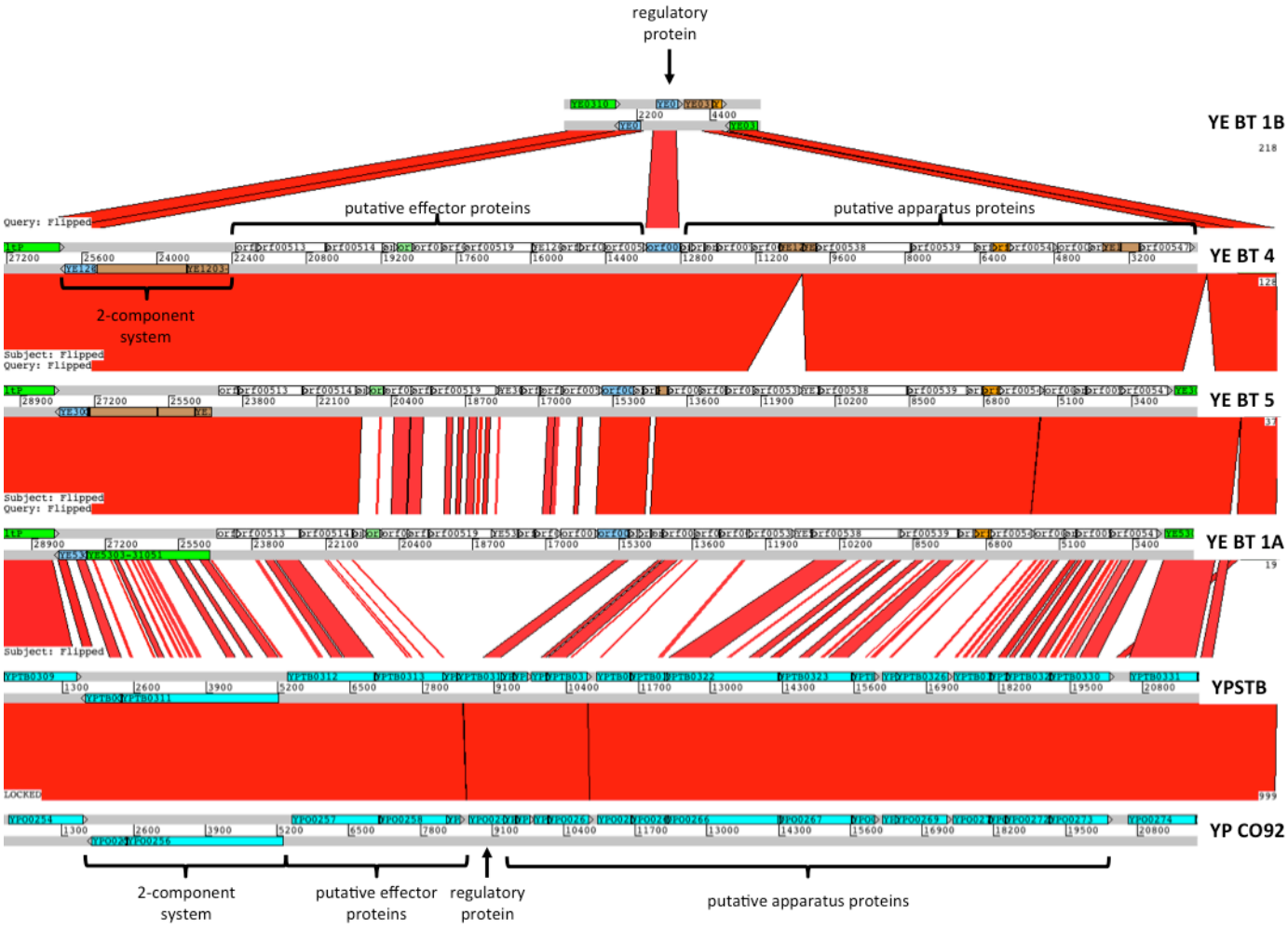


Figure 4.9: Comparison of YGT within several *Yersinia spp.* Four representative *Y. enterocolitica* BT (YE BT 1B, 4, 5, and 1A) are compared to *Y. pseudotuberculosis* (YPSTB: IP32953) and *Y. pestis* (CO92). Effector and apparatus regions are highlighted, as well as the 2-component system.

As mentioned before, the YGT is absent from BT 1B, only three proteins are present: the response regulator (*YE0311*), the AraC-family regulatory protein (*YE0312*) in the central part of the YGT, and an inner membrane protein (*YE0313*) that is missing the N terminus. As the YGT is present in all other Yersiniae, and as the central regulatory protein has been retained, this indicates a possible deletion in at least two separate steps from BT 1B.

The YGT is generally well conserved amongst the non- and low-pathogenic BTs although there is evidence that some genes in YGT are defunct in BTs 4 and 5. Those mutations are different in BTs 4 and 5, with BT 4 having mutations in three apparatus related genes and the deletion of another apparatus protein and BT 5 having only one independent mutation in an apparatus gene. Moreover the sensor kinase of the 2-component system has a stop codon introduced in BT 4 and several frameshifts in BT 5. The inner membrane protein (*YE0313*) is missing the central part in BTs 2, 3, and 4, and is only functional presumably in BT 1A and BT 5. Some parts of the effector protein region show reduced sequence similarity between non- and low-pathogenic BTs.

A comparison of YGT in several *Yersinia* spp. to *Y. enterocolitica* shows that parts of the effector protein region have been deleted in *Y. pseudotuberculosis* and *Y. pestis* (Figure 4.9). Preliminary studies (Golubov, Heesemann and Rakin, 2003; Pujol and Bliska, 2003; Balada-Llasat and Mecsas, 2006) have been undertaken on YGT in *Y. pestis* and *Y. pseudotuberculosis* and have shown that this system plays no obvious role in causing disease, but functionality in *Y. pestis* and *Y. pseudotuberculosis* is probably compromised due to the absence of the effector region. The distribution of YGT within the genus suggests that this system is more likely to be required for colonisation outside a human host. Initial investigation into the phylogeny of the T3SS (Foultier, et al., 2001) is in accordance with the finding that this T3SS is most closely related to SPI-2 of *Salmonella*.

4.2.2.2 Variation of the Plasticity Zone

Another large-scale region of variation distinguishing different BTs is the plasticity zone (PZ). The PZ was first described in the high-pathogenic BT 1B strain 8081 and constitutes a region, which contains several operons associated with increased pathogenic potential. The region shows signs of having been formed through several independent acquisition events, hence being termed plasticity zone (Thomson, et al., 2006). A comparison of this region with the other BTs confirms this locus to be highly variable between BTs (Figure 4.10). The plasticity zone in the high-pathogenic BT shows the largest size. The acquisition of the *Yersinia* adhesion pathogenicity island YAPI is possibly specific for strain 8081 where it was originally defined, as it was not observed consistently in other BT 1Bs (Thomson, et al., 2006; Carniel, Guilvout, and Prentice, 1996). Furthermore BT 1B acquired the *ysa* T3SS as well as the general secretion pathway *yst1*. At the same locus, the non-pathogenic BT possesses heat

shock proteins, hypothetical proteins, and an additional *lac* operon. No equivalent region is present in the low-pathogenic BTs. All BTs have an iron transporter and operons for hydrogenase 2 biosynthesis.

The second flagella cluster Flag-2 is only retained by the low-pathogenic BTs and has been independently lost in the non- and the high-pathogenic lineages as indicated by the presence of different fragments of Flag-2. This region is also marked by the presence of a transposase in BT 4 likely to have been responsible for the loss of Flag-2 and additionally the deletion of the *fes/fep* operon. Both high- and low-pathogenic BTs then have outer membrane efflux proteins, whereas the low-pathogenic BTs show a different set of membrane proteins. These membrane protein operons have no specifically assigned functions so far. As mentioned before, the tigh-adherence operon *tad* is present in all biotypes.

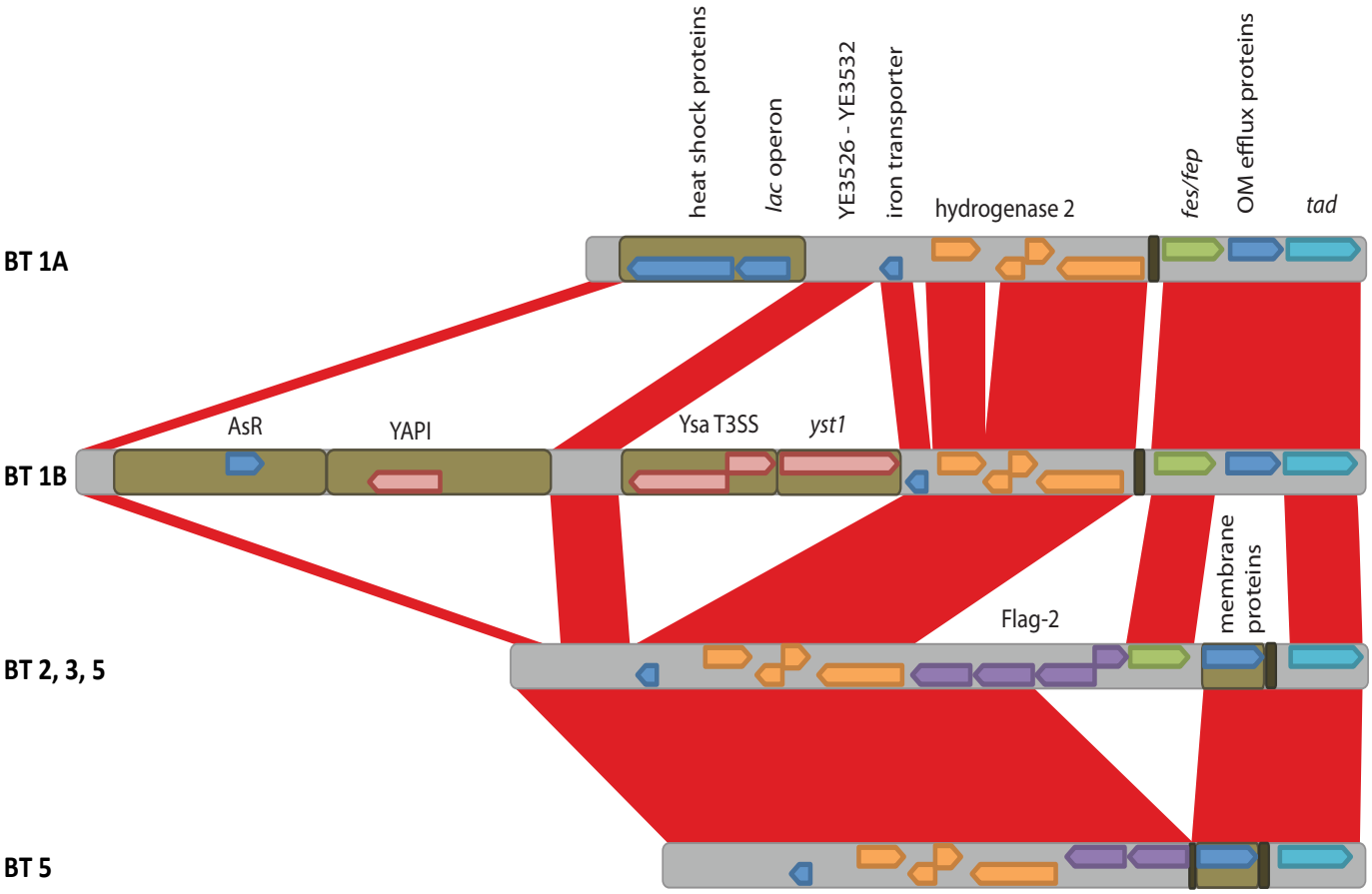


Figure 4.10: Arrangement of the plasticity zone in the different biotypes. Non-pathogenic BT – 1A, high-pathogenic BT – 1B, low-pathogenic BT – 2,3,4,5. Red – colinear homology, modified from an ACT comparison. Light brown – potential inserted region. Dark brown – deleted region.

4.2.2.3 Lipopolysaccharide operons

The lipopolysaccharide (LPS) of *Y. enterocolitica* has been proposed as a supporting virulence factor as it confers resistance to polymyxin and antimicrobial peptides and has been associated with virulence in the mouse model (Zhang, et al., 1997; Skurnik, et al., 1999; Skurnik, et al., 2007). The LPS structure for serotypes O:8, O:3 and O:9 is known, and a schematic is shown in Figure 4.11.

All *Y. enterocolitica* serotypes share – with minor variations – the same lipid A and inner core structures (Müller-Loennies, et al., 1999; Oertelt, Lindner, Skurnik and Holst, 2001). It is known that the outer core is absent from the high-pathogenic BT 1B O:8 strains but shared in the low-pathogenic BTs O:9 and O:3 (Skurnik and Bengoechea, 2003). The O-antigens differ in sugar composition and in linkage of those sugars. The O:8 antigen is a heteropolymeric pentasaccharide in which three saccharides are linked in a chain with a further two saccharides branching off (Zhang, et al., 1997). In O:3 and O:9 O-antigens the chain consists of unbranched homopolymers (Zhang, Toivanen and Skurnik, 1996; Skurnik and Bengoechea, 2003). The outer core of O:3 and O:9 serotype strains consists of a branched heteropolymer (Reeves, et al., 1996b; Skurnik, et al., 1999). It has been proposed that the outer core presents an ancestral O-antigen and that it has been supplemented by later acquisition of homopolymeric O-antigens in the low-pathogenic BTs (Müller-Loennies, et al., 1999; Skurnik, et al., 1999; Sirisena and Skurnik, 2003).

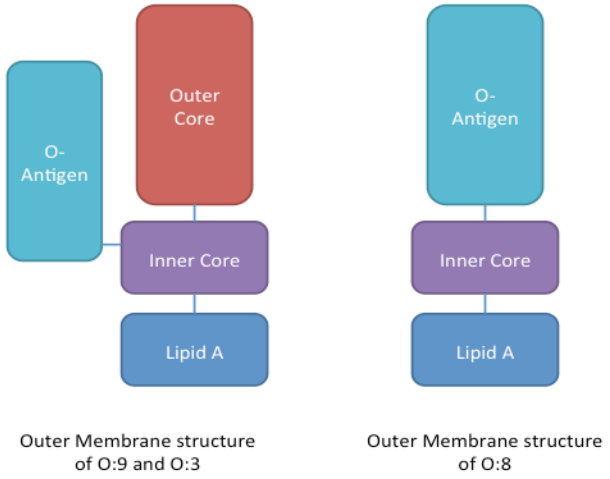


Figure 4.11: Lipopolysaccharide structure of *Y. enterocolitica* serotypes. Structures with shared functions are highlighted with the same colour. Adapted from Sirisena and Skurnik, 2003, and http://www.hi.helsinki.fi/yersinia/Research1.htm.

It has been shown before that the O:8 encoding locus is in the same genomic context as the outer core operon of serotypes O:3 and O:9 (Zhang, Toivanen and Skurnik, 1996; Skurnik and Bengoechea, 2003; Figure 4.12). The outer core is located beween *gsk* (inosine-guanosine kinase, *YE3069*) and *hemH* (ferrochelatase, *YE3088*). Apart from *gsk* and *hemH*, the UDP-glucose 4-epimerase *galE* has been conserved between the outer core and the O-antigen. The same locus also contains

genes involved in the synthesis of the heteropolymeric lipopolysaccharides of *Y. pseudotuberculosis* and *Y. pestis*, although five genes are inactivated in the latter (Zhang, Toivanen and Skurnik, 1996; Skurnik and Bengoechea, 2003). In a comparison between BT 2 O:9, BT 1B O:8, and the environmental *Yersiniae*, the outer core region could be detected in *Y. frederiksenii*, *Y. kristensenii*, and *Y. rohdei*. The absence of the outer core in the genome data for the other environmental *Yersiniae* could be due to assembly problems, therefore a definitive answer on the distribution of the outer core within the genus is not possible.

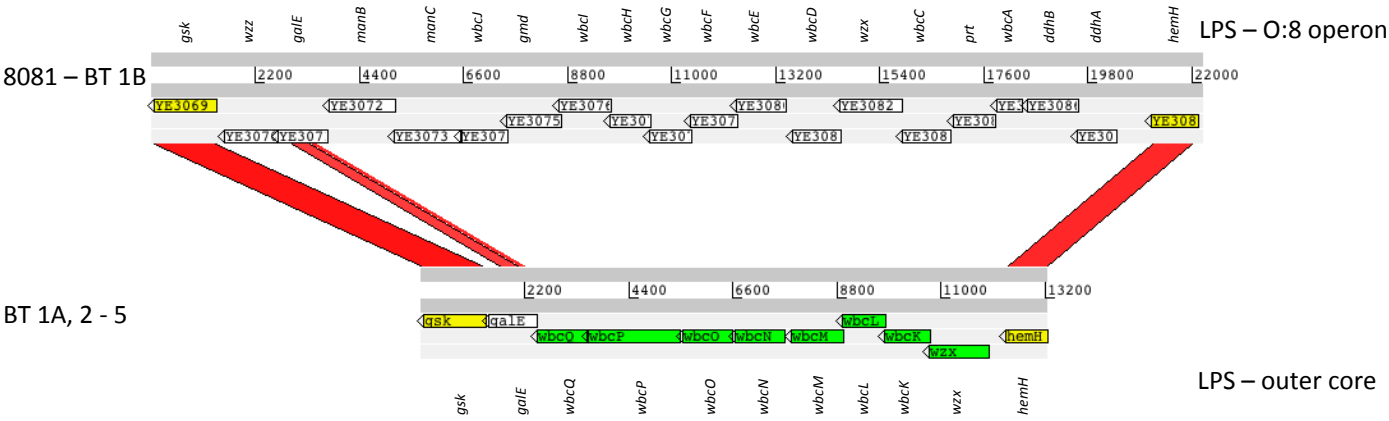


Figure 4.12: Genomic context of LPS O:8 operon and outer core. Both outer core and O:8 operon are located between *gsk* and *hemH*.

In the comparative analysis it became clear that the outer core is shared in low- and non-pathogenic BTs. Considering that BT 1A shows signs of being the historically oldest biotype in that it has fewer IS elements, less genomic reorganization, fewer pseudogenes and an extensive metabolic repertoire as apparent from the unique regions, it is likely that the outer core is the ancestral form. This supports the idea of an exchange of one region for another.

The location of the O-antigen clusters for O:3 and O:9 was unknown, although their operon structure has been investigated in the past (Reeves, et al., 1996b; Skurnik and Bengoechea, 2003; Skurnik, et al., 2007). The O-antigen clusters for O:3 and O:9 are found in the same genomic location between a putative membrane protein (*YE2779*) and *gnd*, a phosphogluconate dehydrogenase (*YE2773*, Figure 4.13).
Both BT 2 O:9 and BT 3 O:9 clusters are identical and are the same O:9 operons described in the past (Skurnik, et al., 2007). There are six CDSs involved in sugar biosynthesis (*manCB, gmd, per, wbcTV*). Four CDSs are encoding transferases for

linking the sugars (*wbcUW galUF*). *galU* and *galF* are conserved with the O:3 operon, and are also found in the other biotypes. Two CDSs are then encoding the transport proteins (*wzm, wzt*).

The O:3 operon is larger than the O:9 operon. An IS element for IS1668 has inserted upstream of the locus, and a remnant for IS1400 can be found in the central part of the operon. The LPS O:3 operon of BT 4 contains one pseudogene, a hypothetical protein carrying a stop codon.

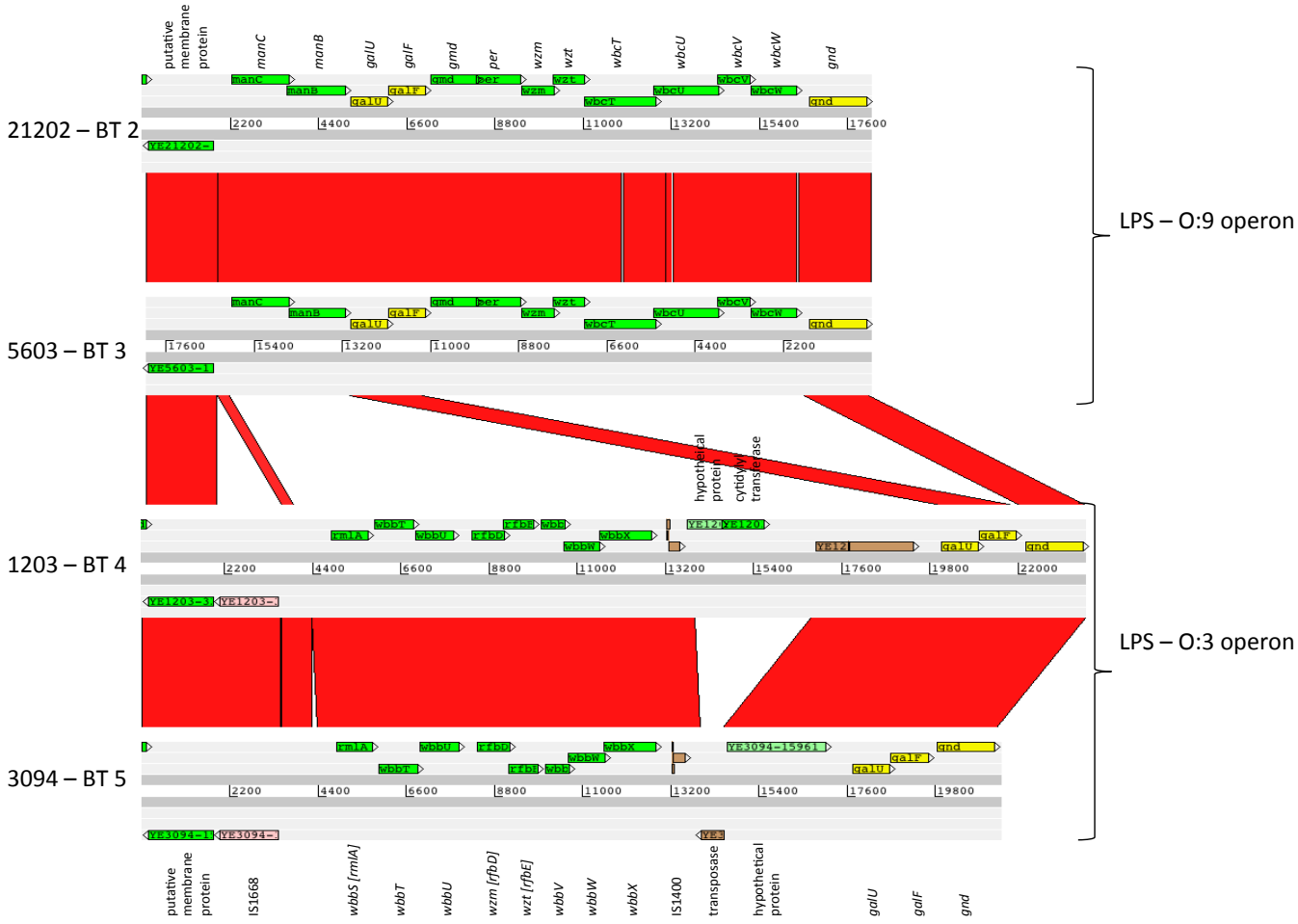


Figure 4.13: Genomic context of O:9 and O:3 operons.

The exact serotype of the reference BT 5 isolate is unknown, but the serotype typically described for this biotype is often a mixture of O-antigens 2a, 2b, and 3 (Bottone, 1997; Bottone, 1999; Bartling, et al., 2004). The operon found in BT 5 is similar in sequence to the O:3 operon of BT 4 and is thus thought to contribute to the serotype of BT 5. There are however differences to the standard O:3 operon. A hypothetical protein and a cytidylyl transferase are absent, possibly due to the action of transposases, as remnants of these are found upstream and downstream of these

CDSs. The hypothetical protein, which has a nonsense mutation in BT 4, is not mutated in BT 5. These differences could be an explanation for the difference in serotyping. Both O:3 and O:2a,2b,3 homopolymeric repeat units are reported as 6-deoxy-L-altrofuranose, with the difference for O:2a,2b,3 being that they have O-acyl side groups (Reeves, et al., 1996b; Bruneteau and Minka, 2003).

Both O:3 and O:9 O-antigen operons encode for homopolymeric polysaccharides and share the same biosynthesis mechanism. The sugars are assembled to chains and then transported across the membrane using ATP-binding cassette (ABC) transporters (Zhang, et al., 1997; Samuel and Reeves, 2003). In the new nomenclature for bacterial polysaccharide synthesis genes, these are designated as *wzm* and *wzt* although they do not necessarily share nucleotide homology (Reeves, et al., 1996a,b; Samuel and Reeves, 2003). The synthesis of heteropolymeric polysaccharides as are found in O:8 differs in that each subunit is assembled and then translocated by a flippase *wzx* before the units are polymerized by Wzz (Zhang, Toivanen and Skurnik, 1996; Reeves, et al., 1996a,b; Zhang, et al., 1997; Samuel and Reeves, 2003).

The structure of LPS O:5 has not been published yet, although a putative cluster has been deposited in GenBank for BT 1A O:5 strain T83 (accession number AY653208; Tennant, S.M., Joe, A. and Robins-Browne, R.M., 2004). The putative O-antigen operon in BT 1A strain 5303 is located between the glutathione reductase *gor* (*YE4057*) and the autotransporter protein YapE (*YE4059*) (Figure 4.14). A comparison between the two BT 1A strains showed a high degree of sequence similarity in this cluster. BT 3 O:5,27 strain 14902 has a similar operon, but in a different genomic context. The putative O-antigen operon is located between *YE0502* and *YE0513* and is surrounded by IS elements and fragments of transposases, possibly explaining the change in location. This operon is also missing two hypothetical proteins, a DNA binding protein, and an alcohol dehydrogenase of the presumed O:5 cluster, which could explain the mixed serotyping reaction.

The sugars of the O:5 serotype have been described as homopolymers of rhamnose (Gorshkova, Kalmykova, Isakov and Ovodov, 1986), explaining the presence of two polysaccharide export proteins that constitute the ABC transport of the homopolymeric chain. Homopolymeric lipopolysaccharide operons are generally smaller in size than heteropolymeric operons, as a higher complexity in the number of sugars and their linkages calls for more genes (Reeves, et al., 1996a). With a size

of ~18 kb it compares well to the other homopolymeric clusters of O:3 (18 kb) and O:9 (14 kb).

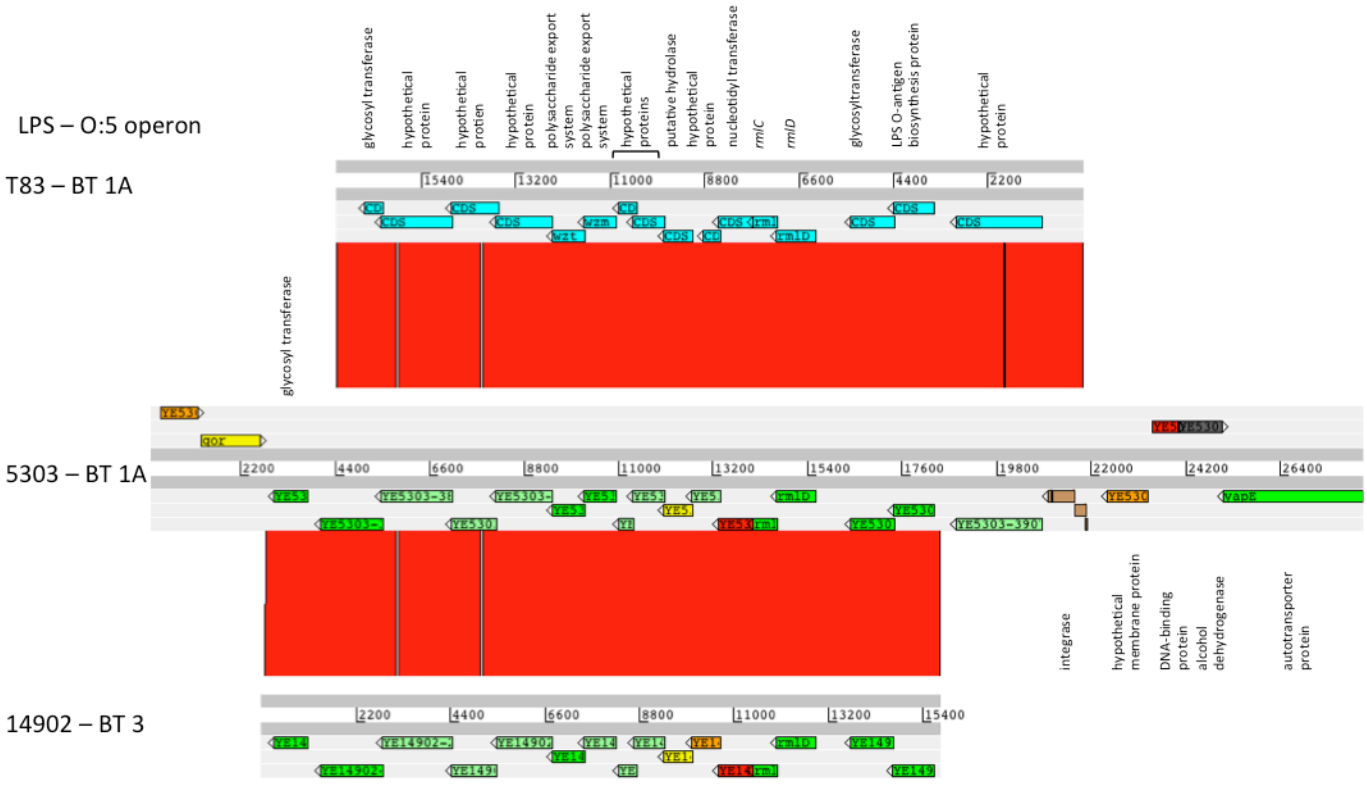


Figure 4.14: Putative O:5 LPS operon. The putative O:5 operon has been deposited in GenBank for BT 1A strain T83, but has not been published yet (accession number AY653208). The presumed region in BT 1A strain 5303 is presented. A truncated version of this operon is found BT 3 O:5,27, but in a different genomic context.

Three genes can be identified that resemble glycosyl transferases, and two rhamnose processing genes, *rmlCD*, are present. Together with the putative export system, the three main classes of genes of processing and linking sugars as well as sugar pathways involved in O-antigen synthesis are thus present (Reeves, et al., 1996a,b). The sugar composition of O:5,27 is very similar to O:5. The LPS structure for O:5,27 also contains rhamnose but is O-acylated (Gorshkova, Kalmykova, Isakov and Ovodov, 1986; Bruneteau and Minka, 2003). This explains the close relatedness of the O-antigen synthesis operons.

4.2.3 Pseudogene distribution amongst the *Y. enterocolitica* biotypes

Pseudogenes can be defined as CDSs that show mutations that may inhibit the expression of a fully functional protein. Mutations include the truncation of the protein due to deletion events of the termini or other parts of the protein and the disruption

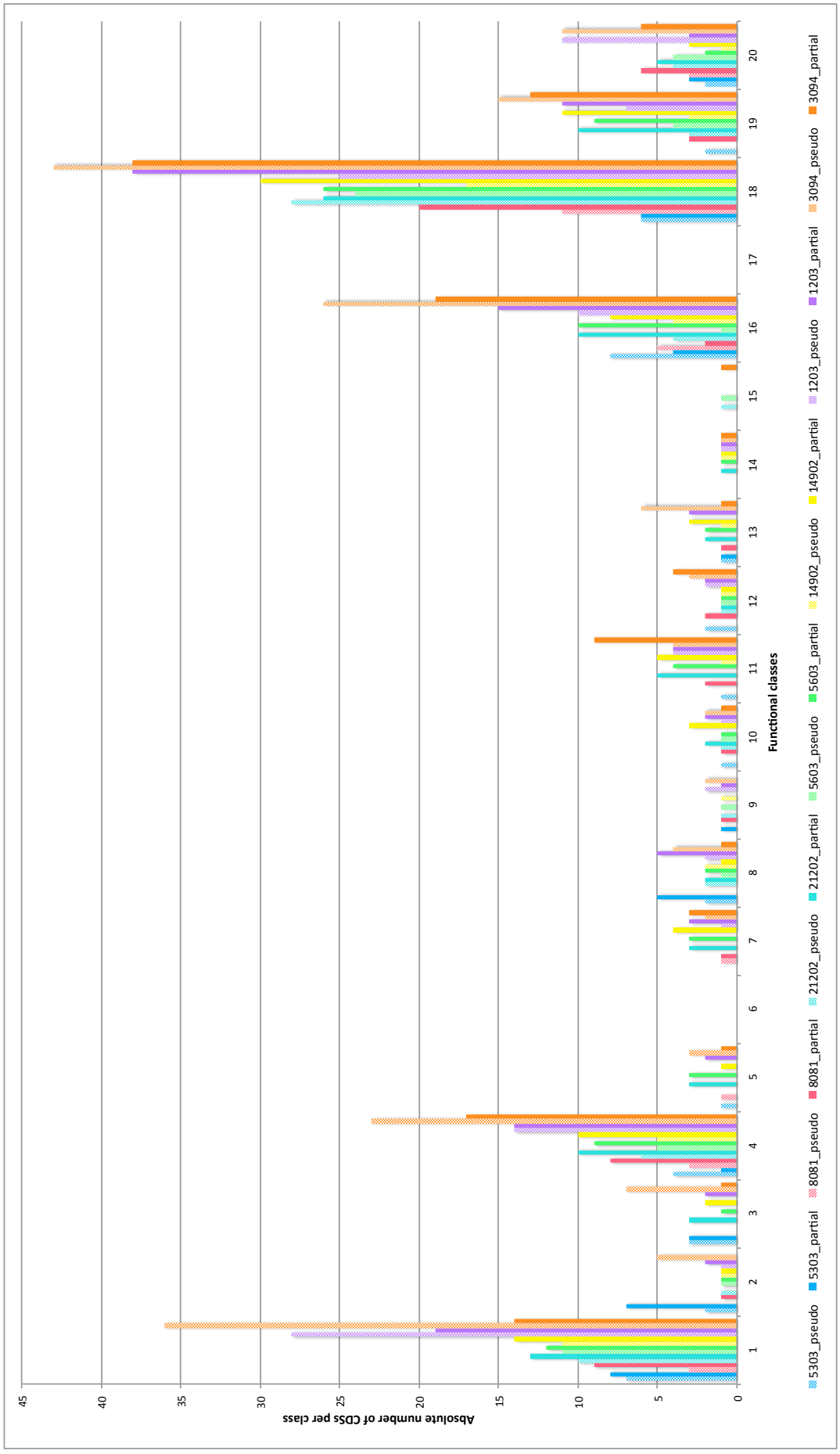
due to the insertion of an IS element into the reading frame. This first category can also be described as partial proteins, or fragments/remnants. These might be viewed as truly non-functional proteins. In a second category, the pseudogenes show mutations such as frameshifts and introduced stop codons that interrupt transcription. In this case the complete reading frame is still present and it might be argued that stop codons are read through on some occasions and that in the case of homopolymeric runs the RNA polymerase might easily slip into the other reading frame resulting in a full-length protein. Two examples in the given dataset are *YE3376*, peptide chain release factor 2, which carries a frameshift mutation in all orthologs, but has been shown to be functional, and *YE4135*, the major subunit of formate dehydrogenase, which has a read-through of the opal (UGA) stop codon in all orthologs. To take these things into consideration, pseudogenes and partial genes have to be compared separately.

A list of all pseudogenes found in the *Y. enterocolitica* reference genomes and their orthologs can be found in the appendix (Table 9.7, p. 227). In addition to the above described *YE3376* and *YE4135*, only two CDSs are present as pseudogenes in all biotypes. *YE0416* is a conserved hypothetical protein, and *YE2410* a putative oxidoreductase.
As seen in Table 3.3, there is a difference between the biotypes in the total number of pseudogenes. BT 1A and 1B have the lowest number of pseudogenes, 81 and 84 respectively, whereas the low-pathogenic bio-serotypes show more than 140 pseudogenes each. BT 4 O:3 (strain 1203) and BT 5 O:2a,2b,3 (strain 3094) show a marked increase in pseudogenes with 235 and 322 pseudogenes in total, respectively. This amounts to 6.0% and 7.9% of the total number of CDSs in each genome. In contrast, only about 2% of the genomes of BT 1A and 1B (1.9%, 2.1%) show signs of genome decay. BT 2 and 3 possess less than 4% of CDSs with potentially impaired function (BT 2 O:9 4.0%, BT 3 O:9 3.6%, BT 3 O:5,27 3.7%). This trend of increased genome decay towards BT4 and 5 can be observed when only considering partial CDSs, too. Both BT show about 130 CDSs that are non-functional.

Figure 4.15 shows the split of pseudogenes and partial genes for each of the reference strains according to function. One can see that some classes are absent

because they constitute essential cell functions. These are genes involved in cell division (6) and ribosome constituents (17).

The highest number of pseudogenes and partial genes in hypothetical proteins (1) is found in BTs 4 and 5. Hypothetical proteins are putative open reading frames with either homologs in *E. coli* or other bacterial species with unassigned functions or do not have any homologs. Their transcriptional and functional status is thus unknown. As the contribution towards cell survival under various conditions still needs to be assessed, the impact of the pseudogenes in this category is uncertain.
When looking at cell processes (2), it is striking that BT 1A has the highest number of partial genes compared to the other BTs. These are found in chemotaxis and motility genes of flagellar cluster 1 and 2. The four copies of the flagellin gene *fliC* of Flag-1 are missing 40-70 amino acids, and Flag-2 is only present in remnants. Yet, this strain has been studied and was shown to be motile, thus the mutations seen in Flag-1 do not have an impact on function (McNally et al., 2007). This might be different in BTs 4 and 5. BT 4 has lost half of the Flag-2 region, and both BTs show mutations and gene deletions in Flag-1 and -2. A similar inactivation of the flagellar clusters is apparent in *Y. pestis* as an adaptation to a niche where motility is not necessary (Parkhill et al., 2001a).



Groups corresponding to classification scheme (Table 9.1): 1 – unknown function, no known homologs, 2 – cell processes, 3 – protection responses, 4 – transport/binding proteins, 5 – adaptation, 6 – cell division, 7 – macromolecule metabolism, 8 – macromolecule synthesis, modification, 9 – metabolism of small molecules, 10 – biosynthesis of cofactors, carriers, 11 – central intermediary metabolism, 12 – degradation of small molecules, 13 – energy metabolism, carbon, 14 – fatty acid biosynthesis, 15 – nucleotide biosynthesis, 16 – cell envelope, 17 – ribosome constituents, 18 – foreign DNA, 19 – regulation, 20 – not classified

Figure 4.15: Pseudogene and partial gene distribution in the *Y. enterocolitica* reference genomes.

BT 1A also shows a high number of pseudogenes and partial genes in protection responses (3) when compared to the other BTs. Some of these CDSs relate to adhesion/hemagglutinin proteins unique to BT 1A, an enterotoxin-like protein of *Y. pestis*, and a hemolysin paralog of BT 1B. Furthermore, BT 1A has a stop codon mutation in *umuD* (*YE1733*), a UV protection protein. There is a functional unique copy of *umuD* elsewhere in the genome though. Interestingly, in this second operon, *umuC* carries a frameshift mutation. A drug resistance operon shared in BTs 1A and 1B has been deleted from the low-pathogenic BTs. These retain only fragments of two proteins encoded in this region (*YE0443, YE0445*). A stop codon has been introduced into the BT 1A homolog of *YE0444*, hence the functionality of this multidrug efflux protein is compromised. Some mutations are found in the *tc* PAI. The BT 2 O:9 strain has a frameshift mutation in *tccC2* due to the fact that this CDS links two contigs. Therefore it could be functional, but sequence information is missing. Irrespective of this, two additional, functional copies of *tccC* are present upstream. BTs 4 and 5 also show mutations in *tc* PAI. Both have mutations in the regulator *tcaR2* and the protein *tcaC'(2)* in common, and BT 5 additionally has mutation in *tcaB2*. BT 1B has no mutations in protection responses.

The lowest number of pseudogenes in transport and binding proteins (4) is given in BT 1A. The highest number of pseudogenes is given in BTs 4 and 5. Transport proteins are involved in sugar uptake and metabolism, and a higher number of pseudogenes in pathogenic BTs could reflect the reduced metabolic capabilities apparent in the biotyping scheme. With respect to adaptation (5), both BT 1A and BT 1B only have one partial gene. The acid shock protein Asr (*YE0933*) is missing 15 aa in a repeat region, resulting in a shortened protein. This same mutation is evident in the low-pathogenic BTs. Both BT 2 and BT 3 O:9 isolates also show a truncation of another copy of this protein (*YE2086*) which is functional in the other BTs. The pseudogenes in BT 5 are located in the haemin storage operon. A stop codon has been introduced in *hmsF* (*YE2483*), and *hmsH* (*YE2484*) shows a frameshift mutation in a homopolymeric run. In BT 4 this last protein also has a frameshift, with 10 amino acids missing.

BT 1A shows no mutations in macromolecule metabolism (7). A beta-glucosidase (*YE1236A*) has been truncated in both high- and low-pathogenic BTs due to a deletion event. The other pseudogene in BT 1B is a putative protease (*YE1406*). A putative DNA-binding protein (*YE1874*) and a metalloprotease (*YE4052*) are commonly fragmented in the low-pathogenic BT. BTs 4 and 5 again share additional

pseudogenes. A putative serine protease (*YE1389*) absent from BT 1A has a stop codon introduced in BT 4, and several stop codons plus a frameshift in BT 5. Further more in BT 5, there is an exported zinc protease (*YE0320*) with a frameshift in a homopolymeric run.

The partial genes in macromolecule synthesis and modification (8) found in BT 1A are mostly in CDSs that are unique to this strain. Annotation is thus based on comparison with homologs in other bacterial species and functionality is unknown. This applies also to the pseudogenes in BT 4; these are also not found in the backbone of *Y. enterocolitica* CDSs but within the accessory genome. The inactivation of those genes might present niche adaptation. One partial gene in BT 1A, *ampM* (*YE3285*) a methionine aminopeptidase, is also mutated in BT 5. Interestingly, BT 1B shows no mutations in this class.

There are only very few pseudogenes and partial genes in the metabolism of small molecules (9). In BT 1A, this is due to an additional copy of *trpA* involved in tryptophan synthesis. This protein is missing 160 amino acids at the N terminus, but due to a functional gene copy present elsewhere in the genome, this should have no impact on metabolism of tryptophan. In BT 1B the inactivated gene is a unique possible homoserine synthase (*YE1737*) that has been truncated by the insertion of an upstream IS element. A common pseudogene in the low-pathogenic BTs is a frameshift in the glutamate synthase precursor *gltB* (*YE3735*). One of the pseudogenes in BT 4 is in the tryptophanase *tnaA* (*YE0650*). The protein could be responsible for a negative phenotype for indole production in the biotyping reaction, but does not explain why BTs 3 and 5 show a negative phenotype as well.

The group of biosynthesis of cofactors and carriers (10) shows a marked increase in partial genes in BT 3 O:5,27 strain 14902. All pathogenic BTs carry pseudogenes in genes encoding the metabolism of the cofactors cobalamin and cobinamide, *cbiG* (*YE2719*) and *cobC* (*YE2708*). The exact effect of these mutations is unknown as there are potential alternative pathways. BT 5 also possesses two more pseudogenes in this pathway affecting the cobinamide kinase CobU and a membrane protein (*YE2724*).

There are few mutations in central intermediary metabolism (11) in the non- and high-pathogenic BTs. BT 1A has one mutation in *pldB* (*YE0207*), a lysophospholipase. A stop codon has been introduced terminating the protein 7 amino acids early. An impact of this mutation is equivocal. All BTs except BT 1A show various degrees of fragmentation and remnants of *YE2586A* encoding a CoA

transferase familiy protein. All low-pathogenic BTs also only have the 3' end of *ssuD* (*YE1950*) encoding an alkanesulfonate moooxygenase. Additionally pseudogenes are found in BT 5 in the backbone of *Y. enterocolitica*. Both *amn* (*YE2052*, AMP nucleosidase) and *ushA* (*YE3066,* UDP sugar hydrolase) are missing the 3' end, *amn* also shows a frameshift mutation.

BT 1A shows two pseudogene mutations in the degradation of small molecules (12). An opal UGA stop codon has been introduced in *idh* (*YE4026*, m-inositol dehydrogenase). The other mutation is shared in the non- and low-pathogenic BTs and regards the xylose kinase *xylB* (*YE4123*). Stop codons have been introduced towards the C terminal end of the protein. Both *idh* and *xylB* could still be functional, as the full coding sequence is still present. *xylB* could be linked to xylose metabolism, but the xylose acid production in the biotyping test for *Y. enterocolitica* does not agree with the mutation pattern. BTs 1A, 1B, 2 and 3 are positive for xylose acid production, only BT 4 is negative, and BT 5 variable. The genetic basis for the biotyping reaction is unknown. An operon involved in N-acetyl-D-galactosamine metabolism has been deleted from BT 1B. This is evident in the gene remnants encoding for AgaS (*YE2449*) and AgaY (*YE2450*), which are missing the N and C terminus, respectively. BT 5 also shows mutations in this operon, with frameshifts present in *agaS* and *agaW.*

In carbon energy metabolism (13), most mutations again are evident in the low-pathogenic BTs, although BT 1A has the highest number of total CDSs in this group. The mutations present in BT 1A are outside of the core genome for *Y. enterocolitica*. The mutation present in BT 1B is due to a deletion event of a region involved in cytochrome C type biogenesis. Only a remnant of the nitrate reductase *nrfG* (*YE1598*) is present. The low-pathogenic BTs also show mutations in this cluster in *nrfF* and BT 5 also in *nrfE*, thus a fully functional operon might only be present in BT 1A. The carbonic anhydrase *cah* is absent from BT 1B and has been disrupted in BTs 2, 3, and 4 by a transposase. It is only functional in BTs 1A and 5. BT 5 again has additional mutations in pathways involving NADH-quinone reductase (*ngrC, YE3218*), formate dehydrogenase (*fdhF, YE2810*), tetrathionate reductase (*ttrA, YE1617*), and cytochrome C type proteins (*napC, YE1158*).

The non- and high-pathogenic BTs have no mutations in proteins involved in fatty acid biosynthesis (14). The low-pathogenic BTs have retained a remnant of the Yersinia genomic island 2 (*fabF, YE0911*). The N terminus has been deleted possibly due to the presence of a transposase. No mutations are evident in BTs 1A, 1B, 3

O:5,27, and 4 for nucleotide biosynthesis (15). The two O:9 isolates have a stop codon in the cytosine deaminase *codA* (*YE3964*). BT 5 has a N-terminal deletion in a protein involved in purine metabolism.

Most of the CDSs mutated in BT 1A in the cell envelope (16) relate to putative exported proteins. The cellulose synthase 1 catalytic subunit gene *bcsA* has an opal stop codon introduced at the 5' end, thus the function is questionable. The pseudogenes found in the low-pathogenic BTs are often shared and related to exported proteins without specific functions.

The highest number of pseudogenes and partial genes is found in genes encoding products that belong to the functional group of "foreign DNA" (18). This group contains IS elements and phage proteins. These were not corrected or re-checked for their mutations, and the true numbers are probably lower than stated here. Moreover, especially the low-pathogenic BTs show an expansion of IS elements, therefore the number of CDSs contained in this class is higher than in BT 1A and 1B. Even though the number of pseudogenes is corrected for the total number of CDSs per group, the fact that these CDSs were not re-checked increases their overall contribution.

Most of the pseudogenes in the low-pathogenic BTs involved in regulation (19) are shared in all BTs, and a function has only been retained in the high- and non-pathogenic BTs. For example, the regulator of the tetrathionate reduction operon, *ttrS* (*YE1614*), has a stop codon introduced in the O:9 BTs 2 and 3, and a frameshift in BT 5. As mentioned earlier, BT 5 has an additional mutation in this operon.

Unclassified and putative assignments (20) constitute proteins of unknown function and preliminary assignments, similar to hypothetical proteins (1). It appears that the number of CDSs is again highest for pseudogenes found in BT 4 and 5.

To summarize there are certain trends within the pseudogenes. The three lineages of high-, low-, and non-pathogenic BTs described for the shared genome content are also visible when comparing the pseudogenes. It appears that especially BTs 4 and 5 have acquired a considerable number of pseudogenes and partial genes. The predominant groups (Figure 4.16) for pseudogenes in BT 1A are cell processes, macromolecule synthesis and modification, and cell envelope proteins. Pseudogenes in the pathogenic BTs relate to transport/binding proteins, proteins related to the cell envelope, and regulatory proteins. BTs 4 and 5 additionaly show pseudogenes in

degradation of small molecules, which could reflect the reduced metabolic potential visible in the biotyping scheme.

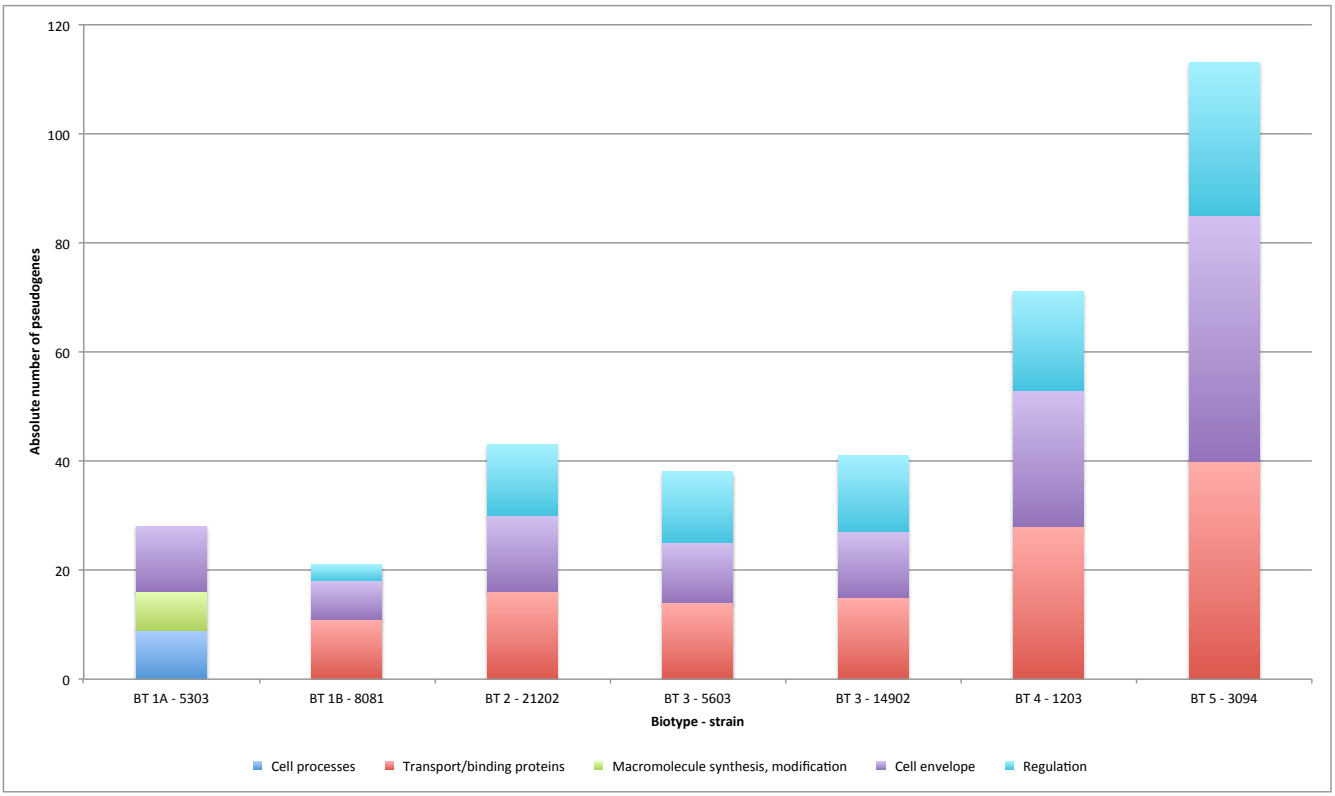


Figure 4.16: Top 3 pseudogene categories for *Y. enterocolitica* BTs. Absolute numbers are given only for top 3 categories, disregarding possible pseudogenes in other categories.

Fewer pseudogenes are observed in the non- and high-pathogenic BTs than in the low-pathogenic BTs. The non-pathogenic BT has more pseudogenes than partial genes, whereas it is the other way round in the high-pathogenic BT. Hence, the high-pathogenic BT shows a higher degree of genome decay than the non-pathogenic BT. Quite often the pseudogenes in the low-pathogenic BTs are shared in all of those isolates. BTs 4 and especially 5 have additional mutations. The mutations observed in the same genes in BTs 4 and 5 are not identical, but appear to have arisen independently.

4.2.4 The role of IS elements and recombination in *Y. enterocolitica* biotypes

IS elements are believed to have played an essential part in the evolution of *Y. pestis* from *Y. pseudotuberculosis* through transposition into coding genes and through genomic rearrangements between IS elements (Parkhill, et al., 2001a; Deng, et al., 2002; Chain, et al., 2004). During sequencing and assembly of the *Y. enterocolitica*

biotypes, it became clear that there was a predominance of one particular IS element (IS1667). The total breakdown of all IS elements is given in Table 4.2.

Table 4.2: IS elements in the *Y. enterocolitica* BT

IS name	BT1B – 8081*	BT1A – 5303	BT2 – 21202	BT3 – 5603	BT3 – 14902	BT4 – 1203	BT5 – 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

*taken from Thomson et al., 2006

There is an uneven distribution of IS elements in the BTs. The reference genome for BT 1A shows none of the typical IS elements found in the other BTs. A search in other BT 1A isolates showed that this might be atypical and copies of IS1329 and IS1667 could be present in BT 1A. Additionally, IS1222 is present, but not consistently. There are a number of unclassifed IS elements which have homologies to ISEhe3 and IS621 found in *E. coli*, transposases similar to IS116/IS110/IS902 of *Stenotrophomonas*, and transposases similar to IS2 and IS3, but these have not been characterized in more detail here.

The IS elements of BT 1B belong to a wide range of different families, with no more than 8 copies of any single IS element. In total, BT 1B carries 60 IS elements. A similar number of IS elements can be found in the low-pathogenic BTs 2 – 5, but here the distribution is skewed towards IS1667, which is present in more than 30 copies per strain. Other IS elements present are IS1328, IS1400, IS1668, and IS1669.

The position of IS1667 (Figure 4.17) indicates that it might be mediating intragenomic recombination, as it is often found at borders of transpositions and inversions.

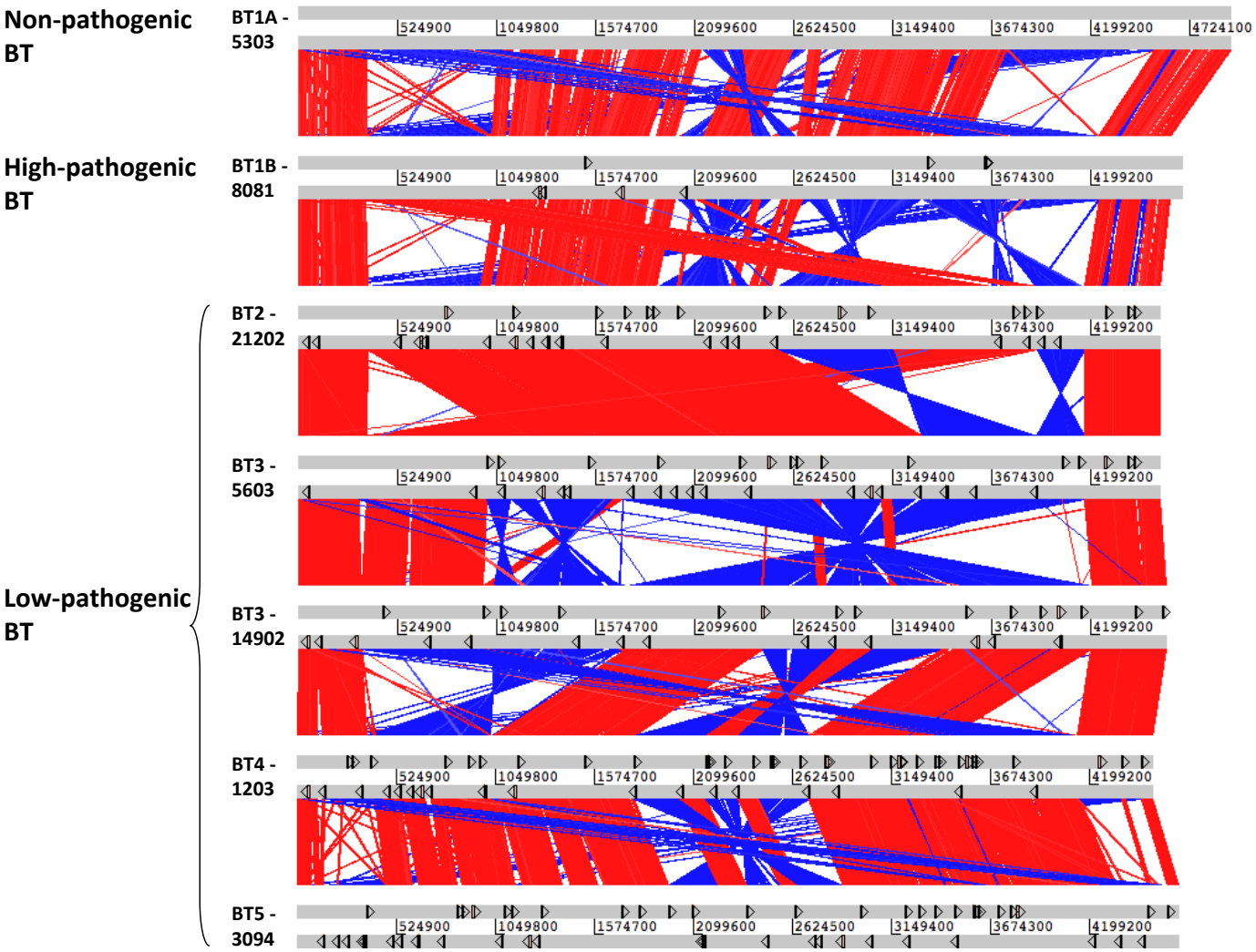


Figure 4.17: Position of the predominant IS element IS1667 in the reference genome assemblies. Genomes are compared in a pair-wise fashion. Red bars – co-linear homology >999 bp, blue bars – inverted homology >999 bp. Contigs are not highlighted. Arrows indicate position and orientation of the IS1667.

As is the case in *Y. pestis*, IS elements can also be shown to promote genome instability via recombination. There are examples of IS elements which have inserted into CDSs and thus disrupt their expression in the low-pathogenic BTs. One example mentioned above is the carbonic anhydrase *cah*, which has been disrupted by a transposase in BTs 2, 3, and 4. Another good example as well is found in BT4 where transposase action lead to the loss of half of the flagellar cluster and the iron-uptake operon, as described before.

As can be seen from Figure 4.17 there are some regions in the *Y. enterocolitica* reference genomes that are not to affected by IS mediated recombination or disruption. These regions retain the same gene arrangements althought their position respective to each other can vary. This concept has been introduced in *Y. pestis* by Deng, et al. (2002), and was used to describe syntenic blocks. Later Chain, et al.

(2004) tried to deduce the number of recombinatorial events in *Y. pestis* since the divergence from *Y. pseudotuberculosis* through the investigation of those syntenic blocks.

This concept can also be applied in *Y. enterocolitica*, and 29 syntenic blocks could be identified through the comparison of the seven reference genomes (Table 4.3).

Table 4.3: Syntenic blocks identified in the reference genomes of *Y. enterocolitica*.

Syntenic block	start position†	end position†	CDS start†	CDS end†	size	# CDS
1	0	107030	YE0001	YE0097	107030	97
2*	108300	191330	YE0099	YE0160	83030	61
	191500	202555	YE0161	YE0168	11055	7
	282830	313630	YE0169	YE0273	30800	104
3	319700	353900	YE0274	YE0303	34200	29
4	362987	874170	YE0307	YE0747	511183	440
5	879520	980830	YE0754	YE0852	101310	98
6	1039400	1251800	YE0911	YE1115	212400	204
7	1264600	1802600	YE1124	YE1602	538000	478
8	1802900	1991400	YE1603	YE1798	188500	195
9	2027865	2087750	YE1846	YE1901	59885	55
10	2108500	2142600	YE1933	YE1962	34100	29
11	2154500	2325300	YE1975	YE2124	170800	149
12	2326000	2406630	YE2125	YE2196	80630	71
13	2407700	2446400	YE2198	YE2234	38700	36
14	2447240	2499700	YE2237	YE2285	52460	48
15	2554700	2591553	YE2365	YE2402	36853	37
16	2602500	2630230	YE2409	YE2439	27730	30
17	2640950	2668185	YE2454	YE2477	27235	23
18	2668285	2709585	YE2478	YE2513	41300	35
19	2709700	2800900	YE2514	YE2602	91200	88
20	2854263	3294700	YE2636	YE3022	440437	386
21	3313400	3544778	YE3048	YE3248	231378	200
22	3552700	3610900	YE3252	YE3308	58200	56
23	3610900	3712864	YE3309	YE3399	101964	90
24	3726200	3761260	YE3413	YE3447	35060	34
25	3897560	3933900	YE3581	YE3609	36340	28
26	3960000	4238800	YE3631	YE3882	278800	251
27	4245400	4464400	YE3883	YE4083	219000	200
28	4504400	4561500	YE4118	YE4162	57100	44
29	4561800	4615788	YE4163	YE4215	53988	52

† labelling respective to BT 1B strain 8081; * 7 CDSs present a clean transposition in BT 5 due to the action of transposases; colours correspond to syntenic blocks highlighted in Figure 4.18.

The syntenic blocks vary in size and gene content, ranging from 23 CDSs (17) up to 478 CDSs (7). It is possible that with sequencing of further *Y. enterocolitica* genomes this number could change, but considering the divergence between these isolates gives confidence in the assignment of syntenic blocks. An example for the variable positions of the syntenic blocks is given in Figure 4.18.

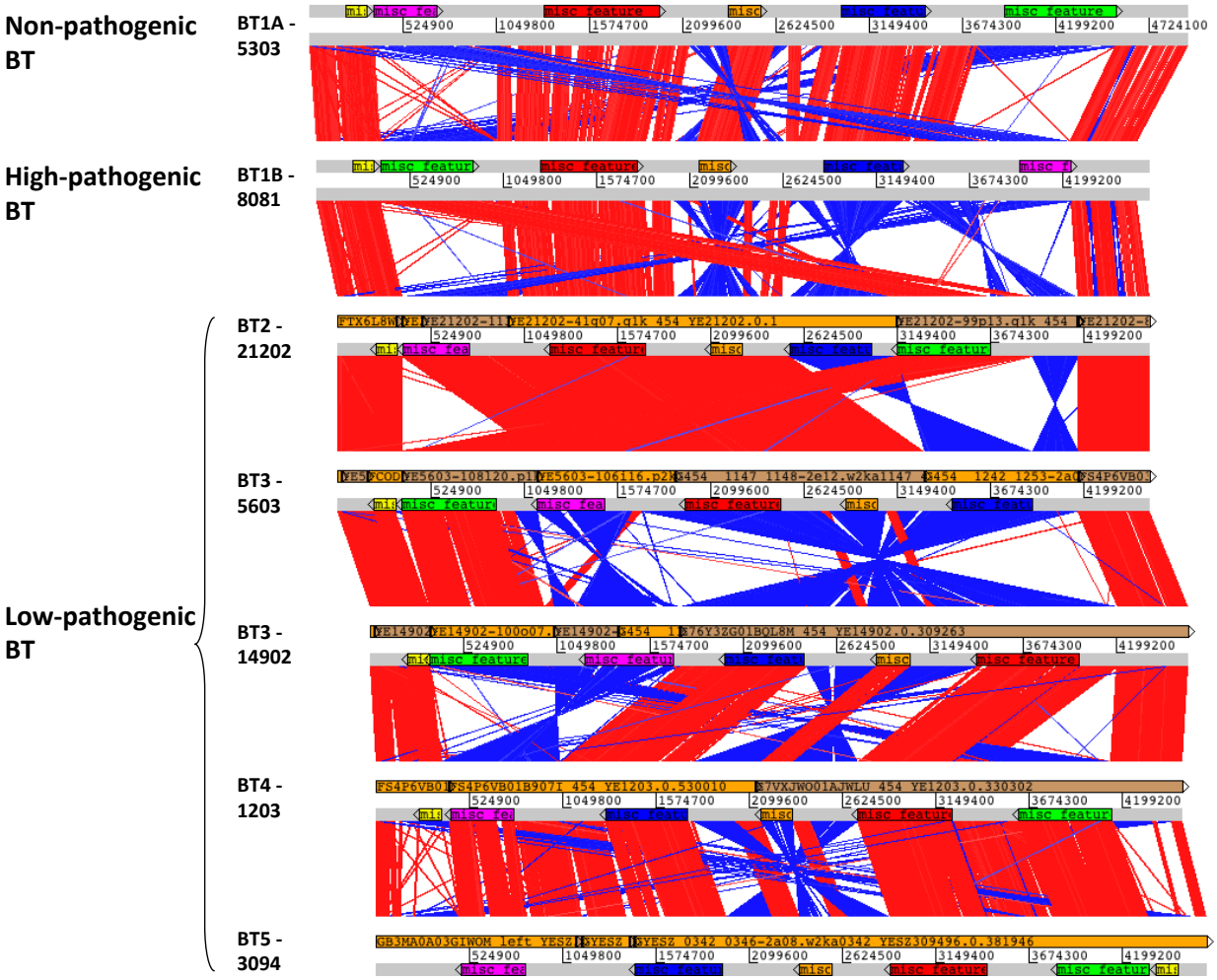


Figure 4.18: Exemplar position of six syntenic blocks. Six syntenic blocks are highlighted in yellow, pink, red, orange, blue, and green in the respective reference genomes to illustrate their variable positions within the genomes relative to the sequence start. Colours correspond to syntenic block colours in Table 4.3. Contigs are highlighted in orange and brown in the low-pathogenic BTs.

As one can see none of the regions appear in the same order and orientation as in any of the other genomes. BT 1A and BT 2 do have the same order in the six syntenic blocks observed, but the blue and green block show inversions which are not easily visible as the syntenic blocks have been artificially allocated to the reverse strand to allow for visualization of the contigs on the forward strand.

The identification of syntenic blocks could help in the design of a multilocus sequencing typing (MLST) scheme. One of the underlying principles for MLST is that the genes should be evenly distributed around the chromosome to exclude that the genes are linked with respect to evolutionary changes and recombination events (Maiden, et al., 1998). In chosing genes located within major syntenic blocks one can ensure that the genes are unlinked, as the blocks are conserved in structure. Despite genomic rearrangements due to recombination that change the position of the blocks, the syntenic blocks should cover different parts of the chromosome and guarantee an even spread.

4.3 Discussion

4.3.1 High-, low-, and non-pathogenic biotypes have distinct genomic features

Through the analysis of orthologous gene sets it is apparent that high-, low-, and non-pathogenic BTs are very distinct in their genomic content. High- and non-pathogenic BTs each have a considerable number of unique CDSs, whereas the low-pathogenic BTs appear as a tight cluster of isolates. The regions specific to the low-pathogenic BTs are shared in all of the isolates, and unique regions only relate to independently acquired phages and LPS operons. It is important to note that BT 2 and BT 3 are genetically indistinguishable from each other. BT 4 has lost the *fes-fep* iron acquisition operon and half of the second flagellar cluster, and this has been confirmed in larger strain collections for *fes-fep* operon (Schubert, Fischer and Heesemann, 1999; Thomson, et al., 2006). The loss of half of the flagellar cluster has not been reported before, and could only be detected through in-depth comparison between all of the reference genomes. BT 5 is signified by massive genome decay. The genetic basis for the biotyping reactions could only be determined for trehalose and sorbose, both pseudogenes in BT 5.

It was interesting to see how closely related high- and non-pathogenic BTs appear to be with respect to shared gene content. On the other hand, the non- and low-pathogenic BTs also do have a lot in common. In fact, these two groups incorporate more CDSs than the supposedly more important group of pathogenic BTs. A common gene set of pathogenic BTs is reduced to the virulence plasmid pYV and the attachment/invasion locus protein Ail. There are also no signatures for human-pathogenic isolates or for animal isolates. A similar finding has been made when comparing pathogenic and commensal *E. coli* strains (Chaudhuri, et al., 2010). The authors found only 13 CDSs that resembled potential virulence factors absent from the commensal strain. They suggest that this absence of pathogenic signature indicates a phylogenetic heterogeneity of enteroaggregative *E. coli* (Chaudhuri, et al., 2010). It will be investigated later if this case is similar in *Y. enterocolitica* (Chapter 7), but the split into high- and low-pathogenic BTs due to heterogeneity resembles the findings in *E. coli*.

Many genes and operons have been suggested as “virulence markers” or “virulence genes” of pathogenic *Yersiniae* in the past, only later for homologs to be found in environmental species as well. As classic example of this is the presence of hitherto

considered *Yersinia* virulence genes in BT 1A including the invasion gene *invA*, the tight adherence operon and two YGI (Pierson and Falkow, 1990; Thomson, et al., 2006). Similarly, the host-responsive element *rscBACR* (Young and Miller, 1997; Nelson, Young and Miller, 2001), thought to be involved in the dissemination of *Y. enterocolitica* in the spleen, is found in all BTs. These discoveries may help explain why BT 1A is increasingly being associated with mild human intestinal infectious disease. The only robust markers for pathogenicity identified in BT 1B relate to the HPI, Ysa T3SS, and *yst-1.*Some important regions and islands of *Y. enterocolitica* and their distribution within the species are shown in Table4.4.

Table 4.4: Distribution of important regions and islands of *Y. enterocolitica* in reference genomes.

	BT 1A	BT 1B	BT 2	BT 3	BT 4	BT 5
YGI-1 (*tad*)	+	+	+	+	+	+
YGI-2	+	+	-	-	-	-
YGI-4	+	(+)	-	-	-	-
Ni-Co efflux	+	+	-	-	-	-
Ag-Cu efflux	+	-	-	-	-	-
tc PAI	-*	-	+	+	+	+
Flag-1	+	+	+	+	+	+
Flag-2	-	-	+	+	(+)	+
Outer core LPS	+	-	+	+	+	+
YGT	+	-	+	+	+	+
Ysa T3SS	-	+	-	-	-	-
pYV	-	+	+	+	+	+
fes-fep	+	+	+	+	-	+
ail	-	+	+	+	+	+
HPI	-	+	-	-	-	-
yst-1	-	+	-	-	-	-
yst-2	+	+	+	+	+	+
YAPI	-	+†	-	-	-	-
inv/rovA	+	+	+	+	+	+
myf fibrillar operon	+	+	+	+	+	+
hydrogenase-2,4	+	+	+	+	+	+

* only described in one BT 1A isolate so far (T83, Tennant, Skinner, Joe and Robins-Browne, 2005); † not consistently present in all BT 1B (Thomson et al., 2006); (+) only truncated versions present

The regions that were thought to be involved in niche adaptation to the human gut now need to be re-evaluated for their contribution to disease. The distribution of putative virulence factors should also be examined in larger strain collections of *Y. enterocolitica* as well as environmental *Yersiniae*.

4.3.2 Metabolism, pseudogenization and IS elements are important evolutionary forces in *Y. enterocolitica*

A theme emerging from the comparison of the different BTs is the potential difference in metabolic properties. Regions that are unique to BT 1A, regions shared between BT 1A and 1B, and regions common to BT 1A and 2 – 5 relate to metabolic pathways and transport operons. Similarly, the comparison of the pseudogenes found that especially the BTs 4 and 5 show an increase in pseudogenes in operons associated with metabolism and transport. This trend of reduced metabolic flexibility is also apparent in the biotyping scheme. BT 1A is positive for all of the reactions, then BT 1B follows, and a reduced number of positive reactions is observed in the low-pathogenic BTs, finally arriving at BT 5 which is negative or variable for all but one of the reactions (Wauters, Kandolo and Janssens, 1987).

Metabolism is also a driver in the host-adaptation in other pathogenic bacteria (Rohmer, Hocquet and Miller, 2011). Metabolic streamlining to a specific niche can be specifically observed in obligate intracellular organisms for instance when comparing *Mycobacterium tuberculosis* and *Mycobacterium leprae* (Cole, et al., 2001). Here it was found that *M. leprae* shows signs of extensive genome decay in form of accumulation of a large number of pseudogenes. Predominant functional groups of pseudogenes are, similar to *Y. enterocolitica*, small molecule catabolism, energy metabolism, regulatory proteins, synthesis and modification of macromolecules, cell envelope proteins, and transport proteins. Similar trends of pseudogene acquisition alongside host or niche restrictions are also observed in *Salmonella enterica* serovars Tyhpi (Parkhill, et al., 2001b) and Gallinarum (Thomson, et al., 2008). Again, targets for pseudogenization are metabolic pathways. Interestingly, there are certain pathways found to be targets of pseudogene formation in *Y. pestis*, *S.* Typhi, and *S.* Gallinarum (Parkhill, et al., 2001a,b; Thomson, et al., 2006, 2008) that are also affected by mutations in the potentially highly successful, more host-restricted lineages of *Y. enterocolitica* (BTs 1B, 4, 5). These relate to cobalamin biosynthesis, tetrathionate respiration, and mutations in flagella operons. *Y. ruckeri* was also found to be deficient in cobalamin biosynthesis and tetrathionate respiration (Chen, et al., 2010). These could potentially be pathways essential in environmental survival and which are lost upon host adaptation due to maintenance cost. For tetrathionate respiration it has also been suggested that the loss of the pathway could be advantageous for growth outside the intestine (Thomson, et al., 2008; Rohmer, Hocquet and Miller, 2011). A recent study (De Paepe, et al., 2011)

investigated the effect of colonization of germ-free mouse intestines with *E. coli* and found a diversification process. The original *E. coli* population split into separate lineages, one acquiring mutations in the flagellar *flhDC* operon, another showing mutations leading to a maltose-negative phenotype. This microevolution is potentially driven by the presence of bile salts (De Paepe, et al., 2011). The authors also suggest that maintenance of flagellar operons might require extensive energy, and that functionality in the gut is dispensible. A similar dynamic could explain the parallel loss of flagellar operons in other pathogenic bacteria.

It is not known though if the regions uniquely acquired by BT 1A or the regions showing pseudogenes in the pathogenic BTs do in fact have a phenotype associated with them. This will be investigated further in Chapter 6. Pseudogenes are also said to be lost rapidly in populations as their truncated transcript could have negative effects on the cell (Kuo and Ochman, 2010). It could thus be that a lot of the pseudogenes in BTs 4 and 5 have been acquired recently and thus have not been completely deleted yet, or that in fact these pseudogenes still fulfil a function and thus are retained. Again, confirmation of pseudogene status is subject to further phenotypic experiments.

IS elements also present an important evolutionary driving force. The expansion of IS1667 in the low-pathogenic BTs mirrors the overall expansion of IS elements in *Y. pestis*. A similar finding was made in *Burkholderia mallei*. This horse-pathogen evolved out of the environmental *B. pseudomallei*. It was shown that there has been an expansion of three different IS elements in *B. mallei* since its divergence from *B. pseudomallei* (Song, et al., 2010). These have been attributed to genome decay and reduction in overall genome size. Interestingly, the authors found that a large number of IS elements were conserved with respect to genomic location in several unrelated *B. mallei* genomes indicating that these might have been acquired during evolution of this lineage (Song, et al., 2010). Following geographic spread of the bacteria, there have been few further expansions and transpositions of the IS elements. This can also be observed to some respect in the low-pathogenic BTs. Amongst the orthologous genes of the low-pathogenic BTs, there are 35 transposases and IS elements, of which 10 belong to IS1667. On the other hand, transposition appears to be an ongoing process after the separation into different bio-serotypes. As an example, some transposase are further shared between BTs 4 and 5, and BT 5 possesses 35 unique transposases (12 IS1667). Active IS elements

have also been observed in *Y. pestis* when one genomic region was found to be present in two orientations suggesting a high genome fluidity (Parkhill, et al., 2001a). A proteomic study also found unique peptides corresponding to active transposases of IS285, demonstrating that genome recombination is an ongoing process in *Y. pestis* (Payne, Huang and Pieper, 2010).

4.3.3 Differences in niche adaptation

It is apparent that the non-pathogenic BT 1A is suitably adapted for environmental survival. It possesses a large number of operons associated with transport and metabolism of various sugars and other metabolites.

The pathogenic BTs lack most of these properties and seem to have chosen a different approach. The high-pathogenic BT seems to have found a unique niche that allowed it to get in contact with other bacteria to exchange pathogenicity islands, thus acquiring the HPI and the Ysa T3SS. In turn, it loses the YGT and regions associated with generalist survival. The YGT is present in all *Yersiniae* but shows signs of fragmentation specifically in more pathogenic lineages. *Y. pestis* and *Y. pseudotuberculosis* are missing parts of the effector region, explaining why an initial investigation into functionality showed no role in pathogenicity (Pujol and Bliska, 2003; Balada-Llasat and Mecsas, 2006). This cluster has also been deleted from the high-pathogenic BT 1B and shows signs of pseudogenization in BTs 4 and 5, which are potentially the most host-adapted low-pathogenic BTs. BT 4 is isolated from animals and the environment as well, but presents the predominant bio-serotype in human infections in Europe (McNally, et al., 2004), suggesting some degree of host-adaptation over the other low-pathogenic BTs. This inactivation of YGT in host-adapted lineages and the distribution in the environmental *Yersiniae* indicate that the YGT possibly has a role in environmental survival rather that in causing human disease. YGT resembles the SPI-2 T3SS of *Salmonella*. This system was identified together with the SPI-1 T3SS following signature-tagged mutagenesis and isolation of attenuated clones in a mouse infection model (Shea, Hensel, Gleeson and Holden, 1996). SPI-2 was later implied to be involved in intracellular pathogenesis and survival within macrophages (Hensel, 2000). Recently though it has also been associated with survival inside amoebae and it has been suggested that the mechanisms for survival inside amoebae is the same as for macrophages (Bleasdale, et al., 2009). Insect and protozoa could present an important environmental niche not only for *Salmonella* but also for *Yersinia*. Further

investigations could be carried out to determine a role in colonization of amoeba or insects.

The low-pathogenic BTs have kept the second flagella cluster and possess the *tc* PAI involved in insect toxicity. Instead of a large metabolic repertoire required in the general environment, the low-pathogenic BTs appear to have adapted to an insect and amoebae niche. *Salmonella* has been described as containing two flagella cluster as well, and it was hypothetized that the second cluster might act as a spare tyre in that the loss results in host restriction whereas a retention enables colonization of environmental niches (McQuiston, Fields, Tauxe and Logsdon, 2008). On the other hand a second flagellar cluster might enable phase variation and thus immune evasion, especially in a new niche (McQuiston, Fields, Tauxe and Logsdon, 2008). No phase variation has been described in *Yersinia* thus far, and the fact that both flagella clusters are deactivated in *Y. pestis* might argue the point that Flag-2 is involved in environmental survival in the low-pathogenic BT, possibly due to their reduced metabolic capacities. It has been suggested that a second flagella cluster supports the search for an insect vector (Bresolin, Trček, Scherer and Fuchs, 2008). Mutations in BTs 4 and 5 on the other hand again possibly reflect host restriction to a niche other than insects, similar to findings made in other pathogenic bacteria, namely *Salmonella* Typhi and Gallinarum as well as *Y. pestis* (Thomson, et al., 2008; Parkhill, et al., 2001a).

An important part in infection is immune evasion and how the bacteria present themselves to a host through outer membrane proteins. Both core and outer antigen are recognized by the adaptive immune response and trigger the host complement cascade that results in attraction of phagocytes and lysis of bacterial cells, and are also important in bacterial cell membrance integrity and resistance (Prior, et al., 2001; Skurnik, Venho, Bengoechea and Moriyón, 1999). It is therefore interesting to observe the differences in LPS makeup between high-pathogenic BTs on the one hand and non- and low-pathogenic BTs on the other hand. Because the pathogenic and non-pathogenic *Y. enterocolitica* lineages exhibit a large number of different LPS structures, full protective immunity might only be acquired over time upon infection with each lineage. This would impact on vaccine strategies, but are of low impact due to the usually self-limiting nature of the disease.

The fact that the outer core is shared in non- and low-pathogenic BTs indicates that this might be the ancestral LPS structure of *Y. enterocolitica*. It is interesting to see

that the high-pathogenic BT 1B abandoned the outer core in favour of a heteropolymeric operon, whereas the low- and non-pathogenic BTs seem to have acquired additional homopolymeric operons. The LPS structure of the high-pathogenic BT thus resembles more the structure of *Y. pseudotuberculosis* and *Y. pestis* (Skurnik and Bengoechea, 2003). It has been suggested that an ancestor of the pathogenic *Yersiniae* had acquired a common ancestral version of this heteropolymeric O-antigen (Zhang, Toivanen and Skurnik, 1996). Considering that the non-pathogenic BT exhibits signs of being the historically oldest lineage, and that it shared the outer core biosynthesis operon with the low-pathogenic BTs, this scenario is unlikely. It is more likely that the outer core presents the ancestral form in *Y. enterocolitica* and that it has been replaced in BT 1B upon colonization of a new niche. So far, nothing is known about the genetic organization and location of the O-antigen operons in the environmental *Yersiniae*, but it has been suggested that the outer core is also present in the environmental species (Skurnik, et al., 1999). A quick search on the published genomes (Chen, et al., 2010) confirmed the presence of outer core in *Y. frederiksenii, Y. kristensenii* and *Y. rohdei*, but due to possible assembly problems, no definitive answer on the other isolates is possible. Thus the directionality of the outer core is loss in the high-pathogenic BT. It also has to be considered though that the O-antigens undergo horizontal gene transfer as well. This has been well documented in *E. coli* and *Salmonella* as well as for *Y. pseudotuberculosis* (Samuel and Reeves, 2003; Skurnik and Bengoechea, 2003).

In conclusion, the comparative analysis carried out enabled a deeper look into what is shared and what is different in *Y. enterocolitica* BTs. This detailed knowled is specifically important with respect to truncation and mutations of regions in view of BTs 4 and 5. A pathogenic core of genes shared exclusively in high- and low-pathogenic BTs as such does not exist except for the virulence plasmid pYV and the adhesion/invasion locus protein Ail. The presence of some genes that were thought to be virulence determinants in the past in BT 1A could explain why it can be associated with disease in humans. The lack of markers for increased pathogenic potential such as the HPI explains the lowered pathogenicity of BTs 2 – 5 in a mouse infection model. The regions unique to low-paths indicate a potential environmental niche associated with insects. Evolutionary driving forces for pathogenesis can be defined as IS elements and pseudogenization as well as metabolic properties.

5 Plasmid Variety and Variability in *Y. enterocolitica* Biotypes

5.1 Introduction

The only common “virulence factors” shared in pathogenic BTs of *Y. enterocolitica* are the chromosomally encoded *ail* as well as the virulence plasmid pYV. It was shown in the past that functionality of pYV is conserved, despite marked sequence differences and size variations between plasmids of high- and low-pathogenic BTs of *Y. enterocolitica* as well as *Y. pseudotuberculosis* (Portnoy, Moseley and Falkow, 1981; Portnoy and Falkow, 1981; Heesemann, et al., 1983; Portnoy, et al., 1984; Kapperud and Nesbakken, 1987; Fukushima, Gomyoda, Aleksic and Tsubokura, 1993). Variability was also shown in the sequence, orientation and position of some genes as well as gene content in general (Roggenkamp, et al., 1997; Boland, Havaux and Cornelis, 1998; Snellings, Popek and Lindler, 2001; Wagner, et al., 2009; Foultier and Cornelis, 2003; Oberhettinger, et al., 2011). The plasmids can consequently be considered as a "closely related family of plasmids" rather than a uniform molecule (Portnoy, Moseley and Falkow, 1981; Portnoy and Falkow, 1981).

A variety of other plasmids are found in the genus *Yersinia*. Analysis of plasmids in *Y. enterocolitica* strains showed that strains often harbour more than one plasmid of different types and sizes (Kapperud and Nesbakken, 1987). Those plasmids vary in sequence similarity. Plasmids have also been the object of interest in the non-pathogenic BT1A with a considerable number of isolates containing plasmids of varying sizes (Lewin, et al., 1996). Although BT 1A is considered to be non-pathogenic due to the lack of the virulence plasmid, this study showed that some of the plasmid-bearing strains had atypical biotyping reactions; they exhibited autoagglutination, slow and weak Congo Red uptake, and were also resistant to serum (Lewin, et al., 1996).

The aim of this chapter is to investigate the differences in virulence plasmids found in the pathogenic *Y. enterocolitica* biotypes. Differences found will be investigated in terms of distribution in high- and low-pathogenic BTs and are also compared to *Y. pseudotuberculosis* and *Y. pestis*. The plasmid repertoire of *Y. enterocolitica* will also be considered as several unique plasmids were identified during sequencing.

5.2 Results

5.2.1 Genetic Organization of the virulence plasmid pYV in *Y. enterocolitica* BT

General properties of the sequenced virulence plasmids in the reference genomes are given in Table 5.1.

Table 5.1: General properties of the sequenced pYV.

Property	BT1B O:8 pYVe8081	BT2 O:9 pYVe21202	BT3 O:9 pYV5603	BT3 O:5,27 pYVe14902	BT4 O:3 pYVe1203	BT5 O:2a,2b,3 pYVe3094
Plasmid size (bp)	67,721	75,103	75,107	72,395	72,413	72,824
GC content (%)	43.92	44.16	44.16	44.24	43.98	44.40
Number of CDS	90	90	92	88	86	91
Coding density (%)	73.0	70.6	74.0	71.6	71.9	78.5
Average gene size (bp)	618	671	664	659	678	647
w/o pseudogenes (bp)	677	698	678	682	694	665

One can see that the plasmids from the low-pathogenic BT range between 72.4 and 75.1 kb, and thus are considerably larger than the plasmid from the high-pathogenic BT, which is only 67.7 kb. The GC content of the plasmids ranges from 43.92 to 44.4%. Number of CDS, coding density and average gene size can vary, with no clear pattern for high- or low-pathogenic strains. 59 CDSs are conserved between all pYVs, and a further 22 CDSs are present in pYVs of low-pathogenic BTs. The conserved CDSs are part of the Yop T3SS machinery and regulatory proteins and are encoded within the virABGC operons. This can also be seen in the overall organization of the virulence plasmid (Figure 5.1). From the plasmid comparison one highly conserved region is obvious, which corresponds to the type 3 secretion system machinery. This region shows no inversions or deletions at all, but appears to be essentially the same in gene composition and orientation in all biotypes. In addition the plasmids from the low-pathogenic biotypes present only minor variations in gene content, relating to fragments or pseudogenes of transposases, and show few regions of rearrangement. The orientation of an IS element near the origin can vary and is absent in BT 4.

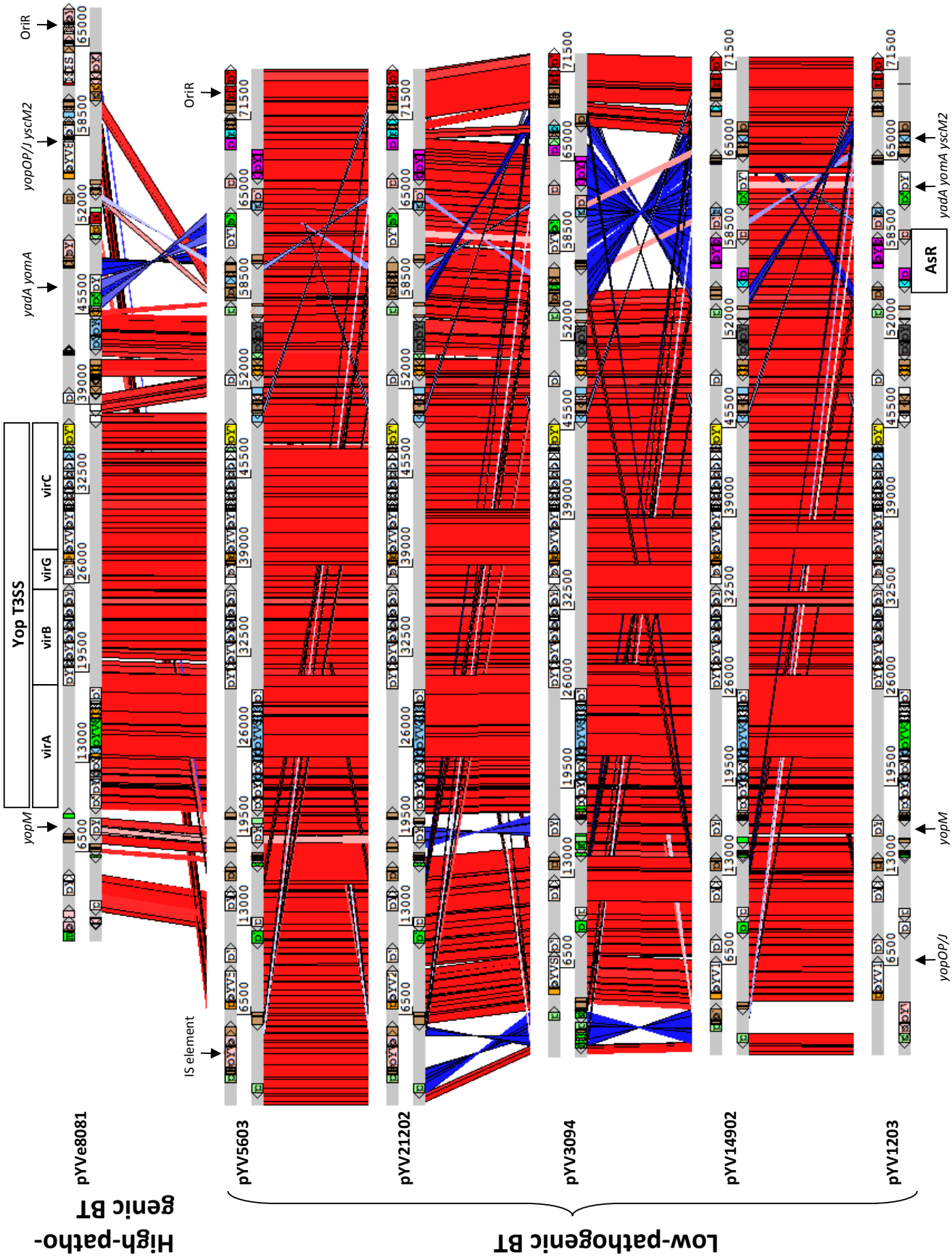


Figure 5.1: Alignment and comparison of the sequenced pYV. The plasmid from the high-pathogenic biotype is presented at the top, with the low-pathogenic biotype pYVs aligned below. Regions of specific interest are labelled; OriR – origin of replication.

Smaller differences are seen in the orientation of *yopM* and *yscM2*. These do not show any clear biotype-specific pattern, and this is likely to be explained by the transposase fragments and pseudogenes that surround both genes. The gene *yopM* is found in the anticlockwise orientation in pYVe8081, pYV21202, and pYV5603, but has the opposite, clockwise orientation in pYV3094, pYV14902, and pYV1203. *yscM2* found in pYV14902 and pYV1203 is in the opposite orientation to all other pYVs. This indicates that the rearrangments probably happened independently over time, mediated by adjacent transposases.
The genes for the targeted effector proteins YopO and YopP/J present the only translocations visible in the virulence plasmids. *yopOP/J* are located upstream of the origin of replication in the high-pathogenic BT, but are located downstream of it in the low-pathogenic BTs. Transposase remnants are present adjacent to these genes, and thus have probably facilitated the translocation. In *Y. pestis* and *Y. pseudotuberculosis*, *yopOP/J* are found downstream of the origin of replication, thus resembling the pYV of the low-pathogenic BTs (Snellings, Popek and Lindler, 2001).

It is also obvious that the origin of replication represented at the end of the alignments is different in plasmids of low- and high-pathogenic biotypes. This difference has been noted before in hybridization experiments (Portnoy and Falkow, 1981). The origin of replication of pYV8081 has been allocated to the IncL/M incompatibility group (Snellings, Popek and Lindler, 2001), and it has been noted that this is a feature of other pYVs from high-pathogenic BTs as well (Snellings, Popek and Lindler, 2001; Oberhettinger et al., 2011). Sequence similarity of the origin of replication in high-pathogenic pYVs to pYVs from low-pathogenic BTs is very low with 16% whereas protein identity is higher than 92% for the rest of the plasmid (Snellings, Popek and Lindler, 2001). Conversely, the origins of replication of both pYVs of low-pathogenic *Y. enterocolitica* and of *Y. pseudotuberculosis* as well as pCD1 of *Y. pestis* are similar and belong to the IncFIIA incompatibility group (Snellings, Popek and Lindler, 2001).

A major difference in gene content between pYVs of high- and low-pathogenic BTs is the presence of an arsenic resistance operon in the low-pathogenic BTs. The orientation of this region and with it the position of *yomA*, *yadA*, and *yscM2* plus fragmented transposases can vary amongst the biotypes.

The arsenic resistance operon has been described as part of the transposon Tn2502 (Neyt, Iriarte, Thi and Cornelis, 1997). It contains a total of six CDSs. The first CDS is the gene for transposase *tnpA*. The second CDS *tnpR* encodes the resolvase. The other four CDSs make up the arsenic resistance operon, containing an arsenite-inducible transcriptional repressor regulator ArsR, the transmembrane channel ArsB for export of arsenite, and the arsenate reductase ArsC catalyzing the conversion of arsenate to arsenite, which is then exported. The fourth CDS, *arsH*, is of unknown function, but is essential in the full resistance towards arsenite and arsenate. The presence of this transposon only in plasmids of low-pathogenic strains has been reported previously (Neyt, Iriarte, Thi and Cornelis, 1997). As the transposase is presumed to be non-functional in the described plasmid, it has been proposed that the acquisition event is ancient, and might be related to the worldwide spread of those biotypes where the arsenic resistance is thought to be of importance in environmental survival. In the past arsenic has been used as a growth factor in pigs. As pigs are thought to be one of the main reservoirs of *Y. enterocolitica*, their treatment with arsenic could have selected for plasmids carrying the operon (Neyt, Iriarte, Thi and Cornelis, 1997). Alternative arsenic resistance operons however are also found in BT 1B and BT 1A. The high-pathogenic BT 1B carries one unique copy inside the plasticity zone. Another copy outside the plasticity zone is orthologous to a copy in the non-pathogenic BT 1A. BT 1A also has a unique additional arsenic resistance operon. The chromosome or the plasmid can consequently carry arsenic resistance operons. pYV-independent arsenic resistance has been reported in the high-pathogenic BT 1B strain 8081 (Neyt, Iriarte, Thi and Cornelis, 1997), but nothing is known about functionality of arsenic resistance operons found in the non-pathogenic BT 1A.

In the presented strains it seems as if the region for the transposon extends beyond that described by Neyt, Iriarte, Thi and Cornelis (1997). Not only the Tn2502 region changes orientation in different plasmids, but also *yscM2, yadA* and *yomA* plus fragmented transposases surrounding *yscM2*. It can be argued that only Tn2502 might change orientation if it was mobile, as opposed to the whole region between positions 58.5 kb and 71.5 kb flipping, as seen here. There is no current evidence or experimental data whether these transposition events are ancient or still undergoing recombination in current cultures.

5.2.2 Generic Map of pYV

Figure 5.2 summarises all the genetic variations between pYV plasmids.

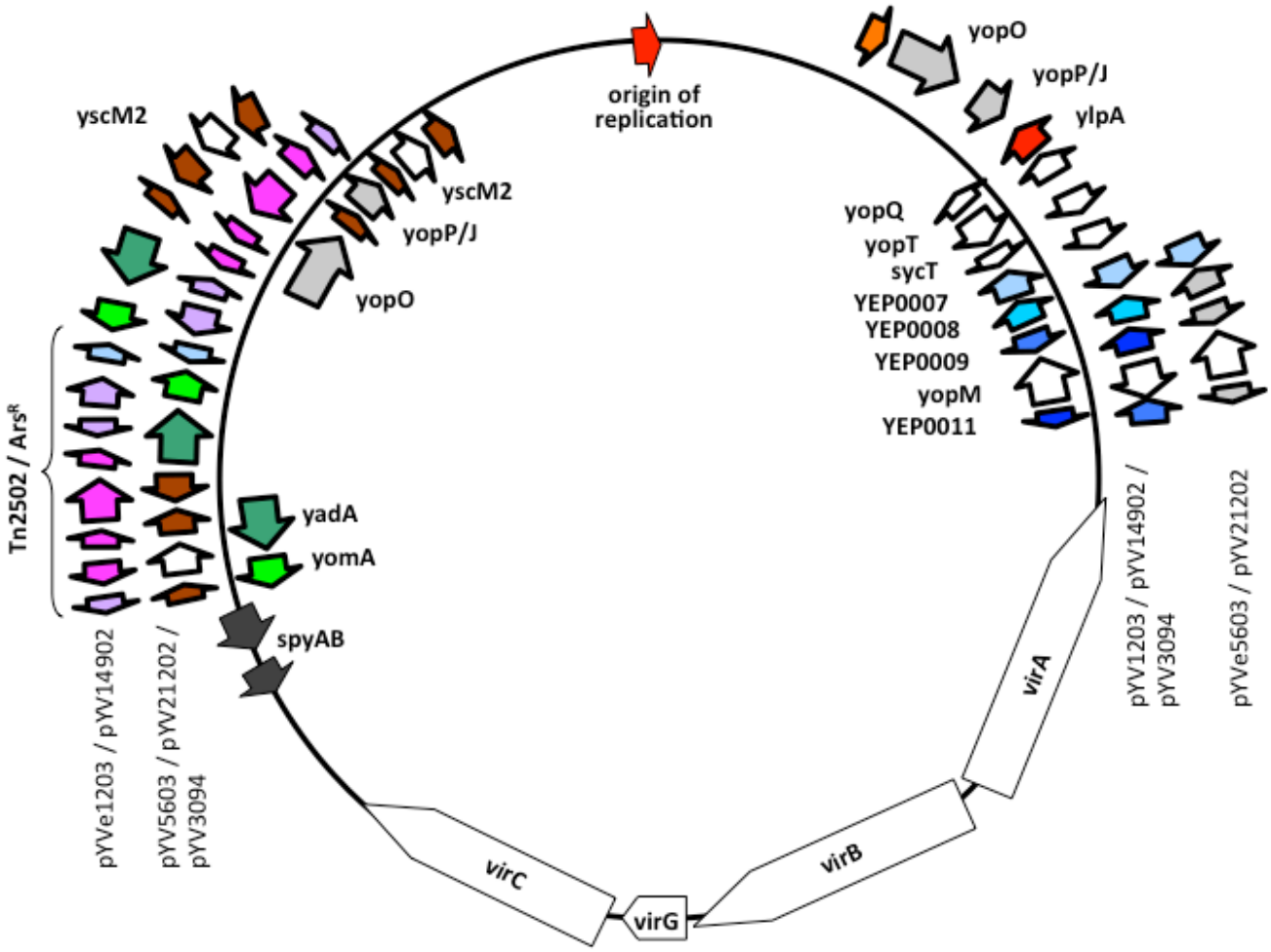


Figure 5.2: Backbone of pYV and variable regions in different *Y. enterocolitica* biotypes. The unvariable backbone of the virulence plasmid is presented on the circle and consists of the operons virABGC. Inside the circle the genes of the high-pathogenic BT pYVe8081 are depicted, outside the circle are the genes of the low-pathogenic BT pYVs. Orientation of following genes is variable: *yopM, yadA, yomA, yscM2*. Position of *yopO, yopP/J* is variable. Position of *yadAyomA* varies due to insertion of arsenic resistance cassette (As^R).

The T3SS machinery and regulatory proteins (virABGC) is the backbone of the virulence plasmid and comprises about half of the plasmid. The other regions are variable. In case of the origin, the region appears to have been swapped for another region in high- or low-pathogenic biotypes. It could be argued that the T3SS machinery, being the backbone of pYV, was inserted into different, pre-existing plasmids in high- and low-pathogenic BTs. This could explain the difference in origin of replication. An argument against this is that there are numerous CDSs outside the T3SS encoding effector proteins, which then would have had to be introduced individually. A more likely scenario is the acquisition and subsequent exchange of the

origin of replication possibly after a co-integration event with another plasmid in high-pathogenic BTs as suggested by Snellings, Popek and Lindler (2001).

Reciprocal fasta searches and analysis of orthologous gene was used to define the core and variable pYV genes comparing the reference plasmid sequences: 38 CDSs of a core set of 59 CDS belong to the T3SS region. In the accessory genes, 22 CDSs are found commonly in the low-pathogenic biotypes. Most of these low-pathogen specific genes are transposase fragments and pseudogenes, *ylpA*, CDSs that represent an alternative origin of replication and the arsenic resistance operon. 28 CDS are unique to pYVe8081, and account for CDSs of the origin of replication and transposase fragments and pseudogenes.

5.2.3 Variability in pYV core gene sequence

5.2.3.1 *yopM*

Heterogeneity has been reported for some of the genes encoded on the virulence plasmid. YopM is one of the T3SS effector proteins. Its size can vary between 367 and 515 amino acids, as the central part contains repeat domains of 60 amino acids, with the number of repeats present leading to a difference in size (Boland, Havaux and Cornelis, 1998; Snellings, Popek and Lindler, 2001). This heterogeneity though does not appear to make a difference in function and does not change the virulence potential of strains expressing an altered YopM (Boland, Havaux and Cornelis, 1998). There is also no apparent link to bio-serotype, as two BT 1B O:8 strains were found to encode YopM proteins of different lengths (Boland, Havaux and Cornelis, 1998; Foultier and Cornelis, 2003). In the plasmids sequenced in this study, the same length of *yopM* as in pYVe8081 was observed.

5.2.3.2 *lcrV*

Another protein known to be polymorphic is LcrV (Foultier and Cornelis, 2003; Roggenkamp, et al., 1997). This protein is part of the T3SS machinery and encodes a translocator (Foultier and Cornelis, 2003). As LcrV is part of the apparatus of the T3SS and expressed on the cell surface where it is referred to as the V-antigen (Roggenkamp, et al., 1997). Immunization of mice with V-antigen sera conferred protection against infection with *Y. enterocolitica* (Roggenkamp, et al., 1997). Comparison of LcrV sequences between serotype O:8 isolates and *Y. pestis* and *Y. pseudotuberculosis* showed that a 9-amino-acid insertion is present in some BT

1B O:8 isolates (Roggenkamp, et al., 1997; Foultier and Cornelis, 2003). Yet, similarly to YopM, LcrV variation is not linked to bio-serotype as other O:8 isolates were found with sequences identical to those in *Y. pseudotuberculosis*. A comparison of LcrV sequences of the reference strains of this study to exemplar strains studied in the past (Roggenkamp, et al., 1997) is shown in Figure 5.3. As was shown before, LcrV of BT 1B O:8 strain WA314 is found to be 9 amino acids longer than BT 1B O:8 strain 8081. There is also variation in LcrV sequences of O:9 isolates. The two isolates sequenced in this study show 5 additional amino acids directly upstream of the WA314 insertion, which corresponds to amino acid position 225 to 232 of *Y. pseudotuberculosis*.

```
YPS-YPI     MIRAYEQNPQHFIEDLEKVRVEQLTGHGSSVLEELVQLVKDKNIDISIKYDPRKDSEVFANRVITDDIELLKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMA
YE-8081     MIRAYEQNPQHFIEDLEKVRVEQLTGHGSSVLEELVQLVKDKKIDISIKYDPKKDSEVFAERVITDDIELLKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMA
YE-WA314    MIRAYEQNPQHFIEDLEKVRVEQLTGHGSSVLEELVQLVKDKKIDISIKYDPKKDSEVFAERVITDDIELLKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMA
YE-W22703   MIRAYEQNPQHFIEDLEKVRVEQLTGHGSSVLEELVQLVKDKKIDISIKYDPKKDSEVFADRVITDDIELLKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMA
YE-21202    MIRAYEQNPQHFIEDLEKVRVEQLTGHGSSVLEELVQLVKDKNIDISIKYDPRKDSEVFANRVITDDIELLKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMA
YE-5603     MIRAYEQNPQHFIEDLEKVRVEQLTGHGSSVLEELVQLVKDKNIDISIKYDPRKDSEVFANRVITDDIELLKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMA
YE-Y96      MIRAYEQNPQHFIGDLEKVRVDQLTGHGSSVLEELVQLVQDKKIDISIKYDPQKDSEVFADRVITDDIELLKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMA
YE-14902    MIRAYEQNPQHFIEDLEKVKVEQLTGHGSSVLEELVQLVKDKKIDISIKYDPKKDSEVFADRVITDDIELLKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMA
YE-Y527     MIRAYEQNPQHFIGDLEKVRVEQLTGHGSSVLEELVQLVQDKKIDISIKYDPKKDSEVFADRVITDDIELLKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMA
YE-1203     MIRAYEQNPQHFIEDLEKVRVEQLTGHGSSVLEELVQLVKDKKIDISIKYDPKKDSEVFANRVITDDIELLKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMA
YE-SZ3094   MIRAYEQNPQHFIEDLEKVRVEQLTGHGSSVLEELVQLVKDKKIDISIKYDLKKDSEVFADRVITNDIELLKKILAYFLPEDAILKGGHYDNQLQNGIKRVKEFLESSPNTQWELRAFMA

YPS-YPI     VIHFSLTADRIDDDILKVIVDSMNHHGDARSKLREELAELTAELKIYSVIQAEINKHLSSGGTINIHDKSINLMDKNLYGYTDEEIFKASAEYKILEKMPQTTI-----QX--------X
YE-8081     VMHFSLTADRIDDDILKVIVDSMNHHGDARSKLREELAELTAELKIYSVIQAEINKYLSNSGTINIHDKSINLMDKNLYGYTDEEIFKASAEYKILEKMSQTTI-----KX--------X
YE-WA314    VMHFSLTADRIDDDILKVIVDSMNHHGDARSKLREELAELTAELKIYSVIQAEINKHLSSSGTINIHEKSINLMDKNLYGYTDEEIFKASAEYKILKKMPQTTI-----KDDELHEVGVI
YE-W22703   VMHFSLTADRIDDDILKVIVDSMNHHGDARGKLREELAELTAELKIYSVIQAEINKHLSNSDTINIHDKSINLMDKNLYGYTDEEIFKASAEYKILEKMPQTTI-----KX--------X
YE-21202    VMHFSLTADRIDDDILKVIVDSMNHHGDARSKLREELAELTAELKIYSVIQAEINKHLSNSGTINIHDKSINLMDKNLYGYTDEEIFKASAEYKILEKMPQTTVPPIPGNX--------X
YE-5603     VMHFSLTADRIDDDILKVIVDSMNHHGDARSKLREELAELTAELKIYSVIQAEINKHLSNSGTINIHDKSINLMDKNLYGYTDEEIFKASAEYKILEKMPQTTVPPIPGNX--------X
YE-Y96      VMHFSLTADRIDDDILKVIVDSMNHHGDARGKLREELAELTAELKIYSVIQAEINKHLSNSDTINIHDKSINLMDKNLYGYTDEEIFKASAEYKILEKMPQTTI-----KX--------X
YE-14902    VMHFSLTADRIDDDILKVIVDSMNHHGDARSKLREELAELTAELKIYSVIQAEINKHLSNSGTINIHDKSINLMDKNLYGYTNEEIFKASAEYKILEKMSQTTI-----KX--------X
YE-Y527     VMHFSLTADRIDDDILKVIVDSMNHHGDARSKLREELAELTAELKIYSVIQAEINKHLSNSDTINIHDKSINLMDKNLYGYTNEEIFKASAEYKILEKMSQTTI-----KX--------X
YE-1203     VMHFSLTADRIDDDILKVIVDSMNHHGDARSKLREELAELTAELKIYSVIQAEINKHLSNSDTINIHDKSINLMDKNLYGYTNEEIFKASAEYKILEKMSQTTI-----KX--------X
YE-SZ3094   VMHFSLTADRIDDDILKVIVDSMNHHGDARSKLREELAELTAELKIYSVIQAEINKHLSNSDTINIHNNSINLMDKNLYGYTNEEIFKASAEYKILNNMSQTTT-----KDKX------X

YPS-YPI     GETEKKIVSIKNFLESEKKRTGALGNLKDSYSYNKDNNELSHFATTCSDKSRPLNDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDTSGK*
YE-8081     GETEKKIVSIKNFLESENKRTGALGNLKNSYSYNKDNNELSHFATACSDKSRPLNDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDT--R*
YE-WA314    AGAEKQIVSIKNFLESENKRTGALGNLKDSYSYNKDNNELSHFATACSDKSRPLNDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDT--RL
YE-W22703   GETEKKIVSIKNFLESENKRTGALGNLKDSYSYNKDNNELSHFATTCSDKSRPLNDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDT--R*
YE-21202    NGSEKKIVSIKNFLESENKRTGALGNLKDSYSYNKDNNELSHFATTCSDKSRPLNDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDT--R*
YE-5603     NGSEKKIVSIKNFLESENKRTGALGNLKDSYSYNKDNNELSHFATTCSDKSRPLNDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDT--R*
YE-Y96      GETEKKIVSIKNFLESENKRTGALGNLKDSYSYNKDNNELSHLGTTCSDKSRPLNDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDT--R*
YE-14902    GETEKKIVSIKNFLESENKRTGALGNLKDSYSYNKDNNELSHFATTCSDKSRPLNDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDT--R*
YE-Y527     GETEKKIVSIKNFLESENKRTGALGNLKDSYSYNKDNNELSHLGTTCSDKSRPLNDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDT--R*
YE-1203     GETEKKIVSIKNFLESENKRTGALGNLKDSYSYNKDNNELSHFATTCSDKSRPLNDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDT--R*
YE-SZ3094   GKTEKKIVSIKNFLESENKRTGALNNLKESYSYNKDNNELSHFATTCSDKSRPLNDLVSQKTTQLSDITSRFNSAIEALNRFIQKYDSVMQRLLDDT--R*
```

Figure 5.3: Protein sequence alignment for LcrV. YPS-YPI – *Y. pseudotuberculosis* YPIII, YE – *Y. enterocolitica*, 8081 and WA314 – BT 1B O:8, W22703, 21202, 5603, and Y96 – BT 2 O:9, 14902 and Y527 – BT 3 O:5,27, 1203 – BT 4 O:3, SZ3094 – BT 5 O:2a,2b,3.

The difference in sequence between American O:8 isolates and *Y. pestis*, *Y. pseudotuberculosis*, and European *Y. enterocolitica* isolates has been implicated

in plague epidemiology (Roggenkamp, et al., 1997). The authors suggest that the antibodies raised against the *Y. enterocolitica* V-antigen are cross protective in rodents to *Y. pestis* infection. Since the American O:8 isolates express a different antigen it is thought that this cross protection is not given, explaining the plague foci found in the Americas (Roggenkamp, et al., 1997). Undoubtedly the insertion of additional amino acids can cause a conformational change in the protein and thus leads to the expression of an altered epitope, but this variation is not found in all O:8 isolates. BT 1B strain 8081 was isolated from North America and shows high LcrV sequence identity with *Y. pseudotuberculosis* and *Y. pestis* and no amino acid insertions, suggesting that this isolate should theoretically confer immunity against *Y. pestis*. The variation in the O:9 isolates would presumably not give protective immunity against plague, yet these isolates are very common in Europe without associated plague foci. An effect on infectivity, virulence potential or severity of disease has not been correlated with a changed LcrV epitope.

5.2.3.3 *yscP*

YscP forms an essential part of the type III secretion machinery. Deletion mutants showed that abolishing the complete gene leads to a loss of secretion of Yops, and a functional protein is necessary for the delivery of the effectors into target cells (Stainier, et al., 2000). It was noted as well that it shows more variability than any other protein in the T3SS between *Y. pseudotuberculosis*, *Y. pestis*, and *Y. enterocolitica* (Stainier, et al., 2000). YscP from *Y. enterocolitica* has additional 60 amino acids that appear to be part of a repeat compared to *Y. pseudotuberculosis* and *Y. pestis*. Later investigations showed that YscP controls the length of the injectisome needles, acting like a ruler to determine optimum length (Journet, Agrain, Broz and Cornelis, 2003). The N and C terminal part of the protein are essential for length control in general, as deletion mutants showed a wide range of needle sizes. Deletions or insertions into the central part of the protein resulted in proportionally shorter or longer needles, therefore being the functional domain responsible for fixed needle length (Journet, Agrain, Broz and Cornelis, 2003). Additionally, YscP acts as a specificity switch signalling the apparatus to stop producing needle proteins and change over to late substrates (Journet, Agrain, Broz and Cornelis, 2003; Wagner, et al., 2009). YscP is anchored to the base and the growing tip of the needle so that it is stretched alongside the needle during assembly. The content of helix makers such as alanine and the overall length of the protein YscP determine the needle length

(Wagner, et al., 2009). Once YscP reaches a breaking point when stretched to its maximum, it detaches and signals back so that late injectisome molecules are produced (Wagner, et al., 2009).

A comparison between YscP protein sequences previously investigated (Foultier and Cornelis, 2003; Wagner, et al., 2009; Oberhettinger, et al., 2011) and sequences generated here shows that YscP contains three different repeat units (Figures 5.4 and 5.5). These repeats are modules that can be duplicated and reorganized in different strains. The red repeat contains 13 amino acids, green 25 amino acids, and blue 46 amino acids, identical to repeats described before (Wagner, et al., 2009). The region of variability begins at amino acid position 219 respective to BT 1B strain YE-8081 (Figure 5.4).

At the beginning of the region of variability there is a first red repeat in all strains except for one BT 1B isolate that shows a 24 amino acid deletion. The red repeat is followed by the first green repeat that is present in all strains except for BT 5. Instead of a green repeat, one BT 5 strain shows blue and red repeat units. The next repeats are the first blue and the second red repeats in all low-pathogenic BTs and three BT 1B isolates. The BT 3 O:5,27 isolate YE-14902 shows an in-frame deletion of 6 bp (2 amino acids). The three BT 1B isolates are A11/86, WAT-988, and WA314, which then also have a second green repeat. The region of variability ends with a second blue repeat in all strains.

The protein sequences of the strains are shown in Figure 5.5 to highlight the modular nature of YscP.

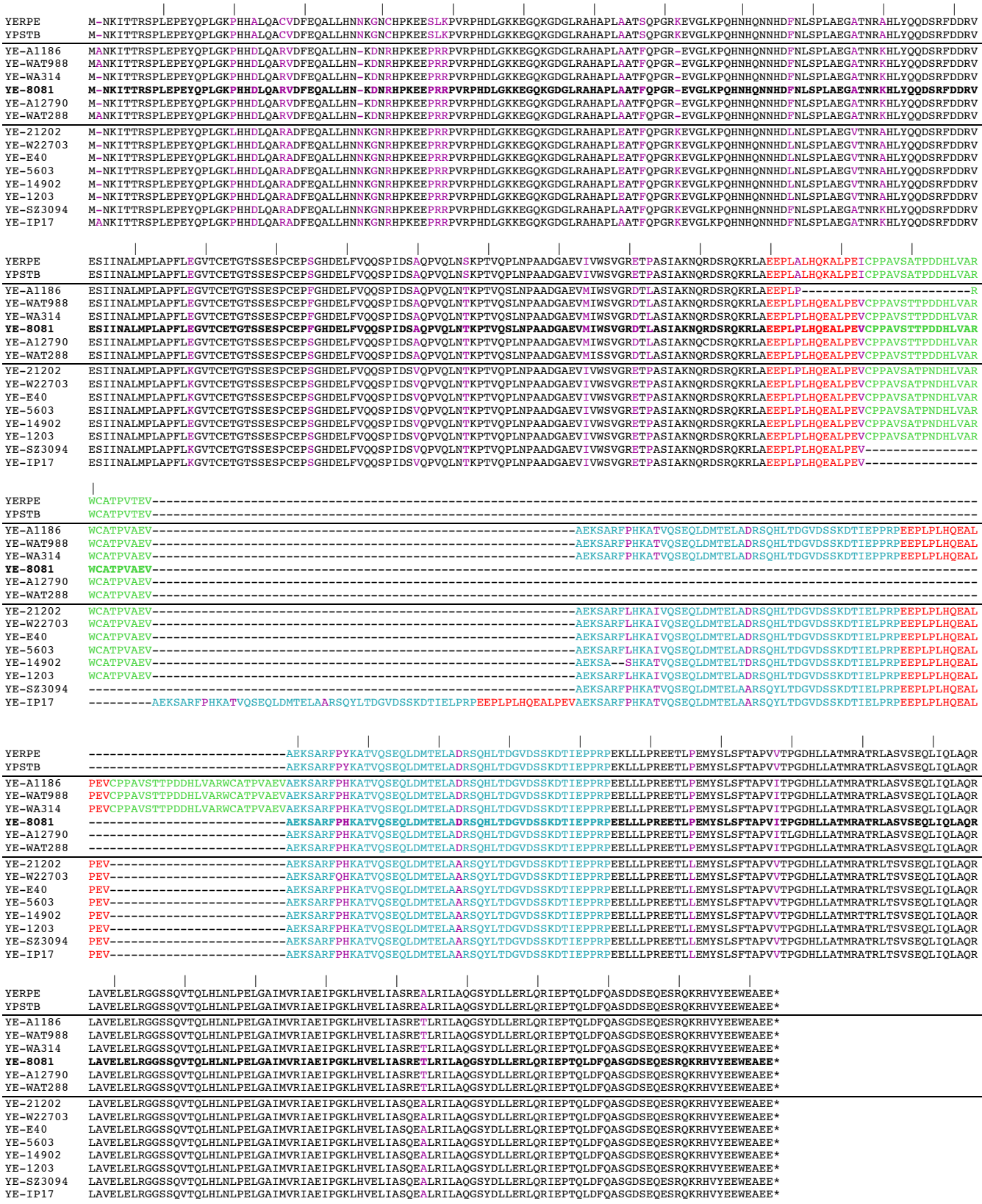


Figure 5.4: Protein sequence alignment of YscP protein sequences of different *Yersinia* spp. Repeats in amino acid sequences are highlighted in red (13 amino acids), green (25 amino acids), and blue (46 amino acids). Variations in amino acids are highlighted in purple. Vertical bars indicate every 10 amino acids respective to YE-8081 (bold sequence). YERPE – *Y. pestis* strain IP275, YPSTB – *Y. pseudotuberculosis* strain YPIII, YE- – *Y. enterocolitica*, BT 1B: A1186 (O:21), WAT988 (O:4), WA314 (O:8), 8081 (O:8), A127/90 (O:8), WAT288 (O:13), BT 2: 21202, E40 (O:9), BT 3: 5603 (O:9), 14902 (O:5,27), BT 4: 1203 (O:3), BT 5: SZ3094 (O2a,2b,3), IP17 (O:2,3).

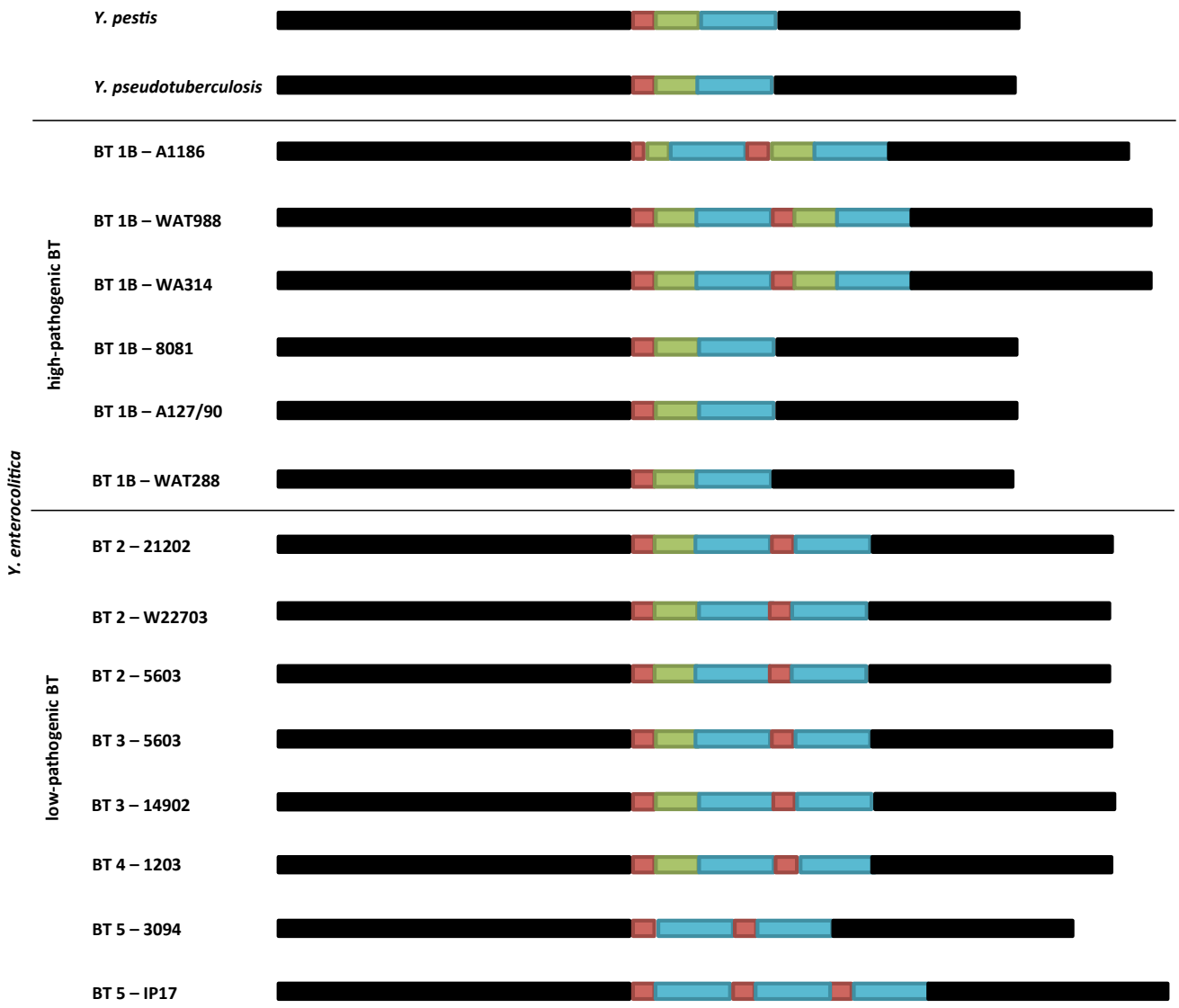


Figure 5.5: Modular nature of YscP in different *Yersinia* spp. Conserved sequences are shown in black, repeat sequences are red (13 amino acids), green (25 amino acids), and blue (46 amino acids). Strains are the same as in Figure 5.4.

From Figure 5.5 one can see that there are certain themes concerning the makeup of YscP in *Yersinia* spp. First of all, *Y. pestis* and *Y. pseudotuberculosis* show an identical YscP protein with a single repeat of each the red, green, and blue module. For the high-pathogenic BT 1B of *Y. enterocolitica*, there are two options. The first option is to have the same YscP sequence as *Y. pestis*. Presumably evolved from this is the second option of duplicating the red-green-blue modules. Strain A1186 shows a deletion of the parts of the red and green modules. This could have happened after duplication of the whole repeat region, as it is otherwise identical to strains WAT988 and WA314. In the low-pathogenic BTs there are also two possibilities. For BTs 2, 3, and 4, the red and blue modules are duplicated. For BT 5,

only the red and blue modules are present, and can vary in copy number. A caveat for BT 5 has to be added, as so far only these two isolates have been studied and it is not known whether this is a common phenomenon for BT 5 isolates. It is also unknown whether the differences observed correlate to a change in pathogenicity and whether a change in needle length affects efficiency of effector translocation.

5.2.4 Plasmid Variety: Unique plasmids

In the *Y. enterocolitica* strains that were sequenced, four unique plasmids were identified apart from the virulence plasmid. BT 2 strain 21202 and BT 5 strain 3094 both had large additional plasmids, and biotype 1A, though lacking the virulence plasmid, showed the presence of two small plasmids.

5.2.4.1 Unique plasmids in the reference strain for BT1A: pSR2-1A and pSR3-1A

Two unique plasmids are present in the reference strain 5303 for BT 1A, which were named pSR2-1A and pSR3-1A. pSR2-1A (Figure 5.6) has a size of 8 kb. The GC content is 48.0% and therefore remarkably close to the chromosomal GC content of strain 5303 (48.42%). The coding percentage is 81.3%. A total of 12 CDSs were identified, but function could only be assigned to four of these. The other eight CDSs are currently hypothetical proteins with uncharacterized homologs in other bacterial species other than *E. coli*. These homologs are either distantly related or are not well characterized, thus a function cannot be allocated. It contains three genes *mobABC* that encode for mobility of the plasmid. There are no transfer genes encoded on this plasmid, which means that this plasmid can be co-mobilized if another plasmid with *tra* genes for conjugational transfer is present.

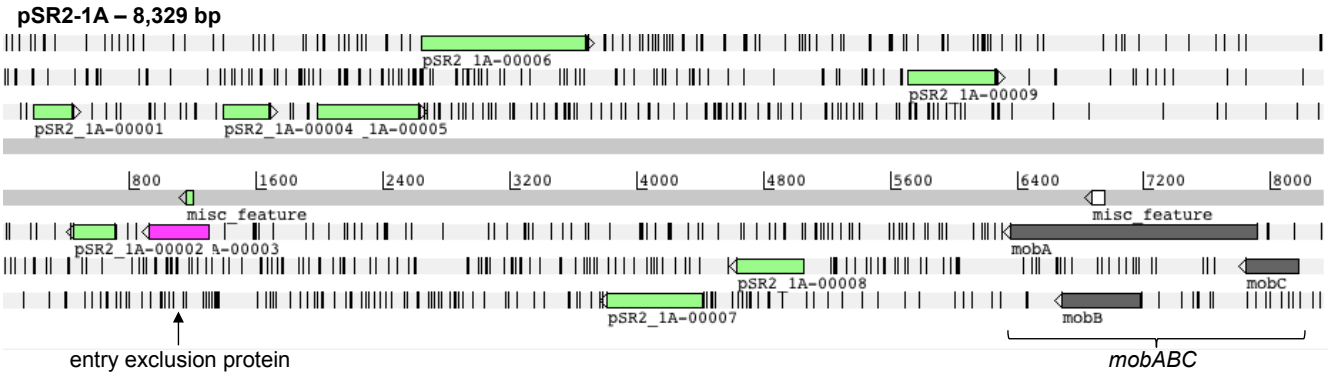


Figure 5.6: Linear plasmid map of pSR2-1A found in *Y. enterocolitica* BT 1A O:5 strain 5303.

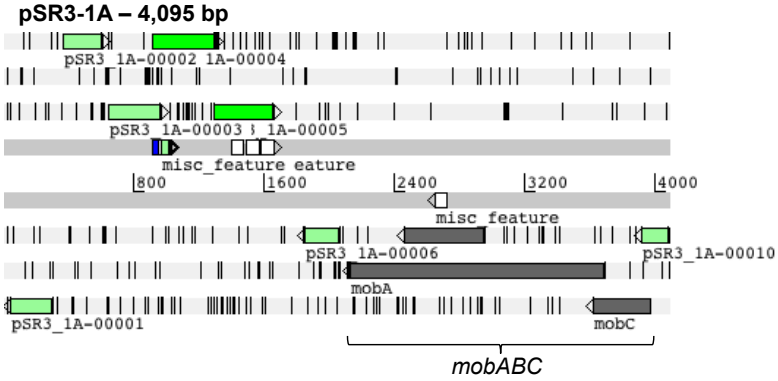


Figure 5.7: Linear plasmid map of pSR3-1A found in *Y. enterocolitica* BT 1A O:5 strain 5303.

Similarly, pSR3-1A contains *mobABC*, but no *tra* genes (Figure 5.7). This plasmid only has a size of 4 kb. The GC content of pSR3-1A is 53.92%. This is higher than the GC content of pSR2-1A as well as the chromosome of strain 5303. The coding percentage is 104.3%. It is likely that this is an overestimation as half of the putative CDSs are of hypothetical nature.

Both plasmids appear to be cryptic, as they do not contain any genes of assigned functions. pSR3-1A contains two genes which might encode for membrane proteins, as one of them has a SignalP-domain, and the other shows potential transmembrane domains.

Two cryptic plasmids, pYe4449-1 and pYe4449-2, described in another BT 1A strain (Lepka, et al., 2009) show no similarity to the plasmids described here apart from the mobilization region *mobABC*. These plasmids add to the overall gene pool in *Y. enterocolitica*, especially the non-pathogenic BT, but their function and impact is unknown.

5.2.4.2 Plasmids in atypical BT 1A/1B strains

It is know that there are some isolates of *Y. enterocolitica* that biochemically type as BT 1B but which lack all of the determinants of high-pathogenic strains such as the virulence plasmid and the high-pathogenicity island. A plasmid profile of potential strains matching these criteria was established using the approach of Kado and Liu (1981; Figure 5.8).

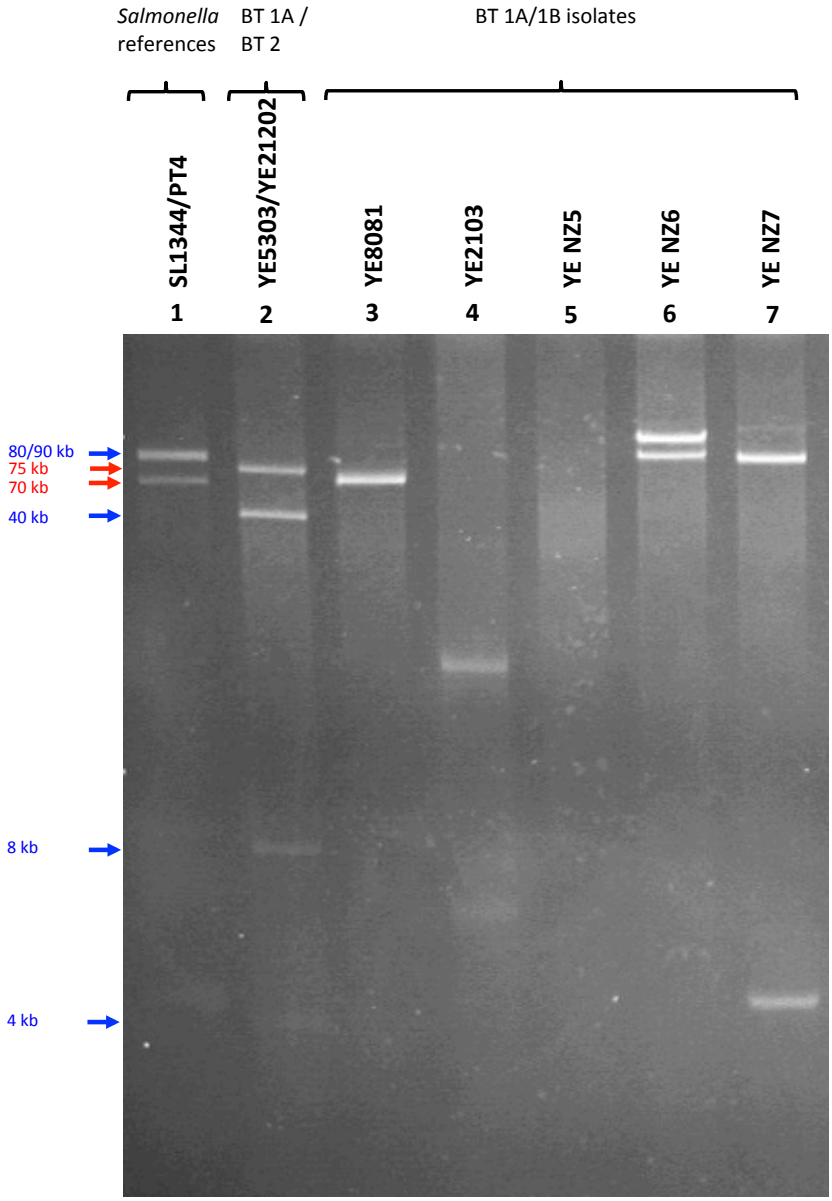


Figure 5.8: Plasmid profile of atypical BT 1A/1B isolates. Lane 1: Mixture of *Salmonella* plasmids of isolates SL1344 and PT4. Lane 2: Mixture of YE5303 and YE21202 plasmids (pSR1-2, pYV, pSR2-1A, pSR3-1A). Lane 3-7: BT 1A/1B isolates. YE8081 – BT 1B O:8, YE2103 – BT 1B O:19, NZ5-7 – BT 1B, no serotype information. Red arrows indicate size of pYV.

Plasmid size can be estimated by the use of reference plasmids of known size, presented here in lane 1 and 2. In lane 1 there is a mixture of two *Salmonella* strains.

SL1344 has two large plasmids of 80 and 90 kb that form a single band on this gel, and further smaller plasmids are not visible. PT4 has a plasmid of about 70 kb size. Lane 2 has a mixture of plasmids from strain 5303 (pSR2-1A and pSR3-1A) and strain 21202 (pYV and pSR1-2). The high-pathogenic BT 1B reference strain 8081 is shown to demonstrate the size difference between high- and low-pathogenic pYVs.

YE2103 is an atypical BT 1B strain that was isolated during the IID study in the UK (Tompkins, et al., 1999). It was obtained from a GP control, i.e. an asymptomatic person matched in age and sex to a case subject. It was non-invasive in tissue culture (A. McNally, personal communication) and it was shown in a microarray study that it is closely related to non-pathogenic isolates (Howard, et al., 2006). One can see that this isolate does not possess the virulence plasmid pYV or any other plasmid of similar size. Instead, it has two smaller plasmids, one of ~20 kb, the other ~6 kb. As this strain was not included in the sequencing project (Chapter 7), no information on the nature of these plasmids is currently available.

NZ5–7 were isolated from humans by collaborators in New Zealand (M. DuFour) and are lacking serotype information. NZ5 shows no presence of plasmids. NZ6 contains two plasmids of 75 kb and ~95 kb. NZ7 also has two plasmids, one of 75 kb, the other of 4 kb. From the information that these isolates were typed as BT 1B, it was initially assumed that the 75 kb plasmids of NZ6 and NZ7 are pYV. Two of these strains, NZ5 (NZ523784) and NZ7 (NZ761915), were sequenced as part of a strain collection (Chapter 7). Sequencing of NZ7 revealed that pYV is not present. Both NZ5 and NZ7 phylogenetically group with non-pathogenic BT 1A isolates, and HPI, Ysa T3SS, and YAPI could not be detected. Whilst nothing can be deduced for NZ6, it is clear that NZ5 and NZ7 have been mistyped as BT 1B and that the presence of a large plasmid of similar size to pYV may have influenced this decision in NZ7. It confirms that large plasmids can be present in non-pathogenic strains and that the presence of large plasmids does not indicate the presence of pYV.

5.2.4.3 Unique plasmid in the reference strain for BT2: pSR1-2

BT 2 O:9 strain 21202 carries a unique plasmid of 42 kb which was named pSR1-2. The plasmid has a GC content of 52.2%, which is higher than the chromosomal GC content of strain 21202 (48.1%). This would indicate an acquisition of this plasmid from a donor outside the genus *Yersinia*, because the average GC content is 48-49% (Sprague and Neubauer, 2005). 54 CDSs have been identified (coding percentage

87.8%). It belongs to the IncN incompatibility group, and a map is shown in Figure 5.9.

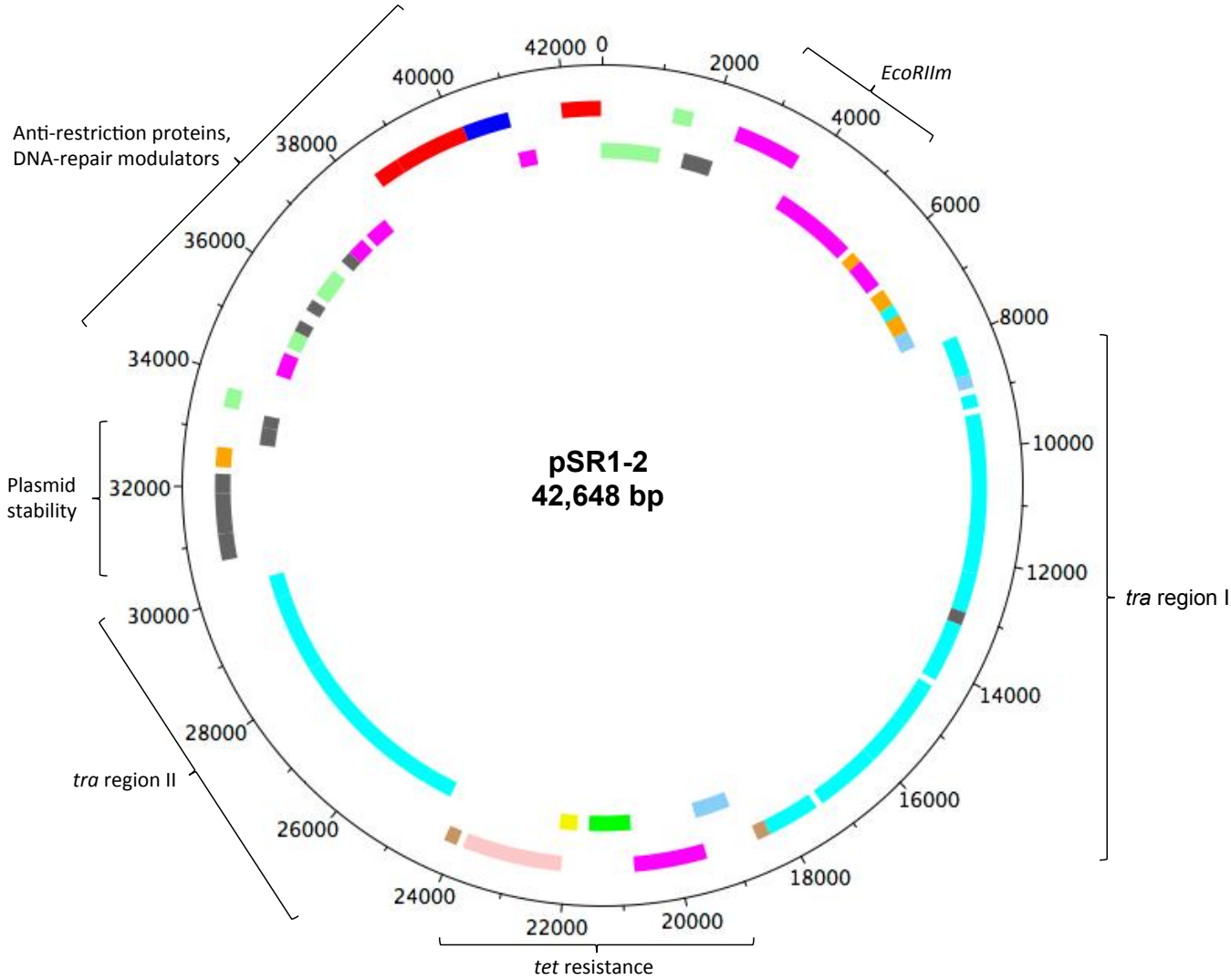


Figure 5.9: Circular plasmid map of pSR1-2 found in *Y. enterocolitica* BT 2 O:9 strain 21202. Regions of interest are highlighted.

pSR1-2 carries *stbABC* involved in plasmid stability, as well as anti-restriction proteins, DNA-repair modulators, and an *EcoRIIm* restriction modification enzyme. In contrast to pSR2-1A and pSR3-1A, there are only five hypothetical genes. The plasmid contains *tra* genes for pilus formation and conjugational transfer. These genes are split into two operons. In between these two operons there is a tetracycline resistance cassette. The functionality of this antibiotic resistance cassette has been established by growth on LB agar containing 10, 20, and 50 µg/ml tetracycline. Compared to the other reference sequence strains, only BT 2 O:9 strain 21202 was able to grow on this medium. Five other BT 2 strains and one BT 3 O:9 strain

(YE21402) tested negative for growth on LB supplemented with tetracycline. The same BT 2 strains were examined for the presence of plasmids (Figure 5.10).

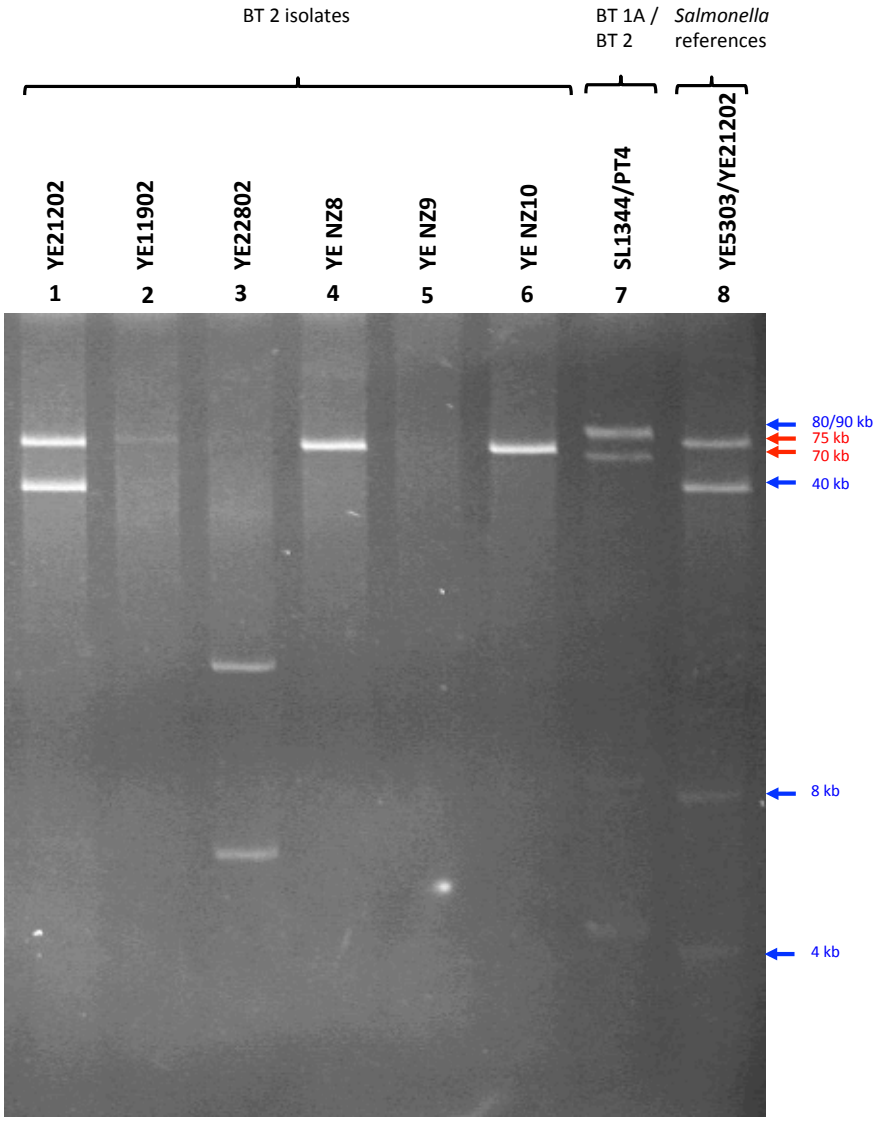


Figure 5.10: Plasmid profile of BT 2 isolates. Lane 1-6: BT 2 isolates. YE21202, YE11902, YE22802 – BT 2 O:9, NZ8-10: BT 2, no serotype information. Lane 7: Mixture of *Salmonella* plasmids of isolates SL1344 and PT4. Lane 8: Mixture of YE5303 and YE21202 plasmids (pSR1-2, pYV, pSR2-1A, pSR3-1A). Red arrows indicate size of pYV.

Strain 21202 shows the clear presence of pYV at 75 kb and pSR1-2 at 40 kb. YE11902 has a plasmid of a size corresponding to pYV although in this extraction the band is very faint. YE22802 does not contain pYV but has two smaller plasmids of ~20 kb and ~6 kb. NZ8 (NZ832501) and NZ10 (NZ1032126) both are positive for the virulence plasmid, and NZ9 (N2982399) does not carry any plasmids.

Similar to arsenic, tetracycline is used as a growth promoter in pigs. The BT 2 O:9 strain 21202 is a pig isolate, whereas the other tested strains are of other origins. This might explain the presence of a tetracycline-carrying plasmid in this strain. The

BT 3 O:9 pig isolate was not examined for the presence of pSR1-2, as it was sensitive to tetracycline. Sequencing of other isolates of BT 2 O:9, BT 3 O:9, BT 2 O:5,27, and BT 3 O:5,27 (Chapter 7) showed that only one other isolate potentially carries a copy of pSR1-2. That respective strain YE0403 is of BT 3 O:5,27 and was isolated from human faeces. The acquisition of pSR1-2 in 21202 therefore might present a unique event not shared by other isolates of the same niche.

pSR1-2 shows overall similarity to pMAK-2, a plasmid found in *Salmonella* Dublin (accession number AB366441.1; Akiba, 2007). A comparison between these two plasmids is shown in Figure 5.11.

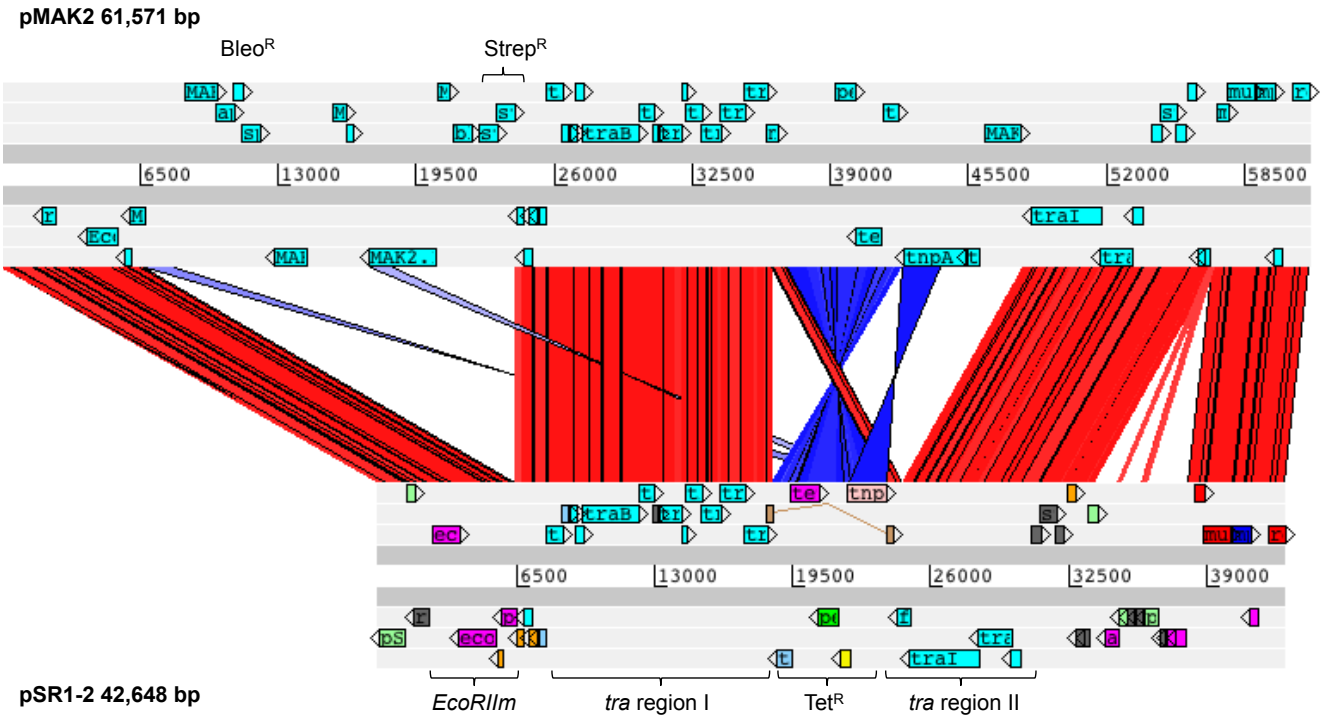


Figure 5.11: Comparison between *Salmonella* Dublin pMAK-2 and *Y. enterocolitica* pSR1-2. BleoR – bleomycin resistance, StrepR – streptomycin resistance, TetR – tetracycline resistance.

pMAK-2 is larger in size than pSR1-2, and contains additional genes encoding for bleomycin and streptomycin resistance. Tetracycline resistance is shared in both plasmids, but is inverted in pSR1-2.

Salmonella Dublin is predominantly found in cattle, and tetracycline resistance carried on plasmids has been described before (Frech and Schwarz, 1998). It is unknown whether there is a shared niche between *S.* Dublin and *Y. enterocolitica* that could facilitate the specific transfer of tetracycline resistance plasmids.

5.2.4.4 Unique plasmid in the reference strain for BT5: pSR4-5

BT 5 O:2a,2b,3 strain 3094 also carries a unique plasmid which has been named pSR4-5. The GC content of the plasmid is lower than the chromosome (42.2% and 48.2%, respectively). This is a large plasmid of 107 kb (Figure 5.12). 108 CDSs were identified (coding percentage 66.3%), of which 42 are hypothetical genes. Its replication protein could not be assigned to any incompatibility group. Similar to pSR1-2, this plasmid carries its own *tra* genes encoding a type IV pilus for conjugational transfer.

pSR4-5 also carries a putative hemolysin and its activator protein, a putative toxin/antitoxin system, an endonuclease and a transglycosylase, a putative invasion protein, and proteins involved in plasmid partitioning.

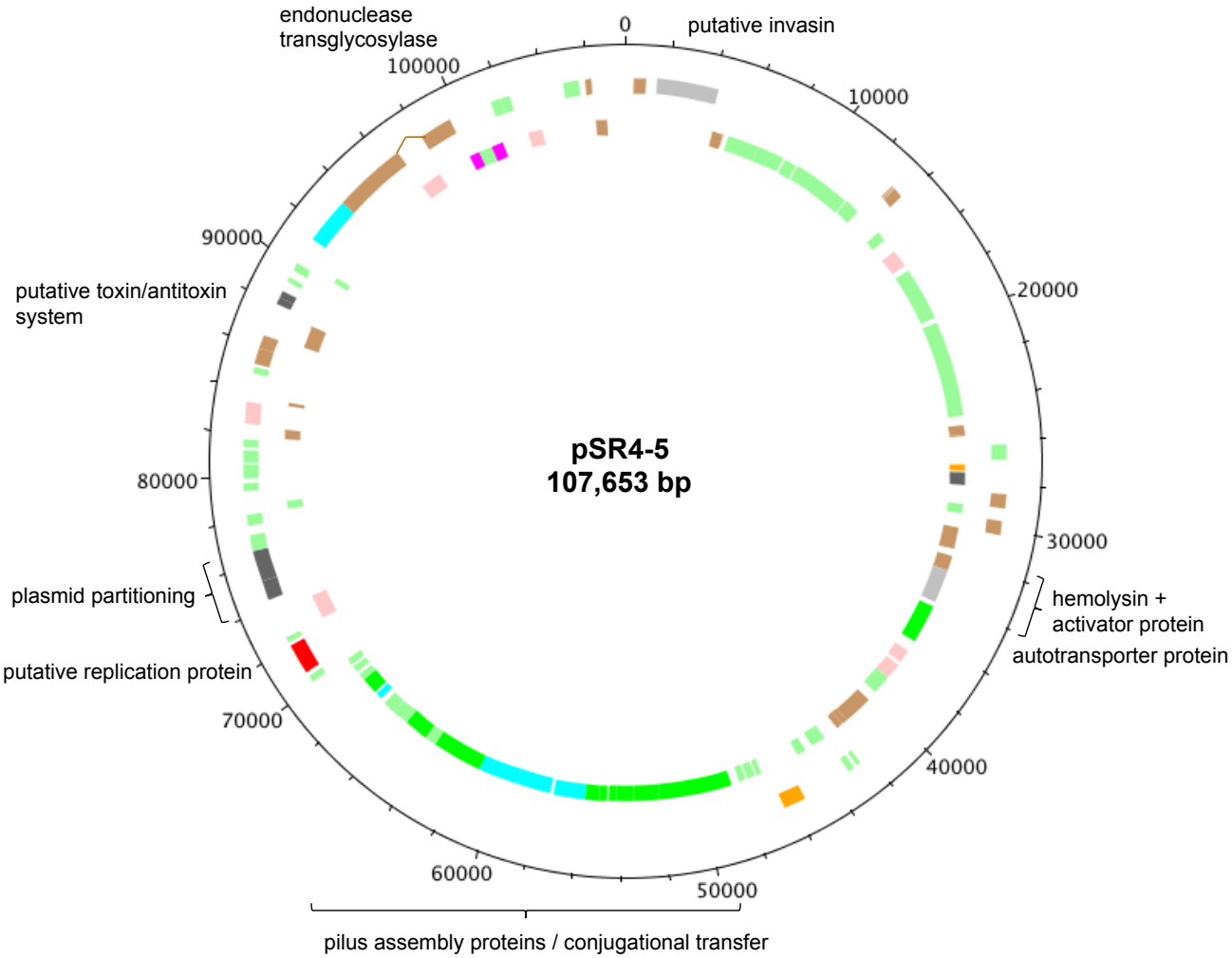


Figure 5.12: Plasmid map of pSR4-5 as found in *Y. enterocolitica* BT 5 O:2a,2b,3 strain 3094. Regions of interest are highlighted.

pSR4-5 was also found to be present in two other BT 5 isolates sequenced in a larger strain collection. A similar plasmid, pYE854, has been isolated from a non-

pathogenic BT 1A strain (Hertwig, Klein, Hammerl and Appel, 2003; Hammerl, et al., 2008). A comparison between these two plasmids is shown in Figure 5.13.

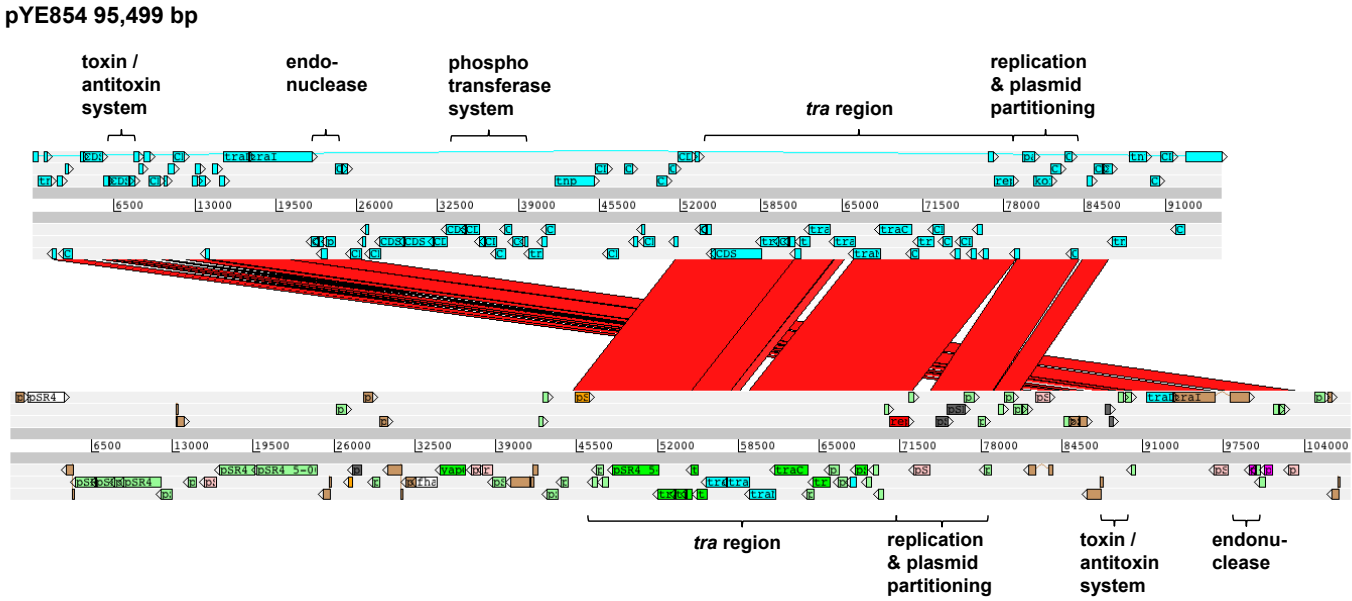


Figure 5.13: Comparison between *Y. enterocolitica* pYE854 and pSR4-5. Regions of interest are highlighted.

Both plasmids have about 60 kb in common, including the conjugative transfer region, as well as replication and plasmid partitioning region. Additionally, pYE854 carries a phosphotransferase system potentially involved with the transport and metabolism of sugars, although functionality could not be shown (Hammerl, et al., 2008).

pYV854 was shown to be able to transfer not only itself but also mobilize the virulence plasmid pYV in experimental studies (Hammerl, et al., 2008). This is important, as pYV itself is non-conjugative. Interestingly, pYE854 was isolated from a non-pathogenic BT 1A strain, which is known never to carry the virulence plasmid (Cornelis, et al., 1987; Lepka, et al., 2009; Wren 2003). The study suggests though that self-transmissible plasmids are common amongst *Yersiniae* as they also found a plasmid related to pYE854 which could also mobilize pYV (Hammerl, et al., 2008).

pYE854 potentially belongs to an unknown incompatibility group as it showed no homology with the known incompatibility groups (Hammerl, et al., 2008). Furthermore co-carriage of pYE854 with plasmids of various incompatibility groups was shown (Hammerl, et al., 2008). As pSR4-5 shares the replication protein with pYE854, they might belong to the same incompatibility group.

5.3 Discussion

5.3.1 The virulence plasmid differs in high- and low-pathogenic BTs

The virulence plasmids of high- and low-pathogenic BTs differ in two important points. The low-pathogenic BTs carry an additional arsenic resistance operon, and the origin of replication differs between high- and low-pathogenic BTs.

Arsenic resistance appears to be an important theme in *Y. enterocolitica*. The low-pathogenic BTs carry an operon on the virulence plasmid as part of the transposon Tn2502. The high- and non-pathogenic BTs on the other hand show the presence of operons for arsenic resistance on the chromosome, one of which is orthologous between the two BTs. In the last century and largely before World War II, arsenic has been used as a therapeutic agent in veterinary and human medicine, in particular in pigs (Neyt, Iriarte, Thi and Cornelis, 1997). This is specifically relevant as pigs are considered one of the main hosts outside the human and an important source of infections. Arsenic resistance thus might promote survival in an important niche. The presence of the arsenic resistance on pYV together with the fact that the low-pathogenic pYVs show few variations in gene content that only relate to IS elements can explain why human case and pig isolates showed identical restriction enzyme patterns (Kapperud and Nesbakken, 1987; Fukushima, Gomyoda, Aleksic and Tsubokura, 1993).

To carry virulence and resistance genes together on a plasmid is unusual (Neyt, Iriarte, Thi and Cornelis, 1997). It is also interesting because the virulence plasmid is known to be easily lost under non-selective conditions (Heesemann, et al., 1983). It has been suggested that the acquisition of arsenic resistance coincided with the use of arsenic compounds, which then lead to a global spread of a low-pathogenic clone (Neyt, Iriarte, Thi and Cornelis, 1997). The population structure and relatedness of the low-pathogenic BTs cannot be deduced from a virulence plasmid and will be investigated in Chapter 7. Given the fluid nature of plasmid maintenance, arsenic resistance might also be viewed as a dispensable trait in low-pathogenic BTs.

Analysis of the origin of replication showed dissimilarity between high- and low-pathogenic BTs. They carry different proteins of distantly related replicons. It has been proposed that the two plasmids are of an independent origin (Portnoy, et al., 1984; Snellings, Popek and Lindler, 2001). The selective pressure is on the type III secretion system and the secreted effector proteins; these build the backbone of the

plasmid (Portnoy, et al., 1984). The rest of the plasmid shows divergent, independent evolution (Snellings, Popek and Lindler, 2001). Different plasmids, one carrying the calcium-dependence region, the other carrying a different origin of replication, might have been cointegrated to form the present variety of pYVs (Portnoy, et al., 1984; Snellings, Popek and Lindler, 2001). It could be argued that the T3SS has integrated into pre-existing plasmids as a mobile genetic element. High- and low-pathogenic BTs might have carried different plasmids into which the T3SS inserted, explaining the difference in the origin of replication. The fact that more than 75% of the plasmid sequence is shared, that the only region of difference is the origin, and that important effector proteins are located outside of the conserved T3SS operons argue against this hypothesis though. The effector proteins located outside the T3SS operons also show some degree of conservation with respect to gene order, and their varying location is explained by the presence of transposases rather than further independent acquisition events. The difference observed in the origins argues for a single source for the T3SS, and that there is a distinct repertoire of plasmids (Portnoy and Falkow, 1981). The pYV should therefore be rather designated as a family of plasmids rather than a clonal molecule that shows no variation (Portnoy and Falkow, 1981). It is difficult to say, which of the plasmids came first, and how they were passed on to other lineages, and how they evolved further. The events leading to the family of virulence plasmids have happened a long time in the past, and intermediates that are potentially found in environmental bacteria would be needed to shed light on the source and evolutionary direction of pYV.

It has been argued that the degree of divergence between plasmids from *Y. pseudotuberculosis/Y. pestis* and *Y. enterocolitica* BT 1B is similar to the divergence found in the chromosome (Portnoy and Falkow, 1981). Together with 16S analysis and biochemical typing, this has probably led to the common conception of evolution of pathogenic *Yersiniae* through a single acquisition event of the virulence plasmid and subsequent speciation into *Y. pseudotuberculosis* and *Y. enterocolitica* (Carniel, 2002; Wren, 2003). The divergence between the virulence plasmid though is only based on hybridization values, and presumably focuses on the conserved T3SS backbone. Generally the virulence plasmids of *Y. enterocolitica* appear to be hybrids. Sometimes the variations observed in the high-pathogenic pYV more closely resemble the organization and single amino acid changes of *Y. pseudotuberculosis* and *Y. pestis*, at other times the low-pathogenic BTs are closer to *Y. pestis*, for example in the origin of replication. Chapter 7 will investigate the relationship

between the *Yersiniae* more thoroughly, and the consequences for plasmid acquisition.

Plasmid variability has also been reported in *Y. pestis* (Eppinger, et al., 2010). Usually, the two *Y. pestis* unique plasmids pMT1 and pPCP1 are highly conserved. Eppinger, et al. (2010) describe a chimeric plasmid pMT-PCP consisting of a cointegrate of a single copy of pMT1 and two copies of pPCP1 in *Y. pestis* biovar Pestoides strain Angola. Some other Pestoides isolates were shown to lack one of the two plasmids, indicating that virulence plasmids play an important role in disease but are not predominant (Eppinger, et al., 2010). Just like in *Y. enterocolitica*, virulence is more than the acquisition and maintenance of a single plasmid and also involves the interplay with chromosomally encoded factors (Revell and Miller, 2001).

5.3.2 Sequence variation in single genes: *lcrV*, *yopM* and *yscP*

Although the T3SS forms a highly conserved part of the virulence plasmid, there are a few variations detected in single genes. These differences could be of physiological interest if they are predominantly present in certain lineages.

The epitope LcrV is also known as the V antigen of pathogenic *Yersiniae* (Roggenkamp, et al., 1997). Research has shown that there are two variants (Roggenkamp, et al., 1997). The first variant is found in *Y. pestis*, *Y. pseudotuberculosis* and in European *Y. enterocolitica* strains as well as the biotype 1B strain 8081. The second variant has an additional 9 amino acids in the hypervariable region of the protein and is mostly found in American BT 1B isolates of *Y. enterocolitica* (Roggenkamp, et al., 1997). Protective immunity is only given for the respective antigen, cross-reactivity *in vivo* is low; this might help explain why there are natural niches for plague in the Americas, whereas in Europe the wild rodent population is immunized due to *Y. enterocolitica* infection resulting in antibody production against LcrV (Roggenkamp, et al., 1997).

Here, a third variation with 5 additional amino acids present in the low-pathogenic BTs 2 and 3 O:9 is described. As this variation possibly changes the protein structure, a protective immunity would not be given after exposure to this antigen. Yet, BTs 2 and 3 are very common in Europe. A physiological role of LcrV as plague protection is therefore questionable as the presence of the three variants of LcrV does not correlate with the known plague foci. It is unknown though whether this third variant is commonly found in BT2 and 3 isolates, as only two strains were examined

here. It could also be that these isolates show mutations independent to the rest of the lineage.

YopM is the only protein of the Yop regulon that contains leucine-rich repeats. It could be shown that the copy number of the repeat region containing 60 aa can vary between different strains (Boland, Havaux and Cornelis, 1998). This phenomenon is not linked to a specific biotype, as biotype 1B plasmids pYVe8081 and pYVeA127/90 have different sizes (Boland, Havaux and Cornelis, 1998). In the presently investigated strains, no sequence variation compared to BT 1B was detected. The size and orientation of YopM does not have any visible effect on its function. pYVe8081 and pYVeA127/90 have not only proteins of different sizes, the orientation of said gene is also opposite on these plasmids; it is in clockwise orientation on pYVeA127/90 and in counter clockwise orientation on pYVe8081 (Foultier and Cornelis, 2003; Snellings, Popek and Lindler, 2001).

The needle protein YscP was found to be 60 amino acids shorter in pYVe8081 than in pYVe227 (BT 2) (Snellings, Popek and Lindler, 2001), which made the protein from pYVe8081 more closely related to the ones from *Y. pestis* and *Y. pseudotuberculosis* plasmids. It was shown that YscP functions as a molecular ruler when the needle of the T3SS is formed (Journet, Agrain, Broz and Cornelis, 2003; Wagner, et al., 2009). As the needle is built with YscR building blocks, YscP is stretched alongside the growth of the needle, until a certain breaking point. YscP then has a second function in signalling back to the T3SS to produce late injectisome molecules (Wagner, et al., 2009). The functionality of YscP depends both on the length of the protein (Journet, Agrain, Broz and Cornelis, 2003) and on the helical content determined by amino acid content (Wagner, et al., 2009). It was shown that needles were longer when encoded by a longer CDS, but also when helix maker alanine was substituted by helix breakers proline and glycine (Wagner, et al., 2009). The needle length might relate to the effectiveness of protein transfer into the target cell.
Here, three different repeats were identified. The red repeat contains 13 amino acids, the green repeat 25 amino acids, and the blue repeat 46 amino acids. *Y. pestis*, *Y. pseudotuberculosis*, and *Y. enterocolitica* BT 1B had a single red-green-blue repeat only, except for some BT 1B strains that showed two full red-green-blue repeats. The low-pathogenic BTs had one red-green-blue repeat and an additional

red-blue repeat, and the low-pathogenic BT 5 strains only showed red-blue repeats in variable number. This has not been correlated with any change in mechanism or outcome of disease so far.

5.3.3 Plasmid Repertoire in Yersinia

Several unique plasmids were detected in the sequenced reference strains. The two plasmids of the non-pathogenic BT 1A are most likely cryptic, as they are small and only potentially encode for few proteins of undefined nature. Large plasmids can also be present in non-pathogenic BTs as demonstrated by the atypical BT 1B/1A strains. The presence of large plasmid is not indicative for pathogenic potential of a strain and might confuse the distinction of high- and non-pathogenic BT. These large plasmids warrant further investigation into the genes encoded and their biological relevance. The unique plasmid of BT 2 strain 21202 carries an antibiotic resistance cassette for tetracycline. Similar to arsenic, this antibiotic has been extensively used in pigs in the past, and might explain the acquisition of this plasmid. Although an initial search was limited to isolates available in the strain collection at Nottingham Trent University, the hypothesis of presence of pSR1-2 in pig isolates of BT 2 could not be validated and was not further pursued. Compared with larger strain collection it now seems more likely that pSR1-2 is a unique acquisition event in BT 2 O:9 strain 21202. The BT 5-unique plasmid could be potentially involved in conjugative transfer of the virulence plasmid pYV as it shows homology with another plasmid in which such a mobilization event was proven (Hammerl, et al., 2008). The unique plasmids constitute an extended gene pool for *Y. enterocolitica* and selection is driven by environmental factors such as the pig niche.

6 Metabolic profiling of the *Y. enterocolitica* biotypes

6.1 Introduction

6.1.1 Biotyping and speciation based on biochemical properties

Biochemical tests such as sugar utilization and fermentation and testing for the presence of enzymes like ornithine decarboxylase traditionally differentiate bacterial species. Given the heterogeneity within *Y. enterocolitica* and the differences in pathogenic potential, this species was further subdivided into several biotypes (see Table 1.2; Wauters, Kandolo and Janssens, 1987; Bottone, 1997). Biotyping is reasonably robust, and reflects the human pathogenic potential. There have been reports however of strains with atypical biotyping reactions (Guiyoule, et al., 1998) as well as high-pathogenic strains that lack the typical invasive phenotype, and which cluster with non-pathogenic BTs in microarray and AFLP analysis (Fearnly, et al., 2005; Howard, et al., 2006; McNally, et al., 2006; personal communication M. Prentice). This variability arises because biotyping is affected by the media used for culturing of the strain before typing and is subject to incubation time and temperature (Wauters, Kandolo and Janssens, 1987; Bottone, 1999; Cornelis, et al., 1987; Farmer, et al., 1992; Stock, Henrichfreise and Wiedemann, 2002; personal communication E. Carniel).

Metabolism is now being recognized as an important factor in colonization not only in the environment but also in human, animal or insect hosts (Rohmer, Hocquet and Miller, 2011). Transition from environmental lifestyle to a host-adapted niche imposes challenges on bacteria in terms of pH, oxygen levels, and nutrient availability such as free iron. An extensive metabolic repertoire is needed to accommodate these changes. Conversely certain metabolic functions might become redundant with limitation to highly specific niches (Rohmer, Hocquet and Miller, 2011). For *Y. pestis*, such a change in lifestyle to become reliant on an insect vector and a warm-blooded host has been investigated in the past, and additional nutrient requirements are known (Burrows and Gillett, 1966). Less is known about *Y. enterocolitica* metabolism and whether some biotypes have adapted to certain niches.

6.1.2 Phenotypic microarray technology

Generally, biotyping is based on a handful of metabolic properties of a cell, and only presents a fraction of the metabolic profile. The phenotypic microarray technology produced by Biolog offers a wider range of metabolites that can be tested in an automated fashion. An overview of the workflow is shown in Figure 6.1 and 6.2.

Bacteria grown on solid medium are suspended in a minimal, defined medium (Bochner, Gadzinski and Panomitros, 2001; Bochner, 2009). The phenotypic microarray is carried out in 96-well plates, with each of the wells containing a different nutrient. For the carbon plates, the minimal medium supplies sulphur, phosphate and other essential nutrients apart from carbon, which will be supplied separately in each of the wells (Bochner, Gadzinski and Panomitros, 2001; Bochner, 2009). For the nitrogen plates, all nutrients except nitrogen are contained in the minimal medium, and so forth.

The phenotypic microarray is based on cellular respiration reflecting cell activity (Figure 6.1; Bochner, Gadzinski and Panomitros, 2001). Transport and metabolism of nutrients results in respiration through the production of energy along the electron transport chain.

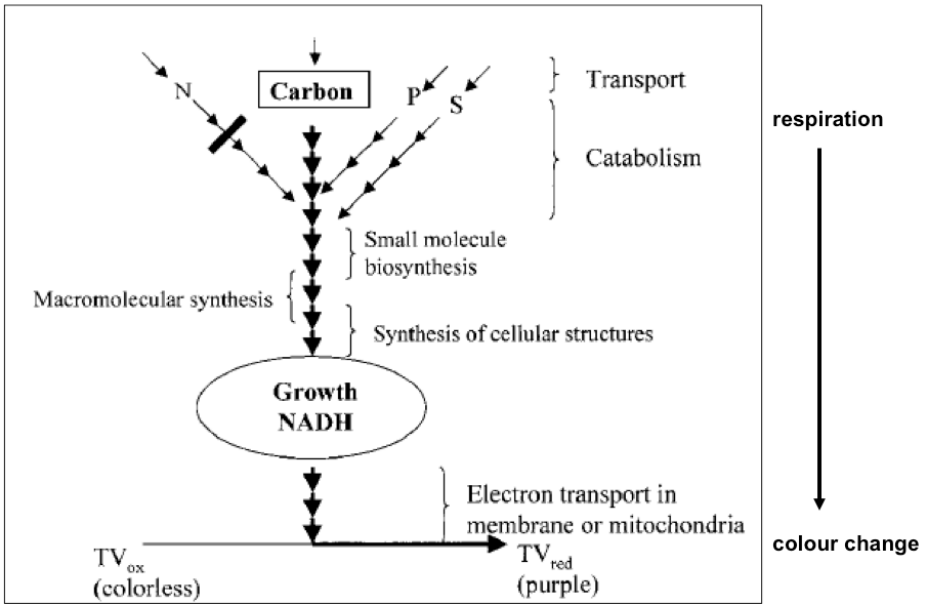


Figure 6.1: Respiration is coupled to conversion of tetrazolium violet dye [taken from Bochner, Gadzinksi and Panomitros, 2001].

Respiration is not necessarily coupled to growth and can therefore also detect phenotypes that do not produce growth but respiration only (Bochner, Gadzinski and Panomitros, 2001; Bochner, 2009). A tetrazolium violet dye added to the reaction is reduced in the presence of electrons causing a colour change from colourless to

purple (Bochner, Gadzinski and Panomitros, 2001). An increased rate of conversion leads to a quicker colour change, because the dye reduction is proportional to the electron flow (Bochner, 2009).

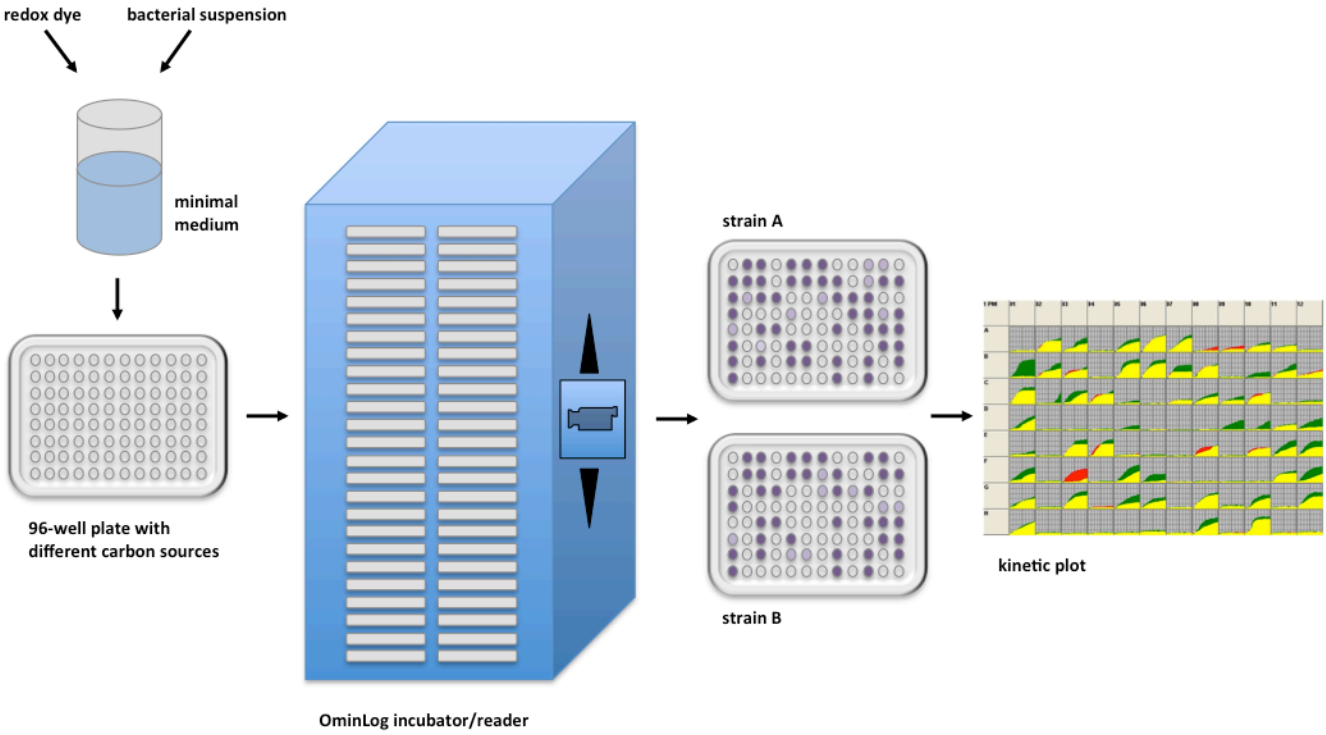


Figure 6.2: Workflow of phenotypic microarray [adapted from Bochner, Gadzinski and Panomitros, 2001; Bochner, 2003]. Redox dye and bacterial suspension are added to the minimal medium. This is used to inoculate 96-well plates containing different nutrient sources in each of the wells. The plates are placed in the OmniLog incubator/reader, which takes up to 50 plates at one time. The camera incorporated in the machine takes readings of the plates. Respiration results in a purple colour change in active wells. Different strains can be compared visually and more in detail in the kinetic plot, showing strain A in green, strain B in red, and the overlay in yellow.

If either transport of the nutrients or metabolism is dysfunctional, there will be no electron flow and consequently no reduction of the dye. Therefore metabolic activity can be monitored and quantified through purple colour production. Accumulation of the dye is harmless to the cells and the signal is intensified as respiration continues (Bochner, Gadzinski and Panomitros, 2001). The 96-well plates are placed in the OminLog instrument, which acts as an incubator and measures dye production in parallel (Figure 6.2; Bochner, Gadzinski and Panomitros, 2001). Data is recorded every 15 minutes over 48 hours.

The phenotypic microarray can be used instead of an API test to identify bacterial species and is potentially more versatile as it can test for 95 different substrates simultaneously (Miller and Rhoden, 1991). It has also been used to study microbial

communities and the changes in composition upon incubation with different carbon sources (Smalla, et al., 1998).

Chosen for this study were two phenotypic microarray plates with carbon sources, PM1 and PM2A, one phenotypic microarray plate with nitrogen sources, PM3B, one mixed phenotypic microarray plate with phosphorus and sulphur sources, PM4A, and phenotypic microarray plates PM9 and 10 which test for osmolytes and pH. For each strain, triplicates were carried out at optimum growth temperature of 28°C and at human body temperature of 37°C. Data will mainly be presented on PM1-4 at 28°C.

The phenotypic microarray was used to establish the metabolic profile of each sequenced reference strain. The differences observed in the genomic analysis should be reflected in the metabolism of each strain. The phenotypic microarray should aid annotation and establish direct links between phenotype and genotype.

6.2 Results

6.2.1 Metabolic activity in *Y. enterocolitica* biotypes at optimum temperature

Given the apparent role of metabolism in determining the breadth of host range and pathogenic potential and the targeted loss of function as presented in the pseudogenes, the reference strains were assayed using the phenotypic microarray for their ability to utilise different sources of carbon, nitrogen, sulphur, and phosphate at optimum temperature of 28°C. The signal values are given in the appendix Table 9.6 (p. 242) and the rearranged layout for positive wells plus significant changes towards BT 1B strain 8081 are listed in Table 9.7 (p. 254).

The heat map in Figure 6.3A shows the signal values corresponding to the intensity of utilization for the different nutrient sources. The three replicates are shown for each strain, with blue indicating utilization and red absence of utilization. Figure 6.3B shows the change towards BT 1B, with yellow indicating a significantly stronger activity and blue a significantly reduced activity compared to the high-pathogenic BT. Even from an overview it is apparent that the non-pathogenic BT 1A shows a wider metabolic activity with the highest number of positiv signals (Figure 6.3A). This is consistent with the biotyping scheme in which this BT is positive for the fermentation of all the sugars tested and maintains functionality of metabolism. The metabolic activity detected is often also significantly higher than in BT 1B (Figure 6.3B), suggesting a more effective metabolism. BT 2 O:9 strain 21202 and BT 3 O:9 strain 5603 were shown to be nearly identical with respect to genomic content and pseudogenes. Looking at their metabolic activity (Figure 6.3A) but specificatlly the change towards BT 1B (Figure 6.3B) one can see that they exhibit very similar profiles. This confirms their genetic relatedness, and also suggests that there is little difference between BTs 2 and 3 if they are of the same serotype.

Comparing the low- and high-pathogenic BT to the non-pathogenic BT, both have lost metabolic pathways with two extreme cases of loss of function presented by BT3 O:5,27 and BT5. BT 3 O:,27 shows marked reduction of nutrient utilization in nitrogen and phosphorus sources. The metabolism of all nutrient sources appears to be greatly reduced in BT 5, with only clear positive signals in carbon sources and hardly any signals in nitrogen, phosphorus and sulphur sources. The biotyping scheme shows an indication for this as well, as BT5 is mostly negative for the fermentation of sugars tested.

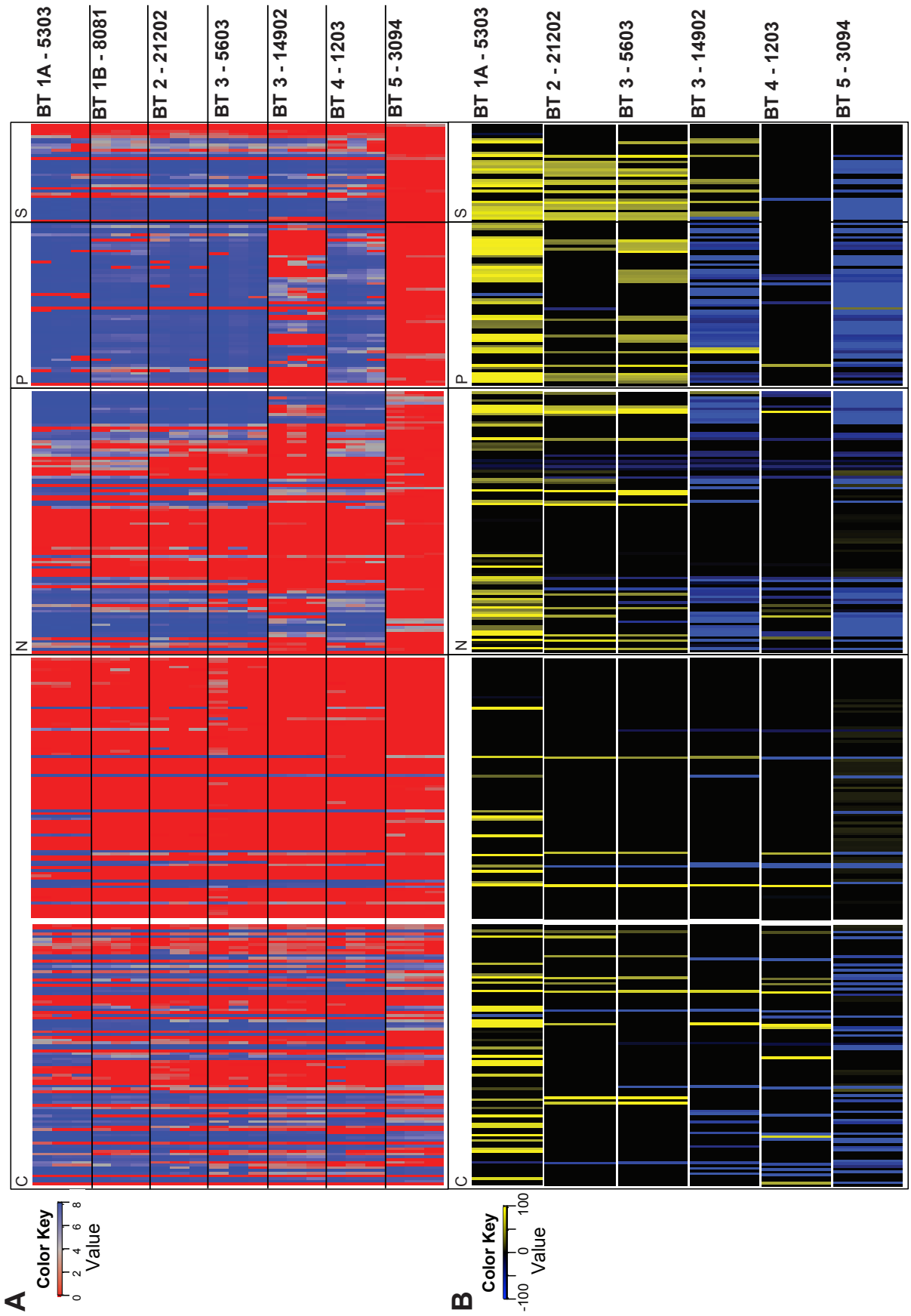


Figure 6.3: Signal values (A) and significant changes towards BT 1B (B) depicted as heatmaps. Triplicates are shown for the signal values. C – carbon sources (190), N – nitrogen sources (95), P – phosphorus sources (59), S – sulphur sources (35).

Determining positive and negative signals, it was also examined whether the change observed posed a significant difference towards the BT 1B reference strain 8081 (Figure 6.3B, Figure 6.4 and 5 for absolute numbers).

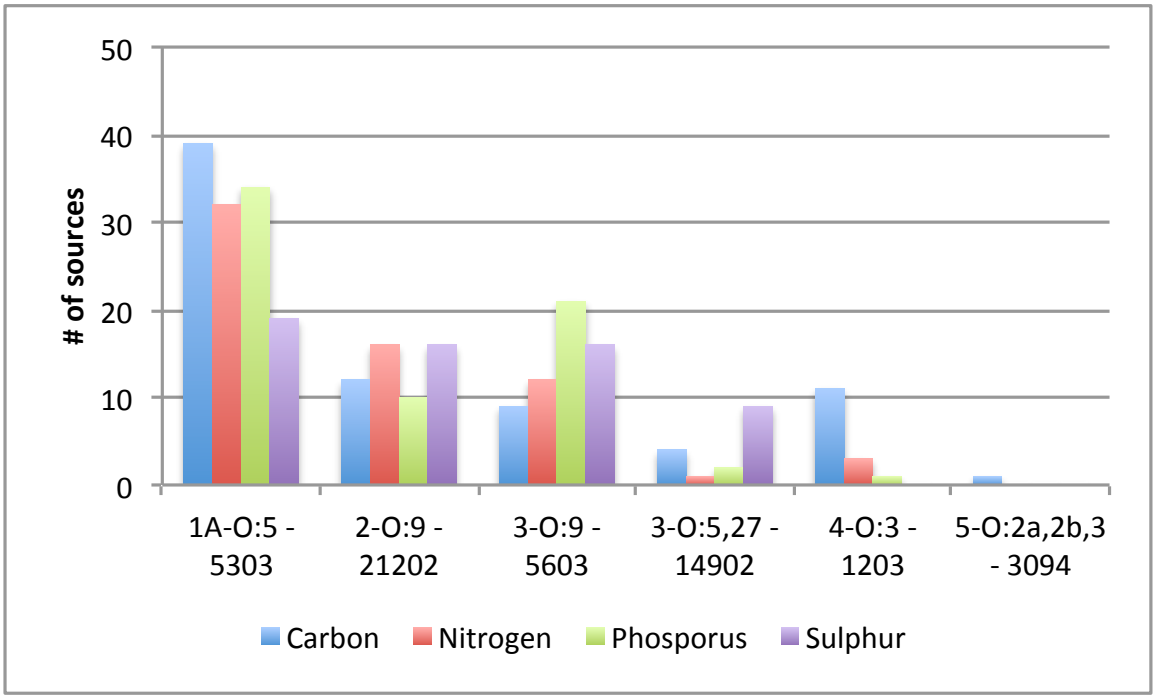


Figure 6.4: Number of sources that are expressed significantly higher than in BT 1B strain 8081.

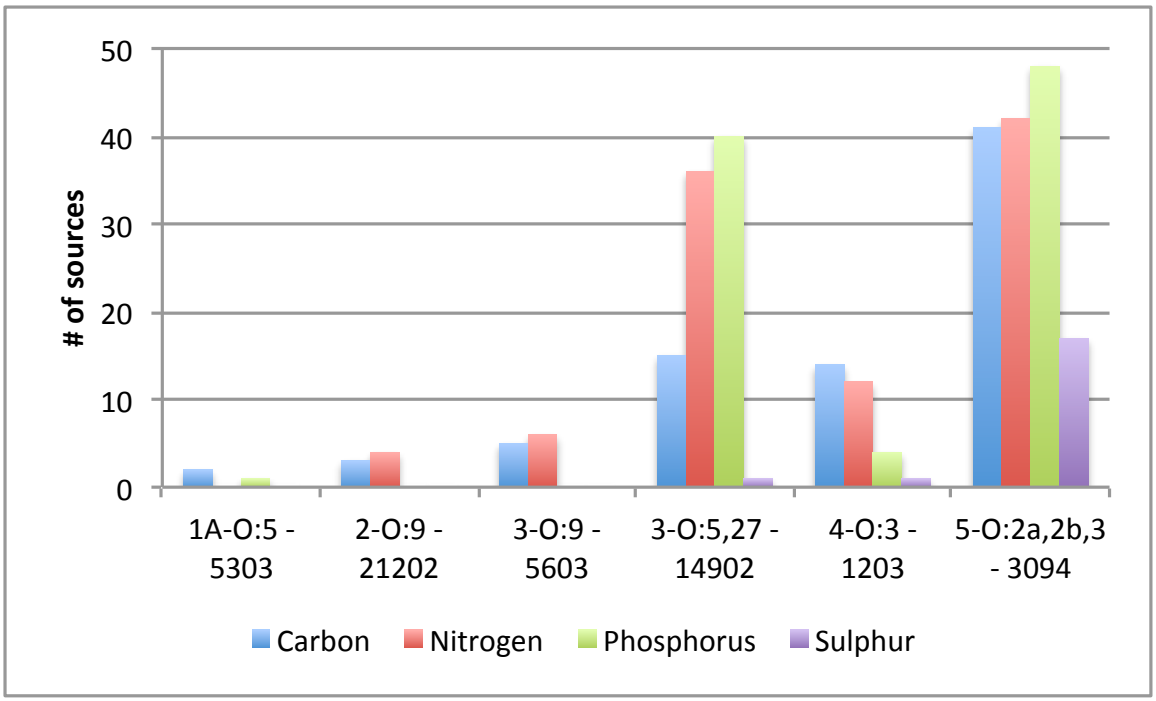


Figure 6.5: Number of sources that are expressed significantly lower than in BT 1B 8081.

The non-pathogenic BT1A not only shows a wider range of metabolic activity, this activity is also significantly higher than in BT 1B. In contrast there are few sources that show a significantly lower activity, and two out of these three relate to sources that are not utilized in BT 1A. Both BT 2 and 3 O:9 isolates show a significantly higher activity in some sources, but also show a significantly lower activity especially in carbon and nitrogen sources. The other three BTs show only a selected number of sources which are utilized significantly higher than in BT 1B. BT 3 O:5,27 shows a significant reduction in nitrogen and phosphorus sources as was already apparent from the heatmap. BT 4 shows some reduction in carbon and nitrogen sources, similar to the ST O:9 isolates. The absence of metabolic activity in BT 5 is seen in the high number of significantly reduced sources in all nutrients.

6.2.1.1 Metabolism of carbon sources

Looking in more detail at the carbon sources (phenotypic microarray plates PM1 and 2A), a total of 27 carbon sources (14.2% of total of 190 tested carbon sources) were found to be utilised in all biotypes, with a further 14 sources (7.4%) utilised by all but BT 5 (Table 9.7). In line with the genomic observations there are a significant number of additional metabolic capacities shared by BTs 1A and 1B strains including the ability to use D-alanine, L-asparagine, dextrin, succinamic acid, and L-histidine as sole carbon sources. Also reflected in the genetic makeup of the BT 1A is a greater metabolic capacity than all the other biotypes with 15 carbon sources, which are utilized only by BT1A. In addition there are a further 12 carbon sources that BT1A shares in combination with one or more of the BT, whereas there are only two carbon sources that are present in any of the other BTs and not in 1A. These two carbon sources are glycogen, which is only utilized in BT 2 O:9 strain 21202, and m-inositol which is metabolized in all BTs except BTs 1A and 5.

With the exception of the two mentioned carbon sources there are no other carbon sources unique to the high- and low-pathogenic BTs that are not present in BT 1A. BT 1A is an environmental organism well adapted to a number of niches. Together with the fact that it has the lowest number of pseudogenes, this suggests that BT 1A is the oldest lineage from which the other BTs evolved through loss of metabolic functions.

6.2.1.2 Utilization of essential nutrients nitrogen, phosphorus and sulphur

The utilization of the essential nutrients nitrogen, phosphorus, and sulphur is tested for in phenotypic microarray plates PM3B and 4A. BT1A, 1B, and 2-4 have 13 nitrogen sources (13.7% of total of 95 sources), 2 phosphorus sources (3.4% of total 59 sources), and 14 sulphur sources (40.0% of total 35 sources) in common. No metabolic activity can be seen for BT5 for any of these sources.

Similar to the observations made for carbon sources, there are two sources which are exclusively utilized in BTs 1A and 1B. Both BTs metabolize D-asparagine and L-phenylalanine as sole nitrogen sources. There are seven nitrogen sources, two phosphorus sources and one sulphur source, which are uniquely metabolised in BT 1A. Further 27 nitrogen sources, 44 phosphorus sources, and 4 sulphur sources show shared metabolism in BT 1A and any of the other BTs, underlining the increased metabolic flexibility of the non-pathogenic BT. There are three nitrogen sources, seven phosphorus, and 2 sulphur sources absent from BT 1A. These are listed in Table 6.1.

Table 6.1: Utilization of N, P, and S sources absent from BT 1A.

	Nutrient source	BT 1A 5303	BT 1B 8081	BT 2 21202	BT 3 5603	BT 3 14902	BT 4 1203	BT 5 3094
N	D,L-a-amino-caprylic acid	–	–	+	–	–	–	–
N	L-tryptophan	–	+	–	–	–	–	–
N	Xanthine	–	+	–	–	–	–	–
P	Adenosine-5'-monophosphate	–	+	+	+	–	+	–
P	D-glucose-6-phosphate	–	+	+	+	–	+	–
P	Cytidine-3'-monophosphate	–	+	+	–	–	–	–
P	Cytidine-5'-monophosphate	–	+	–	+	–	+	–
P	Uridine-5'-monophosphate	–	+	–	+	–	+	–
P	O-phospho-L-tyrosine	–	–	+	+	–	–	–
P	Adenosine-2',3'-cyclic monophosphate	–	–	–	+	–	–	–
S	Hypotaurine	–	–	–	+	+	–	–
S	D,L-ethionine	–	–	–	+	–	–	–

Adenosine, cytidine, and uridine phosphates are involved in purine and pyrimidine metabolism. Their physiological relevance in BT 1A is unknown. An impairement of purine and pyrimidine metabolism is rather apparent in BT 3 O:5,27 and BT 5, as these also lack the utilization of guanosine and thymidine phosphates as well as further adenosine, cytidine, uridine mono- and cyclic phosphates.

As was seen earlier, BT 3 O:5,27 has a defect in nitrogen metabolism. This relates to amino acids amongst others. Three of the amino acids which BT 3 O:5,27 cannot metabolize are L-arginine, L-glutamic acid, and L-glutamine. These may be grouped together in the glutamate family. Apart from these, the other nine amino acids and di-amino acids cannot be grouped chemically or physiologically.

6.2.1.3 Biotyping reactions

The sugars used for differentiation of *Yersiniae* and for biotyping of *Y. enterocolitica* are also found on the phenotypic microarray. A comparison of results is given in Table 6.2.

Table 6.2: Comparison of biotyping and phenotypic microarray reactions.

Substrate	Biotyping result *	Phenotypic microarray substrate (PM plate and position)	Result	Agreement biotyping and phenotypic microarray
Glucose	All BT +	A-D-Glucose (PM1 C09)	All BT +	Yes
Sucrose	All BT +	Sucrose (PM1 D11)	All BT +	Yes
Rhamnose	All BT –	L-Rhamnose (PM1 C06)	All BT –	Yes
Raffinose	All BT –	D-Raffinose (PM2A D01)	BT 1A +, BT 1B, 2, 3, 4, 5 –	No
Melibiose	All BT –	D-Melibiose (PM1 C11)	All BT –	Yes
Cellobiose	All BT +	D-Cellobiose (PM1 F11)	All BT +	Yes
Sorbose	All BT +, except BT 5 –	L-Sorbose (PM2A D04)	All BT +, except BT 5 –	Yes
Salicin	BT 1A +	Salicin (PM2A D02)	BT 1A +	Yes
Xylose	BT 1A, 1B, 2, 3 +, BT 4 –, BT 5 V	D-Xylose (PM1 B08)	BT 1A, 1B, 2, 3 +, BT 4 –, BT 5 +	Yes
Trehalose	BT 5 –	D-Trehalose (PM1 A10)	BT 5 –	Yes
Inositol	All BT +	M-Inositol (PM1 F03)	BT 1A –, BT1B, 2, 3, 4 +, BT 5 –	No

*adapted from Bottone, 1997, and Wauters, Kandolo and Janssens, 1987. + positive, – negative, V variable.

There are only two sugars in which the biotyping reaction and the phenotypic microarray disagree.

Raffinose utilization is supposed to be absent from *Y. enterocolitica* but in the phenotypic microarray BT 1A gives a positive signal. Looking at the absorbance measurements, one can clearly see that BT 1A metabolizes raffinose, although an activity can only be detected about 28 hours after the start of the incubation (Figure 6.6).

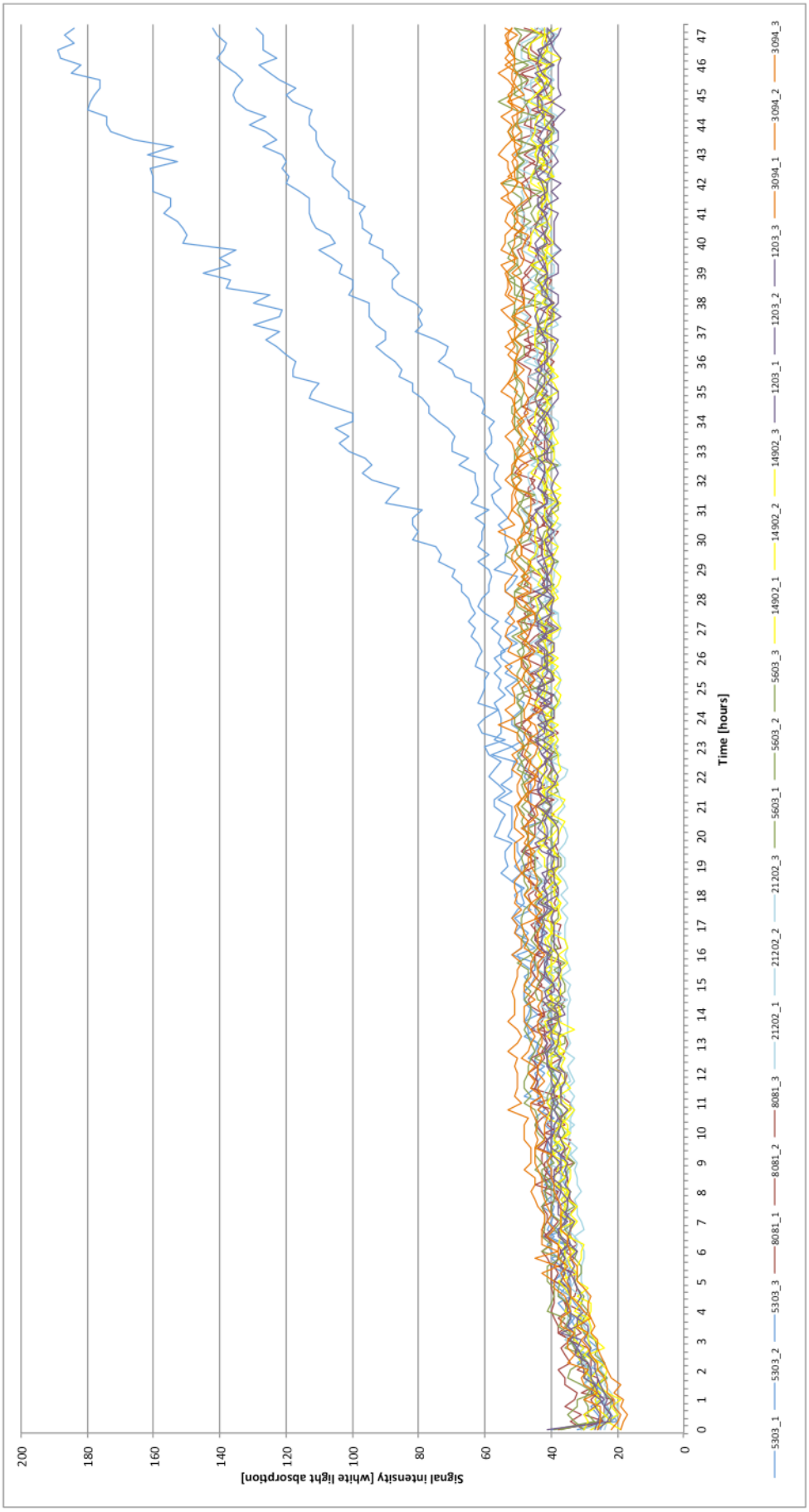


Figure 6.6: Kinetic plots for raffinose in *Y. enterocolitica* biotypes. Triplicates are shown. BT 1A / 5303 – blue; BT 1B / 8081 – red; BT 2 / 21202 – light blue; BT 3 / 5603 – green; BT 3 / 14902 – yellow; BT 4 / 1203 – purple; BT 5 / 3094 – orange.

It has been reported that raffinose might be metabolized if the *lac* permease is present conferring lactose uptake as well (Cornelis, Luke and Richmond, 1978). BT 1A does utilize α-D-lactose as well as lactulose, and a copy of *lacIZY* is present encoding the lactose operon repressor, β-glucosidase, and permease respectively (*YE5303-03921 - 41*). Copies of *lacI* (*YE2591*) and *lacZ* (*YE2592*) are carried in all *Y. enterocolitica* biotypes but a permease is absent. The presence of transposases in some of the strains could potentially indicate a loss of the gene mediated by recombination. A definitive genomic basis for the raffinose-positive phenotype could not be established but it could be that the *lacIZY* operon is involved in the metabolism of this sugar. Another possibility for a different test result is that the mode of the test is different, i.e. utilization under aerobic conditions versus anaerobic fermentation.

The second difference between biotyping reaction and phenotypic microarray result is inositol. All BTs are normally positive for the utilization of this sugar, but the phenotypic microarray only showed metabolism in BT 1B, 2, 3 and 4. This will be explained in 6.2.2.5.

6.2.2 Differences relating to genetic differences

The general trends observed both in the genome comparisons and the analysis of the pseudo- and partial genes highlighted metabolic regions and transport proteins (Chapter 4). BT 1A was found to have additional operons relating to the uptake and conversion of sugars such as D-arabitol, and BT 5 was shown to contain mutations in core genes that might relate to the biotyping reactions.

These trends are reflected in the metabolic profiling. In the following, some examples of linking genotypes to specific phenotypes are given. This task is most straightforward when considering carbon metabolism, because well characterized examples from *E. coli* and *Salmonella* can be found. Nitrogen, phosphorus and sulphur metabolism on the other hand is less clear and the metabolites cannot be tied into metabolic pathways easily. This is due to the fact that some metabolites are not naturally occuring intermediates in standard metabolic pathways and constitute synthetic nutrients not found in the environment. Their metabolism is unknown and their physiological relevance is unidentified.

Other difficulties presented in analysing the phenotypic microarray relate to the fact that genomic data alone is insufficient for allocation of pathways, as homology to known enzymes does not indicate specificity of a transporter or an enzyme, as for

example ABC transporters might transport a range of metabolites from amino acids to sugars. Additionally, the mutations giving rise to the phenotype might be acquired independently in different genes of a pathway. If then a pathway is not clearly defined in terms of regulation, associated transporters, secondary pathways, etc., a clear link between phenotype and genotype is difficult to establish and requires additional experiments with targeted knock-out mutants. The metabolites tested in the phenotypic microarray also do not present complete pathways, but rather individual points in larger networks. It is therefore not possible to obtain a continuous picture of active or disrupted pathways. Nevertheless it is possible to obtain an overall impression of metabolic capacities of a cell.

6.2.2.1 L-arabinose

One clear example for linking phenotype and genotype can be seen in L-arabinose metabolism (PM1 A02). This is a carbon source metabolized in all BTs, and the respective enzymes are encoded by the *araABFGHC* operon (Figure 6.7).

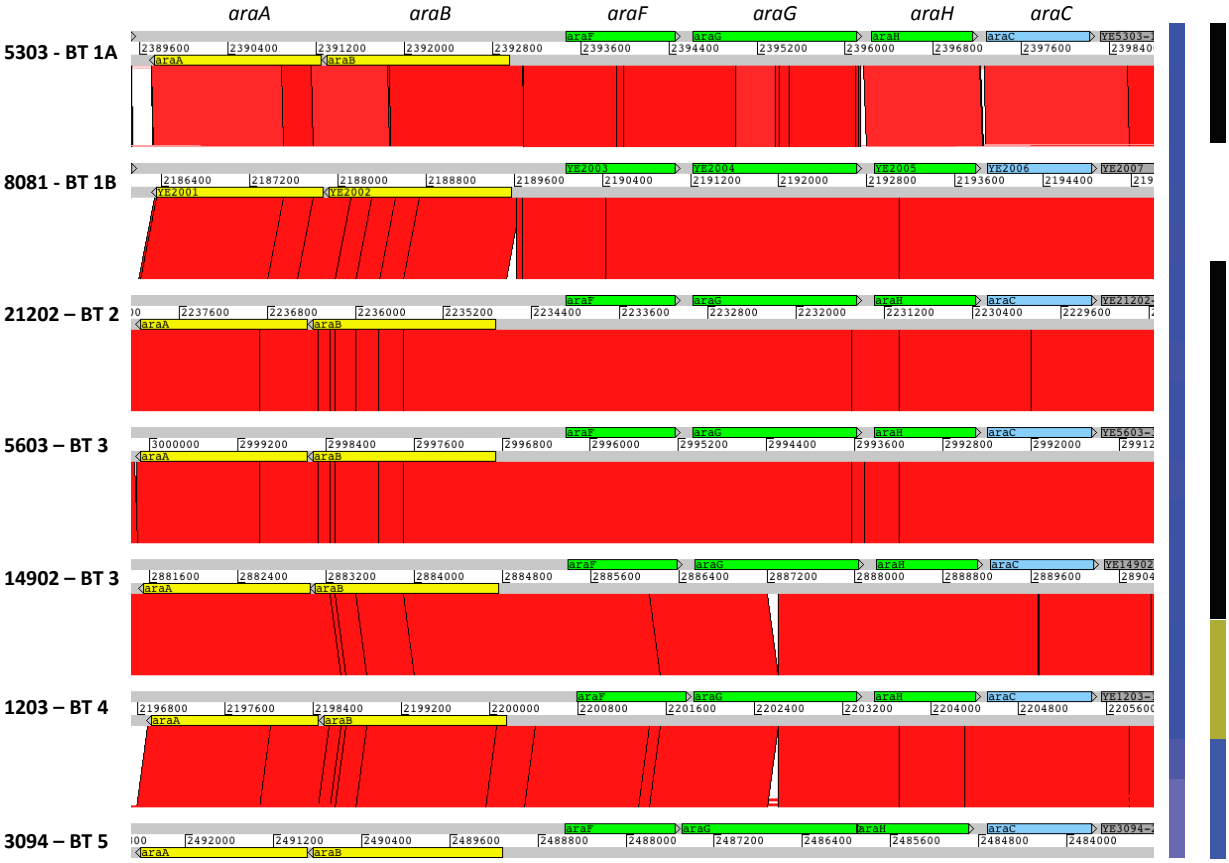


Figure 6.7: Operon for L-arabinose utilization contrasted with signal values and significant changes compared to BT 1B. Utilization (blue +, red −), significant change compared to BT 1B (yellow +, black none, blue −).

The transport proteins are AraFGH, *araAB* encode for an isomerase and a L-ribulokinase, respectively, and AraC is the regulatory protein for this operon. Although there appear to be slight variations between and within genes (BT 4 deletion in *araG*), these do not have an effect as utilization could be shown in all BTs. The signal values for all BTs are positive, although there is a significantly higher intensity in BT 4 and a significantly lower intensity for BT 5 compared to BT 1B. Looking at the absorbance recordings over time (Figure 6.8), one can see that BT 5 shows utilization of L-arabinose, but only to a limited extent. In contrast, BT 4 is very efficient at using L-arabinose. There appears to be a slight delay of 4 hours compared to BTs 1A, 1B, 2, and 3 in the onset of utilization, but then BT 4 outperformes the other BTs in all three replicates.

Other carbon sources that are metabolized in all BTs and for which unambiguous operons exist are N-acetyl-D-glucosamine (*nagEBACD*), D-galactose (*galDBAM*), D-mannose (*manZYXA*), glycerol (*glpKFTABCR*), D-glucuronic acid (*uxaC*), D-mannitol (*mtlRDA*), D-fructose (*fruBKAR*), and L-serine (*sdaBCA*).

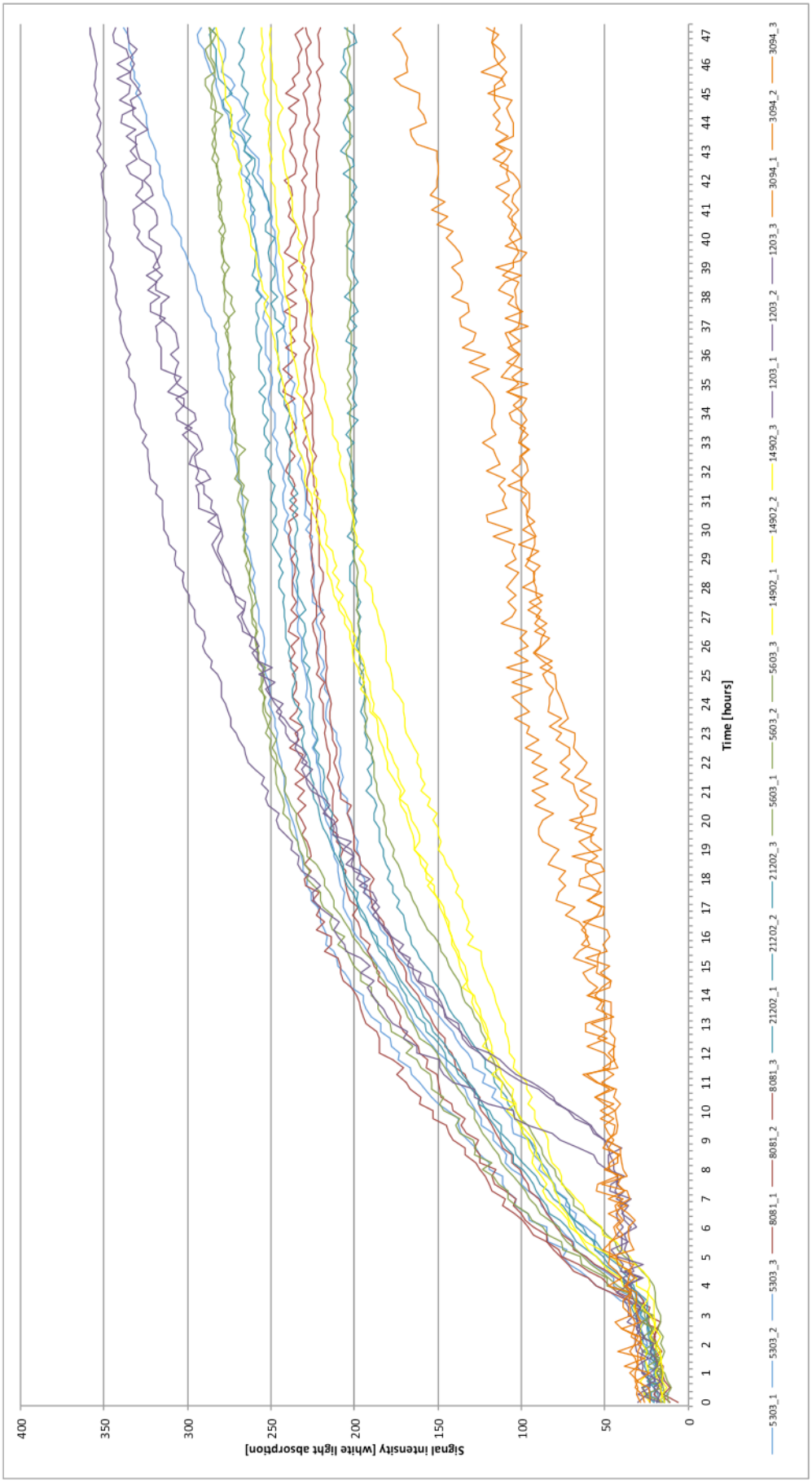


Figure 6.8: L-arabinose utlization kinetic plots over time for *Y. enterocolitica* biotypes. Triplicates are shown. BT 1A / 5303 – blue; BT 1B / 8081 – red; BT 2 / 21202 – light blue; BT 3 / 5603 – green; BT 3 / 14902 – yellow; BT 4 / 1203 – purple; BT 5 / 3094 – orange.

6.2.2.2 D-trehalose and D-sorbose

Examples for mutations specific for BT 5 are found in D-trehalose and D-sorbose, which are also used in biotyping. The respective tests in the phenotypic microarray are found in PM1 A10 and PM2A D04.

The sorbose operon is shown in Figure 6.9A.

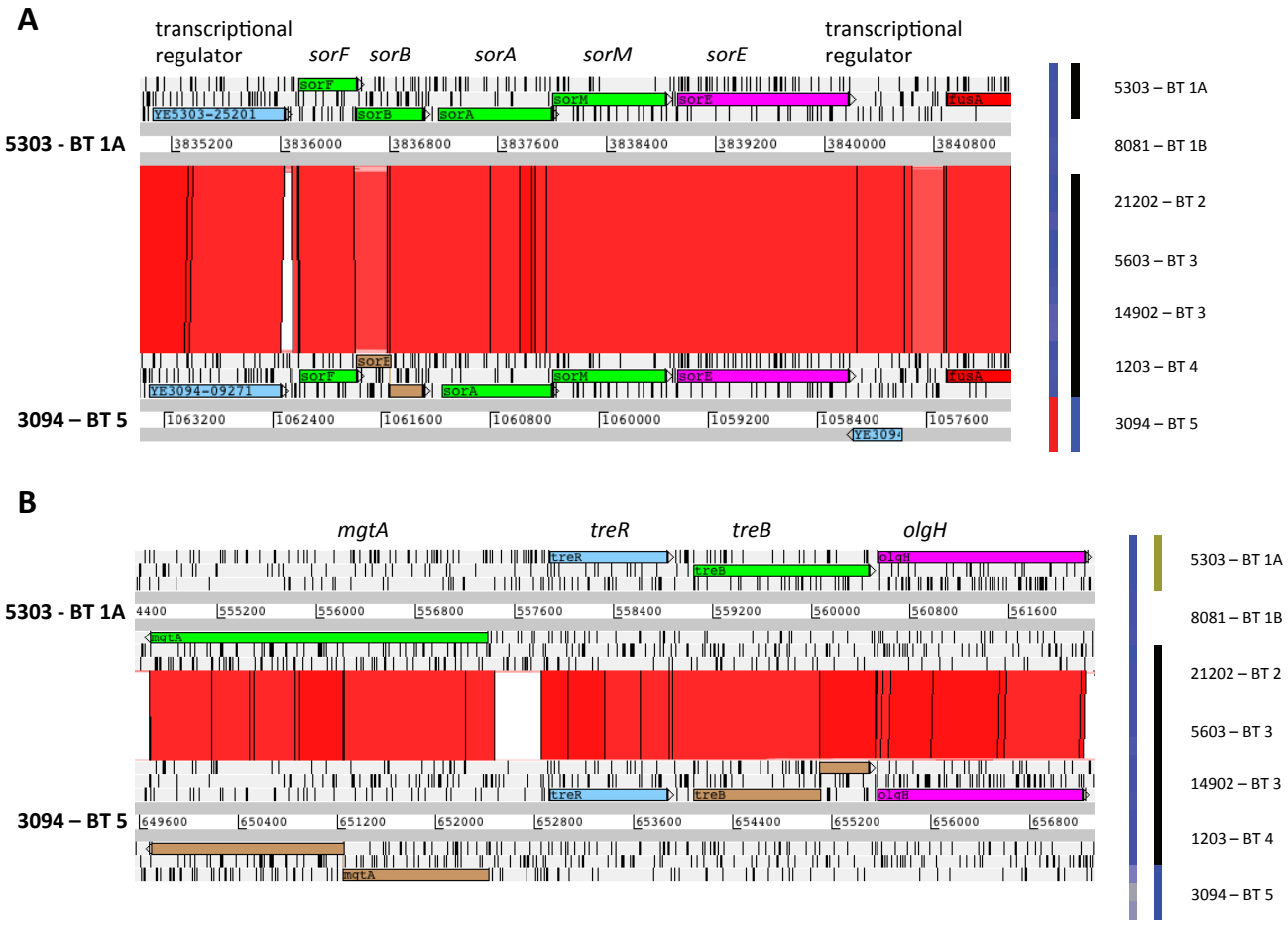


Figure 6.9: Mutations specific for biotyping of BT 5 contrasted with signal values and significant changes compared to BT 1B. Utilization (blue +, red −), significant change compared to BT 1B (yellow +, black none, blue −). A – Sorbose operon. B – trehalose operon.

The sorbose-specific phosphotransferase uptake system (PTS) is made up of several components encoded by *sorFBAM*. SorE is a L-sorbose-1-phosphate reducatase. There are two transcriptional regulatory proteins upstream and downstream of the operon but it is not known whether any of these or other regulators are involved with governing the expression of this operon. BT 5 is the only biotype with a mutation in this operon. It is missing a TTTA repeat motif resulting in a frameshift mutation in *sorB*, one of the PTS components. The signal values reflect this mutation. BT 5 is the only biotype without metabolic activity in this well, and this reduction is also

significant compared to BT 1B. BT 1A shows a significantly higher utilization than BT 1B.

The trehalose operon is depicted in Figure 6.9B. Similar to the sorbose operon, BT 5 shows a frameshift in the PTS system TreB resulting from an insertion of one A. The trehalose operon repressor TreR and the trehalose-6-phosphate hydrolase OlgH remain intact. Upstream of the trehalose operon is a magnesium transporter encoded by *mgtA*, which also specifically shows a frameshift mutation in BT 5. There is a further gene possibly encoding a component of the PTS involved in trehalose transport (*YE3690*). This gene is disrupted by the presence of a transposase in BT 5, indicating further mutations in trehalose metabolism.

6.2.2.3 N-acetyl-D-galactosamine

The utilization of N-acteyl-D-galactosamine (PM2A B01) is encoded by the *aga* operon (Figure 6.10).

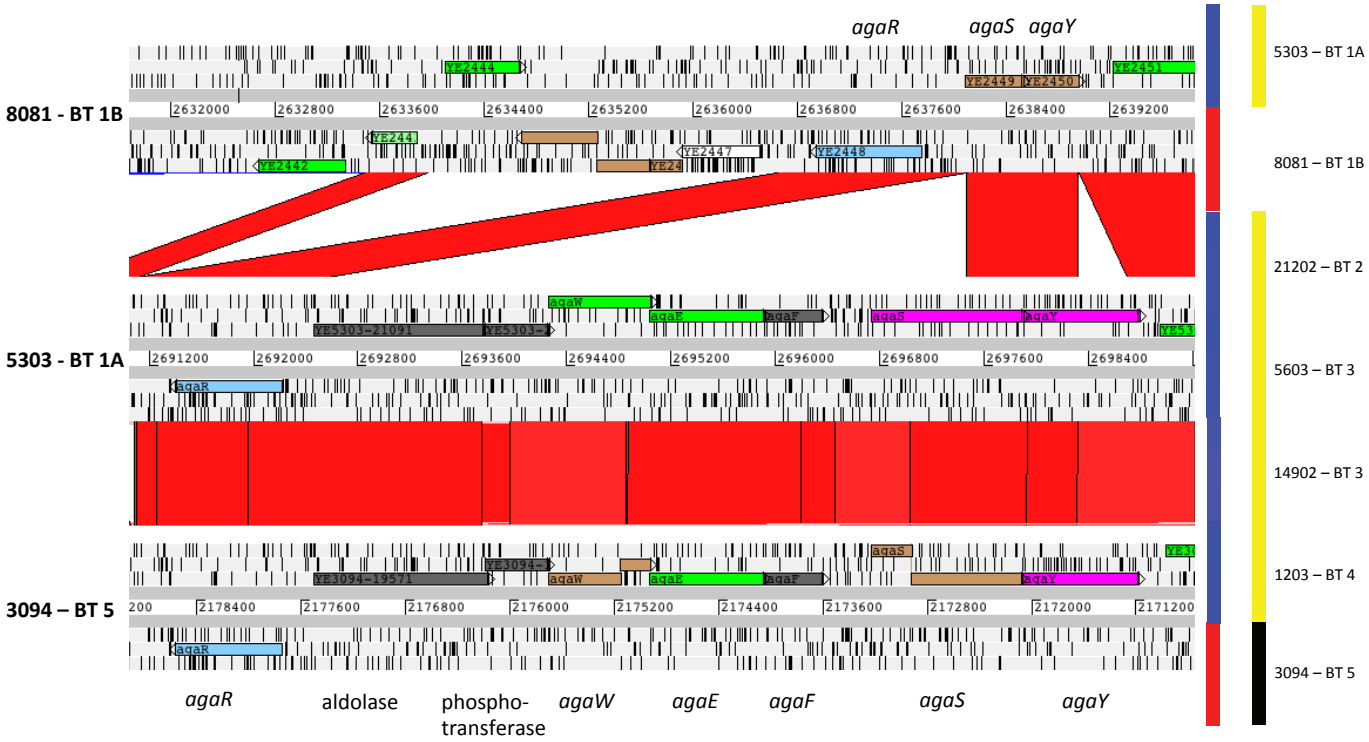


Figure 6.10: Operon structure for utilization of N-acetyl-D-galactosamine in representative BTs contrasted with signal values and significant changes compared to BT 1B. Utilization (blue +, red −), significant change compared to BT 1B (yellow +, black none, blue −). BT 1B – deletion of most of the operon; BT 1A – representative for BTs 1A, 2, 3, and 4, fully functional; BT 5 – operon present, but mutations inhibit functionality.

The operon starts with a DeoR-family regulatory protein AgaR. This is present in all BTs. The D-tagatose-1,6-bisphosphate aldolase, N-acteylgalactosamine-specific phosphotransferase, and the phosophotransferase system encoded by *agaWEF* have been deleted from BT 1B. BT 5 is the only BT to show a frameshift in a homopolymeric run of Gs in *agaW*. The tagatose-6-phosphate ketose/aldose isomerase AgaS, and the D-tagatose-1,6-bisphosphate aldolase AgaY have been retained as fragments in BT 1B. BT 5 has a 4 base deletion resulting in a frameshift mutation in *agaS.* Both mutations in BT 5 render the utilization of N-acetyl-D-galactosamine non-functional, as can be seen in the signal values. The biotypes metabolising this carbon source show a significantly higher expression than BT 1B.
A highly conserved, homologous operon is found in *E. coli* (Reizer, et al., 1996) and its functionality was confirmed for the phosphotransferase system of N-acetyl-D-galactosamine (Brinkkötter, Klöß, Alpert and Lengeler, 2000). Functionality of this cluster has been independently studied in a BT 4 O:3 strain (Batzilla, et al., 2011b). Complementation studies showed that this operon could confer the ability to grow on N-acetyl-D-galactosamine onto BT 1B strains (Batzilla, et al., 2011b), confirming that this cluster is responsible for the phenotype.

6.2.2.4 D-arabitol

The utilization of D-arabitol (PM2A B06) is unique to the BT 1A strain studied here. The utilization could be linked to a unique region contained in the BT 1A genome (Figure 6.11).
D-arabitol utilization is encoded for by a mannitol dehydrogenase DalD and a xylulose kinase AtlK. A transcriptional regulatory protein of the DeoR2 family guards the expression of the operon, and AtlT constitutes the D-arabitol membrane transporter. The gene *acpD* for an acyl carrier protein phosphodiesterase is located upstream of the operon. This region has been deleted from the pathogenic BTs. BT 1B retains a fragment of the regulatory protein, but the operon is completely absent from the low-pathogenic BTs.
A similar, conserved D-arabitol operon structure has been described in *Klebsiella pneumoniae* and function confirmed (Heuel, Turgut, Schmid and Lengeler, 1997).

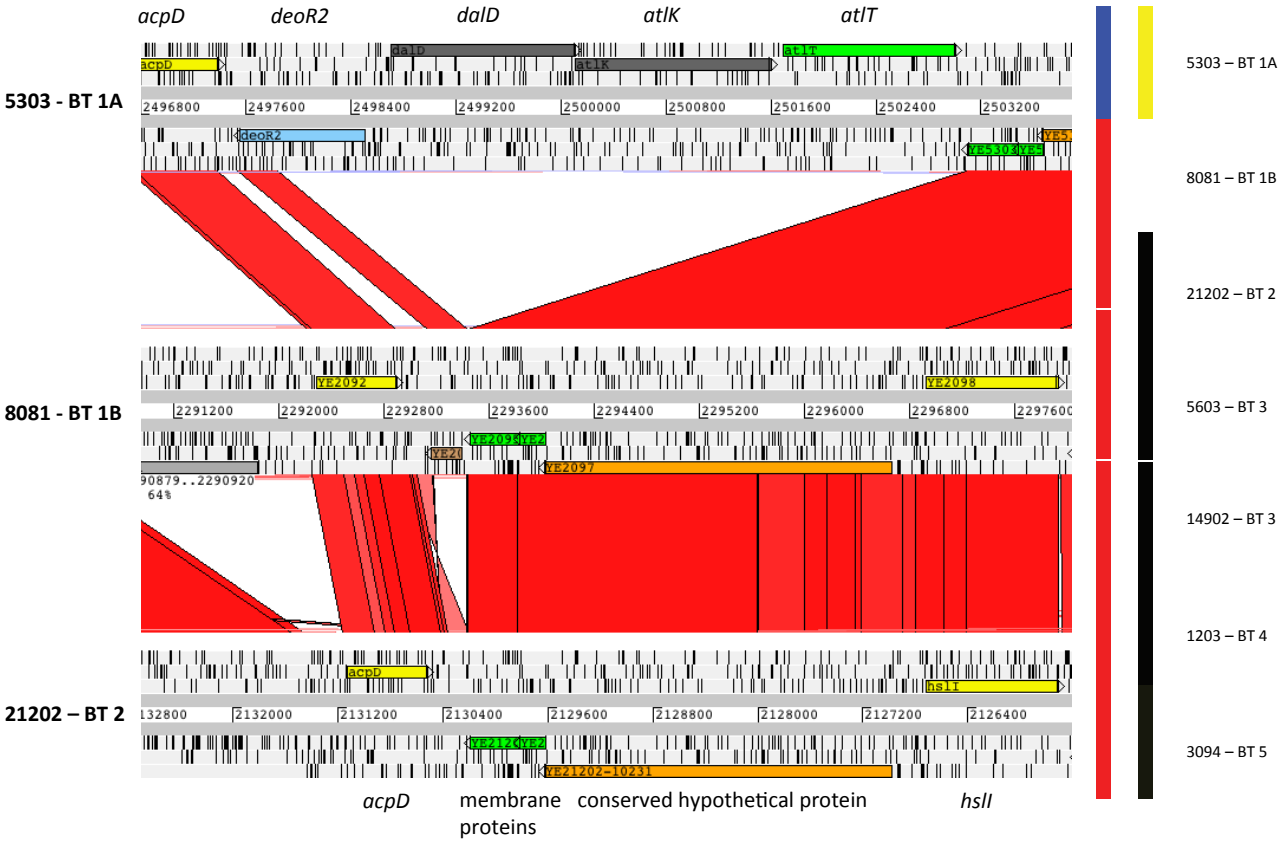


Figure 6.11: Unique operon for D-arabitol utilization in BT 1A, contrasted with signal values and significant changes compared to BT 1B. Utilization (blue +, red −), significant change compared to BT 1B (yellow +, black none, blue −). BT 2 as representative for low-pathogenic BTs.

D-arabitol has been linked to the non-pathogenic biotype before in comparison to BT 4 O:3 (Shehee and Sobsey, 2004). A positive D-arabitol phenotype is not exclusive to non-pathogenic BTs though (Fantasia Mazzotti, et al., 1985; Fantasia Mazzotti, Mingrone and Martini, 1987), as some BT 1A strains can be negative, just as some pathogenic BT 3 strains were found to metabolize D-arabitol.

6.2.2.5 m-inositol

Acid production from inositol is one of the biotyping reactions for which all *Y. enterocolitica* BT are supposed to be positive. The phenotypic microarray also tests for this in form of m-inositol (PM1 F03). As could be seen from the signal values, with the exception of BT 1A and BT 5, all BTs are positive for the utilization of this carbon source. Looking at the absorbance plot over time (Figure 6.12), one can see some differences in the kinetic plots.

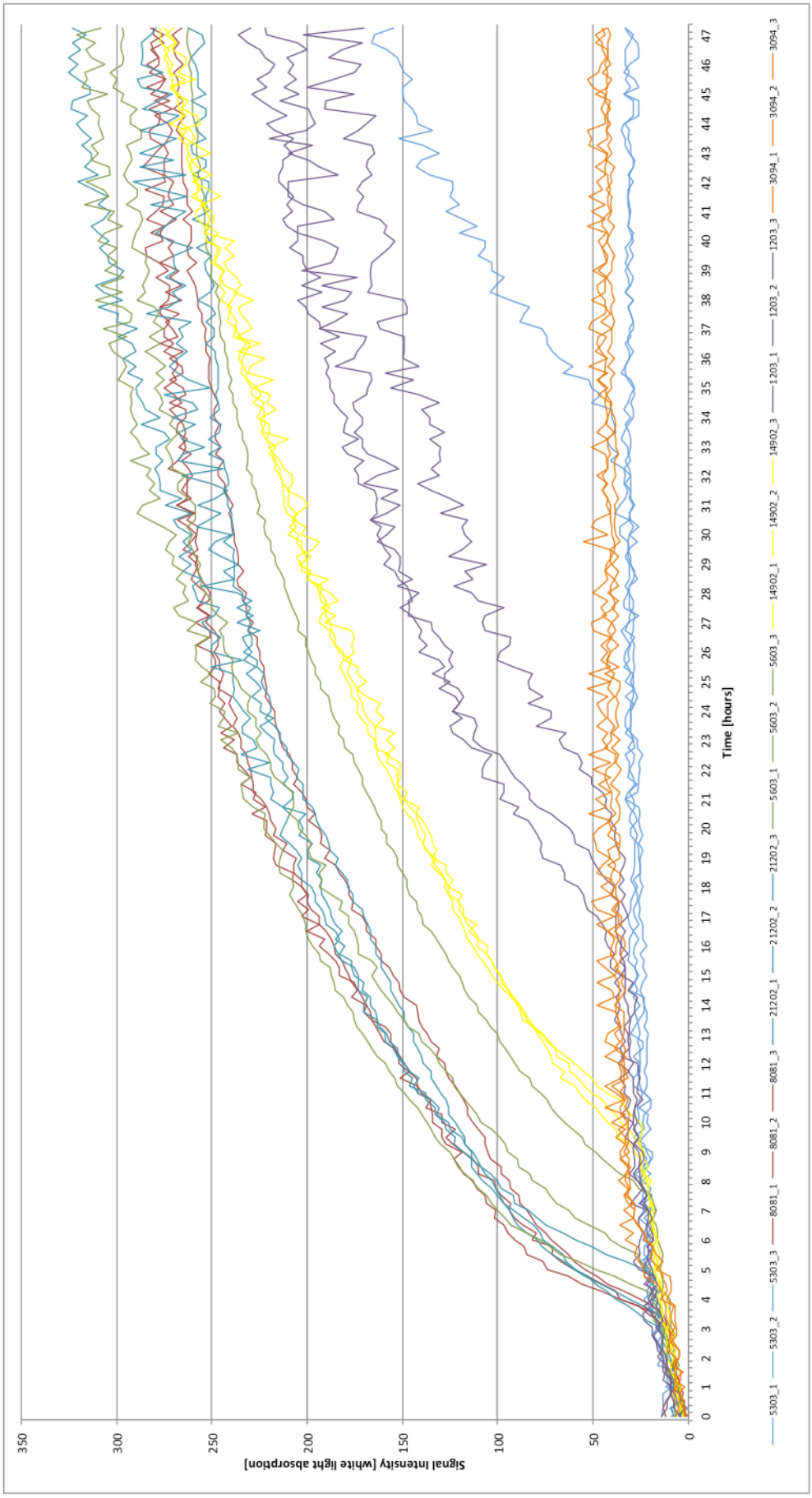


Figure 6.12: Absorbance curves over time for m-inositol utilization. Triplicates are shown. BT 1A / 5303 – blue; BT 1B / 8081 – red; BT 2 / 21202 – light blue; BT 3 / 5603 – green; BT 3 / 14902 – yellow; BT 4 / 1203 – purple; BT 5 / 3094 – orange.

The strains 8081, 21202, and 5603, representing biotypes 1B and 2/3 O:9, start to show activity after about 4 hours. BT 3 O:5,27 strain 14902 starts 6 hours later, 10 hours into the incubation time. BT 4 O:3 strain 1203 shows signs of metabolic activity around 18 hours after the start of the incubation. BT 5 strain 3094 shows no activity apart from background noise, but one of the replicates for BT 1A strain 5303 shows activity after 35 hours.

Looking into the genomic data, no basis for the phenotype in BT 5 could be established. In BT 1A, the myo-inositol dehydrogenase gene *idh* shows a point mutation (Figure 6.13).

Figure 6.13: Point mutation in the *idh* gene is the basis for BT 1A negative m-inositol phenotype. Utilization (blue +, red −), significant change compared to BT 1B (yellow +, black none, blue −). Mutation in BT 5 unknown.

The sequence TGG encoding for tryptophan has changed to TGA. The translation into UGA results in a premature stop codon, inhibiting the expression of a fully functional protein. Therefore the utilization of m-inositol is not possible in this strain. The UGA stop codon is peculiar though, as it has been shown to encode for the unusual amino acid selenocysteine occasionally. *Y. enterocolitica* does have a

selenocysteine tRNA, so it could be that the stop codon is either read through or decoded by selenocysteine, giving a fully functional protein, albeit in lower concentrations than normal. This would then result in a delayed response to utilization of the carbon source, explaining the metabolic activity observed in one of the BT 1A strain 5303 replicates. In a larger strain collection, this point mutation was shown to be specific for BT 1A strain 5303, and not representative for BT 1A strains in general. It also highlights the problem of strain-to-strain variation in biotyping following random inactivating mutations that lead to misidentification of isolates and potential misrepresentation of their pathogenic potential.

6.2.3 Metabolic activity of *Y. enterocolitica* biotypes at human body temperature

The phenotypic microarray was mainly evaluated at 28°C. This represents the optimum growth temperature of *Y. enterocolitica*. The metabolic profile for 37°C is given in Figure 6.14. One can see a dramatic overall reduction in metabolism in all BTs. BT 5 shows hardly any metabolic activity at all, having lost most of the few carbon sources it was utilizing at 28°C. In the other BTs, a specific temperature effect can be observed in nitrogen and phosphorus sources. A marked reduction of sulphur metabolism can be observed in BT 4. Generally, the non-pathogenic BT 1A still shows the highest metabolic flexibility as it shows more metabolically active nutrient sources.

This severe change in metabolism with temperature shift was unexpected as *Y. enterocolitica* causes human infections and thus thrives at 37°C. It could be that oxygen levels or the presence of additional nutrients play a role in this temperature effect. The basis is most likely less genetic in terms of presence or absence of operons, but likely to be of regulatory nature.

The effect of temperature change from 28°C to 37°C would warrant further in depth experiments such as a transcriptional response following temperature shift.

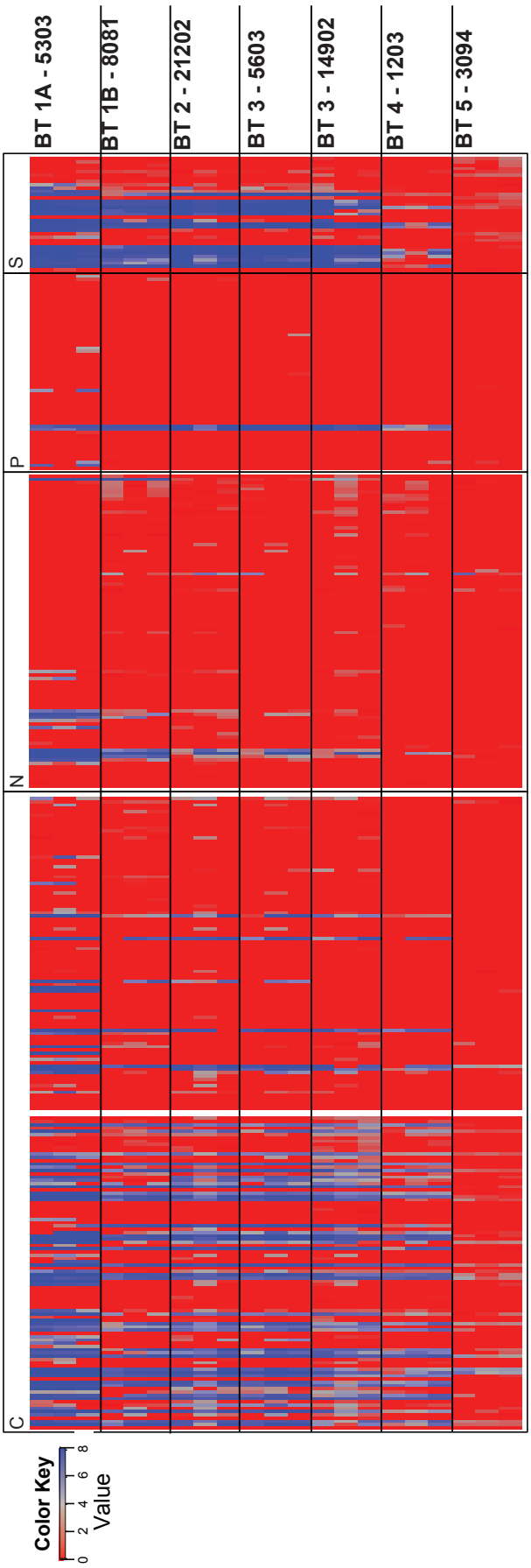


Figure 6.14: Signal values depicted as heatmap. Triplicates are shown for the signal values. C – carbon sources (190), N – nitrogen sources (95), P – phosphorus sources (59), S – sulphur sources (35).

6.3 Discussion

6.3.1 Metabolic profiling of *Y. enterocolitica* confirms genomic analysis

The phenotypic microarray mirrors both the observations made in the genomic analysis in terms of increased genome decay and loss of metabolic functions in BTs 4 and 5, as well as the biotyping scheme in that a reduction of metabolism can be observed (Figure 6.15).

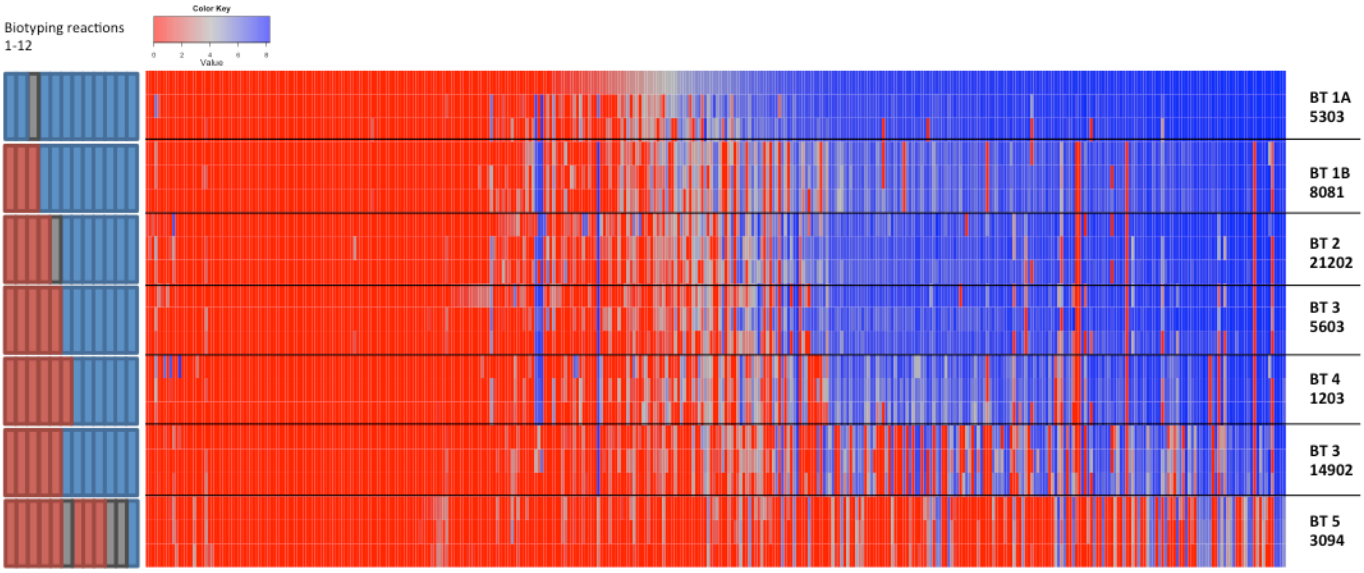


Figure 6.15: Comparison of biotyping reactions and phenotypic microarray. The 12 biotyping reactions correspond to the order given in Table 1.2. The phenotypic microarray data has been rearranged for the first replicate of BT 1A from negative to positive reactions irrespective of plate and well position. The other replicates are linked to the BT 1A ordering so that differential utilization of nutrients is reflected.

The non-pathogenic, environmental BT 1A shows the ability to use a wider repertoire of nutrient sources that it can utilize. Of the metabolic capacities it shares with other BTs, it can metabolize these better or more efficiently resulting in higher signal intensity. The high-pathogenic BT 1B and the low-pathogenic BTs 2 O:9, 3 O:9, and 4 O:3 show some reduction in metabolism compared to BT 1A. A marked reduction in metabolic flexibility can be observed in two strains. BT 3 O:5,27 is deficient specifically in nitrogen and phosphorus sources. This strain type is of interest in that it is predominantly associated with asymptomatic carriage in livestock, and is less frequent in human diarrheal samples (McNally, et al., 2004). This isolate specifically was isolated in cattle and represents the UK O:5,27 isolates. The reduction in metabolism therefore could be a result of adaptation to a cattle host. The metabolism of BT 5 is greatly reduced, very little metabolic activity can be observed in carbon

sources and no utilization of nitrogen, phosphorus, or sulphur sources is detectable. It might thus be no coincidence that BT5 is the rarest of all biotypes. It is referred to as the “hare-biotype”, and has not been isolated from humans (Swaminathan, Harmon and Mehlman, 1982; Wuthe and Aleksic, 1997; Bartling, et al., 2004). It is possible that BT5 has adapted to a very narrow niche that makes survival outside this niche near impossible. It may even present a dying lineage. This resembles the extreme loss of function and degenerate genome of another pathogenic *Yersinia*: *Y. pestis*. It was shown that *Y. pseudotuberculosis* and *Y. enterocolitica* can both grow on minimal medium at 28°C with few extra nutrients, but that *Y. pestis* needs further additives for adequate growth (Burrows and Gillett, 1966; Perry and Fetherston, 1997; Navid and Almaas, 2009). This is due to the biphasic lifecycle of *Y. pestis*, being present either in a warm-blooded host or in the flea. *Y. pestis* is not subjected to environmental metabolic challenges, and at the same time the bacterium depends on the host to provide for nutrients (Navid and Almaas, 2009). Both situations lead to accumulation of mutations in metabolic pathways. The genus *Yersinia* thus presents two examples of extreme niche adaptation to warm-blooded animals – *Y. pestis* and *Y. enterocolitica* BT5 – that have evolved independently. It is possible that BT 5 would have needed additives of amino acids or other nutrients to improve metabolic activity. This could potentially be uncovered by incubation on PM 5, which tests for common biosynthetic pathways via the utilization of nutrient supplements. An addition of a nutrient supplement though would have influenced the kinetic plots making a direct comparison to the other BTs impossible.

In contrast to the biotyping scheme, the phenotypic micorarray also highlighted some nitrogen, phosphorus, and sulphur sources that are absent from BT 1A. These could potentially be used for differentiation resulting in an improved biotyping scheme or maybe the development of new media. Unusual sources might present a possible advantage to also distinguish *Y. enterocolitica* from other enterobacteriaceae. These sources would need to be screened in a wider BT selection.

Metabolic properties thus play an important role in niche adaptation for *Y. enterocolitica*. BT1A is the metabolically most diverse biotype reflecting its ubiquitous nature and wide range of niches. High- and low-pathogenic biotypes have lost some metabolic flexibility. Especially in the predominant animal isolate BT3 O:5,27 and the rare BT5 there are signs of extreme niche adaptation, reflected in a significant loss of metabolic activities.

6.3.2 Linking genomic and phenotypic differences

The phenotypic microarray has been originally designed to study isogenic mutants, that is the wild-type strain against a single-gene mutant (Bochner, Gadzinski and Panomitros, 2001; Bochner, 2003; Zhou, Lei, Bochner and Wanner, 2003; Johnson, et al., 2008). This way very clean phenotypes could be observed. For instance, a wild-type strain of *E. coli* was compared to a *malF* mutant, and it could be shown that the metabolism of maltose, maltotriose, and dextrin is coupled to the functionality of that gene (Bochner, 2003). Here, several strains are compared at the same time, with differences in gene content and pseudogene distribution. BT 5 for example differs in more than one pseudogene compared to its nearest neighbour BT 4, therfore it is difficult to assign phenotypes to a single gene or even operons. It is possible for alternative secondary transporters or enzymes to complement mutations, obscuring possible links. Often, genes thought to be involved in certain pathways are either not mutated in any BT although differential expression is detected or mutations are found in the "wrong" BT, i.e. a BT utilizing the source, whereas no mutation is observed in a phenotype-negative BT. In addition to the principle of "One transporter – one substrate", one needs to consider "One transporter – several substrates" as well as "Several transporters – one substrate". Only very clear-cut phenotypes such as the maintenance of the D-arabitol operon in BT 1A can be observed and annotated unambiguously. For most cases of BT 1A-unique regions, there are indications of encoding metabolic pathways, but either no high-homologous hits in other bacteria to justify a clearer enzyme assignment or no good reference to base the annotation on. In the phenotypic microarray, there are new indications of possible pathways encoded, but no indication if several pathways could be activated by one and the same operon.

A study used the phenotypic microarray to detect changes in metabolism caused by mutations in 2-component systems (Zhou, Lei, Bochner and Wanner, 2003). 2-component systems consist of a sensor kinase and a response regulator. The sensor protein is a membrane-associated protein which detects environmental signals and passes them on to the response protein which is either a transcription factor or acts on sigma factors to modulate gene expression. Zhou, Lei, Bochner and Wanner (2003) confirmed links between NtrBC to nitrogen metabolism and PhoRB to phosphorus metabolism. No mutations in these or other 2-component systems implicated with phenotypic changes are mutated in any of the *Y. enterocolitica* biotypes except *creC* (*YE3383*) in BT 4. This component system is thought to be

involved in carbon/energy metabolism (Chaudhuri, et al., 2010), but a direct link cannot be established in the *Y. enterocolitica* dataset. BTs 4 and 5 show mutations in regulatory proteins, including potential 2-component systems, thus these could be the reason for the widespread metabolic defects visible especially in BT 5. The 2-component systems of *Y. enterocolitica* as well as of *Y. pestis* and *Y. pseudotuberculosis* have not been as extensively characterized as in *E. coli* and regulatory relationships are theoretical and deduced from DNA homology (Marceau, 2004). It is possible that either the *E. coli*-homologous systems in *Y. enterocolitica* fulfil a different role, or that one of the uncharacterized 2-component systems is in fact taking the place of one of the well-established *E. coli* 2-component systems.

The phenotypic microarray has highlighted general metabolic capacities of each strain, and has lowered the possibilities of nutrients that could be encoded for in the unique regions or the pseudogenes affected. In this way, the phenotypic microarray facilitates future annotation.

Variability of phenotypic tests has also been described in *Y. pestis*. Most notably is rhamnose fermentation. Classically, *Y. pestis* is negative for rhamnose fermentation; a positive phenotype is associated only with Biovar Pestoides (Achtman, et al., 2004). Rare rhamnose-positive mutants are observed in *Y. pestis* after prolonged incubation (Englesberg, 1957a,b). The rhamnose operon *rhaTRSBAD* is present in all *Y. pestis*, and BetB, a betaine-aldehyde dehydrogenase, and RhaS, a transcriptional regulator, have been suggested as the genetic basis for this phenotype (Navid and Almaas, 2009; Eppinger, et al., 2010). This ability of *Y. pestis* for phenotypic gain of function is known as meiotrophy, and has also been documented for ammonia assimilation, urease, production of *inv*, fermentation of melibiose, and biosynthesis of certain amino acids (Brubaker, 1991). The respective gene regions are presumably present and inactivated by reversible mutations or highly regulated (Brubaker, 1991). This variability also makes assignments of phenotype to genotype difficult.

Achtman, et al. (2004) demonstrated that biovar designation does not reflect the genetic relationship in *Y. pestis*. The distinction of *Y. pestis* into different biovars is based on fermentation patterns of glycerol, rhamnose, melibiose, and nitrate reduction and characterizes the three classic biovars Orientalis, Medievalis, and Antiqua, as well as atypical biovars Pestoides and Microtus. Rhamnose, as described above, is known to be variable and strains can show a positive phenotype

upon prolongued incubation. The phylogenetic relationship in *Y. pestis* also showed that the inability to metabolize sugars or reduce nitrate can arise independently in several lineages and can thus have a different genetic basis (Achtman, et al., 2004). The same biovar might therefore be allocated to very different lineages. The grouping in *Y. pestis* was revised on the basis of evolutionary relationships rather than metabolic traits, which do not have the correlation with evolutionary origin (Achtman, et al., 2004). It is even suggested that biovars and metabolic traits are unsuitable for distinction into taxonomic groups (Achtman, et al., 2004). Given the metabolic variety, variable biotyping reactions, and inactivating mutations as found in BT 1A strain 5303 in the utilization of *m*-inositol, the grouping of *Y. enterocolitica* might need to be re-evaluated likewise, to move away from phenotypic variability.

6.3.3 The influence of temperature on metabolism in *Y. enterocolitica*

It is known that *Y. enterocolitica* displays differential expression of genes upon changes in temperature. Certain genes thought to be involved in environmental survival as well as insect pathogenicity such as genes in the *tc* PAI and motility and chemotaxis genes as well as genes potentially involved in nitrogen utilization could be shown to be upregulated in low-temperature conditions of 10°C (Bresolin, Neuhaus, Scherer and Fuchs, 2006b). Genes such as *inv* and *ail* involved in pathogenicity in a human host are upregulated when the bacteria are exposed to 37°C corresponding to human body temperature (Straley and Perry, 1995). Most famous is the low-calcium response coupled with temperature increase that triggers the expression of the Yop T3SS on the virulence plasmid (Straley and Perry, 1995; Bochner, 2009; Brubaker, 1991). The optimum temperature of *Yersinia* though is 25 - 28°C (Cornelis, et al., 1987).

In the first set of experiments conducted at optimum growth temperature *Y. enterocolitica* showed a wide range of metabolic activities that were dependent on the bio-serotype. Repeating these experiments at 37°C, one could have expected a possible upregulation of nutrient utilization in pathogenic biotypes as a response to conditions similar to those found in the human body. The results however showed a severe reduction in metabolic activity specifically in nitrogen and phosphorus sources. BT 4 also had a reduction of sulphur metabolism. This might have several explanations. First of all, it could be that the human host provides specific trace elements or other nutrients that are essential for growth under high temperatures. As these are absent under the conditions tested (only L-cystine was added as a reduced

sulphur source at both temperatures) this could explain the absence of metabolic activity. Secondly, it could also be that certain metabolic pathways are restricted for use under anaerobic conditions. The 96-well plates do not allow for much oxygen circulation, but can still be considered rather aerobic. For the anaerobe *Streptococcus pneumoniae* it could be shown that sealing of the 96-well plates in bags under anaerobic conditions and subsequent incubation had a great influence on experimental outcome (T. Feltwell, personal communication).

For all *Yersinia spp.* additional nutrient requirements were demonstrated upon temperature shift (Burrows and Gillett, 1966). Defined minimal medium complemented with sodium metabisulphite and amino acids isoleucine, valine, cystine, methionine, and phenylalanine was enough for growth of *Y. pestis* at 28°C, but had to be further complemented with glutaminc acid and thiamine and replacing cystine with thiosulphate. *Y. pseudotuberculosis* and *Y. enterocolitica* also had additional requirements at 37°C, although to a lower degree than *Y. pestis* (Burrows and Gillett, 1966; Brubaker, 1991). It is entirely possible that addition of other nutrients at 37°C could increase the metabolic activity and possibly show the activation of sources specific to pathogenic BTs. Unpublished results indicate that calcium might play an important role in adaptation to higher temperatures (Bochner 2009; Holtz-Morris, 2005; Fodor, et al., 2005), but unfortunately a direct comparison with the data obtained for *Y. enterocolitica* is not possible. Other signals could include magnesium concentration, pH, or acidic conditions (Bochner, 2009). One microarray study showed that following temperature shift specific metabolic pathways are differentially regulated (Motin, et al., 2004). They could demonstrate that a full citric acid cycle is present at 37°C, whereas nitrogen and amino acid metabolism appear downregulated. These differences are not as clear in *Y. enterocolitica* and might be due to differences in the culture media used thus supplying nutrients that influence metabolic state. Not all metabolic pathways or all intermediates of those pathways are present on the phenotypic microarray either, making a full assessment of metabolic state difficult. A further experiment could be carried out, testing growth at 37°C first on PM 5 for nutrient supplements. Depending on the outcome, essential nutrients could be added to the inoculation medium and the experiment repeated at 37°C. Alternatively, the experiment could be repeated with the present medium, but under anaerobic conditions. Either way, only one of the parameters should be altered at any point to allow for comparison and determine which trigger is decisive.

It might not only be the acquisition of beneficial factors enabling survival / potential pathogenesis in the human body that is important, it might also be the loss of metabolic flexibility that drives the colonization of new niches and associated putative virulent behaviour. To fully assess all aspects of *Y. enterocolitica* biology, one further interesting experiment to carry out would be to repeat the phenotypic microarray at 10°C. This would complement the current data set of optimum temperature and human body temperature with temperature found under environmental conditions and conditions associated with potential insect hosts.

Generally, the phenotypic microarray results reflected the regions of difference and pseudogenization noticed in the reference genomes. Comparing the results with traditional biotyping, there were only two discrepancies found, one of which could be related back to a unique mutation in the reference strain chosen for BT 1A. Prediction of metabolic operons from the microarray data is restricted by annotation, i.e. some annotation is potentially too general to relate an operon to a specific metabolic function. The other way round, annotation is limited by the lack of clear phenotypic results that can be unambigiously linked to metabolic pathways. The phenotypic microarray data can be used in future for directed experiments to link phenotype and genotype, and might help identification of new pathways such as the utilization of N-acetyl-D-galactosamine.

7 Phylogeny of the Genus *Yersinia* and the species *Y. enterocolitica*

7.1 Introduction

From the data presented thus far it appears that the driving forces of pathogen evolution are the same in *Y. enterocolitica*, *Y. pseudotuberculosis* and *Y. pestis*. It is thought that the pathogenic *Yersiniae* evolved out of a common ancestor that acquired the virulence plasmid pYV, enabling the coloniziation of new niches such as warm-blooded animals including humans (Carniel, 2002; Wren, 2003). *Y. enterocolitica* and *Y. pseudotuberculosis* then evolved separately, and *Y. pseudotuberculosis* gave rise to the *Y. pestis* biovars. The evolution of *Y. pestis* has been studied in detail (Achtman, et al., 1999; Achtman, et al., 2004; Morelli, et al., 2010) and is characterized by the expansion of IS elements leading to wide-ranged genome rearrangements and subsequent gene loss (Parkhill, et al., 2001a).
Unlike the *Y. pseudotuberculosis/pestis* cluster the evolutionary dynamics of *Y. enterocolitica* are understudied and knowledge is currently limited to microarray and amplified fragment length polymorphism studies (Fearnley, et al., 2005; Howard, et al., 2006). As seen in the previous chapters, *Y. enterocolitica* presents as a heterogeneous groups of organisms with very clear distinctions in gene content. It is necessary to establish the evolution of the *Y. enterocolitica* lineages to obtain a trajectory for acquisition and loss events of genomic islands, and clarify the presence of remnants of genes and islands and their potential influence on the evolution of pathogenesis in *Y. enterocolitica*.

This chapter aims to elucidate the evolutionary relationships within *Y. enterocolitica* as well as with respect to the whole genus *Yersinia* and to investigate the inter-relationships between the different lineages. This should help to identify any shared ancestry or patterns of evolution for the human pathogenic lineages of this genus.
Sequencing more than one genome for each lineage will also clarify the status of reference sequences amongst a larger strain collection. The assumptions made for the reference genomes will be integrated with their respective lineage and confirm their role as representatives of the respective BT.

7.2 Results

7.2.1 Phylogeny of the species *Y. enterocolitica*

In addition to the seven reference genomes a further 89 *Y. enterocolitica* isolates were sequenced (Table 9.1, p. 204). These were contributed by collaborators from Europe and New Zealand and sequenced using Illumina technology. The strains represent human case isolates as well as isolates from humans without disease association, animal isolates, and environmental isolates. Twenty-four BT 1A, seven BT 1B, thirteen BT 2, twenty-two BT 3, twenty-one BT 4, and two BT 5 isolates were included. Only two BT 5 isolates were sequenced, as this is a rarely isolated biotype.

The phylogeny of the *Y. enterocolitica* species was obtained from genome-wide single nucleotide polymorphism (SNP) distribution by mapping against the BT1B reference strain 8081. The four prophage regions were excluded from the analysis as they would result in false-positive SNPs from distantly related phages present in other BTs. This work was carried out in close collaboration with Dr Simon R. Harris and Dr Thomas R. Connor at the Wellcome Trust Sanger Institute.

It is evident from Figure 7.1 that there are three different *Y. enterocolitica* lineages corresponding to non-, high-, and low-pathogenic BTs. The different lineages show considerable differences in the SNP variation across the tree. This is also evident from the number of SNPs in the different lineages (Table 7.1). Individual biotype 1A strains are differentiated by up to 32,000 SNP sites. Since most of these isolates came from passive sampling from abattoirs and some clinical studies, this diversity is likely to be a significant underestimate of the genuine diversity of this clade as true environmental isolates such as strains isolated from water are currently underrepresented. Within this non-pathogenic clade there are five non-BT 1A strains. Two of these strains were typed as BT 1B and were donated to the study from collaborators in New Zealand. Both strains were isolated from humans and had no serological information.

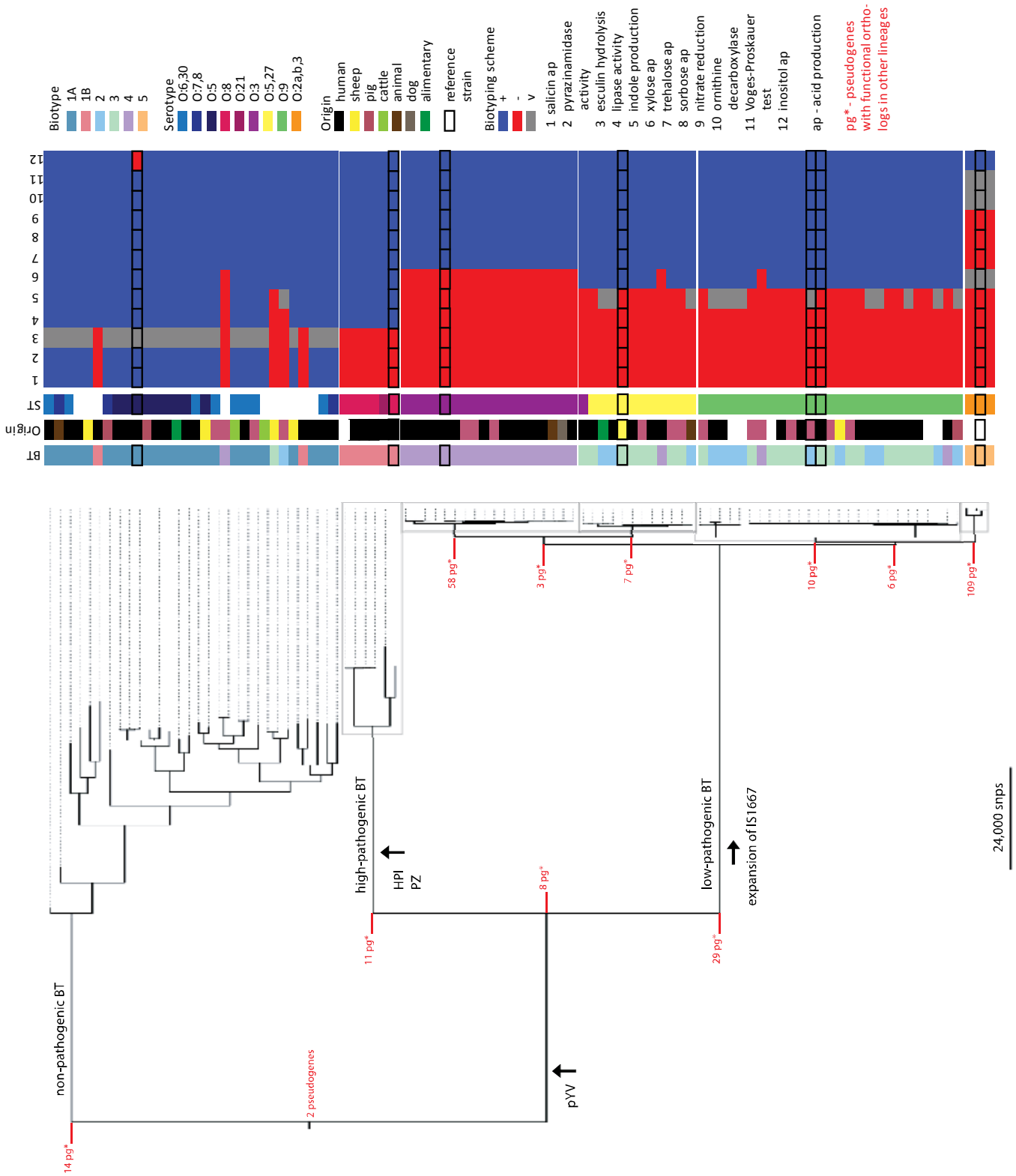


Figure 7.1: The maximum likelihood phylogenetic tree of *Y. enterocolitica* based on the SNP differences across the whole genome excluding laterally acquired phages. The tree highlights the key acquisition events of the virulence plasmid pYV in the pathogenic lineages of Y. enterocolitica, and the acquisition of markers of high-pathogenic BT. The low-pathogenic BTs are marked by an expansion of the IS element IS1667. The number of pseudogenes that are shared within specific lineages and that have functional orthologs in other lineages is given (pg*). The different lineages are characterized by biotype (BT), biological origin, and serotype (ST). The biochemical tests used for biotyping of Y. enterocolitica are presented for each strain.

Table 7.1: Numbers of SNPs between different lineages of *Y. enterocolitica*.

	BT 1A	BT 1B	BT 2/3 O:9	BT 2/3 O:5,27	BT 4	BT 5
BT 1A	0 – 31,690					
BT 1B	87,039 – 117,383	11 – 29,511				
BT 2/3 O:9	79,461 – 115,724	63,727 – 92,073	0 – 5,582			
BT 2/3 O:5,27	81,599 – 115,965	64,399 – 92,699	1,607 – 7,265	1 – 250		
BT 4	80,828 – 115,651	63,808 – 91,888	4,075 – 8,202	4,637 – 5,685	8 – 882	
BT 5	84,994 – 114,528	72,361 – 89,850	5,406 – 9,691	7,503 – 10,734	8,211 – 11,461	82 – 2,883

Blue – range of SNPs within a BT/ST lineage; green – range of SNPs between BT 1A and other lineages; red – range of SNPs between 1B and other lineages; yellow – range of SNPs within the low-pathogenic BTs.

It has been noted before that BT 1A strains can be mistyped as BT 1B strains, but they commonly lack the virulence factors associated with the high-pathogenic BT (personal communication A. McNally, M. Prentice; Fearnley, et al., 2005). Analysis of these two strains showed that they do not carry the virulence plasmid, and do not have the Ysa T3SS, or the HPI. On the contrary, these two strains have operons for the utilization of D-arabitol and N-acetyl-D-galactosamine, and they also possess the YGT.

Three other isolates also do not group according to their biotyping reaction but cluster with BT 1A. These are a BT 3 O:5,27 isolate, one BT 4 O:3 strain, and a BT 2 strain. The BT 2 strain, YE228/02, is known to give variable biotyping reactions, and which consequently might explain the discrepancy between typing and clustering of this particular strain (personal communication A. McNally, T. Cheasty). Sequences of the other isolates were mapped to the respective serotype operon but did not yield any homologous hits, suggesting difficulties in typing these strains.

Although there is a reduced number of true 1B isolates, like BT 1A, they also show significant genetic divergence with up to 30,000 SNPs defining this lineage. In contrast it is evident from Figure 7.1 that despite the large number of samples, the average divergence between BT 2 – 5 isolates is much lower than the BT 1A and BT 1B (~12,000 SNPs), resulting in very short terminal branch lengths within this clade. As the included strains represent a wide temporally and geographically dispersed collection the possibility of this feature being due to a sampling artefact is unlikely.

Another explanation is that the low genetic diversity is a signature consistent with a relatively recent population bottleneck.

While it is apparent from Figure 7.1 that there is no obvious clustering of the isolates according to human or animal origin, the phylogeny of the pathogenic *Y. enterocolitica* lineages is fully congruent with serotype and partially with biotype. Although there is some variation in the results of these biochemical tests they appear largely phylogenetically robust: BT 1A isolates show a much higher metabolic potential whilst, even from these limited biochemical tests, it is clear that metabolic flexibility is sequentially reduced when moving from the high pathogenic 1B lineages across the low pathogenic lineages (Figure 7.1). This trend of reduced metabolic flexibility was observed on a much larger scale in the phenotypic microarray (Chapter 6) and is consistent with an increased number of pseudogenes in the low-pathogenic BTs and an expansion of IS1667. The observed variation in the biochemical tests confusing the assignments of BT2, BT3 and BT4 isolates is largely explained by differences in the ability to catabolise D-xylose and the production of indole. The latter test is known to show a high degree of methodological variation. In particular BTs 2 and 3 are distinguishable only by a variable indole reaction, and Chapter 4 illustrated that the two O:9 isolates are identical apart from three transposases and two putative phage proteins. Both BTs even present the same pseudogene mutations with very few exceptions. Phylogenetically BTs 2 and 3 intermingle and isolates are only clearly separated by serotype.

For the reference strains, the number of pseudogenes in each branch with functional orthologs in other lineages has been highlighted on the tree. The pseudogenes defining a lineage are listed in Table 7.2.

Table 7.2: Pseudogenes in the core genome of *Y. enterocolitica* lineages.

ID	1B 8081	1A 5303	2 O:9 21202	3 O:9 5603	3 O:5,27 14902	4 O:3 1203	5 3094	product	mutation
YE0416	Δ	Δ	Δ	Δ	Δ	Δ	Δ	conserved hypothetical protein	stop codon
YE2410	Δ	Δ	Δ	Δ	Δ	Δ	Δ	putative oxidoreductase	FS deletion
YE0207	+	Δ	+	+	+	+	+	putative lysophospholipase	stop codon
YE0480	+	Δ	+	+	+	+	+	putative exported protein	stop codon
YE0688	+	Δ	+	+	+	+	+	putative type II secretion system protein	FS deletion
YE0972	+	Δ	+	+	+	+	+	conserved hypothetical protein	stop codon
YE1432	+	Δ	+	+	+	+	+	hypothetical protein	FS deletion
YE1733	+	Δ	+	+	+	+	+	putative UV protection and mutation protein	stop codon
YE1743	+	Δ	+	+	+	+	+	putative AraC-family transcriptional regulatory protein	stop codon
YE1759	+	Δ	+	+	+	+	+	conserved hypothetical protein	disrupted by IS element
YE2179	+	Δ	+	+	+	+	+	hypothetical protein	stop codon
YE2521	+	Δ	+	+	+	+	+	flagellin	deletion
YE3354	+	Δ	+	+	+	+	+	putative oxidoreductase	disrupted by IS element
YE4026	+	Δ	+	+	+	+	+	myo-inositol 2-dehydrogenase	opal stop codon
YE4074	+	Δ	+	+	+	+	+	cellulose synthase 1 catalytic subunit	opal stop codon
YE4185	+	Δ	+	+	+	+	+	Probable two-component sensor kinase	stop codon
YE0522	Δ	+	Δ	Δ	Δ	Δ	Δ	putative exported protein	deletion N terminus
YE1236A	Δ	+	Δ	Δ	Δ	Δ	Δ	6-phospho-beta-glucosidase	deletion
YE1400	Δ	+	Δ	Δ	Δ	Δ	Δ	Putative type I secretion protein	FS mutation
YE2586A	Δ	+	Δ	Δ	Δ	Δ	Δ	putative coA transferase family protein	FS mutation
YE2708	Δ	+	Δ	Δ	Δ	Δ	Δ	α-ribazole-5'-phosphate phosphatase	deletion
YE2719	Δ	+	Δ	Δ	Δ	Δ	Δ	conserved hypothetical protein	deletion
YE3341	Δ	+	Δ	Δ	Δ	Δ	Δ	prepilin peptidase	FS insertion
YE0270	Δ	+	+	+	+	+	+	conserved hypothetical protein	stop codon
YE0936	Δ	+	+	+	+	+	+	putative MFS family transport protein	FS deletion
YE0944	Δ	+	+	+	+	+	+	hypothetical protein	stop codon, FS mutation
YE1406	Δ	+	+	+	+	+	+	putative protease	FS mutation
YE1488	Δ	+	+	+	+	+	+	putative inner membrane protein	FS mutation
YE1598	Δ	+	+	+	+	+	+	nitrite reductase gene remnant	deletion
YE1728	Δ	+	+	+	+	+	+	putative sugar permease	FS deletion
YE1729	Δ	+	+	+	+	+	+	putative sucrose operon repressor	FS deletion
YE2450	Δ	+	+	+	+	+	+	putative D-tagatose-1,6-bisphosphate aldolase subunit	deletion C terminus

YE2640 *	Δ	+	+	+	+	+	+	putative membrane protein	opal stop codon
YE3168	Δ	+	+	+	+	+	+	conserved hypothetical protein	disrupted by IS element
YE3180	Δ	+	+	+	+	+	+	phosphate binding protein	stop codon
YE0324	+	+	Δ	Δ	Δ	Δ	Δ	putative ATP/GTP-binding protein	deletion C terminus
YE0443	+	+	Δ	Δ	Δ	Δ	Δ	multidrug efflux protein	deletion C terminus
YE0445	+	+	Δ	Δ	Δ	Δ	Δ	probable outer membrane efflux lipoprotein	deletion N terminus
YE0483	+	+	Δ	Δ	Δ	Δ	Δ	putative LysR-family transcriptional regulatory protein	deletion
YE0807	+	+	Δ	Δ	Δ	Δ	Δ	putative transcriptional regulatory protein	deletion
YE0911	+	+	Δ	Δ	Δ	Δ	Δ	3-oxoacyl-[ACP] synthase II	deletion N terminus
YE0934	+	+	Δ	Δ	Δ	Δ	Δ	hypothetical protein	deletion N terminus
YE0935	+	+	Δ	Δ	Δ	Δ	Δ	AraC-family transcriptional regulator	deletion N terminus
YE0941	+	+	Δ	Δ	Δ	Δ	Δ	AraC-family transcriptional regulator	FS deletion
YE1005	+	+	Δ	Δ	Δ	Δ	Δ	phospholipase A	FS insertions
Ye1165	+	+	Δ	Δ	Δ	Δ	Δ	putative LyxR-family transcriptional regulatory protein	deletion
YE1166	+	+	Δ	Δ	Δ	Δ	Δ	putative LysR-family transcriptional regulatory protein	FS mutation, deletion C terminus
YE1251	+	+	Δ	Δ	Δ	Δ	Δ	putative regulatory protein	deletion N terminus
YE1252	+	+	Δ	Δ	Δ	Δ	Δ	putative exported protein	deletion
YE1409	+	+	Δ	Δ	Δ	Δ	Δ	putative prophage repressor protein	stop codon
YE1610	+	+	FS deletion	FS deletion	FS deletions	FS deletions	stop codon	putative MFS-family transport protein	independent mutations
YE1874	+	+	Δ	Δ	Δ	Δ	Δ	putative DNA-binding protein	deletion
YE1905	+	+	Δ	Δ	Δ	Δ	Δ	hypothetical protein	stop codon
YE1944	+	+	FS insertion	FS insertion	FS insertion, stop codon	FS insertion, stop codon	stop codon	putative membrane permease	independent mutations
YE1950	+	+	Δ	Δ	Δ	Δ	Δ	alkanesulfonate monooxygenase	deletion
YE2286	+	+	Δ	Δ	Δ	Δ	Δ	acyl-[ACP]--UDP-N-acetylglucosamine o-acyltransferase	deletion
YE2603	+	+	Δ	Δ	Δ	Δ	Δ	putative sugar kinase	deletion C terminus
YE2609	+	+	Δ	Δ	Δ	Δ	Δ	ABC transporter, substrate binding component	deletion
YE2812	+	+	Δ	Δ	Δ	Δ	Δ	putative formate transporter	deletion
YE3630	+	+	Δ	Δ	Δ	Δ	Δ	putative cation efflux	deletion

								system protein	
YE3694	+	+	Δ	Δ	Δ	Δ	Δ	Putative transcriptional regulator	deletion, FS mutation
YE3735	+	+	Δ	Δ	Δ	Δ	Δ	glutamate synthase [NADPH] large chain precursor	FS deletion
YE3989	+	+	Δ	Δ	Δ	Δ	Δ	conserved hypothetical protein	deletion N terminus
YE4052	+	+	Δ	Δ	Δ	Δ	Δ	metalloprotease	deletion
YE1410	+	+	+	+	Δ	Δ	+	probable phage anti-termination protein Q	deletion C terminus
YE1971	+	+	+	+	FS deletion	opal stop codon	+	probable membrane transport protein	independent mutations
YE0924	+	+	Δ	Δ	+	+	Δ	glycine betaine/L-proline transport system permease	deletion
YE1614	+	+	stop codon	stop codon	+	+	FS insertion	tetrathionate reduc-tase complex: sensory transduction histidine kinase	independent mutations
YE1746	+	+	Δ	Δ	+	+	Δ	hypothetical protein	disrupted by IS element
YE2036	+	+	disrupted by IS element	disrupted by IS element	+	+	deletion N terminus	conserved hypothetical protein	independent mutations
YE4106	+	+	FS insertion	FS insertion	+	+	deletion N terminus	Hypothetical ROK family protein	independent mutations

+ – functional ortholog; Δ – mutation; column mutation lists shared mutational inactivation, further mutations possible for individual isolates; independent mutations are listed within the table.

There are two pseudogenes which are shared in all BTs and which show the same mutational inactivation. The non-pathogenic BT 1A has acquired 14 pseudogenes. There is no predominant class among those pseudogenes, four are in hypothetical proteins and two are in regulatory proteins. All pathogenic lineages share seven pseudogenes in the core genome. These have been acquired in single events before the split of high- and low-pathogenic lineages because the same mutations are discernible. It has to be mentioned though that some of these genes have acquired further mutations following the split of the lineages. The high-pathogenic BT has 12 pseudogenes with functional orthologs in the other lineages, three of which are in hypothetical proteins, and four are involved in metabolism and transport.

A high accumulation of pseudogenes is apparent in the low-pathogenic BTs, again indicating a potential bottleneck evolution of this lineage. The low-pathogenic BTs have 29 pseudogenes of which only two have been acquired independently. These independent mutations target a transport protein and a membrane permease. The

shared pseudogenes are predominantly found in transport proteins (4 pseudogenes), proteins involved in metabolism (6 pseudogenes), and regulatory proteins (8 pseudogenes). This supports the reduced metabolic flexibility observed in the phenotypic micorarray for these biotypes. The branch leading to the O:3 and O:5,27 lineages has accumulated only two further pseudogenes, one in a shared phage, the other acquired independently in a membrane protein. Five pseudogenes are found in the branch leading to the O:9 and O:2a,2b,3 lineages. Of these, only two are shared. These are present in a transport permease and a hypothetical protein. The independent mutations are found in the histidine kinase controlling the tetrathionate respiration operon, and in hypothetical proteins. The two O:9 isolates show identical mutational inactivations, but these mutations were acquired independently in BT 5.
A significant expansion of core pseudogenes in the terminal branches of the low-pathogenic BTs is visible in BT 4 O:3 (58 pseudogenes) and BT 5 O:2a,2b,3 (109 pseudogenes). These two isolates can be considered the two most host-adapted lineages of *Y. enterocolitica*. BT 4 O:3 is the predominant isolate found in human cases, and BT 5 is responsible for disease in hares. BT 5 is very niche-restricted. After accumulation of a massive amount of pseudogenes, it appears to struggle to survive outside its host and is very rarely isolated in general. The increased amount of pseudogenes also again reflects the loss of metabolic flexibility detected in the phenotypic micorarray (Chapter 6).

As mentioned before (Chapter 4), some trends found within the pseudogenes also translate in a broader context. Cobalamin biosynthesis, tetrathionate respiration, and the flagella cluster are not only affected by mutations in *Y. enterocolitica* lineages, but are also targets of pseudogene formation in *Y. pestis*, *S.* Typhi, and *S.* Gallinarum (Parkhill, et al., 2001a,b; Thomson, et al., 2006, 2008). Fish pathogen *Y. ruckeri* has also been described as lacking cobalamin biosynthesis and tetrathionate respiration (Chen, et al., 2010). Evolution from environmental survival to host adaption has probably driven the loss of these properties.

It has been proposed to sub-speciate BT 1B into *Y. enterocolitica* subspecies *enterocolitica* and BT 1A, 2-5 into *Y. enterocolitica* subspecies *palearctica* (Neubauer, et al., 2000). This distinction of two subspecies is based on variations in 12 SNPs over the full length of the 16S rRNA gene (1637 bp). However, this differentiation is not supported by our data. Based on SNP distribution in all core

genes three distinct lineages can be observed. Very specifically, the proposed subspecies *Y. enterocolitica palearctica* is based on the worldwide distribution of those isolates but their very distinct genomic content argues against this. Obviously, the resolution power is much greater using the full genome information as opposed to a single, highly conserved gene.

Judging solely from our tree, it appears that the virulence plasmid pYV has been acquired in the pathogenic lineages of *Y. enterocolitica*. Signifying the split into high-pathogenic BTs is the uptake of the HPI and operons within the plasticity zone. Indicative for the low-pathogenic BTs is the expansion of IS1667. From this tree, the non-pathogenic BT seems to be the oldest lineage. The second flagella cluster Flag-2, which is only consistently present in the low-pathogenic BTs, has been lost independently in high- and low-pathogenic BTs. The shape of the tree dictates this, but different gene fragments of Flag-2 in those BTs are also suggestive of independent loss events. Further to this, there are some BT 1A isolates that have retained Flag-2, but there is no link to a specific biological origin.

7.2.2 Phylogeny of the genus *Yersinia*

To determine sequence based phylogeny of the *Yersinia* genus we examined a representative sample of pathogenic isolates composed of *Y. pseudotuberculosis* (33 in total; 29 sequenced in this study) and *Y. pestis* (using the published genomes of CO92/Parkhill, et al., 2001; KIM5/Deng, et al., 2002; Yp91001/Song, et al., 2004) as well as the relatively understudied lineage of *Y. enterocolitica* (98 isolates). To explore the large number of environmental *Yersinia* spp., 81 isolates from all species except the recently described *Y. entomophaga* (Hurst, et al., 2011) and *Y. nurmii* (Murros-Kontiainen, et al., 2010a) were included. These isolates are listed in Table 9.2 (p. 206) in the appendix. In total 199 *Yersinia* isolates were sequenced which had been conventionally typed using biochemical and serological techniques. Some isolates that had been assigned to species but showed atypical biochemical profiles were included as well.

As the genus *Yersinia* shows a very high sequence divergence, an approach based on mapping sequence reads to gain information on the whole genome could not be carried out. As seen in *Y. enterocolitica*, even within a single species of *Yersinia* > 250, 000 SNPs differentiate members of that species. Instead, the syntenic blocks identified in *Y. enterocolitica* (Chapter 4) were used to sample a set of 85

housekeeping genes common to all members of the *Yersiniae*. The housekeeping genes were identified on the basis of showing 10-25% SNP divergence between *Y. pestis* and *Y. enterocolitica*. This work was carried out in collaboration with Miquette Hall, NTU. The genes identified are given in Table 2.3 and their locations with respect to the BT 1B reference genome for strain 8081 is shown in Figure 7.2

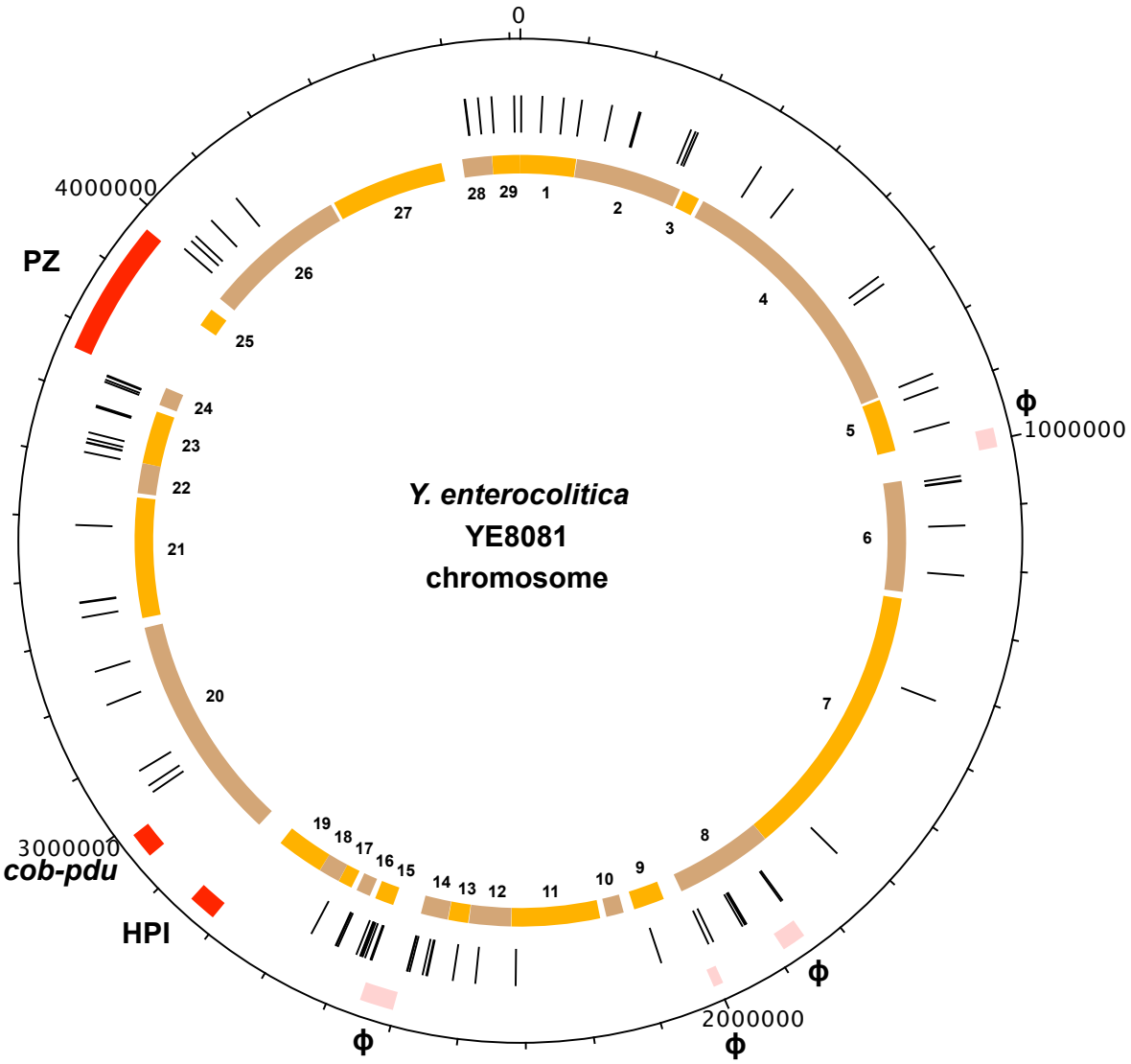


Figure 7.2: Distribution of the 85 housekeeping genes used for construction of the phylogenetic relationship within the genus *Yersinia*. The first circle shows regions naturally excluding the presence of housekeeping genes, such as phage, and pathogenicity islands. In the second circle, each bar represents a single housekeeping gene. The third circle highlights the 29 syntenic blocks in alternating colours of orange and brown (see Table 4.3).

There are some syntenic blocks that are not represented with genes (#3, 10, 18, 19, 22, 25, 27). This can be due to several reasons. First of all, the blocks were identified purely on the basis of being a stretch of DNA sequence that did not show any

recombination in the different *Y. enterocolitica* BT. Some blocks, like #3 and #10, are small in size and only contain ~30 CDSs. This restricts the number of possible genes, and the genes contained in these blocks might not be housekeeping genes to start with, or do not comply with the required divergence levels. The absence of housekeeping genes in larger regions such as #27 could be explained by the fact that housekeeping genes were identified, but they could not be obtained from sequence reads for *Y. ruckeri*. This species is arguably part of the genus Yersinia (Sulakvelidze, 2000) and notably very distantly related to the other members of the genus. There were occasions when housekeeping genes were easily identified in all species except *Y. ruckeri* and thus had to be excluded from an initial set of potential genes.
Region #25 corresponds to a conserved region within the plasticity zone, but does not contain any housekeeping genes.

Bayesian Analysis of Population Structure (BAPS) was used to divide the population into groups of shared sequence similarity. Phylogenetic analyses are usually based on a concatenated sequence. In contrast, BAPS compares individual genes and thus can look at a broader level of relationships, including potential recombination events and linkage between genes. This BAPS clustering resolved the genus into 12 distinct species clusters (SC; Figure 7.3). Definition of *Yersinia* spp. has been based on biochemical differences, which have been overlayed onto the tree. It is obvious that some *Yersinia* spp. show a similar biochemical profile, yet are distantly related. This is the case for *Y. frederiksenii*. The type strain is part of SC 7, but a distantly related group can be found in SC 2, together with isolates designated as *Y. massiliensis*. This phenomenon has been noted before (Sulakvelidze, 2000; Virdi and Sachdeva, 2005; Merhej, et al., 2008). On the other hand, SC 3 groups three species, *Y. aleksiciae*, *Y. bercovieri*, and *Y. mollaretii*, with similar biochemical profiles together. *Y. mollaretii* and *Y. bercovieri* were originally classified as *Y. enterocolitica* BT 3A and 3B, respectively (Sulakvelidze, 2000). They were later speciated based on differential tests of sorbose, sucrose, indole, and inositol (Sulakvelidze, 2000). Here, they are once again grouping together in one heterogeneous lineage.

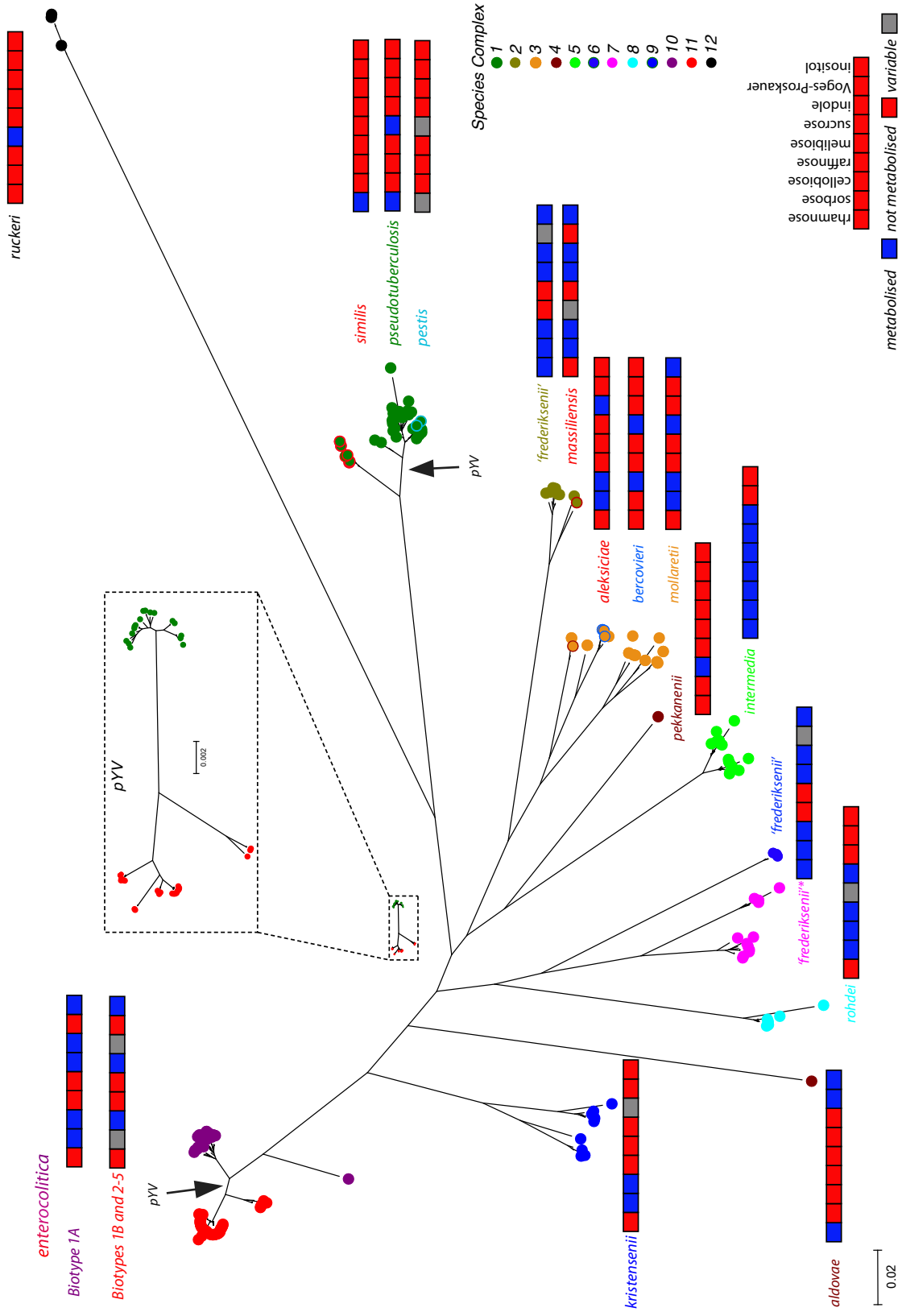


Figure 7.3: The phylogeny of the genus *Yersinia* and the virulence plasmid pYV. The maximum likelihood phylogenetic tree of the genus Yersinia is based on the concatenated sequence of 85 housekeeping genes. Current species assignments (colour circle borders) are contrasted with the species complexes (coloured circles) as allocated by BAPS analysis and the biochemical typing scheme used for speciation. There are three separate *Y. frederiksenii* lineages, labelled 'frederiksenii', thus subject to renaming to differentiate them. The type strain (Chen, et al., 2010) is depicted by *. An arrow shows the two independent acquisition events of the virulence plasmid pYV. The maximum likelihood phylogenetic tree of pYV is based on the whole sequence of the plasmid mapping against pYVe8081. The colours are the same as for the species complexes in which the plasmid is detected.

Y. pestis and *Y. pseudotuberculosis* form a tight, almost indistinguishable group which is consistent with previously published results (Achtman, et al., 1999; Kotetishvili, et al., 2005). They form a species complex with *Y. similis*. This non-pathogenic species was recently split off from *Y. pseudotuberculosis* on the basis of a melibiose-negative phenotype and differences in the 16S rRNA gene sequence (Sprague, Scholz, Amann and Busse, 2008). Furthermore, isolates of *Y. similis* do not possess the virulence plasmid pYV (Sprague, Scholz, Amann and Busse, 2008). Based on clustering analysis, it shows significant sequence similarity to be still considered part of the *Y. pseudoutberculosis/pestis* species cluster.

The other environmental species occupy branches between the *Y. pseudotuberculosis* complex and *Y. enterocolitica*, which is positioned diametrically opposed to *Y. pseudotuberculosis* at the other side of the tree. The only exception to this is *Y. ruckeri*, which forms an independent, long branch, and is clearly distantly related to the other members of the genus *Yersinia*. As was seen in Figure 7.1, *Y. enterocolitica* forms three separate lineages corresponding to non-, high-, and low-pathogenic BTs. As suggested from the genomic data, BT 1A is indeed the oldest lineage. One early branching isolate potentially reflects the diversity inherent in this environmental lineage. The pathogenic lineages are signified by the acquisition of the virulence plasmid after their separation from the non-pathogenic BT 1A.

Two new species have been recently described as members of the genus *Yersinia*: *Y. entomophaga* and *Y. nurmii* (Hurst, et al., 2011; Murros-Kontiainen, et al., 2010a). These isolates were not included in the sequencing collections, as they were too recently suggested. Both are described as being closely related to *Y. ruckeri*. This is interesting, as *Y. ruckeri* presents the most outlying lineage of the genus, and so far no intermediates to other lineages have been identified. The description of two such species now raises the question whether these are the same species described by two research teams independently and where they can be positioned with respect to the other lineages identified.

For these two species, only the sequences of the four housekeeping genes *glnA*, *gyrB*, *hsp60*, and *recA* are currently available. Although these genes might not be informative representatives to re-establish the phylogeny of the genus, as especially *recA* shows a high level of conservation, in combination with the known lineages and

their evolutionary history it is possible to establish the position of *Y. entomophaga* and *Y. nurmii* (Figure 7.4).

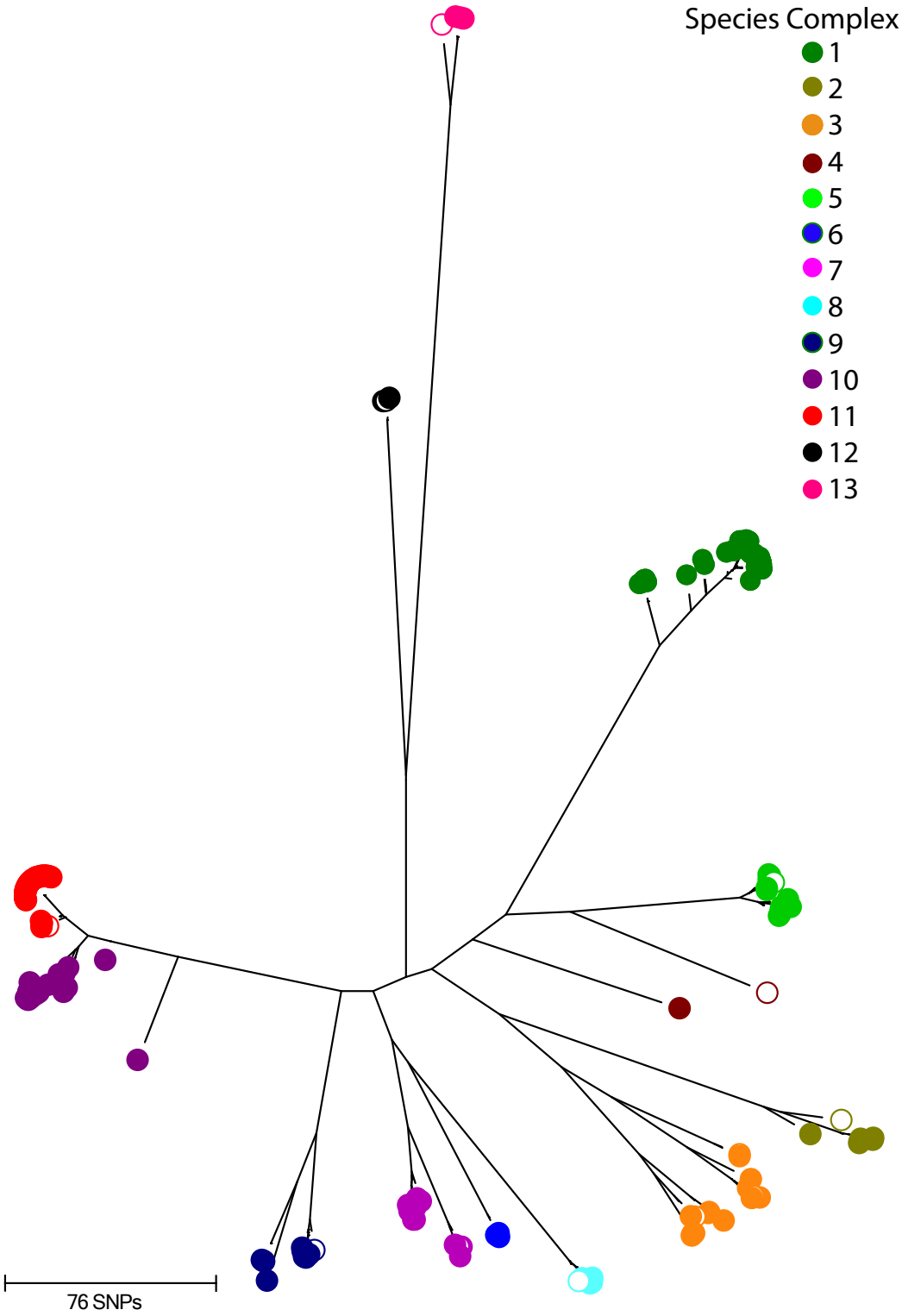


Figure 7.4: The phylogeny of the genus *Yersinia* based on *recA*, *glnA*, *gyrB*, and *hsp60*. The maximum likelihood phylogenetic tree of the genus Yersinia is based on the concatenated sequence of four housekeeping genes. The same species complexes (coloured circles) as allocated by BAPS analysis (Figure 7.3) are highlighted. The type strains are depicted by open circles. *Y. entomophaga* and *Y. nurmii* form SC13, with *Y. entomophaga* shown as open circle.

From Figure 7.4 one can see that the branching order differs in these four hosuekeeping genes from the 85 genes, yet the exact same species clusters are formed and can be used to establish the dimension of divergence or clonality of *Y. entomophaga* and *Y. nurmii*. It is obvious that *Y. entomophaga* and *Y. nurmii* form a single lineage. Whilst no clustering analysis is available, comparison with the diversity of other lineages makes it likely that these form a single species complex SC13.

The 17 officially described and recognized species of *Yersinia* therefore from 13 different species cluster. Of the old species definition, eight designations do not agree with the species complexes, and taxonomy and nomenclature will need to be revised to clarify the nature of the isolates.

7.2.3 Phylogeny of the genus *Yersinia* based on 16S rDNA sequences

The independent evolution of pathogenic lineages in *Yersinia* has not been noticed before, but it was assumed due to their pathogenic potential in humans that they shared a common pathogenic predecessor. The evolutionary distance of environmental, "*Y. enterocolitica*-like" species (Sulakvelidze, 2000) appears to have been hugely underestimated as well. The study of only few genes or of genes that are not phylogenetically informative or robust might have skewed the image of an evolutionary relationship as well. One example for this is the 16S rRNA gene (Figure 7.5).

Although this representation of the 16S rRNA genes in the genus is based on corrected Illumina reads and not on finished Sanger sequencing, the sequences for the representative type strains of the environmental *Yersiniae* fall together with other isolates of their species cluster (Figure 7.5). This indicates that the sequencing data is comparable to the Sanger sequencing approach. Overall, there is very little variation within the 16S gene. 95.69% of the gene, or 1398 out of 1461 bases investigated, are invariable and phylogenetically non-informative. One highly variable region is detected between bases 429 and 449. Based on this data set, one can see that the phylogeny based on 16S rDNA is in disagreement with the 85 housekeeping genes. There is still an indication of clustering according to species cluster, and the *Y. enterocolitica* and *Y. pseudotuberculosis/pestis* lineages are still at opposite ends of the tree. Notably though, in *Y. enterocolitica* pathogenic and non-pathogenic lineages intermingle. The distinction of two subspecies was based on variation in 12 bases only (Neubauer, et al., 2000), and the corrected Illumina data might not reflect

this, yet these subspecies are invalid nonetheless with respect to the whole genome phylogeny.

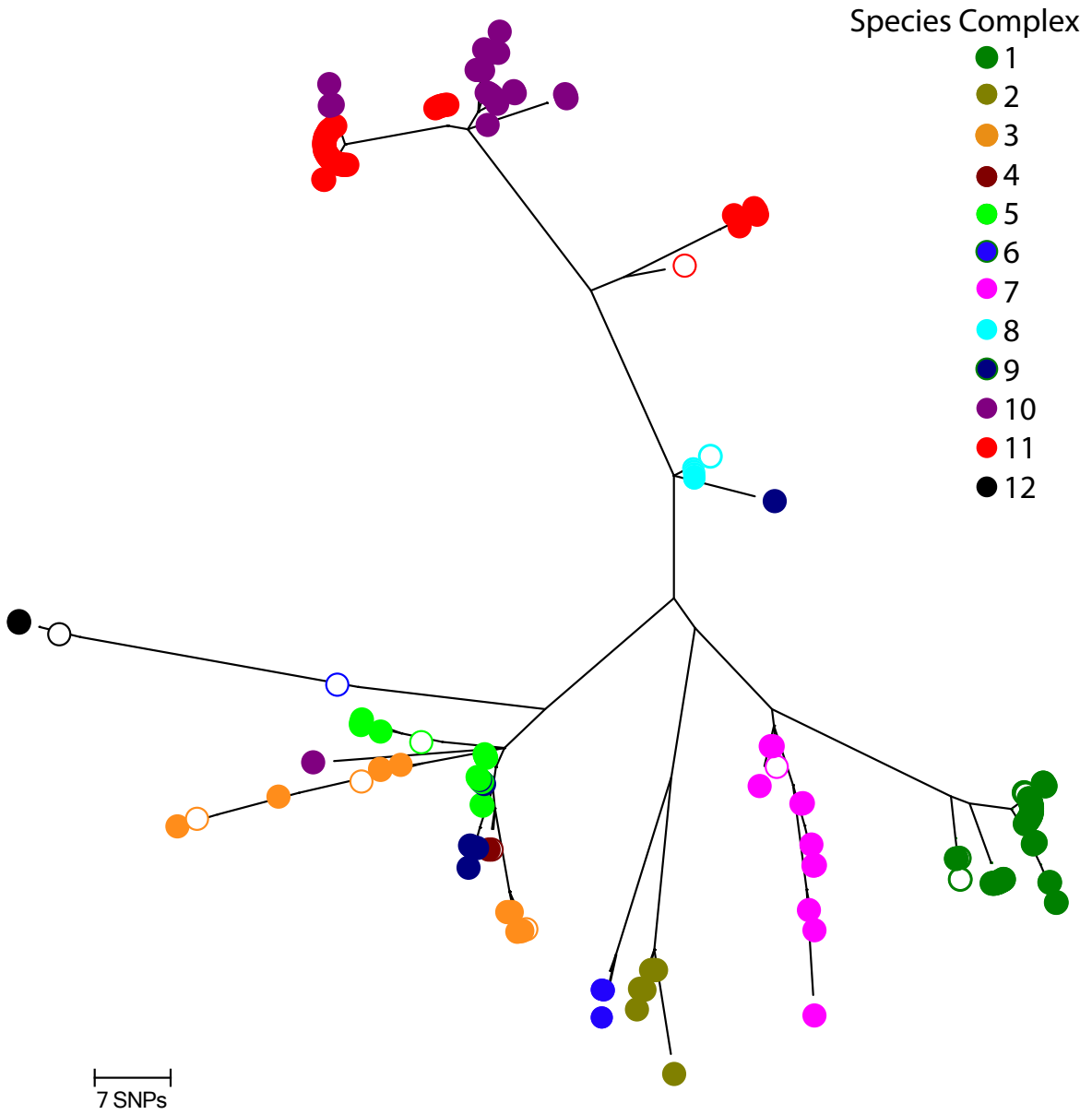


Figure 7.5: The phylogeny of the genus *Yersinia* based on 16S rDNA sequences. The maximum likelihood phylogenetic tree of the genus Yersinia is based on 1461 bases of the 16S rRNA gene. The species complexes (coloured circles) as allocated by BAPS analysis are shown. The type strains sequenced in previous studies are depicted by open circles.

The clusters formed are the same clusters originally described by Ibrahim, et al. (1993 and 1997). The environmental *Yersiniae* generally form a heterogeneous cluster. The branching order of individual lineages that was seen in the phylogeny of 85 housekeeping genes cannot be observed, *Y. frederiksenii* has changed its position to be closer to *Y. pseudotuberculosis/pestis*, and some isolates of

Y. mollaretii/bercovieri/aleksiciae as well as *Y. intermedia* are found on two different branches.

The phylogenetic analysis of 85 housekeeping genes therefore provides much greater resolution of the genus than is possible using the 16S gene sequence, which fails to adequately resolve the environmental *Yersinia* strains. The limitations of species delineation (and assignation) by biochemical tests and 16S sequence illustrates the necessity of using a finer scale approach to distinguish between members of the genus, and the need for thorough assessment of the genomic signature of novel species currently being reported in the *Yersinia* genus.

7.2.4 Acquisition of the virulence plasmid pYV in pathogenic *Yersiniae*

Also shown in Figure 7.3 (inset) is the phylogeny of the virulence plasmid pYV. The acquisiton of this plasmid is thought to be a key element in the evolution of pathogenic *Yersiniae* lineages. So far, the plasmid has only been found in pathogenic *Yersinia* spp., and this is confirmed in the present strain collection as pYV was only detected in the *Y. pseudotuberculosis*/*Y. pestis* cluster as well as the pathogenic *Y. enterocolitica* BTs 1B and 2 – 5. The previous model for the evolution of *Yersinia* postulated a single acquisition event of pYV in a pathogenic predecessor and subsequent loss in the non-pathogenic *Y. enterocolitica* BT 1A (Figure 1.2; Portnoy and Falkow, 1981; Carniel, 2002; Wren, 2003). From Figure 7.3 it is evident that this is unlikely. With the environmental isolates positioned between the pathogenic lineages a more parsimonious justification is the acquistion of pYV in two separate events by the pathogenic lineages. Further support to this is given in the evolutionary distance between virulence plasmids that does not equal the distance of *Y. enterocolitica* to the *Y. pseudotuberculosis*/*Y. pestis* cluster. Drawn to scale, the diversity within the plasmids is considerably less than in the genome (Figure 7.3).

Looking more in detail at the plasmid phylogeny (Figure 7.6) one can see that it relates to the phylogeny seen in the *Y. enterocolitica* lineages.

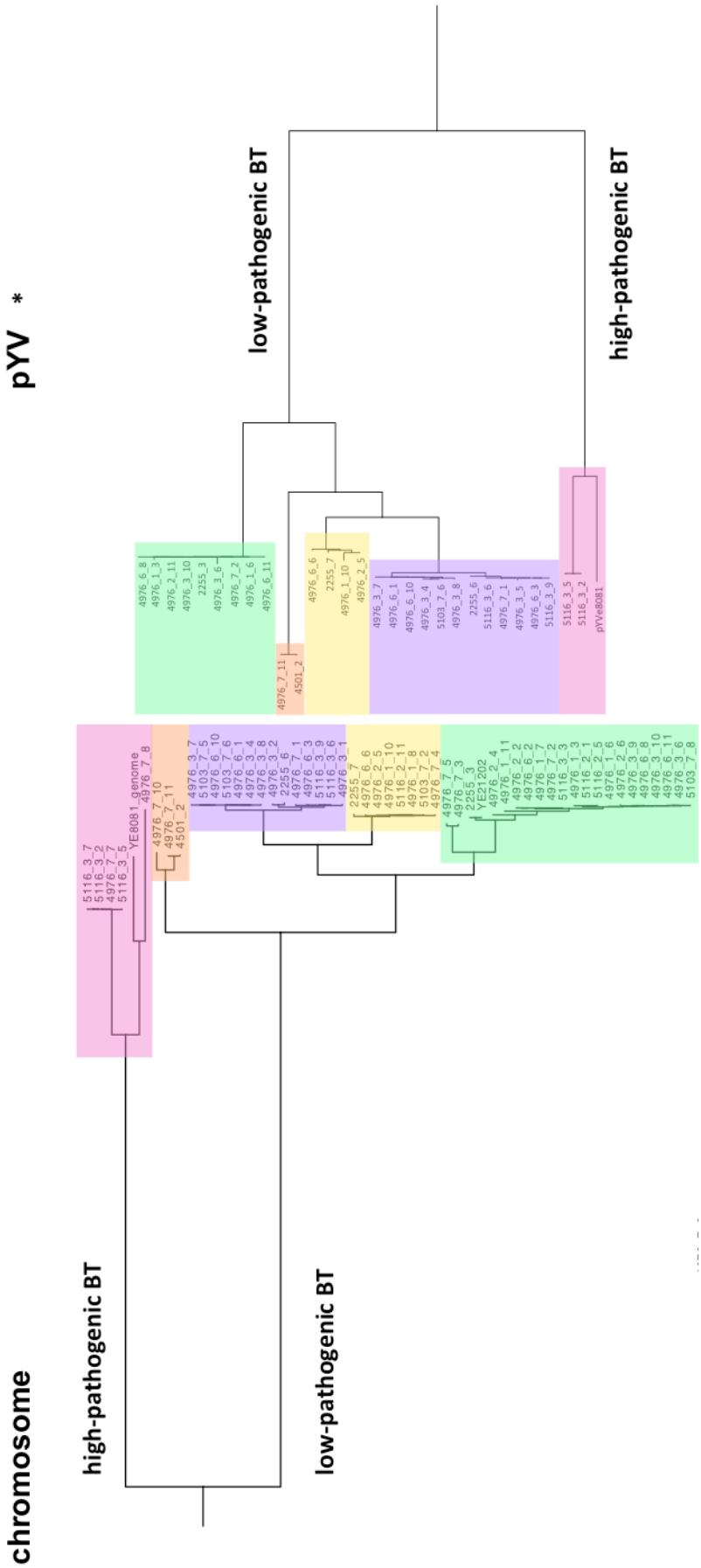


Figure 7.6: Phylogeny of the plasmid pYV compared to the phylogeny of the chromosome in *Y. enterocolitica*. * not drawn to scale of each other. Red – BT 1B O:8; orange – BT 5 O:2a,2b,3; purple – BT 4 O:3; yellow – BT 2/3 O:5,27; green – BT 2/3 O:9.

The different lineages of the virulence plasmid exactly reflect the lineages in the pathogenic BT, although the branching order is slightly different. This is consistent with a single acquisition event, but also indicates that the plasmid then co-evolved with a specific lineage and was not transferred between bio-serotypes. It would indicate that the pathogenic strains have adapted to specific niches, which were enabled by the genes encoded on the virulence plasmid. The low-pathogenic BTs show signs of evolving into specific ecotypes, with no exchange of genetic material between them, be it LPS operons, or the virulence plasmid.

The acquisition of the virulence plasmid therefore has added significantly to the pathogenic potential of two different pathogenic species, in independent events and in different genetic backgrounds.

7.3 Discussion

7.3.1 The evolution of pathogenic *Y. enterocolitica* lineages

The phylogeny of *Y. enterocolitica* shows that there are three very distinct lineages. BT 1A is the oldest lineage, from which the pathogenic lineages have evolved through the acquisition of the virulence plasmid on at least one independent occasion. The non-pathogenic BT shows heterogeneity, as indicated by the large number of SNPs between individual isolates resulting in long branch lengths. Due to predominant sampling in a clinical setting and during an abbatoir study, this divergence is likely to be an underestimate. Analysis of the other sequenced isolates might yield a considerable number of unique CDSs other than phage, and the pan-genome of BT 1A can be considered open.

The position of BTs 1A at the root of the *Y. enterocolitica* tree explains some of the features of the reference genome. When considering regions of difference between BT 1A and the other BTs, it is now apparent that loss in the pathogenic lineages is a driver for evolution rather than acquisition events in the non-pathogenic BT. Inspection for example of the D-arabitol cluster (Chapter 6) shows that most of it has been lost during the evolution of the pathogenic lineages, and that the last gene remnant was lost without trace in the low-pathogenic BTs, whilst the remnant remains in BT 1B. This trajectory also helps to explain the differences in pseudogenes and remnants of the second flagella cluster Flag-2; the operons have been lost independently in BT 1A and 1B, and the genome decays has happened in different genes in the cluster. A search for Flag-2 genes in the *Y.* enterocolitica strain collection showed that some BT 1A isolates even retain the complete cluster. It is not known, however, whether the clusters are still functional, yet their frequent loss in individual branches indicates a dispensable nature.

The pathogenic BTs have a common evolutionary path, acquiring seven pseudogenes, before splitting into high- and low-pathogenic BTs. The high-pathogenic BT is marked by the acquisition of previously mentioned markers of isolates with increased pathogenicity, the HPI, and the Ysa T3SS in the plasticity zone. Considering that only five strains were sequenced, there is still considerable diversity amongst this BT, although it is not comparable with the heterogeneity in BT 1A. As for the distribution of the pathogenicity islands, the larger strain collection confirmed the Ysa T3SS and the HPI as specific for BT 1B of *Y. enterocolitica*,

whereas the YAPI appears to be an acquisition specific to strain 8081. This was found before (Thomson, et al., 2006), and it is similar to *Y. pseudotuberculosis*, which shows a variable distribution of the HPI and YAPI in different combinations (Fukushima, et al., 2001; Eppinger, et al., 2007).

The low-pathogenic BTs have evolved through genome rearrangements and gene loss, as indicated by the increased number of IS1667 and the marked pseudogenization. The phylogeny of the low-pathogenic BTs of *Y. enterocolitica* is strictly based on serotype. Serotypes O:9, O:5,27, O:3, and O:2a,2b,3 each form their own branch. This is surprising, as the LPS operons can be exchanged in other bacteria, for instance *Streptococcus pneumoniae*, and serotype switching may occur so that an evolutionary signal based on serotype is not visible (Croucher, et al., 2011). In *Y. enterocolitica*, the different lineages strictly adhere to serotype, with no indication of serotype recombination. This is also different from the population structure of *Y. pseudotuberculosis* where recombination is frequent, and where parts of the serotype operons appear to undergo exchange and recombination with other clusters (Ch'ng, et al., 2011). The population structure of the low-pathogenic *Y. enterocolitica* BTs suggest niche adaptation following a bottleneck leading to specific ecovars. This niche could involve different insect vectors suggested by the presence of insecticidal toxin complexes.

One caveat has to be added – the serotypes investigated in this study present the dominant, most prevalent serotypes. The low-pathogenic BTs may exhibit other serotype as well, but these are infrequent and not associated with disease (McNally, et al., 2004; Bottone, 1999). It would be interesting to see where they appear on a phylogenetic tree of the low-pathogenic BTs.

Also in contrast to *Y. pseudotuberculosis*, the differentiation of BTs in *Y. enterocolitica* into high-, low-, and non-pathogenic lineages appears to be ancient, with a clear phylogenetic signal and long branches. Although *Y. pseudotuberculosis* was divided in high- and low-pathogenicity groups as well, these do not present in specific lineages in a MLST tree (Ch'ng, et al., 2011). This population appears to be more dynamic in terms of exchange of pathogenicity islands, whereas *Y. enterocolitica* seems to have a long history of adaptation, with recent clonal expansion of the low-pathogenic BTs. The short branches in the low-pathogenic BTs, and the fact that unique regions of the reference genomes pertain to phage, this might suggest a closed pan-genome for BTs 2 – 5. This means that further sequence

analysis is unlikely to uncover new pathogenicity islands, similar to the closed pan-genome of *Y. pestis* (Eppinger, et al., 2010). This is in contrast to the non-pathogenic BT 1A, and possibly the high-pathogenic BT, which show a reasonable diversity to suggest a pan-genome that is still open for new discoveries.

7.3.2 The genus *Yersinia* and the virulence plasmid pYV – independent evolution of pathogenicity

The phylogeny of the whole genus *Yersinia* and the virulence plasmid pYV do not agree with the current evolution of pathogenic lineages of *Yersinia*. The pathogenic lineages are positioned diametrically at opposite ends of the genus tree. The most parsimonious explanation for this is the independent acquisition of the virulence plasmid on at least two separate occasions; this presents 2+ events as opposed to a single gain in the last common ancestor of *Yersinia* and 10+ loss events for each environmental lineage. Further to this, not only is the virulence plasmid not present in any of the environmental lineages, the evolutionary distance between the virulence plasmids is too recent to compete with the ancient lineages presented in the analysis of the chromosome. It could be argued that the virulence plasmid is under strict selection and therefore shows less sequence variation. Yet, pYV is only essential inside a warm-blooded host, and the life cycle of *Yersinia* includes environmental and insect stages. The housekeeping genes on the other hand are under constant selection no matter the environmental condition as they perform essential function for cellular survival. Their divergence should thus give a more accurate estimate and scale of the relationship within the genus.

Interestingly, the virulence plasmid also shows signs of co-evolution with specific lineages in *Y. enterocolitica*. This indicates that the acquisition was a key event that led to a marked change in lifestyle and niche. It implies that the exchange of mobile genetic elements is not important in the low-pathogenic BTs, possibly due to the adaptation to specific niches. This would also explain the strict adherence to serotype in the phylogenetic grouping.

Considering the genus as a whole, it is apparent that both the human pathogenic lineages share patterns of evolution. Their development following the acuqisition of the virulence plasmid is characterised by functional gene loss. In one lineage, this lead to the emergence of the acutely pathogenic *Y. pestis*, and in the other lineage to the low-pathogenic *Y. enterocolitica* biotypes. Functional gene loss is also observed

in other human pathogens, such as *M. tuberculosis* and *Salmonella*, and can thus be considered an emerging pattern on the road to human pathogenicity.

The evolution of *Y. pestis* through gene loss, genomic rearrangements, expansion of IS elements, and metabolic streamlining is exactly mirrored in the low-pathogenic BTs of *Y. enterocolitica*, especially BTs 4 and 5. BT 4 is ubiquitous, but is the biotype that is predominantly isolated from human cases in mainland Europe (McNally, et al., 2004; Batzilla, et al., 2011), which indicates some degree of host adaptation, and BT 5 is the hare biotype.

7.3.3 Differentiation of species in the genus *Yersinia*

Several papers have already pointed out that commerically available biochemical tests are insufficient to distinguish *Y. enterocolitica* from environmental, *Y. enterocolitica*-like species (Stock, Henrichfreise and Wiedemann, 2002; Merhej, et al., 2008). Especially with identification of new species in the genus, the list of possible phenotypic tests is growing and might become unmanagle, confusing, and impractical. A more useful approach is needed, that unambiguously speciates isolates, without dependency on phenotypes that might vary with reagents, incubation conditions, and are subject to random inactivating mutations.

The 16S rDNA gene that has been commonly used to establish a tree of life is unsuitable for speciation in *Yersinia*. It has been noted before that the 16S rDNA is only useful up to the genus level, but that sequences most often are too conserved to distinguish individual species (Drancourt and Raoult, 2002; Janda and Abbott, 2007). This is the case for example in *Staphylococcus*, where for example *rpoB*, the RNA polymerase gene, and others are used instead of 16S gene to speciate isolates (Drancourt and Raoult, 2002; Ghebremedhin, Layer, König and König, 2008; Janda and Abbott, 2007). It has also been shown that, despite the high degree of conservation, and despite the essentiality of the small ribosomal subunit, an exchange of 16S rDNA genes between different species is possible and that active recombination between species does not exclude the 16S gene (Asai, Zaporojets, Squires and Squires, 1999; Schouls, Schot and Jacobs, 2003). Adding to this the fact that the 16S rDNA gene is commonly present in several copies in a single genome (Schouls, Schot and Jacobs, 2003; Janda and Abbott, 2007), and that sequence variation – however small – are known, makes the 16S rDNA inapt to *Yersinia* speciation.

With the 85 housekeeping genes presented here, the development of an MLST scheme reflecting the evolution of the genus is a self-evident proposal. Improvement on the MLST scheme proposed by Kotetishvili, et al. (2005) is currently under way by Miquette Hall to establish a set of seven genes that can be used to differentiate *Yersinia* isolates exactly, and that will also allow for future identification of potential new species. The *Yersinia* taxonomy has to be updated to incorporate the findings of this study, and an MLST approach circumvents the reliance on biochemistry.

8 Overall Discussion and Future Work

8.1 Analysis of reference genomes and plasmids

The sequencing of the reference genomes highlighted the fact that a single genome might not be representative of a whole species, depending on the heterogeneity of said species. Although the general genome composition is similar in all *Y. enterocolitica* BTs with respect to overall size, coding density, and functional classes of CDSs, significant differences can be observed on closer inspection relating to pseudogenes and regions that are absent or inactivated in certain lineages.

All BTs have about 80% of their CDSs in common, signifying the *Y. enterocolitica* core functions. Amongst these are the flagella cluster Flag-1, the tigh-adherence operon, and both hydrogenase operons. A considerable number of CDSs is shared in BT 1A and BT 1B as well as BT 1A and the low-pathogenic BTs, indicating differential loss after the split of the pathogenic BTs. The non- and high-pathogenic BTs share two genomic islands, YGI-2, encoding a glycolipoprotein, and YGI-4, an integrated plasmid, as well as regions relating to resistance to heavy metals nickel, cobalt, and arsenic. The non- and low-pathogenic BTs share the *Yersinia* Genus T3SS, the outer core as part of their LPS, and metabolic properties. On the other hand, only 40 CDSs are shared in all pathogenic BTs apart from the virulence plasmid pYV. Most of these are related to hypothetical or conserved hypothetical proteins and have therefore no allocated or confirmed function in a role in disease, except for the adhesion invasion locus *ail*. Therefore there is no conserved pathogenic core present in high- and low-pathogenic BTs apart from *ail* and the virulence plasmid pYV.

The non-pathogenic BT has unique regions involved in arsenic, silver and copper resistance as well as other metabolic operons. The HPI, YAPI, general secretion pathway *yst1* and Ysa T3SS were confirmed to be unique to the high-pathogenic BT. The low-pathogenic BTs share the *tc*PAI and a second flagella cluster. The low-pathogenic BTs are further signified by the expansion of IS1667 leading to wide-ranged genome rearrangements.

The virulence plasmid pYV shows considerable variation regarding its organization, and should therefore rather be considered a family of closed related plasmids as suggested by Portnoy, Moseley and Falkow (1981). The Yop T3SS forms the

conserved part of the plasmid, whereas other CDSs can change orientation and position. Nevertheless, the overall functionality of the virulence plasmid is preserved. Smaller variations concerning YscP have been correlated with needle length, but there are no indications that BTs with different needle length show a difference in efficiency of protein transfer into the target cell.
The variations seen between pYV plasmids of high- and low-pathogenic BTs suggest divergent evolution of the virulence plasmid in *Y. enterocolitica*. The origins of replication in low-pathogenic BTs is the same as in *Y. pestis* and *Y. pseudotuberculosis*, which could mean that the high-pathogenic pYV evolved through co-integration with a plasmid of a different incompatibility group (Snellings, Popek and Lindler, 2001). The exact evolutionary history and potential exchanges of pYV between *Y. pseudotuberculosis/pestis* and *Y. enterocolitica* BTs cannot be established, unless a similar, intermediate plasmid is discovered in an environmental bacterium outside the *Yersinia* genus.

The unique plasmids add to the overall genetic pool of *Y. enterocolitica*. Whilst the plasmids sequenced from BT 1A are cryptic, pSR1-2 is potentially important for tetracycline resistance in the pig host, and pSR4-5 might be able to mobilize the virulence plasmid based on homology with a similar conjugative plasmid.

8.2 Analysis of phenotypic microarray

A recurring theme in regions of difference and pseudogene targets are metabolic operons. Metabolism is both a driver and a consequence of host or niche adaptation; it is a driver because the search for nutrients drives bacteria into exploring new environments, and it is a consequence as subsequent metabolic streamlining locks bacteria in a niche. The best example for this are *Y. enterocolitica* BT 5 and *Y. pestis*, which exhibit reduced ability to survive outside their respective hosts.

Owing to the fact that BT 1A is a ubiquitous, environmental organism, it can utilize a wide range of different metabolites, and is adaptable to very different conditions. This provides a survival advantage in the environment, as conditions change with season and with respective current environment such as aquatic or soil environments that pose very different challenges on a bacterium. The unique operons involved in metabolism of nutrients carried in the non-pathogenic BT confirm this.

The increased number of pseudogenes observed in the low-pathogenic BTs correlates with the decreased metabolic flexibility that is apparent in the phenotypic microarray. The pathogenic BTs have lost part of their metabolic flexibility upon adaptation to human hosts and warm-blooded animals. Both human and animal hosts provide a stable surrounding compared to the outside environment regarding temperature and available nutrients. On the other hand, nutrients might be available that are not found outside the gut environment and that require specialized transport and metabolic operons. Additionally, the availability of other nutrients may be limited and targeted acquisition systems are needed. Examples in the latter category are iron and heme acquisition systems, for example encoded on the high-pathogenicity island of BT 1B, that allow for capturing and utilization of this trace element. The phenotypic microarray thus far only highlighted the loss of functions in the pathogenic BTs rather than the gain of utilizing new nutrients compared to the environmental BT 1A. An in-depth analysis of the 37°C data might yield results, but it is possible that the conditions tested do not activate the utilization of those nutrients. Considering the conditions found in the gut, oxygen starvation and change in pH might be important signalling factors that could be investigated in the future using the phenotypic micorarray.

8.3 Revised model for the evolution of pathogenicity in the genus *Yersinia*

The themes of gene loss, expansion of IS elements, metabolic streamlining and acquisition of mobile genetic elements in the high- and low-pathogenic BTs is reminiscent of the evolutionary path of *Y. pestis*. To investigate the relationships between different *Yersinia* species, 85 housekeeping genes were chosen for comparison. This extensive set of genes should give a robust representation of this diverse genus. The phylogeny of the genus *Yersinia* showed that the humanpathogenic lineages evolved independently out of the environmental species, yet employing the same driving forces of “add, stir, and reduce” that have been suggested for *Y. pestis* (Wren, 2003).

The comparison of the phylogeny of the chromosome of *Yersinia* spp. against the phylogeny of the virulence plasmid pYV showed that the virulence plasmid has been acquired in at least two instances and more recently than the divergence of the *Y. pseudotuberculosis* and *Y. enterocolitica* lineages. This is in disagreement with the

notion that the pathogenic *Yersiniae* had a common, pathogenic precursor that acquired the virulence plasmid in a single event.
It also needs to be considered that the environmental species form very distinct lineages from *Y. enterocolitica* and should therefore not be considered '*Y. enteroclitica*-like' species purely based on the fact of similar biochemical typing reactions.

The model for the evolution of pathogenicity in the genus *Yersinia* needs to be revised and one suggestion is presented in Figure 8.1.
From the ancestral *Yersinia*, several lineages evolved. On opposite ends of the spectrum of diversity are the pathogenic lineages, with the environmental, supposedly non-humanpathogenic species situated in between.
The humanpathogenic lineages have evolved independently, yet employing the same mechanisms. Both have acquired novel genetic material that distinguishes them from environmental species. For *Y. pseudotuberculosis*, the hemin storage locus and the high-pathogenicity island play an important role. The distinguishing features for *Y. enterocolitica* are less clear; in the past the enterotoxin *yst* has been suggested (Wren, 2003). Research is currently underway to determine the distribution of virulence factors within the genus, which will also establish descriptors for *Y. enterocolitica*. For both pathogenic lineages, the acquisition of the virulence plasmid has been crucial in opening a new niche in a warm-blooded host. In *Y. enterocolitica*, this gain has happened after the split of the non-pathgenic, environmental BT 1A. Clustering based on BAPS suggests BT 1A as a separate species cluster, and given the very different nature and environment of BT 1A, this is a reasonable suggestion. Although BT 1A is considered non-pathogenic, it is isolated from healthy and ill individuals (Bottone, 1999; Tennant, Grant and Robins-Browne, 2003; Tennant, Skinner, Joe and Robins-Browne, 2005). Research shows though that the ability to cause disease and the mechanism to do so differs markedly from the other pathogenic *Y. enterocolitica* BT, in that survival and hibernation inside macrophages might play a role (McNally, et al., 2006). On the other hand, the pathogenic BTs are not signified by gaining a large number of CDSs, and the non-pathogenic BT shares a considerable number of operons with both high- and low-pathogenic BTs respectively.

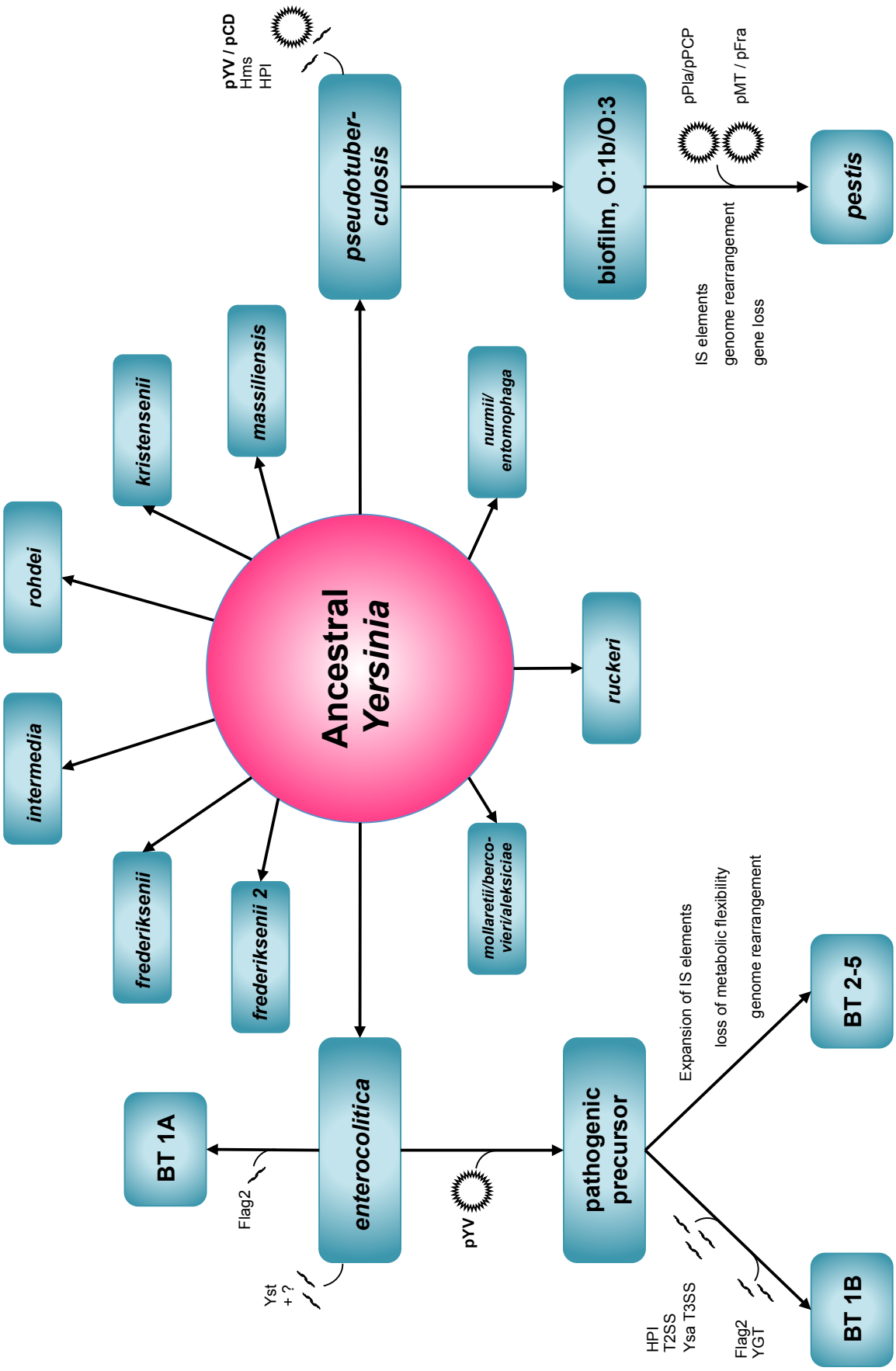


Figure 8.1: Revised model of the evolution of pathogenicity in the genus *Yersinia*.

8.4 Future Work

"As our circle of knowledge expands, so does the circumference of darkness surrounding it." – Albert Einstein

Now that the evolutionary relationships between the *Yersinia* spp. have been established, and it was shown that a biotyping approach might not be as suitable for identification as molecular techniques, research is currently underway to determine a set of seven genes for the use in MLST. The genes will be chosen to reflect the lineages identified using BAPS clustering, and will improve speciation in the genus *Yersinia*. This study should also include the putative species cluster 13, which contains the newly described *Y. entomophaga* and *Y. nurmii*, to confirm the initial observation with respect to their relationship to *Y. ruckeri* and as a single species cluster.

A way of improved identification within *Y. enterocolitica* could be the design of a PCR-based approach testing for the presence of O-antigens associated with pathogenic isolates. As only very specific bio-serotype combinations are known to be causing disease, a PCR reaction for genes specific to those serotypes in combination with biochemical data could improve identification of pathogens. The drawbacks are the unknown O-antigen operons that can potentially cross-react. Instead, primers could also be designed to amplify biotype-specific regions, such as the HPI in the high-pathogenic BT 1B, or the *tc* PAI in low-pathogenic BTs. The caveat for this approach is that a search for important islands and virulence operons is still outstanding. This should determine whether homologous sequences of certain genes are present in the environmental *Yersinia* species. It would also clarify what defines the species "*enterocolitica*" as opposed to the environmental species.

In terms of identification, the data from the phenotypic microarray could also be used to potentially define metabolic reactions that reliably distinguish pathogenic and non-pathogenic BTs, but the nutrients identified as different between the BTs would need to be tested in a larger strain collection. A further requirement would be a clear link between phenotype and genotype to avoid past problems of meiotrophy and phenotype reversal due to genes inactivated with revocable mutations.

One interesting region identified during the comparison of the reference genomes of *Y. enterocolitica* and beyond is the Yersinia Genus T3SS (YGT). The distribution in all *Yersinia* spp. and the apparently targeted inactivation in humanpathogenic

lineages of *Y. pseudotuberculosis/pestis* and *Y. enterocolitica* BT 1B, 4, and 5 might indicate a possible role in environmental survival. This could include interaction with insects such as greater wax moth *Galleria mellonella* and tobacco hornworm *Manduca sexta*, nematodes such as roundworm *Caenorhabditis elegans* or amoeba.

Another open question concerns the essential genes needed during the infection process. As was seen in the comparative analysis, only 30 CDSs plus the virulence plasmid pYV are shared in the pathogenic *Y. enterocolitica* BTs. It would deepen the understanding of mechanisms of disease to find out about the role of these 30 CDSs in the infection process. One could use RNAseq (Croucher, et al., 2009) to investigate the transcripts of the cell under the experimental condition and give a reflection of active genes and pathways. To complement expression studies, knock-out mutants could be examined for their effect, and an interesting advance is the use of TraDIS (Langridge, et al., 2009), the Transposon directed insertion-site sequencing, in which the whole genome is covered with transposon insertions. Genes that are absent of any insertions are thought to be essential under the condition tested (Langridge, et al., 2009). In common to these methods is their global view of the bacterial cell, in trying to understand larger systems and interactions as a whole instead of single entities. This might be important in the case of high- and low-pathogenic BTs that potentially utilize a different combination of genes to establish disease.

Similarly, the phenotypic microarray layout could be used to identify metabolic operons. In this case, only a subset of carbon sources would be studied for a mutant library of one strain. The genes identified under all conditions would represent the normal respiration of a cell, whereas the absence of transposon insertions in some genes under certain conditions only might represent the genes involved specifically in the metabolism of that nutrient. Similarly, RNAseq could identify genes that are generally expressed in carbon metabolism, and also highlight those genes that are specifically upregulated only under certain carbon sources. This would significantly increase annotation precision, and might identify new proteins or assign functions to hypothetical proteins in metabolic pathways.

Further open questions in the phenotypic microarray relate to the role of oxygen saturation in metabolism at 37°C. The marked reduction of metabolic activity with an increase in temperature might be centrally regulated, and might be triggered in low

oxygen concentration. A complementation of the phenotypic microarray could be to repeat the experiments under anaerobic conditions at 37°C. Given the potential roles of insects as vectors and hosts in the environment it could also be of interest to examine the metabolic activities of the BTs at 10-15°C.

Although the reference genomes of this study have been extensively compared to one another, a comparison to the other recently published genomes would be interesting. The genomes are similar in terms of size, GC content, number of coding sequences, but marked differences were observed in the number of pseudogenes. As the reference genomes presented here are of a high quality draft standard, they could be best compared to the deposited BT 4 O:3 strain Y11 (Batzilla, et al., 2011a,b) of the same quality. The comparison could highlight differences between mainland Europe isolates and UK isolates that relate to their difference in incidence.
In this same area, it would also be interesting to take the SNP data one step further than for establishing an evolutionary relationship and relate the SNPs to a biological function. It could be that certain SNPs define human pathogenic isolates despite them not clustering together due to independent SNPs, especially with respect to comparing non-pathogenic and low-pathogenic isolates. It could also be interesting to investigate the SNPs leading to the definition of a lineage, to see whether certain themes relating to membrane proteins and antigen presentation to the host immune system dominate.

Bottone (1999) considered *Y. enterocolitica* as the “apex of bacterial versatility”, but this may also be applied for the whole genus that produced two independent humanpathogenic lineages and a wide range of environmental, underrepresented species. Even nearly 35 years after Bottone’s first review of *Y. enterocolitica* (1977), *Yersinia* remains an enigmatic and charismatic organism as ever.

9 Appendix

Table 9.1: *Y. enterocolitica* strain collection.

Strain	Biotype	Serotype	Source	country	Lab
209-36/84	1B	O:21	Human	Germany	Fuchs
H1527/93	1A	O:5	Human	Germany	Fuchs
H230/89	3	O:5,27	Human	Germany	Fuchs
H324/78	3	O:9	Pig	nd	Fuchs
H450/87	4	O:3	Human	nd	Fuchs
H469/87	4	O:3	Pig	nd	Fuchs
H608/87	4	O:3	Human	nd	Fuchs
H692/94	2	O:9	nd	Germany	Fuchs
H7580/93	3	O:9	nd	nd	Fuchs
IP10393	4	O:3	clinical (stool)	France	Carniel
IP20322	2	O:5,27	alimentary	Greece	Carniel
IP20851	2	O:9	clinical (stool)	Spain	Carniel
IP21447	2	O:9	pig (stool)	England	Carniel
IP22276	4	O:3	clinical (stool)	Australia	Carniel
IP26249	2	O:5,27	clinical (stool)	France	Carniel
IP26618	1A	O:5	alimentary	Italy	Carniel
IP26656	4	O:3	clinical (stool)	France	Carniel
IP26766	2	O:9	clinical (stool)	France	Carniel
IP29194	3	O:3	clinical (stool)	France	Carniel
IP29228	3	O:5,27	clinical (stool)	France	Carniel
IP29610	4	O:3	clinical (stool)	France	Carniel
N2118599	3	nd	Human	New Zealand	DuFour
N21282756	3	nd	Human blood	New Zealand	DuFour
N21383523	3	nd	Human blood	New Zealand	DuFour
N2153435	1A	nd	Human	New Zealand/Fidji	DuFour
N21572344	4	O:3	Animal	New Zealand	DuFour
N21622789	4	nd	Dog	New Zealand	DuFour
N21873627	4	nd	Human blood	New Zealand	DuFour
N22084574	4	nd	Human blood	New Zealand	DuFour
N2214757	4	nd	Human blood	New Zealand	DuFour
N22232770	4	nd	Human blood	New Zealand	DuFour
N22373947	1A	nd	Sheep	New Zealand	DuFour
N2982399	2	nd	Human blood	New Zealand	DuFour
NZ1032126	2	nd	Animal meat	New Zealand	DuFour
NZ1484423	3	nd	Human	New Zealand	DuFour
NZ198109	4	nd	Human blood	New Zealand	DuFour
NZ28081	1A	nd	Human	New Zealand/Asia	DuFour
NZ523784	1B	nd	Human	New Zealand	DuFour
NZ761915	1B	nd	Human	New Zealand	DuFour
NZ832501	2	nd	Human blood	New Zealand	DuFour
SZ166/96	5	nd	nd	nd	Fuchs

SZ375/04	1B	O:8	Human	Germany	Fuchs
SZ506/04	1B	O:8	Human	Germany	Fuchs
SZ5108/01	1B	O:8	Human	Germany	Fuchs
SZ662/97	1A	O:5	Human	Germany	Fuchs
W22703	2	O:9	Human	Belgium	Fuchs
Y127	2	O:9	nd	nd	Fuchs
Y286	1B	O:8b,e,f,i	nd	USA	Fuchs
YE03/03	4	O:3	Human faeces	UK	McNally
YE04/03	3	O:5	Human faeces	UK	McNally
YE07/03	4	O:3	Human faeces	UK	McNally
YE08/03	4	O:3	Human faeces	UK	McNally
YE09/02	1A	nd	Sheep	UK	McNally
YE09/03	1A	O:6,30	Human faeces	UK	McNally
YE11/03	3	O:9	Human case	UK	McNally
YE111/02	3	O:5,27	Sheep	UK	McNally
YE119/02	2	O:9	Sheep	UK	McNally
YE13/02	1A	O:6,30	Cattle	UK	McNally
YE13/03	1A	O:6,30	Human faeces	UK	McNally
YE17/03	3	nd	Human blood	UK	McNally
YE201/02	4	O:3	Pig	UK	McNally
YE204/02	4	O:3	Pig	UK	McNally
YE205/02	1A	O:6,30	Pig	UK	McNally
YE208/02	1A	O:5	Pig	UK	McNally
YE213/02	4	O:3	Pig	UK	McNally
YE214/02	3	O:9	Pig	UK	McNally
YE215/02	3	O:9	Pig	UK	McNally
YE218/02	3	O:9	Pig	UK	McNally
YE221/02	1A	O:6,30	Pig	UK	McNally
YE226/02	3	O:5,27	Pig	UK	McNally
YE227/02	4	O:3	Pig	UK	McNally
YE228/02	2	nd	Pig	UK	McNally
YE231/02	3	O:5,27	Pig	UK	McNally
YE232/02	3	O:5,27	Pig	UK	McNally
YE237/02	3	O:9	Pig	UK	McNally
YE24/03	3	O:9	Human control	UK	McNally
YE2702	1A	O:5	Sheep	UK	McNally
YE30/03	1A	O:5	Human case	UK	McNally
YE34/03	1A	O:5	Human control	UK	McNally
Ye3502	1A	nd	Human case	UK	McNally
YE38/03	1A	O:6,30	Human case	UK	McNally
YE41/03	1A	O:6,30	Human case	UK	McNally
YE46/02	1A	nd	Cattle	UK	McNally
YE519-36/88	5	nd	nd	nd	Fuchs
YE53/30444	1A	O:7,8	Pig	Germany	Fuchs
YE58/03	3	O:9	Human control	UK	McNally
YE69/03	1A	O:19,8	Human case	UK	McNally

Table 9.2: Worldwide strain collection of *Yersinia spp.*

Strain	Acc. No	Species	Serotype	Source	Country	Lab
ATCC35236	NZ_ACCB00000000	aldovae	nd	Drinking water	Czechoslovakia	n/a
404/81		aleksiciae	O:16	Human stool	Finland	Skurnik
127/84		bercovieri	nd	Human stool	Finland	Skurnik
3016/84		bercovieri	O:58,16	Human stool	Finland	Skurnik
ATCC43970	NZ_AALC00000000	bercovieri	nd	Human stool	France	n/a
FE80217		enterocolitica	nd	Human stool	Finland	Sihvonen
1127		enterocolitica	nd	Human	Ireland	Prentice
58735		enterocolitica	nd	Human	Ireland	Prentice
2/C/53NMD7		enterocolitica	O:9	Pig	Ireland	Prentice
3430		frederiksenii	nd	Human	Ireland	Prentice
112/02		frederiksenii	O:16	Sheep	UK	McNally
120/02		frederiksenii	O:16	Sheep	UK	McNally
22714/85		frederiksenii	nd	Human stool	Finland	Skurnik
28/85		frederiksenii	nd	nd	Finland	Skurnik
3317/84		frederiksenii	O:35	Human stool	Finland	Skurnik
3400/83		frederiksenii	O:16	Human stool	Finland	Skurnik
38/83		frederiksenii	O:48	Human stool	Finland	Skurnik
498/85		frederiksenii	nd	Human stool	Finland	Skurnik
ATCC33641	NZ_AALE00000000	frederiksenii	nd	Sewage	Denmark	n/a
BR166/97		frederiksenii	nd	nd	Finland	Skurnik
FCF208		frederiksenii	O:16	Water	Brazil	Falcao
FCF224		frederiksenii	O:16	Water	Brazil	Falcao
FCF343		frederiksenii	O:2,3	Water	Brazil	Falcao
FCF467		frederiksenii	O:16a,58	Water	Brazil	Falcao
FE80151		frederiksenii	nd	Human stool	Finland	Sihvonen
FE80258		frederiksenii	nd	Human stool	Finland	Sihvonen
FE80988		frederiksenii	nd	Human stool	Finland	Sihvonen
IP23047		frederiksenii	O3	nd	France	Skurnik
IP23698		frederiksenii	O:16	nd	France	Carniel
IP25924		frederiksenii	O:46	nd	France	Carniel
RS-42		frederiksenii	nd	Retail meat	Ireland	Prentice
58735		intermedia	nd	Human	Ireland	Prentice
109/02		intermedia	O:16	Sheep	UK	McNally
141/02		intermedia	rough	Sheep	UK	McNally
182/02		intermedia	O:52,53	Cattle	UK	McNally
821/84		intermedia	O:52,54	Human stool	Finland	Skurnik
9/85		intermedia	O:16,21	nd	Finland	Skurnik
93/02		intermedia	O:4,32	Cattle	UK	McNally
94/02		intermedia	nd	Sheep	UK	McNally
ATCC29909	NZ_AALF00000000	intermedia	nd	Human urine	nd	n/a
BR16597		intermedia	nd	nd	Finland	Skurnik
FCF130		intermedia	O:4,32	Sewage water	Brazil	Falcao
FCF202		intermedia	O:10	Sea water	Brazil	Falcao
FCF335		intermedia	O:48	Polluted water	Brazil	Falcao
FCF84		intermedia	O:4,32	Water	Brazil	Falcao
IP10066		intermedia	O:4,32	nd	France	Carniel
IP10209		intermedia	O:4,32	nd	France	Carniel
R148		intermedia	nd	Mouse	?	Fukushima
Y25		intermedia	nd	nd	?	Fukushima

542		kristensenii	nd	nd	?	Fukushima
119/84		kristensenii	O:12,25	Human stool	Finland	Skurnik
ATCC33638	NZ_ACCA00000000	kristensenii	nd	Human urine	nd	n/a
FCF221		kristensenii	O:61	Water	Brazil	Falcao
FCF324		kristensenii	O:61	Polluted water	Brazil	Falcao
FCF326		kristensenii	O:11,24	Polluted water	Brazil	Falcao
FCF580		kristensenii	O:61	Oyster	Brazil	Falcao
FE80982		kristensenii	nd	Human stool	Finland	Sihvonen
IP24139		kristensenii	Nag	nd	France	Carniel
IP6945		kristensenii	O:16	nd	France	Carniel
OK6311		kristensenii	nd	Marten	?	Fukushima
IP27925		kristensenii	O3	nd	France	Carniel
24070		massiliensis	nd	Human	Ireland	Prentice
107/02		mollaretii	nd	Sheep	UK	McNally
159/02		mollaretii	rough	Sheep	UK	McNally
61/02		mollaretii	O:1,2a,3	Sheep	UK	McNally
64/02		mollaretii	O?	Sheep	UK	McNally
92/84		mollaretii	O:59(20,36,7)	Human stool	Finland	Skurnik
ATCC43969	NZ_AALD00000000	mollaretii	nd	Soil	USA	n/a
FE82747		mollaretii	nd	Human stool	Finland	Sihvonen
IP22404		mollaretii	O3	nd	France	Carniel
IP25089		mollaretii	O20	nd	France	Carniel
A125KOH2		pekkanenii	nd	Lettuce	Finland	Skurnik
CO92	NC_003143	pestis Orientalis	na	Fatal human case	USA	Parkhill
KIM5	NC_004088	pestis Medievalis	na	nd	[Iran]	Perry
Yp91001	NC_005810	pestis Microtus	na	Vole	China	Yang
260		pseudotuberculosis	O:1a	Human	Canada	Fukushima
1231		pseudotuberculosis	O:4b	Rodent	Russia	Fukushima
2888		pseudotuberculosis	O:1a	Bird	Italy	Fukushima
D585		pseudotuberculosis	O:3	Dog	Japan	Skurnik
IH111554		pseudotuberculosis	O:1a	Cat	Finland	Siitonen
IP31758	NC_009708	pseudotuberculosis	O:1b	Human stool	former USSR	Carniel
IP32463		pseudotuberculosis	O:5	Guinea pig	Switzerland	Carniel
IP32544		pseudotuberculosis	O:3	Pig	France	Carniel
IP32670		pseudotuberculosis	O:1	Pig	UK	Carniel
IP32881		pseudotuberculosis	O:2	Monkey	Switzerland	Carniel
IP32921		pseudotuberculosis	O:2	Hare	France	Carniel
IP32938		pseudotuberculosis	O:3	Cattle	Argentina	Carniel
IP32953	NC_006155	pseudotuberculosis	O:1	Human case	France	Carniel
IP33038		pseudotuberculosis	O:1	marsupial	Australia	Carniel
IP33054		pseudotuberculosis	O:2	human	Spain	Carniel
IP33177		pseudotuberculosis	O:1	cabbage	Russia	Carniel
IP33250		pseudotuberculosis	O:3	human	Russia	Carniel
MW109-2		pseudotuberculosis	O:11	Water	nd	Fukushima
N912		pseudotuberculosis	O:2b	Rabbit	China	Fukushima
N916		pseudotuberculosis	O:13	Rat	Finland	Skurnik
No5		pseudotuberculosis	O:2b	Goat	New Zealand	Fukushima
OK5586		pseudotuberculosis	O:3	Racoon dog	Japan	Fukushima
OK6088		pseudotuberculosis	O:10	Racoon dog	Japan	Skurnik
PB1/+	NC_010634	pseudotuberculosis	nd	nd	nd	Nikolich
PT682		pseudotuberculosis	O:2b	Pig	Japan	Fukushima
R819		pseudotuberculosis	O:5b	Mouse	nd	Fukushima

WP-931201		pseudotuberculosis	O:15	Water	Korea	Fukushima
Y428		pseudotuberculosis	nd	Badger	Germany	Scholz
Y716		pseudotuberculosis	O:1a	nd	nd	Scholz
Y718		pseudotuberculosis	O:1b	Human	Germany	Scholz
Y719		pseudotuberculosis	nd	Human	Germany	Scholz
Y722		pseudotuberculosis	O:1	Human	Germany	Scholz
YPIII	NC_010465	pseudotuberculosis	nd	nd	nd	Nikolich
3343		rohdei	nd	Human	Ireland	Prentice
82589		rohdei	nd	nd	Finland	Skurnik
011/02		rohdei	O:46	Sheep	UK	McNally
56/02		rohdei	O?	Sheep	UK	McNally
65/02		rohdei	O?	Sheep	UK	McNally
68/02		rohdei	O:52	Sheep	UK	McNally
ATCC43380	NZ_ACCD00000000	rohdei	nd	Dog faeces	Germany	n/a
ATCC29473	NZ_ACCC00000000	ruckeri	nd	Trout (clinical)	Idaho_USA	n/a
OMBL4		ruckeri	nd	Fish	Finland	Skurnik
RS41		ruckeri	nd	nd	Finland	Skurnik
Y228		similis	O:6	rabbit	Germany	Scholz
Y233		similis	O:5b	Mole	Japan	Scholz
Y252		similis	OL5a	Mole	Japan	Scholz

Table 9.3: Class key

Functional Group	Description	Classification
1 – Unknown function, no homologues	Unknown function, no known homologues	0.0.0
	Conserved in E. coli	0.0.1
	Conserved in other organism than E.coli	0.0.2
2 – Cell processes	Cell processes	1.0.0
	Chemotaxis and mobility	1.1.1
	Chromosome replication	1.2.1
	Chaperones	1.3.1
3 – Protection responses	Protection responses	1.4.0
	Cell killing	1.4.1
	Detoxification	1.4.2
	Drug/analogue sensitivity	1.4.3
	Radiation sensitivity	1.4.4
4 – Transport/binding proteins	Transport/binding proteins	1.5.0
	Amino acids and amines	1.5.1
	Cations	1.5.2
	Carbohydrates, organic acids and alcohols; PTS	1.5.3
	Anions	1.5.4
	Other	1.5.5
5 – Adaptation	Adaptation	1.6.0
	Adaptations, atypical conditions	1.6.1
	Osmotic adaptation	1.6.2
	Fe storage	1.6.3
	Nodulation-related	1.6.4
6 – Cell division	Cell division	1.7.1
	Sporulation, differentiation and germination	1.8.1
7 – Macromolecule metabolism	Macromolecule metabolism	2.0.0
	Macromolecule degradation	2.1.0
	Degradation of DNA	2.1.1
	Degradation of RNA	2.1.2
	Degradation of polysaccharides	2.1.3
	Degradation of proteins, peptides, glycoproteins	2.1.4
8 – Macromolecule synthesis, modification	Macromolecule synthesis, modification	2.2.0
	Amino acyl tRNA synthesis; tRNA modification	2.2.1
	Proteins - translation and modification	2.2.10
	RNA synthesis, modification, DNA transcription	2.2.11
	tRNA	2.2.12
	Basic proteins - synthesis, modification	2.2.2
	DNA - replication, repair, restriction/modification	2.2.3
	Glycoprotein	2.2.4
	Lipopolysaccharide	2.2.5
	Lipoprotein	2.2.6
	Phospholipids	2.2.7
	Polysaccharides - (cytoplasmic)	2.2.8
	Protein modification	2.2.9
9 – Metabolism of small molecules	Metabolism of small molecules	3.0.0
	Amino acid biosynthesis	3.1.0
	Alanine	3.1.1
	Histidine	3.1.10
	Isoleucine	3.1.11
	Leucine	3.1.12

	Lysine	3.1.13
	Methionine	3.1.14
	Phenylalanine	3.1.15
	Proline	3.1.16
	Serine	3.1.17
	Threonine	3.1.18
	Tryptophan	3.1.19
	Arginine	3.1.2
	Tyrosine	3.1.20
	Valine	3.1.21
	Asparagine	3.1.3
	Aspartate	3.1.4
	Chorismate	3.1.5
	Cysteine	3.1.6
	Glutamate	3.1.7
	Glutamine	3.1.8
	Glycine	3.1.9
10 – Biosynthesis of cofactors, carriers	Biosynthesis of cofactors, carriers	3.2.0
	Acyl carrier protein (ACP)	3.2.1
	Pantothenate, CoA?	3.2.10
	Pyridine nucleotide, NAD	3.2.11
	Pyridoxine	3.2.12
	Riboflavin	3.2.13
	Thiamin	3.2.14
	Thioredoxin, glutaredoxin, glutathione	3.2.15
	Biotin carboxyl carrier protein (BCCP)	3.2.16
	Ferredoxin	3.2.17
	Biotin	3.2.2
	Cobalamin	3.2.3
	Enterochelin	3.2.4
	Folic acid	3.2.5
	Heme, porphyrin, cytochrome biogenesis	3.2.6
	Lipoate	3.2.7
	Menaquinone, ubiquinone	3.2.8
	Molybdopterin	3.2.9
11 – Central intermediary metabolism	Central intermediary metabolism	3.3.0
	2'-Deoxyribonucleotide metabolism	3.3.1
	Nucleotide hydrolysis	3.3.10
	Nucleotide interconversions	3.3.11
	Oligosaccharides	3.3.12
	Phosphorus compounds	3.3.13
	Polyamine biosynthesis, Putrescine	3.3.14
	Pool, multipurpose conversions of intermediate metabolism	3.3.15
	S-adenosyl methionine	3.3.16
	Salvage of nucleosides and nucleotides	3.3.17
	Sugar-nucleotide biosynthesis, conversions	3.3.18
	Sulfur metabolism	3.3.19
	Amino sugars	3.3.2
	Amino acids	3.3.20
	Other	3.3.21
	Nitrogen metabolism (urease)	3.3.22
	Entner-Douderoff	3.3.3

	Gluconeogenesis	3.3.4
	Glyoxylate bypass	3.3.5
	Incorporation metal ions	3.3.6
	Misc. glucose metabolism	3.3.7
	Misc. glycerol metabolism	3.3.8
	Non-oxidative branch, pentose pwy	3.3.9
12 – Degradation of small molecules	Degradation of small molecules	3.4.0
	Amines	3.4.1
	Amino acids	3.4.2
	Carbon compounds, Cachetuate catabolism	3.4.3
	Fatty acids	3.4.4
	Other	3.4.5
13 – Energy metabolism, carbon	Energy metabolism, carbon	3.5.0
	Aerobic respiration	3.5.1
	Anaerobic respiration	3.5.2
	Electron transport, cytochromes + redox	3.5.3
	Fermentation	3.5.4
	Glycolysis	3.5.5
	Oxidative branch, pentose pathway	3.5.6
	Pyruvate dehydrogenase	3.5.7
	TCA cycle	3.5.8
	ATP-proton motive force	3.5.9
14 – Fatty acid biosynthesis	Fatty acid biosynthesis	3.6.0
	Fatty acid and phosphatidic acid biosynthesis	3.6.1
15 – Nucleotide biosynthesis	Nucleotide biosynthesis	3.7.0
	Purine ribonucleotide biosynthesis	3.7.1
	Pyrimidine ribonucleotide biosynthesis	3.7.2
	Secondary metabolism	3.8.0
	Polyketide synthases	3.8.1
	Non-ribosomal peptide synthases	3.8.2
16 – Cell envelope	Cell envelope	4.0.0
	Membrane/exported/lipoproteins	4.1.0
	Inner membrane	4.1.1
	Murein sacculus, petidoglycan	4.1.2
	Outer membrane	4.1.3
	Surface polysaccharide & antigens, EPS	4.1.4
	Surface structures	4.1.5
	G+ membrane	4.1.6
	G+ exported/lipoprotein	4.1.7
	G+ surface anchored	4.1.8
	G+ peptidoglycan, teichoic acid	4.1.9
17 – Ribosome constituents	Ribosome constituents	4.2.0
	Ribosomal and stable RNAs	4.2.1
	Ribosomal proteins - synthesis, modification	4.2.2
	Ribosomes - maturation and modification	4.2.3
	Extrachromosomal	5.0.0
18 – Foreign DNA	Laterally acquirred elements	5.1.0
	Colicin-related functions	5.1.1
	Phage-related functions and prophages	5.1.2
	Plasmid-related functions	5.1.3
	Transposon-related functions	5.1.4
	Pathogenicity island-related functions	5.1.5

19 – Regulation	Regulation	6.0.0
	Two component system	6.1.0
	Sensor kinase	6.1.1
	Response regulator	6.1.2
	Sensor kinase/response regulator fusion	6.1.3
	RNA polymerase core enzyme binding	6.2.0
	sigma factor	6.2.1
	anti sigma factor	6.2.2
	anti sigma factor antagonist	6.2.3
	Defined families	6.3.0
	AsnC	6.3.1
	DeoR	6.3.10
	LuxR (GerR)	6.3.11
	MerR	6.3.12
	ArsR	6.3.13
	PadR	6.3.14
	LytR	6.3.15
	AraC	6.3.2
	GntR	6.3.3
	IclR	6.3.4
	LacI	6.3.5
	LysR	6.3.6
	MarR	6.3.7
	TetR	6.3.8
	ROK	6.3.9
	Protein kinases	6.4.0
	Serine/threonine	6.4.1
	Tyrosine	6.4.2
	Others	6.5.0
20 – Not classified	Not classified (included putative assignments)	7.0.0
	DNA sites, no gene product	7.1.1
	Cryptic genes	7.2.1

Table 9.4: Colour-coding key of CDSs.

Colour – code	Colour	Function – Chromosome	Function - Plasmid
0	White	Pathogenicity / adaptation / chaperones	Pathogenicity / adaptation
1	Grey	Energy metabolism (glycolysis, electron transport, etc.)	Partitioning and maintenance
2	Red	Information transfer (transcript-tion/translation + DNA/RNA modification	Plasmid replication
3	Green	Surface (inner membrane, outer membrane, secreted proteins, LPS, etc.)	Surface (inner membrane, outer membrane, secreted proteins, pili, etc.)
4	Dark blue	Stable RNA	Inorganic compound / metal resistance / UV, etc.
5	Bright blue	Degradation of large molecules	Conjugal transfer
6	Bright pink	Degradation of small molecules	Antibiotic resistance, surface exclusion, restriction / modification systems
7	Yellow	Central / intermediary / miscellaneous metabolism	Host metabolism
8	Pale green	Unknown / hypothetical	Unknown / hypothetical
9	Pale blue	Regulators	Regulators
10	Orange	Conserved hypothetical	Conserved hypothetical
11	Brown	Pseudogenes and partial genes (remnants)	Pseudogenes and partial genes (remnants)
12	Pale pink	Phage / IS elements	IS elements
13	Pale grey	Some miscellaneous information, e.g. prosite, but no function	Some miscellaneous information, e.g. prosite, but no function

Appendix 9.5: perl script for identification of sets of orthologous genes in seven reference genomes for *Y. enterocolitica*

```
#!/usr/local/bin/perl -w

use strict;
use Bio::PSU::SeqFactory;

my $file = shift;
my $seqi = Bio::PSU::SeqFactory->make(-file   => $file,
                                      -format => 'embl');
my $obj=Bio::PSU::IO::FTHandler->new;

my (@all_gen_ortho, @gen1_gen2_gen3_gen4_gen5_ortho, @gen1_gen2_gen3_gen4_gen6_ortho,
@gen1_gen2_gen3_gen5_gen6_ortho, @gen1_gen2_gen4_gen5_gen6_ortho, @gen1_gen3_gen4_gen5_gen6_ortho,
@gen2_gen3_gen4_gen5_gen6_ortho, @gen1_gen2_gen3_gen4_ortho, @gen1_gen2_gen3_gen5_ortho,
@gen1_gen2_gen3_gen6_ortho, @gen1_gen2_gen4_gen5_ortho, @gen1_gen2_gen4_gen6_ortho, @gen1_gen2_gen5_gen6_ortho,
@gen1_gen3_gen4_gen5_ortho, @gen1_gen3_gen4_gen6_ortho, @gen1_gen3_gen5_gen6_ortho, @gen1_gen4_gen5_gen6_ortho,
@gen2_gen3_gen4_gen5_ortho, @gen2_gen3_gen4_gen6_ortho, @gen2_gen3_gen5_gen6_ortho, @gen2_gen4_gen5_gen6_ortho,
@gen3_gen4_gen5_gen6_ortho, @gen1_gen2_gen3_ortho, @gen1_gen2_gen4_ortho, @gen1_gen2_gen5_ortho,
@gen1_gen2_gen6_ortho, @gen1_gen3_gen4_ortho, @gen1_gen3_gen5_ortho, @gen1_gen3_gen6_ortho, @gen1_gen4_gen5_ortho,
@gen1_gen4_gen6_ortho, @gen1_gen5_gen6_ortho, @gen2_gen3_gen4_ortho, @gen2_gen3_gen5_ortho, @gen2_gen3_gen6_ortho,
@gen2_gen4_gen5_ortho, @gen2_gen4_gen6_ortho, @gen2_gen5_gen6_ortho, @gen3_gen4_gen5_ortho, @gen3_gen4_gen6_ortho,
@gen3_gen5_gen6_ortho, @gen4_gen5_gen6_ortho, @gen1_gen2_ortho, @gen1_gen3_ortho, @gen1_gen4_ortho, @gen1_gen5_ortho,
@gen1_gen6_ortho, @gen2_gen3_ortho, @gen2_gen4_ortho, @gen2_gen5_ortho, @gen2_gen6_ortho, @gen3_gen4_ortho,
@gen3_gen5_ortho, @gen3_gen6_ortho, @gen4_gen5_ortho, @gen4_gen6_ortho, @gen5_gen6_ortho, @gen1_only_ortho,
@gen2_only_ortho,@gen3_only_ortho, @gen4_only_ortho, @gen5_only_ortho, @gen6_only_ortho, @no_ortho, $feature );

# genome1 first comparator genome (gen1), second comparator genome (gen2) are those you are comparing to your genome $gen1 and
$gen2 relate to the identifiers you have used to store the orthologue information. You must also fill in your generic gename pattern YE****
and the qualifier under which it is held eg locus_tag or systematic_id.
# script extended to generate numbers / groups for comparison of seven different genomes.

my $gen1= "8081" . "_orthologue";
my $gen2= "21202" . "_orthologue";
my $gen3= "5603" . "_orthologue";
my $gen4= "14902" . "_orthologue";
my $gen5= "1203" . "_orthologue";
my $gen6= "3094" . "_orthologue";
my $genid= "YE5303-";
my $genqualifier="locus_tag";
#my $genqualifier="systematic_id";

        while (my $seq = $seqi->next_seq)
            {
                    if (my $fnum = $seq->has_features)
                    {
                       foreach $feature ($seq->features)
                       {
                              if ($feature->key() eq "CDS")
                              {
                                 if ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
                                          $feature->qexists("$gen3") and $feature->qexists("$gen4") and
```

```
                $feature->qexists("$gen5") and $feature->qexists("$gen6")) {
        push (@all_gen_ortho, $feature);

} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
                $feature->qexists("$gen3") and $feature->qexists("$gen4") and
                $feature->qexists("$gen5")) {
        push (@gen1_gen2_gen3_gen4_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
                $feature->qexists("$gen3") and $feature->qexists("$gen4") and
                $feature->qexists("$gen6")) {
        push (@gen1_gen2_gen3_gen4_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
                $feature->qexists("$gen3") and $feature->qexists("$gen5") and
                $feature->qexists("$gen6")) {
        push (@gen1_gen2_gen3_gen5_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
                $feature->qexists("$gen4") and $feature->qexists("$gen5") and
                $feature->qexists("$gen6")) {
        push (@gen1_gen2_gen4_gen5_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen3") and
                $feature->qexists("$gen4") and $feature->qexists("$gen5") and
                $feature->qexists("$gen6")) {
        push (@gen1_gen3_gen4_gen5_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen3") and
                $feature->qexists("$gen4") and $feature->qexists("$gen5") and
                $feature->qexists("$gen6")) {
        push (@gen2_gen3_gen4_gen5_gen6_ortho, $feature);

} elsif  ($feature->qexists("$gen1") and  $feature->qexists("$gen2") and
                 $feature->qexists("$gen3") and $feature->qexists("$gen4")) {
        push (@gen1_gen2_gen3_gen4_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
                $feature->qexists("$gen3") and $feature->qexists("$gen5")) {
        push (@gen1_gen2_gen3_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
                $feature->qexists("$gen3") and $feature->qexists("$gen6")) {
        push (@gen1_gen2_gen3_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
                $feature->qexists("$gen4") and $feature->qexists("$gen5")) {
        push (@gen1_gen2_gen4_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
                $feature->qexists("$gen4") and $feature->qexists("$gen6")) {
        push (@gen1_gen2_gen4_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
                $feature->qexists("$gen5") and $feature->qexists("$gen6")) {
        push (@gen1_gen2_gen5_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen3") and
                $feature->qexists("$gen4") and $feature->qexists("$gen5")) {
        push (@gen1_gen3_gen4_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen3") and
                $feature->qexists("$gen4") and $feature->qexists("$gen6")) {
        push (@gen1_gen3_gen4_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen3") and
```

```
                $feature->qexists("$gen5") and $feature->qexists("$gen6")) {
        push (@gen1_gen3_gen5_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen4") and
                $feature->qexists("$gen5") and $feature->qexists("$gen6")) {
        push (@gen1_gen4_gen5_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen3") and
                $feature->qexists("$gen4") and $feature->qexists("$gen5")) {
        push (@gen2_gen3_gen4_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen3") and
                $feature->qexists("$gen4") and $feature->qexists("$gen6")) {
        push (@gen2_gen3_gen4_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen3") and
                $feature->qexists("$gen5") and $feature->qexists("$gen6")) {
        push (@gen2_gen3_gen5_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen4") and
                $feature->qexists("$gen5") and $feature->qexists("$gen6")) {
        push (@gen2_gen4_gen5_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen3") and $feature->qexists("$gen4") and
                $feature->qexists("$gen5") and $feature->qexists("$gen6")) {
        push (@gen3_gen4_gen5_gen6_ortho, $feature);

} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
        $feature->qexists("$gen3")) {
          push (@gen1_gen2_gen3_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
        $feature->qexists("$gen4")) {
          push (@gen1_gen2_gen4_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
        $feature->qexists("$gen5")) {
          push (@gen1_gen2_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2") and
        $feature->qexists("$gen6")) {
          push (@gen1_gen2_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen3") and
        $feature->qexists("$gen4")) {
          push (@gen1_gen3_gen4_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen3") and
        $feature->qexists("$gen5")) {
          push (@gen1_gen3_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen3") and
        $feature->qexists("$gen6")) {
          push (@gen1_gen3_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen4") and
        $feature->qexists("$gen5")) {
          push (@gen1_gen4_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen4") and
        $feature->qexists("$gen6")) {
          push (@gen1_gen4_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen5") and
        $feature->qexists("$gen6")) {
          push (@gen1_gen5_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen3") and
         $feature->qexists("$gen4")) {
```

```
            push (@gen2_gen3_gen4_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen3") and
        $feature->qexists("$gen5")) {
            push (@gen2_gen3_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen3") and
        $feature->qexists("$gen6")) {
            push (@gen2_gen3_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen4") and
        $feature->qexists("$gen5")) {
            push (@gen2_gen4_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen4") and
        $feature->qexists("$gen6")) {
            push (@gen2_gen4_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen5") and
        $feature->qexists("$gen6")) {
            push (@gen2_gen5_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen3") and $feature->qexists("$gen4") and
        $feature->qexists("$gen5")) {
            push (@gen3_gen4_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen3") and $feature->qexists("$gen4") and
        $feature->qexists("$gen6")) {
            push (@gen3_gen4_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen3") and $feature->qexists("$gen5") and
        $feature->qexists("$gen6")) {
            push (@gen3_gen5_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen4") and $feature->qexists("$gen5") and
        $feature->qexists("$gen6")) {
            push (@gen4_gen5_gen6_ortho, $feature);

} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen2")) {
             push (@gen1_gen2_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen3")) {
            push (@gen1_gen3_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen4")) {
            push (@gen1_gen4_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen5")) {
            push (@gen1_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen1") and $feature->qexists("$gen6")) {
            push (@gen1_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen3")) {
            push (@gen2_gen3_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen4")) {
            push (@gen2_gen4_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen5")) {
            push (@gen2_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen2") and $feature->qexists("$gen6")) {
            push (@gen2_gen6_ortho, $feature);
} elsif ($feature->qexists("$gen3") and $feature->qexists("$gen4")) {
            push (@gen3_gen4_ortho, $feature);
} elsif ($feature->qexists("$gen3") and $feature->qexists("$gen5")) {
            push (@gen3_gen5_ortho, $feature);
} elsif ($feature->qexists("$gen3") and $feature->qexists("$gen6")) {
            push (@gen3_gen6_ortho, $feature);
```

```
                                        } elsif ($feature->qexists("$gen4") and $feature->qexists("$gen5")) {
                                                push (@gen4_gen5_ortho, $feature);
                                        } elsif ($feature->qexists("$gen4") and $feature->qexists("$gen6")) {
                                                push (@gen4_gen6_ortho, $feature);
                                        } elsif ($feature->qexists("$gen5") and $feature->qexists("$gen6")) {
                                                push (@gen5_gen6_ortho, $feature);

                                        } elsif  ($feature->qexists("$gen1") xor $feature->qexists("$gen2")
                                               xor $feature->qexists("$gen3") xor $feature->qexists("$gen4")
                                               xor $feature->qexists("$gen5") xor $feature->qexists("$gen6")) {

                                                if ($feature->qexists("$gen1")) {
                                                   push (@gen1_only_ortho, $feature);
                                                } elsif ($feature->qexists("$gen2")) {
                                                   push (@gen2_only_ortho, $feature);
                                                } elsif ($feature->qexists("$gen3")) {
                                                   push (@gen3_only_ortho, $feature);
                                                } elsif ($feature->qexists("$gen4")) {
                                                   push (@gen4_only_ortho, $feature);
                                                } elsif ($feature->qexists("$gen5")) {
                                                    push (@gen5_only_ortho, $feature);
                                                } elsif ($feature->qexists("$gen6")) {
                                                    push (@gen6_only_ortho, $feature);
                                                }
                                        } else {
                                                   push (@no_ortho, $feature);
                                                }
                                    }
                                  }
                                }
                            }
my $n=0;
my $m=0;
my $l=0;
my $y=0;

#print "*****genes shared with all genomes*******:\n";
foreach my $feature1 (@all_gen_ortho) {
$feature1->qadd('group', 1);
   print $obj->make_feature_block($feature1);}

#print "*****genes shared with genome1 ($gen1), genome2 ($gen2), genome3 ($gen3), genome4 ($gen4) and genome5 ($gen5)*****:\n";
foreach my $feature2 (@gen1_gen2_gen3_gen4_gen5_ortho) {
  $feature2->qadd('group', 2);
   print $obj->make_feature_block($feature2);}

#print "*****genes shared with genome1 ($gen1), genome2 ($gen2), genome3 ($gen3), genome4 ($gen4) and genome6 ($gen6)*****:\n";
foreach my $feature3 (@gen1_gen2_gen3_gen4_gen6_ortho) {
  $feature3->qadd('group', 3);
   print $obj->make_feature_block($feature3);}

#print "*****genes shared with genome1 ($gen1), genome2 ($gen2), genome3 ($gen3), genome5 ($gen5) and genome6 ($gen6)*****:\n";
foreach my $feature4 (@gen1_gen2_gen3_gen5_gen6_ortho) {
```

```
  $feature4->qadd('group', 4);
   print $obj->make_feature_block($feature4);}

#print "*****genes shared with genome1 ($gen1), genome2 ($gen2), genome4 ($gen4), genome5 ($gen5) and genome6 ($gen6)*****:\n";
foreach my $feature5 (@gen1_gen2_gen4_gen5_gen6_ortho) {
  $feature5->qadd('group', 5);
   print $obj->make_feature_block($feature5);}

#print "*****genes shared with genome1 ($gen1), genome3 ($gen3), genome4 ($gen4), genome5 ($gen5) and genome6 ($gen6)*****:\n";
foreach my $feature6 (@gen1_gen3_gen4_gen5_gen6_ortho) {
  $feature6->qadd('group', 6);
   print $obj->make_feature_block($feature6);}

#print "*****genes shared with genome2 ($gen2), genome3 ($gen3), genome4 ($gen4), genome5 ($gen5) and genome6 ($gen6)*****:\n";
foreach my $feature7 (@gen2_gen3_gen4_gen5_gen6_ortho) {
  $feature7->qadd('group', 7);
   print $obj->make_feature_block($feature7);}

#print "*****genes shared with genome1 ($gen1), genome2 ($gen2), genome3 ($gen3) and genome4 ($gen4)*****:\n";
foreach my $feature8 (@gen1_gen2_gen3_gen4_ortho) {
  $feature8->qadd('group', 8);
   print $obj->make_feature_block($feature8);}

#print "*****genes shared with genome1 ($gen1), genome2 ($gen2), genome3 ($gen3) and genome5 ($gen5)*****:\n";
foreach my $feature9 (@gen1_gen2_gen3_gen5_ortho) {
  $feature9->qadd('group', 9);
   print $obj->make_feature_block($feature9);}

#print "*****genes shared with genome1 ($gen1), genome2 ($gen2), genome3 ($gen3) and genome6 ($gen6)*****:\n";
foreach my $feature10 (@gen1_gen2_gen3_gen6_ortho) {
  $feature10->qadd('group', 10);
   print $obj->make_feature_block($feature10);}

#print "*****genes shared with genome1 ($gen1), genome2 ($gen2), genome4 ($gen4) and genome5 ($gen5)*****:\n";
foreach my $feature11 (@gen1_gen2_gen4_gen5_ortho) {
  $feature11->qadd('group', 11);
   print $obj->make_feature_block($feature11);}

#print "*****genes shared with genome1 ($gen1), genome2 ($gen2), genome4 ($gen4) and genome6 ($gen6)*****:\n";
foreach my $feature12 (@gen1_gen2_gen4_gen6_ortho) {
  $feature12->qadd('group', 12);
   print $obj->make_feature_block($feature12);}

#print "*****genes shared with genome1 ($gen1), genome2 ($gen2), genome5 ($gen5) and genome6 ($gen6)*****:\n";
foreach my $feature13 (@gen1_gen2_gen5_gen6_ortho) {
   $feature13->qadd('group', 13);
   print $obj->make_feature_block($feature13);}

#print "*****genes shared with genome1 ($gen1), genome3 ($gen3), genome4 ($gen4) and genome5 ($gen5)*****:\n;
foreach my $feature14 (@gen1_gen3_gen4_gen5_ortho) {
   $feature14->qadd('group', 14);
   print $obj->make_feature_block($feature14);}
```

```
#print "*****genes shared with genome1 ($gen1), genome3 ($gen3), genome4 ($gen4) and genome6 ($gen6)*****:\n";
foreach my $feature15 (@gen1_gen3_gen4_gen6_ortho) {
   $feature15->qadd('group', 15);
   print $obj->make_feature_block($feature15);}

#print "*****genes shared with genome1 ($gen1), genome3 ($gen3), genome5 ($gen5) and genome6 ($gen6)*****:\n";
foreach my $feature16 (@gen1_gen3_gen5_gen6_ortho) {
   $feature16->qadd('group', 16);
   print $obj->make_feature_block($feature16);}

#print "*****genes shared with genome1 ($gen1), genome4 ($gen4), genome5 ($gen5) and genome6 ($gen6)*****:\n";
foreach my $feature17 (@gen1_gen4_gen5_gen6_ortho) {
   $feature17->qadd('group', 17);
   print $obj->make_feature_block($feature17);}

#print "*****genes shared with genome2 ($gen2), genome3 ($gen3), genome4 ($gen4) and genome5 ($gen5)********:\n";
foreach my $feature18 (@gen2_gen3_gen4_gen5_ortho) {
   $feature18->qadd('group', 18);
   print $obj->make_feature_block($feature18);}

#print "*****genes shared with genome2 ($gen2), genome3 ($gen3), genome4 ($gen4) and genome6 ($gen6)********:\n";
foreach my $feature19 (@gen2_gen3_gen4_gen6_ortho) {
   $feature19->qadd('group', 19);
   print $obj->make_feature_block($feature19);}

#print "*****genes shared with genome2 ($gen2), genome3 ($gen3), genome5 ($gen5) and genome6 ($gen6)********:\n";
foreach my $feature20 (@gen2_gen3_gen5_gen6_ortho) {
   $feature20->qadd('group', 20);
   print $obj->make_feature_block($feature20);}

#print "*****genes shared with genome2 ($gen2), genome4 ($gen4), genome5 ($gen5) and genome6 ($gen6)********:\n";
foreach my $feature21 (@gen2_gen4_gen5_gen6_ortho) {
   $feature21->qadd('group', 21);
   print $obj->make_feature_block($feature21);}

#print "*****genes shared with genome3 ($gen3), genome4 ($gen4), genome5 ($gen5) and genome6 ($gen6)********:\n";
foreach my $feature22 (@gen3_gen4_gen5_gen6_ortho) {
   $feature22->qadd('group', 22);
   print $obj->make_feature_block($feature22);}

#print "**********genes shared with genome1 ($gen1), genome2 ($gen2) and genome3 ($gen3)**********:\n";
foreach my $feature23 (@gen1_gen2_gen3_ortho) {
   $feature23->qadd('group', 23);
   print $obj->make_feature_block($feature23);}

#print "**********genes shared with genome1 ( $gen1 ), genome2 ( $gen2 ) and genome4 ($gen4)**********:\n";
foreach my $feature24 (@gen1_gen2_gen4_ortho) {
   $feature24->qadd('group', 24);
   print $obj->make_feature_block($feature24);}

#print "**********genes shared with genome1 ($gen1), genome2 ($gen2) and genome5 ($gen5)**********:\n";
foreach my $feature25 (@gen1_gen2_gen5_ortho) {
   $feature25->qadd('group', 25);
```

```
    print $obj->make_feature_block($feature25);}

#print "**********genes shared with genome1 ($gen1), genome2 ($gen2) and genome6 ($gen6)*********:\n";
foreach my $feature26 (@gen1_gen2_gen6_ortho) {
    $feature26->qadd('group', 26);
    print $obj->make_feature_block($feature26);}

#print "**********genes shared with genome1 ($gen1), genome3 ($gen3) and genome4 ($gen4)*********:\n";
foreach my $feature27 (@gen1_gen3_gen4_ortho) {
    $feature27->qadd('group', 27);
    print $obj->make_feature_block($feature27);}

#print "**********genes shared with genome1 ($gen1), genome3 ($gen3) and genome5 ($gen5)*********:\n";
foreach my $feature28 (@gen1_gen3_gen5_ortho) {
    $feature28->qadd('group', 28);
    print $obj->make_feature_block($feature28);}

#print "**********genes shared with genome1 ($gen1), genome3 ($gen3) and genome6 ($gen6)*********:\n";
foreach my $feature29 (@gen1_gen3_gen6_ortho) {
    $feature29->qadd('group', 29);
    print $obj->make_feature_block($feature29);}

#print "**********genes shared with genome1 ($gen1), genome4 ($gen4) and genome5 ($gen5)*********:\n";
foreach my $feature30 (@gen1_gen4_gen5_ortho) {
    $feature30->qadd('group', 30);
    print $obj->make_feature_block($feature30);}

#print "**********genes shared with genome1 ($gen1), genome4 ($gen4) and genome6 ($gen6)*********:\n";
foreach my $feature31 (@gen1_gen4_gen6_ortho) {
    $feature31->qadd('group', 31);
    print $obj->make_feature_block($feature31);}

#print "**********genes shared with genome1 ($gen1), genome5 ($gen5) and genome6 ($gen6)*********:\n";
foreach my $feature32 (@gen1_gen5_gen6_ortho) {
    $feature32->qadd('group', 32);
    print $obj->make_feature_block($feature32);}

#print "**********genes shared with genome2 ($gen2), genome3 ($gen3) and genome4 ($gen4)*********:\n";
foreach my $feature33 (@gen2_gen3_gen4_ortho) {
    $feature33->qadd('group', 33);
    print $obj->make_feature_block($feature33);}

#print "**********genes shared with genome2 ($gen2), genome3 ($gen3) and genome5 ($gen5)*********:\n";
foreach my $feature34 (@gen2_gen3_gen5_ortho) {
    $feature34->qadd('group', 34);
    print $obj->make_feature_block($feature34);}

#print "**********genes shared with genome2 ($gen2), genome3 ($gen3) and genome6 ($gen6)*********:\n";
foreach my $feature35 (@gen2_gen3_gen6_ortho) {
    $feature35->qadd('group', 35);
    print $obj->make_feature_block($feature35);}

#print "**********genes shared with genome2 ($gen2), genome4 ($gen4) and genome5 ($gen5)*********:\n";
```

```
foreach my $feature36 (@gen2_gen4_gen5_ortho) {
   $feature36->qadd('group', 36);
   print $obj->make_feature_block($feature36);}

#print "**********genes shared with genome2 ($gen2), genome4 ($gen4) and genome6 ($gen6)**********:\n";
foreach my $feature37 (@gen2_gen4_gen6_ortho) {
   $feature37->qadd('group', 37);
   print $obj->make_feature_block($feature37);}

#print "**********genes shared with genome2 ($gen2), genome5 ($gen5) and genome6 ($gen6)**********:\n";
foreach my $feature38 (@gen2_gen5_gen6_ortho) {
   $feature38->qadd('group', 38);
   print $obj->make_feature_block($feature38);}

#print "**********genes shared with genome3 ($gen3), genome4 ($gen4) and genome5 ($gen5)**********:\n";
foreach  my $feature39 (@gen3_gen4_gen5_ortho) {
   $feature39->qadd('group', 39);
   print $obj->make_feature_block($feature39);}

#print "**********genes shared with genome3 ($gen3), genome4 ($gen4) and genome6 ($gen6)**********:\n";
foreach  my $feature40 (@gen3_gen4_gen6_ortho) {
   $feature40->qadd('group', 40);
   print $obj->make_feature_block($feature40);}

#print "**********genes shared with genome3 ($gen3), genome5 ($gen5) and genome6 ($gen6)**********:\n";
foreach  my $feature41 (@gen3_gen5_gen6_ortho) {
   $feature41->qadd('group', 41);
   print $obj->make_feature_block($feature41);}

#print "**********genes shared with genome4 ($gen4), genome5 ($gen5) and genome6 ($gen6)**********:\n";
foreach  my $feature42 (@gen4_gen5_gen6_ortho) {
   $feature42->qadd('group', 42);
   print $obj->make_feature_block($feature42);}

#print "**********genes shared with genome1 ( $gen1 ) and genome2 ( $gen2 ) **********:\n";
foreach my $feature43 (@gen1_gen2_ortho) {
   $feature43->qadd('group', 43);
   print $obj->make_feature_block($feature43);}

#print "**********genes shared with genome1 ($gen1) and genome3 ($gen3)**********:\n";
foreach my $feature44 (@gen1_gen3_ortho) {
   $feature44->qadd('group', 44);
   print $obj->make_feature_block($feature44);}

#print "**********genes shared with genome1 ($gen1) and genome4 ($gen4)**********:\n";
foreach my $feature45 (@gen1_gen4_ortho) {
   $feature45->qadd('group', 45);
   print $obj->make_feature_block($feature45);}

#print "**********genes shared with genome1 ($gen1) and genome5 ($gen5)**********:\n";
foreach my $feature46 (@gen1_gen5_ortho) {
   $feature46->qadd('group', 46);
   print $obj->make_feature_block($feature46);}
```

```
#print "*****************genes shared with genome1 ($gen1) and genome6 ($gen6)*****************:\n";
foreach my $feature47 (@gen1_gen6_ortho) {
   $feature47->qadd('group', 47);
   print $obj->make_feature_block($feature47);}

#print "*****************genes shared with genome2 ($gen2) and genome3 ($gen3)*****************:\n";
foreach my $feature48 (@gen2_gen3_ortho) {
   $feature48->qadd('group', 48);
   print $obj->make_feature_block($feature48);}

#print "*****************genes shared with genome2 ($gen2) and genome4 ($gen4)*****************:\n";
foreach my $feature49 (@gen2_gen4_ortho) {
   $feature49->qadd('group', 49);
   print $obj->make_feature_block($feature49);}

#print "*****************genes shared with genome2 ($gen2) and genome5 ($gen5)*****************:\n";
foreach my $feature50 (@gen2_gen5_ortho) {
   $feature50->qadd('group', 50);
   print $obj->make_feature_block($feature50);}

#print "*****************genes shared with genome2 ($gen2) and genome6 ($gen6)*****************:\n";
foreach my $feature51 (@gen2_gen6_ortho) {
   $feature51->qadd('group', 51);
   print $obj->make_feature_block($feature51);}

#print "*****************genes shared with genome3 ($gen3) and genome4 ($gen4)*****************:\n";
foreach my $feature52 (@gen3_gen4_ortho) {
   $feature52->qadd('group', 52);
   print $obj->make_feature_block($feature52);}

#print "*****************genes shared with genome3 ($gen3) and genome5 ($gen5)*****************:\n";
foreach my $feature53 (@gen3_gen5_ortho) {
   $feature53->qadd('group', 53);
   print $obj->make_feature_block($feature53);}

#print "*****************genes shared with genome3 ($gen3) and genome6 ($gen6)*****************:\n";
foreach my $feature54 (@gen3_gen6_ortho) {
   $feature54->qadd('group', 54);
   print $obj->make_feature_block($feature54);}

#print "*****************genes shared with genome4 ($gen4) and genome5 ($gen5)*****************:\n";
foreach my $feature55 (@gen4_gen5_ortho) {
   $feature55->qadd('group', 55);
   print $obj->make_feature_block($feature55);}

#print "*****************genes shared with genome4 ($gen4) and genome6 ($gen6)*****************:\n";
foreach my $feature56 (@gen4_gen6_ortho) {
   $feature56->qadd('group', 56);
   print $obj->make_feature_block($feature56);}

#print "*****************genes shared with genome5 ($gen5) and genome6 ($gen6)*****************:\n";
foreach my $feature57 (@gen5_gen6_ortho) {
```

```
    $feature57->qadd('group', 57);
    print $obj->make_feature_block($feature57);}

#print "*****************genes shared with genome 1 ( $gen1 ) only*****************************:\n";
foreach  my $feature58 (@gen1_only_ortho) {
    $feature58->qadd('group', 58);
    print $obj->make_feature_block($feature58);}

#print "****************genes shared with genome 2 ( $gen2 )only *****************************:\n";
foreach my $feature59 (@gen2_only_ortho) {
    $feature59->qadd('group', 59);
    print $obj->make_feature_block($feature59);}

#print "*****************genes shared with genome 3 ($gen3) only************************:\n";
foreach my $feature60 (@gen3_only_ortho) {
    $feature60->qadd('group', 60);
    print $obj->make_feature_block($feature60);}

#print "*****************genes shared with genome 4 ($gen4) only************************:\n";
foreach my $feature61 (@gen4_only_ortho) {
    $feature61->qadd('group', 61);
    print $obj->make_feature_block($feature61);}

#print "*****************genes shared with genome 5 ($gen5) only************************:\n";
foreach my $feature62 (@gen5_only_ortho) {
    $feature62->qadd('group', 62);
    print $obj->make_feature_block($feature62);}

#print "*****************genes shared with genome 6 ($gen6) only************************:\n";
foreach my $feature63 (@gen6_only_ortho) {
    $feature63->qadd('group', 63);
    print $obj->make_feature_block($feature63);}

#print "******************Your genome ( $genid ) unique genes *****************************:\n";
foreach my $feature64 (@no_ortho) {
    $feature64->qadd('group', 64);
    print $obj->make_feature_block($feature64);}

#print "*******************************************************************************\n";

#print "there are $n unique ( $genid ) genes \n";
#print "there are $m genes shared with genome1( $gen1 ) and genome2 ( $gen2 )\n";
#print "there are $l genes shared with genome1 ( $gen1 ) only \n";
#print "there are $y genes shared with genome2 ( $gen2 ) only \n";
```

Table 9.6: Matrix of orthologous gene sets.

		5303	8081	21202	5603	14902	1203	3094
	genome1	8081	5303	5303	5303	5303	5303	5303
	genome2	21202	21202	8081	8081	8081	8081	8081
	genome3	5603	5603	5603	21202	21202	21202	21202
	genome4	14902	14902	14902	14902	5603	5603	5603
	genome5	1203	1203	1203	1203	1203	14902	14902
	genome6	3094	3094	3094	3094	3094	3094	1203
groups	**content**	**# CDS**	**# CDS**	**# CDS**	**# CDS**	**# CDS**	**# CDS**	**# CDS**
1	**all_gen_ortho**	**3306**	**3307**	**3307**	**3305**	**3305**	**3307**	**3306**
2	gen1+2+3+4+5	31	32	32	32	33	33	18
3	gen1+2+3+4+6	17	18	18	19	18	0	0
4	gen1+2+3+5+6	0	0	0	0	0	0	0
5	gen1+2+4+5+6	0	0	0	0	1	0	0
6	gen1+3+4+5+6	0	0	121	119	118	119	120
7	gen2+3+4+5+6	118	39	40	40	40	40	42
8	gen1+2+3+4	9	8	8	8	8	1	2
9	gen1+2+3+5	1	1	1	1	0	0	0
10	gen1+2+3+6	2	2	3	2	0	0	0
11	gen1+2+4+5	0	0	0	0	0	0	0
12	gen1+2+4+6	0	0	0	0	0	0	0
13	gen1+2+5+6	0	0	0	0	2	2	2
14	gen1+3+4+5	0	0	2	2	2	2	3
15	gen1+3+4+6	0	0	3	4	3	0	0
16	gen1+3+5+6	0	0	0	0	0	0	0
17	gen1+4+5+6	2	2	0	0	0	0	0
18	gen2+3+4+5	3	4	4	4	4	4	0
19	gen2+3+4+6	3	0	0	0	0	0	0
20	gen2+3+5+6	2	0	0	0	0	1	0
21	gen2+4+5+6	0	1	0	0	1	0	0
22	gen3+4+5+6	0	0	154	153	156	153	155
23	gen1+2+3	0	0	0	0	0	0	0
24	gen1+2+4	0	0	0	0	0	0	0
25	gen1+2+5	0	0	0	0	2	2	0
26	gen1+2+6	0	0	0	0	0	0	0
27	gen1+3+4	0	0	0	1	1	0	1
28	gen1+3+5	0	0	0	0	0	0	0
29	gen1+3+6	0	0	1	1	0	0	0
30	gen1+4+5	2	2	0	0	0	0	0
31	gen1+4+6	0	0	0	0	0	0	0
32	gen1+5+6	0	0	0	0	0	0	0

33	gen2+3+4	0	1	2	2	2	1	3
34	gen2+3+5	0	1	1	1	0	0	0
35	gen2+3+6	0	5	4	5	0	0	0
36	gen2+4+5	0	0	0	0	0	0	0
37	gen2+4+6	0	0	0	0	0	0	0
38	gen2+5+6	0	0	0	0	0	0	0
39	gen3+4+5	0	0	8	8	9	6	24
40	gen3+4+6	0	0	22	24	21	15	13
41	gen3+5+6	0	0	16	16	0	2	0
42	gen4+5+6	0	0	0	1	0	1	1
43	gen1+2	0	0	0	0	0	0	0
44	gen1+3	0	0	0	0	0	0	0
45	gen1+4	0	0	0	0	0	0	0
46	gen1+5	0	0	0	0	0	0	1
47	gen1+6	0	1	0	0	1	0	0
48	gen2+3	0	2	1	1	0	0	0
49	gen2+4	0	0	0	0	0	0	0
50	gen2+5	0	0	0	0	0	0	0
51	gen2+6	1	0	0	0	0	1	1
52	gen3+4	0	0	37	37	36	18	8
53	gen3+5	0	0	16	16	0	0	0
54	gen3+6	0	0	8	8	0	0	0
55	gen4+5	0	0	0	0	0	0	0
56	gen4+6	1	0	0	0	0	0	0
57	gen5+6	0	1	0	0	5	8	8
58	gen1 only	82	82	0	0	15	2	32
59	gen2 only	0	0	0	0	1	1	2
60	gen3 only	0	0	106	106	0	2	0
61	gen4 only	17	0	0	0	0	2	0
62	gen5 only	3	1	2	1	11	11	9
63	gen6 only	37	3	0	0	10	25	26
64	**no ortho**	**728**	**540**	**5**	**0**	**171**	**134**	**274**
	total CDS	**4365**	**4053**	**3922**	**3917**	**3978**	**3893**	**4051**

Table 9.7: Pseudogenes and their orthologs in the *Y. enterocolitica* reference genomes.

Green – functional gene, light brown – pseudogene with frameshift or stop codon, * - UGA / opal stop codon, dark brown – partial gene (missing N, C, both termini or central part or protein sequence), - - no ortholog in this isolate. Numbers given are the unique CDS identifiers. Class assignment as per Table 9.1.

1B - 8081	1A - YE5303	2 - YE21202	3 - YE5603	3 - YE14902	4 - YE1203	5 - YE3094	class	product
YE0010	40761	31211	20091	16071	16741	00091	1.5.3	putative sugar transport system, permease protein
YE0012A	40791	31241	20061	16041	16711	00121	6.1.1	ribose operon repressor
YE0036	43161	31491	39311	38920	00251	-	6.1.1	LuxR family transcription regulatory protein
YE0037	-	31491	39301	38921	02241	00372	6.1.1	LuxR family transcription regulatory protein
YE0048	43071	31591	25351	38821	00351	00471	2.2.3	putative DNA ligase
YE0050	43051	31611	25371	38801	00371	00491	4.1.1	putative membrane protein
-	-	-	-	38721	00471	-	5.1.4	transposase for IS1668
-	-	31841	25601	38561	00671	00841	0.0.2	hypothetical membrane protein
-	-	31851	25611	38551	00691	00861	0.0.2	hypothetical membrane protein
-	-	-	-	-	00981	-	5.1.2	putative phage tail fibre assembly protein
-	-	32081	25831	38331	01201	01091	1.4.1	putative nematicidal protein
-	-	32091	25841	38321	01211	01111	5.1.4	putative transposase
-	-	-	-	-	17351	01131	5.1.4	putative transposase for IS1667
-	-	30481	20801	16791	17441	01221	4.1.0	putative exported protein
YE0098	42491	32111	25861	38291	01251	40591	3.3.0	putative fructose 1,6-bisphosphatase
YE0122	42281	32330	26071	38081	01481	40371	1.5.5	putative TonB dependent receptor protein
YE0123	42271	32331	26081	38071	01491	40361	1.5.5	hemophore hasA
YE0124	42261	32341	26091	-	01501	-	1.5.5	hemophore hasA
YE0129	42211	32391	26141	38021	01561	40311	4.1.1	putative membrane protein
YE0130	-	-	-	-	-	40301	5.1.4	transposase for IS1330
-	-	32671	17501	36731	-	-	1.4.0	putative pyocin S2
-	-	32621	17451	36681	01851	40071	0.0.0	hypothetical protein
YE0158	41021	32721	17551	36781	01951	39991	4.1.0	putative exported protein
YE0159	41031	32731	17561	36791	01961	39971	4.1.3	putative outer membrane usher protein
YE0207	41481	33191	18011	37251	02461	39391	3.3.13	putative lysophospholipase
-	-	33391	18221	37461	02681	39191	0.0.0	hypothetical protein
-	-	-	-	37471	-	-	5.1.4	putative transposase
YE0241	41801	33541	18351	37601	02831	29051	1.5.3	sn-glycerol-3-phosphate-binding periplasmic protein
-	-	-	-	37711	02951	-	5.1.4	putative transposase for IS1667
YE0255	41941	33681	18501	37751	02991	38911	7.0.0	putative DNA recombinatin protein
YE0270	42091	33841	18651	37901	03141	38761	0.0.2	conserved hypothetical protein
YE0277	43621	39341	38561	40431	03431	38681	7.0.0	putative acetyltransferase
YE0313	30741	19591	14011	25121	20801	30861	4.1.1	inner membrane protein
YE0320	30661	19661	14081	25191	20891	37981	2.1.4	probable exported Zinc protease
YE0322	30661	19691	14091	25201	20911	37951	4.1.1	putative inner membrane transprt protein
YE0324	30641	19711	14121	25231	20941	37931	7.0.0	putative ATP/GTP-binding protein
-	30531	-	-	-	-	-	4.1.0	putative exported protein
-	30011	-	-	-	-	-	0.0.2	hypothetical protein
-	30001	-	-	-	-	-	0.0.2	hypothetical protein
-	29921	-	-	-	-	-	2.2.3	putative restriction enzyme beta subunit
-	29911	-	-	-	-	-	2.2.3	putative restriction enzyme alpha subunit
YE0344	-	-	-	-	-	-	5.1.1	putative HlyD-family secretion protein
YE0345	-	-	-	-	-	-	5.1.4	integrase
YE0386	29481	20241	14651	25771	21481 *	37401	7.0.0	putative acyl-CoA dehydrogenase
-	29371	20332	14761	25881	21581	37291	1.6.0	regulator of cell morhogenesis and NO signalling
YE0409	29261	20441	14871	25991	21711	37171	1.1.1	methyl-accepting chemotaxis protein
YE0416	29201	20511	14931	26051	21781	37111	0.0.2	conserved hypothetical protein
YE0443	28931	20781	14192	26312	22071	36842	1.4.3	multidrug efflux protein

YE0444	28921 *	-	-	-	-	-	1.4.3	multidrug efflux protein
YE0445	28911	20791	15201	26321	22081	36841	4.1.3	probable outer membrane efflux lipoprotein
YE0453	28841	20861	15271	26391	22151	36771	0.0.2	conserved hypothetical protein
YE0470	28701	21001	15411	26531	22321 *	36631	3.3.13	PhnN protein
YE0475	28651	21051	15461	26581	22371	36581	3.3.13	PhnI protein
YE0480	28601	21101	15511	26631	22421	36531	4.1.0	putative exported protein
YE0483	28571	21131	15541	26661	22451	36510	6.1.1	putative LysR-family transcriptional regulatory protein
YE0491	28491	21202	15591	26711	22501	36451	1.5.0	ABC transporter, substrate binding protein
-	-	-	-	-	22551	-	5.1.4	putative transposase for IS1667
-	-	-	-	-	22631	-	0.0.1	conserved hypothetical protein
YE0505 *	28391	21301	15691	26811 *	-	36341	5.1.4	integrase
-	-	-	-	-	-	36331	5.1.4	putative transposase
-	-	-	-	-	-	36321	5.1.4	putative transposase
-	-	-	-	26831	-	-	0.0.2	hypothetical protein
-	-	-	-	26851	-	-	5.1.4	integrase
-	-	-	-	26861	-	-	2.1.1	putative type I site-specific deoxyribonuclease
-	-	-	-	26881	-	-	0.0.1	hypothetical protein
-	-	-	-	27061	-	-	5.1.4	putative transposase
-	-	21331	15721	-	-	-	2.2.3	putative type I restriction-modification system
-	-	21361	11271	27071	22691	-	2.2.3	superfamily I DNA helicase
YE0511	-	-	-	-	-	-	5.1.4	putative phage-related protein
YE0517	-	21431	11312	-	-	-	4.1.1	putative membrane protein
-	28141	-	-	-	-	-	0.0.1	conserved hypothetical protein
YE0518	28081	21441	11321	27121	22741	-	1.5.0	putative permease
YE0522	28051	21481	11351	27151	22771	36281	4.1.0	putative exported protein
YE0534	27941	21591	11461	27261	22901	36161	1.5.1	arginine/ornithine antiporter
YE0545	27841	21691	11561	27361	23011	36061	7.0.0	putative oxidoreductase
YE0546	27831	21701	11571	27371	23021 *	36051	0.0.0	hypothetical protein
YE0559	27691	21841	11711	27511	23211	35911	6.1.1	LysR-family regulatory protein
YE0580	27521	22011	11881	27681	23411	35721	6.1.1	transcriptional regulator NadR
YE0583	-	22041	11911	27711	23441	35691	0.0.2	conserved hypothetical protein
YE0591	27421	22121	11991	27791	23531	35601	0.0.2	conserved hypothetical protein
-	27011	22522	12391	28201	23971	35211	5.1.5	putative adhesin/invasin protein
YE0650	26881	22661	12531	28341	24161	35071	3.1.19	tryptophanase
YE0656	26821	22721	12591	28401	24231	35001	7.0.0	putative AMP-binding enzyme-family protein
YE0688	26531	23011	12881	28691	24561	34711	1.5.5	putative type II secretion system protein
YE0695	26471	23071	12941	28751	24621	34651	1.5.1	aromatic amino acid transport protein
-	26422	23131	12991	28801	24681	34591	0.0.2	hypothetical protein
YE0717	26211	23281	13141	28951	24851	34451	0.0.2	conserved hypothetical protein
YE0725	26141	23351	13211	29021	24921	34381	6.1.1	DnaK suppressor protein homologue
YE0787	-	25021	36720	31330	38063	09069	0.0.0	hypothetical protein
YE0789	25511	25041	36701	31311	38061	09071	2.2.9	putative thiol:disulfide interchange protein
YE0791	25491	25061	36681	31291	38041	09081	1.5.0	ABC transporter, ATP-binding component
-	-	25141	36601	-	37961	-	5.1.4	putative transposase for IS1667
YE0799	-	25161	35431	31211	37951	09161	7.0.0	putative aminotransferase
YE0807	25311	25191	35451	31191	37931	09181	6.0.0	putative transcriptional regulatory protein
-	25281	-	-	-	-	-	1.4.0	putative enterotoxin-like protein
YE0813	25221	25251	35511	31131	37871	09251	1.5.3	sorbose-specific PTS uptake system component
-	25191	25272	35541	31101	37841	09281	3.5.0	conserved hypothetical protein
-	24941	25282	35641	31001	37741	09381	3.5.2	anaerobic dimethyl sulfoxide reductase chain A
-	24911	25285	35681	30961	37711	09421	1.3.1	putative oxidoreductase component
-	-	25291	35691	30951	37701	09431	5.1.4	putative transposase for IS1667
YE0840	24701	25421	35801	30841	37581	09531	4.1.1	putative membrane protein
YE0847	24631	25491	35871	30771	37511	09631 *	4.1.0	putative exported protein
-	-	25551	35931	30711	37451	09701	5.1.4	transposase for IS1328

YE0911	06591	00021	38181	19541	03481	32661	3.6.1	3-oxoacyl-[acyl-carrier-protein] synthase II
YE0915	06621	00051	38151	19571	03531	32621	1.5.5	MFS-family transporter
-	-	00071	38121	19601	03561	32601	5.1.4	transposase for IS1328
YE0924	06701	00141	38051	19671	03641	32531	1.5.1	glycine betaine/L-proline transport system permease
YE0933	06780	00221	37951	19771	03731	32441	1.6.1	putative acid shock protein
YE0934	06781	00231	37941	19781	03741	32431	0.0.0	hypothetical protein
YE0935	06791	00241	37931	19791	03751	32421	6.1.1	AraC-family transcriptional regulator
YE0936	06801	00242	37921	19801	03761	32411	1.5.0	putative MFS family transport protein
YE0938	06821	00271	37901	19821	03791	32391	4.1.1	putative membrane protein
YE0941	06851	00301	37871	19861	03801	32351	6.1.1	AraC-family transcriptional regulator
YE0944	06881	00323	37831	19891	03831	32321	0.0.2	hypothetical protein
-	-	00531	22222	20091	04061	32121	5.1.4	transposase for IS1667
YE0972	07131	00601	22161	20161	04131	32051	0.0.2	conserved hypothetical protein
-	-	00651	22121	-	-	-	5.1.2	putative prophage integrase
-	-	00691	22081	-	-	-	0.0.2	hypothetical protein
-	-	00771	22001	-	-	-	2.2.3	putative RNA directed DNA polymerase
-	07391	-	-	-	-	-	5.1.2	prophage regulatory protein
YE0974	-	-	-	-	-	-	5.1.2	prophage integrase
YE1005	07621	00911	21861	20301	04271	31931	5.1.5	phospholipase A
-	-	-	-	-	04301	31901	5.1.4	putative transposase for IS1667
-	-	00931	21841	20321	04311	31891	5.1.4	transposase for IS1328
YE1023	07791	01091	21681	20481	04481	31731	1.5.3	putative gluconate transporer
YE1034	07882	01211	21582	20580	04580	31632	4.1.1	putative exported protein
YE1036	07901	01231	21571	20591	04591	31621	4.1.0	putative lipoprotein
-	08181	01501	21291	20871	04891	31351	1.5.0	putative autotransporter protein
YE1068	08211	01531	21261	20901	04931	31311	4.1.0	putative exported protein
-	-	01631	21161	21001 *	-	31211	5.1.4	putative transposase for IS1667
YE1078	08311	01651	04111	21021	05051	31201	4.1.1	putative membrane protein
YE1085	08381	01721	04041	21091	05141	31131	0.0.0	hypothetical protein
YE1086	08391	01731	04031	21101	05151	31121	5.3.5	putative cytochrome c
-	-	01741	04021	21111	05161	31091	5.1.4	putative transposase for IS1667
YE1101	-	-	-	-	-	-	0.0.0	hypothetical protein
YE1116	08631	01971	03781	21351	05511 05511B	30791	4.1.1	putative inner membrane protein [region flipped, protein torn apart on opposite strands]
YE1117	08641	09181	03771	21361	05501	30781	1.5.2	putative divalent cation transport protein
-	08711	02021	03701	21431	05521	30711	1.1.1	flagellin lysine-N-methylasec EC=2.1.1.-
-	-	02211	03511	21621	05721	30531	0.0.2	hypothetical protein
-	-	02261	03471	21651	05771	30501	5.1.2	transposase for IS1328
-	-	02271	03461	21652	05801	30492	5.1.2	hypothetical phage-related protein
-	-	02301	03431	21691	05821	30481	5.1.4	transposase for IS1668
YE1158	09051	02451	03291	21831	05971	30341	3.5.3	cytochrome C-type protein NapC
-	-	-	-	-	-	30421	5.1.4	putative transposase
Ye1165	09081	02511	03252	21852	06010	30302	6.1.1	putative LyxR-family transcriptional regulatory protein
YE1166	09091	02521	03251	21861	06011	30301	6.1.1	putative LysR-family transcriptional regulatory protein
YE1184	-	-	-	-	-	-	3.3.15	putative acetyltransferase
-	09421	-	-	-	-	-	7.0.0	putative acetyltransferase
YE1199	09581	02721	03071	22051	06191	30111	1.5.0	putative ABC transporter, ATP-binding protein
YE1228	09831	02991	02821	22301	06491	29851	6.1.1	two-component system histidine kinase
YE1234	-	03079	02763	22359	06549	29793	4.1.0	putative lipoprotein
YE1236A	09911	03080	02762	22360	06550	29792	2.1.4	6-phospho-beta-glucosidase
YE1238	09931	03091	02751	22371	06561	29781	1.5.1	putative membrane transport protein
YE1246	10011	03171	02671	22451	06651	29701	1.5.0	MFS-family membrane transport protein
YE1251	10061	03221	02621	22501	06711	29652	6.1.1	putative regulatory protein
YE1252	10071	03231	02611	22511	06721	29651	4.1.0	putative exported protein
YE1261	-	-	-	-	-	-	5.1.2	putative phage integrase

YE1262	10561	03341	02511	22711	06821	29481	4.1.1	putative inner membrane protein
-	-	03361	02501	-	-	29461	5.1.4	putative transposase for IS1667
YE1288	10821	03611	02251	22961	07081	29201	0.0.2	conserved hypothetical protein
YE1298	10921	03711	02151	23061	07191	29091	1.1.1	putative flagellar assembly regulatory protein Flk
YE1304	10981	03771	02091	23121	07291	29031	4.1.1	putative inner membrane protein
YE1305	10991	03781	02081	23131	07301	29021	4.1.1	putative membrane protein
-	10161	-	-	-	-	-	1.4.1	putative hemolysin
-	10211	-	-	-	-	-	0.0.2	hypothetical protein
-	10281	-	-	-	-	-	0.0.1	hypothetical protein
-	10301	-	-	-	-	-	0.0.1	hypothetical protein
-	10321	-	-	-	-	-	1.4.0	putative hemagglutinin/hemolysin
-	10351	-	-	-	-	-	1.4.0	putative adhesin/hemagglutinin
YE1322	-	03941	01921	23291	07471	28851 *	5.1.5	putative RTX-family protein
YE1325	11171	03971	01891	23321	07501	28801	0.0.2	conserved hypothetical protein
-	-	03991 *	01871	23341	07521	28771	5.1.4	putative transposase for IS1667
YE1379	11631	04441	01431	23781	08031	28331	3.2.8	putative O-succinylbenzoic acid--CoA ligase
YE1384	-	-	-	-	-	-	4.1.0	putative exported protein
-	11461	04271	01591	23621	07841	28491	0.0.0	hypothetical protein
YE1389	-	04521	01351	23861	08121	28261	2.1.4	putative serine protease
YE1391	11691	04541	01331	23881	08151	28230	7.0.0	putative iron-sulphur binding protein
YE1400	11821	04631	01231	23991	08251	28131	1.5.5	Putative type I secretion protein
YE1406	11881	04691	01171	24051	08351	28071	2.1.4	putative protease
YE1409	11901	04711	01151	24071	08371	28051	5.1.2	putative prophage repressor protein
YE1410	11911	04721	01141	24081	08381	28041	5.1.2	probable phage antitermination protein Q
-	11941	04733	01122	24100	08401	28011	3.3.0	putative ATP-binding protein
-	11981	-	-	-	-	-	5.1.2	putative bacteriophage coat protein
-	12131	04780	01091	24131	08451	27971	3.5.0	carbonic anhydrase, EC=4.2.1.1
-	-	04821	01061	24141	-	27951	5.1.4	transposase for IS1668
YE1412	-	-	-	-	-	-	5.1.4	transposase for insertion element IS285
-	12071	-	-	-	-	-	5.1.2	hypothetical phage protein
-	12651	-	-	-	-	-	2.2.3	protein umuC
YE1432	12801	04992	00880	24340	08641	27762	0.0.0	hypothetical protein
-	-	05241	00631	24581	08921	27531	4.1.0	putative exported protein
YE1488	13361	05551	00321	40061	09261	27211	4.1.1	putative inner membrane protein
YE1529	-	05881	40091	40382	09611	26881	5.1.4	putative transposase for IS1667
-	-	05891	40101	00001	09621	26871	5.1.4	transposase
-	-	-	-	-	09671	-	5.1.4	putative transposase for IS1667
YE1576	14121	06341	29351	00461	10131	26431	3.6.1	putative 3-oxoacyl-[acyl-carrier-protein]synthase II
-	14371	06565	29101	01121	10411	26171	3.5.3	putative cytochrome C-type biogenesis protein
-	14391	06567	29081	01141	10431	26151	3.5.3	cytochrome c biogenesis protein
YE1598	14401	06568	29071	01151	10441	26141	3.5.2	nitrite reductase gene remnant
-	-	06631	29021	-	10491	26091	5.1.4	putative transposase for IS1667
YE1610	14501	06701	19981	01261	30181	26031 *	1.5.0	putative MFS-family transport protein
YE1614	14541	06751	19931	01311	30121	25971	6.1.1	tetrathionate reductase complex: sensory transduction histidine kinase
YE1617	14571	06781	19901	01341	30091	25921	3.5.2	tetrathionate reductase subunit A
YE1656	-	-	-	-	-	-	5.1.4	putative transposase
YE1696	-	07071	19611	01631	29761	25661	1.5.0	putative ABC transport protein, ATP-binding component
YE1697	-	07081	19601	01641	29751	25651	4.1.0	putative exported protein
YE1698	-	07091	19591	01651	29741	25641	0.0.2	conserved hypothetical protein
YE1700	14841	07111	19561	01681	29701	25621	3.1.5	putative shikimate dehydrogenase
YE1716	15001	07271	19391	01851	29521	25441 *	0.0.2	conserved hypothetical protein
-	-	07381	19281	01951	29391	25331	5.1.4	transposase for IS1668
-	-	07391	19261	01961	29371	25321	5.1.2	putative prophage terminase small subunit
YE1728	15141	07442	19201	02011	29301	25281	1.5.3	putative sugar permease

YE1729	15161	07451	19181	02031	29281	25261	6.1.1	putative sucrose operon repressor (LacI-family regulator)
YE1733	15211	07491	19131	02081	29221	25221	1.4.4	putative UV protection and mutation protein
YE1737	-	-	-	-	-	-	3.1.14	possible homoserine synthase
-	-	07541	19071	02140	29141	25161	0.0.0	hypothetical protein
YE1743	15431	07561	19051	02151	29121	25141	6.1.1	putative AraC-family transcriptional regulatory protein
-	15481	07621	18981	02201	29071	25081	5.1.5	putative cytolethal distending toxin subunit A
YE1746	15741	07591	19011	02191	02981	25091	7.0.0	hypothetical protein
YE1748	15511	07651	18951	02231	29041	25051	1.5.0	putative ABC transporter, ATP-binding protein
YE1759	15611	07731	18871	02311	28921	24971	0.0.1	conserved hypothetical protein
YE1767	15681	07801	18801	02381	28851	24921	4.1.2	peptidoglycan synthetase
Ye1771	15721	07832	18751	02431	28811	24871	4.1.3	ferrichrome receptor protein
YE1803A	-	-	-	-	-	-	5.1.2	IS1329 transposase A
YE1804	-	-	-	-	-	-	5.1.2	terminase, ATPase subunit
YE1821	-	-	-	-	-	-	5.1.4	transposase
YE1823	-	-	-	-	-	-	5.1.4	IS1400 transposase B
YE1824	-	-	-	-	-	-	0.0.2	conserved hypothetical protein
YE1844	21032	12891	15841	07601	-	19611	1.5.0	sodium/glutamate symport carrier protein
-	-	12881	15861	07581	-	19621	5.1.4	transposase for IS1328
-	-	12701	16041	07401	13171	19811	0.0.0	hypothetical protein
-	-	12471	16261	-	27921	20051	5.1.4	transposase for IS1328
YE1845	21031	12461	16271	03231	27901	20071	4.1.1	putative inner membrane protein
YE1848	21001	12431	16301	03261	27871	20101	7.0.0	putative transferase
YE1853	20971	12401	16331	03291	27821	20121	0.0.2	conserved hypothetical protein
YE1855	20951	12381	16351	03311	27801	20151	0.0.2	conserved hypothetical protein
YE1874	20801	12231	16501	03461	27641	20272	2.0.0	putative DNA-binding protein
YE1875	20791	12221	16511	03471	27621	20281	1.5.0	putative transport protein
-	-	-	-	-	27571	-	5.1.4	putative transposase for IS1667
YE1879	20751	12181	16551	03511	27561	20341	1.5.3	putative transport protein
YE1895	20581	12011	16741	03701	27351	20531	3.6.0	methylmalonyl-CoA mutase
YE1897	20561	11991	16761	03721	27331	20561	1.5.2	chelated iron transport system membrane protein
-	-	11641	17111	06441	-	-	5.1.4	transposase
YE1905	20491	09121	35401	03791	27252	20622	0.0.0	hypothetical protein
YE1920	-	-	-	-	-	-	5.1.4	IS1400 transposase B
YE1925	-	-	-	-	-	-	5.1.4	putative transposase
YE1935	-	-	-	-	11451	18411	5.1.4	transposase for IS1330
YE1944	22921	19221	25071	08971	11561 *	18501 *	1.5.0	putative membrane permease
YE1947	22891	19251	25101	08941	11601 *	18541	7.0.0	putative hydrolase
YE1950	22681	19281	25131	08911	11631	18571	3.3.19	alkanesulfonate monooxygenase
YE1951	22851	19291	25141	08901	11651	18581	6.1.1	putative TetR-family transcriptional regulatory protein
YE1953	-	-	-	-	-	-	5.1.2	bacteriophage late control protein
YE1954	-	-	-	-	-	-	5.1.2	bacteriophage major tail sheath protein
YE1960	22801	19341	25191	08851	11711	18631	4.1.0	putative exported protein
-	-	19311	25161	08881	11681	18601	5.1.4	putative transposase
-	-	-	-	-	11671	-	5.1.2	putative phage integrase
-	22711	-	-	-	11811	-	5.1.2	putative phage integrase
-	-	-	-	-	11821	-	5.1.4	putative transposase subunit
-	-	-	-	08691	-	-	5.1.2	putative phage tail fibre protein
-	-	-	-	08391	-	-	5.1.4	putative transposase
-	-	-	-	08351	-	-	5.1.2	putative phage protein
-	-	11941	16811	-	-	-	5.1.4	putative transposase for IS1667 (pseudo)
YE1971	18221	11481	17271	06271	14451 *	23231	1.5.0	probable membrane transport protein
-	-	-	-	-	16171	-	5.1.4	putative transposase for IS1667
YE1984	18351	11331	29911	06131	16161	-	4.1.1	putative membrane protein
Ye1997	18461	11211	30031	06011	16011	22581	3.3.15	probable formate--tetrahydrofolate ligase
YE1998	18471	11201	30041	06001	15991	22571	1.5.5	putative RTX-family toxin transporter

YE1999	18481	11191	30051	05991	15981	22551	5.1.5	hemolysin transport protein
YE2000	18491	11181	30061	05981	15971	22531	1.5.5	putative toxin transport protein
YE2011	18601	11061	30181	05861	15851	22371	1.5.0	PTS system, maltose and glucose-specific IIabc component
YE2012	18611	11051	30191	05851	15841	22361	6.1.1	maltose regulon regulatory protein
-	18621	11042	30201	05841	15831	22341	1.4.2	putative acetyltransferase
-	-	-	-	-	15791	-	5.1.4	putative transposase for IS1667
YE2019	18681	10991	30261	05781	15761	22281	1.4.3	beta-lactamase precursor
YE2025	18741	10931	30331	05711	15681	22211	4.1.1	putative membrane protein
YE2026A	18771	10911	30361	05681	15641	-	0.0.2	conserved hypothetical exported protein
YE2027	18781	10901	30371	05671	15631	22210	4.1.0	putative exported protein
YE2030	18811	10871	30401	05641	-	22200	4.1.0	putative exported protein
YE2035	18821	10841	30421	05621	-	-	1.5.0	putative membrane transport protein
-	-	-	-	-	15461	-	5.1.2	putative phage-related membrane protein
-	-	-	-	-	15451	-	0.0.2	conserved hypothetical protein
YE2036	18841	10831	30441	05601	15431	22191	0.0.1	conserved hypothetical protein
YE2037	18851	10801	30461	05591	15421	22181	4.1.0	putative exported protein
-	18871	10781	30481	05571	15401	22171	0.0.0	hypothetical protein
-	18881	10773	30491	05561	15391	22161	4.1.0	putative exported protein
YE2041	18950	10760	30531	05511	15351	22151	1.5.0	putative amino acid transporter, permease
YE2043	18951	10751	30541	05501	15331	22141	1.5.0	amino acid transport system permease protein
YE2044	18961	10741	30551	05491	-	22131	1.5.0	amino acid-binding periplasmic protein precursor
YE2047	18981	10721	30571	05471	15301	-	7.0.0	putative hydrolase
-	-	10720	30581	05421	-	-	5.1.4	transposase
-	-	-	-	-	15251	-	5.1.4	putative site-specific recombinase
YE2052	19001	10681	30591	05411	15241	22121	3.3.17	putative AMP nucleosidase
YE2055	19031	10651	30621	05381	15211	22091	6.1.1	putative inner membrane regulatory protein
YE2057	19041	10631	39771	05361	15181	-	7.0.0	putative virulence factor
YE2058	19051	10621	39781	05351	15171	22081	7.0.0	putative virulence factor
YE2059	19061	10611	39791	05341	15161	22071	7.0.0	putative virulence factor
YE2060	19071	10671	28131	05311	15121	22041	3.1.11	putative acetolactate synthase large subunit
YE2064	19111	10531	28171	05271	-	-	0.0.0	hypothetical protein
YE2066	19131	10511	28181	05261	15071	-	1.5.0	L-fucose permease
-	-	10501	28191	-	-	22001	5.1.3	transposase
YE2067	19141	10491	28201	05251	15061	21991	3.4.3	putative carbohydrate kinase
YE2069	19151	10481	38211	05241	15051	21981	3.3.13	putative short chain dehydrogenase
YE2073	19191	10441	28251	05201	15001	21931	0.0.2	conserved hypothetical protein
YE2085	19311	10321	28371	05011	14861	21821	7.0.0	putative carboxypeptidase
YE2086	19321	10312	28380	05001	14851	21811	1.6.1	putative acid shock protein
YE2094	19391	-	-	-	-	-	6.1.1	putative transcriptional regulatory protein
YE2099	19441	10211	28471	04901	14741	21721	4.1.0	putative lipoprotein
YE2102	19471	10181	28511	04861	14701	21681	7.0.0	putative pyruvate-flavodoxin oxidoreductase
-	-	10161	28521	04851	14691	-	2.2.7	putative phospholipase
YE2115	19581	10071	28601	04771	14593	21591	0.0.2	conserved hypothetical protein
YE2116	19591	10061	28611	04761	14592	21581	0.0.2	conserved hypothetical protein
-	-	-	-	-	-	21391	5.1.4	putative transposase for IS1667
YE2134	19771	09871	28791	04561	16411	21381	0.0.2	conserved hypothetical protein
YE2138	19811	09831	28831	04521	16461	21341	3.3.15	pyridoxamine kinase
YE2141	19841	09801	28861	04491	16501	21311	4.1.0	putative lipoprotein
YE2152	19951	09691	28981	04371	16631	21201	7.0.0	NADH:flavin oxidoreductase / NADH oxidase family protein
-	-	09671	29011	04351	16661	21181	5.1.4	transposase for IS1668
YE2163	20061	09561	34961	04231	26791	21071	1.5.0	putative transport protein
YE2179	20212	09401	35112	04072	26960	20912	0.0.0	hypothetical protein
YE2181	20231	09381	35131	04061	26971	20891	1.5.5	hemin transport protein
YE2182	20241	09371	35141	04051	26981	20881	1.5.5	hemin-binding periplasmic protein
YE2184	20261	09351	35161	04031	27001	20851	2.2.6	putative lipoate-protein ligase A

-	17921	-	-	-	-	-	3.1.19	tryptophan synthase alpha chain, EC=4.2.1.20
-	-	-	-	-	-	22791	5.1.2	hypothetical phage protein
-	-	-	-	-	-	22801	0.0.2	hypothetical protein
-	-	-	-	-	-	23091	0.0.1	hypothetical protein
						23071	5.1.2	putative exinuclease protein
-	-	11571	17181	06371	14361	23311	0.0.0	hypothetical protein
YE2216A	17891	11761	16991	06561	14131	23491	0.0.2	conserved hypothetical protein
-	-	09081	37151	-	-	-	4.1.0	putative transposase for IS1667
YE2243	17641	09011	37221	06801	13871	23711	6.1.1	probable response regulator
YE2245	17621	08991	37241	06821	13851	23731	0.0.0	hypothetical protein
YE2256	17561	08921	37301	06881	13761	23790	4.1.0	putative exported protein
YE2263	17481	08851	37371	06951	13681	23861	0.0.2	conserved hypothetical protein
-	17161	-	-	-	-	-	5.1.2	phage replication protein
YE2273	17021	08761	37431	07011	13591	23921	0.0.2	conserved hypothetical protein
YE2278	16971	08711	37491	07071	13541	23971	7.0.0	putative phosphodiesterase
YE2279	16961	08701	37501	07081	13531	23981	1.4.3	putative multidrug resistance protein
YE2286	16891	08612	37562	07142	13460	24061	4.1.4	acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamne o-acyltransferase
-	-	08611	37591	03221	13401	24081	5.1.4	transposase for IS1328
-	-	08561	33631	03181	27991	24121	5.1.4	transposase for IS1668
-	-	08161	34061	02751	28411	24531	6.3.11	putative LyxR-family transcriptional regulatory protein
-	-	08131	34091	02711	28451	24571	5.1.2	putative bacteriophage protein
-	-	-	-	02721	28441	-	5.1.4	putative transposase
-	-	08181	34021	02791 *	28361	-	5.1.4	transposase
-	-	-	-	07181	-	-	5.1.4	transposase for IS1328
YE2376	16771	12601	16141	07301	13281	19921	4.1.0	putative lipoprotein
-	-	12981	39541	07751	12811	-	0.0.0	hypothetical protein
-	-	13051	39471	07821	12741	19391	5.1.4	putative transposase subunit
-	-	13071	39461	07831	12731	19381	5.1.4	putative transposase for IS1669
YE2410	16461	13261	36551	07991	12521	19181	7.0.0	putative oxidoreductase
YE2416	16421	13301	36522	08022	12491	19142	4.1.1	putative inner membrane protein
YE2422	16371	13351	36481	08071	12431	19101	7.0.0	putative oxidoreductase
-	-	13661	36181	-	12111	18821	5.1.4	IS1400 transposase A
-	-	-	-	-	12101	-	5.1.2	putative phage major capsid protein
-	-	-	-	-	12091	-	5.1.2	putative phage holin
-	-	13791	36081	-	11971	-	5.1.2	putative variable tail fibre protein
-	-	13801	36071	-	11951	-	5.1.2	putative phage tail fiber assembly protein
-	-	13941	35991	-	-	-	5.1.2	putative bacteriophage tail fiber protein
-	-	13771	36101	-	40661	18771	5.1.2	putative phage baseplate assembly protein
-	-	-	-	-	-	18761	5.1.2	putative phage tail protein
-	-	13951	40131	-	-	18751	5.1.2	putative bacteriophage late gene regulator
-		13961	40132	-	-	18731	5.1.4	putative transposse for IS1667
-	16021	-	-	-	-	-	0.0.1	conserved hypothetical protein
YE2445	-	-	-	-	-	-	5.1.4	transposase
-	21111	12924	15791	07681	12921	19541	1.5.0	putative phoshpotransferase system enzyme subunit
YE2449	21141	12931	15761	07711	12891	19511	3.4.3	putative tagatose-6-phosphate ketose / aldoseisomerase
YE2450	21151	12941	15751	07721	12881	19491	3.4.3	putative D-tagatose-1,6-bisphosphate aldolase subunit
YE2458	21191	13001	39521	07771	12791	19441 *	0.0.2	conserved hypothetical protein
YE2461	21221	13031	39491	07801	12761	19411	3.4.2	imidazolonepropionase
YE2474	21341	13201	39351	07941	12581	19261	6.1.1	putative LysR-family transcriptional regulatory protein
-	-	13211	39352	-	12571	19241	5.1.4	putative transposase for IS1667
YE2483	21341	08501	33691	03121	28051	24181	1.6.3	haemin storage system, HmsF protein
YE2484	21441	08491	33701	03111	28061	24191	1.6.3	haemin storage system, HmsH protein, precursor

YE2487	21181	12991	39531	07761	12801	19451	4.1.0	hypothetical protein
YE2488	21481	08451	33741	03071	28101	24251	6.1.1	putative LysR-family transcriptional regulatory protein
YE2499	21591	08341	33851	02961	28211	24371	4.1.1	putative membrane protein
YE2502	21621	08311	33881	02931	28241	24401	7.0.0	putative class II aldolase-family protein
YE2508	21681	08251	33941	02871	28301	24461	1.5.1	L-asparagine permease
-	21761	-	-	-	-	-	5.1.2	putative phage integrase
-	23022	19111	24961	09081	11422	18381	0.0.2	conserved hypothetical protein
YE2520	23091	19031	24881	09161	11341	18291	2.2.11	flagellin lysine-N-methylase
-	23101	-	-	-	-	-	1.1.1	flagellin
YE2521	23111	19021	24871	09171	11331	YE3904-18281	1.1.1	flagellin
YE2522	23121	19011	24861	09181	-	-	1.1.1	thermoregulated motility protein
YE2523	23131	19001	24851	09191	-	18271	1.1.1	flagellin
YE2531	-	18911	24771	09271	11241	18204	0.0.0	hypothetical protein
YE2535	-	18900	24762	09280	11230	18202	5.1.2	IS1329 transposase B
YE2564	23481	18561	24471	09571	10911	17921	5.1.5	invasin
YE2586A	23691	18351	24261	09780	10651	17691	3.3.0	putative coA transferase family protein
YE2594	-	-	-	-	-	-	5.1.4	transposase for insertion sequence element IS1665
-	-	18281	39880	09831	-	-	5.1.4	transposase
YE2597	23791	18251 *	39861 *	09861	10551	17621	6.1.1	AsnC-family transcriptional regulatory protein
YE2598	23801	18241	39851	09871	10531	-	0.0.1	conserved hypothetical protein
YE2600	23821	18221	39831	09891	10511	17591	3.3.15	probable histidine acid phosphatase
-	23851	18175	34831	09931	30271	17551	4.1.5	putative outer membrane fimbrial usher protein
-	23861	18174	34821	09941	30281	17541	4.1.0	putative exported protein
YE2603	23891	18171	34791	09971	30311	17501	3.4.3	putative sugar kinase
YE2609	23951	18151	34781	09981	30321	17491	1.5.0	ABC transporter, substrate binding component
-	-	-	-	-	30361	-	3.3.0	putative hydrolase
-	-	-	-	-	30401	-	2.2.3	putative antirestriction protein
-	-	-	-	-	30421	-	2.2.3	putative NTPase
YE2611	-	-	-	-	-	-	5.1.2	integrase
YE2624	-	-	-	-	-	-	5.1.4	transposase for insertion sequence element IS1328
YE2640 *	24071	18001	34641	10121	30601	17341	4.1.1	putative membrane protein
YE2644	24091	17971	34621	10141	30631	17321	1.5.3	putative C4-dicarboxylate transporter, periplasmic protein
YE2646	24111	17951	34601	10161	30651	17291	1.5.3	putative C4-dicarboxylate transporter, large subunit
YE2652	24171	17891	34541	10221	30711	17221	3.3.15	citrate lyase acyl carrier protein
YE2654	24191	17871	34521	10241	30731	17201	6.1.1	sensor kinase protein
YE2662	24271	18121	34761	10001	30471	17471	6.1.1	putative two-component response regulator
YE2663	24281	18111	34751	10011	30481	17461	6.1.1	putative virulence sensor protein
YE2666	24311	18081 *	34721	10041	30511	17421	4.1.5	probable fimbrial usher protein
-	-	18201	34851	09911	30251	17571	5.1.4	putative transposase for IS1667
-	-	-	-	-	-	17191	5.1.4	putative transposase
YE2670	24351	-	-	10341	30841	17081	6.1.1	putative GntR-family regulatory protein
YE2671	24381	17771	34431	10351	30851	17072	0.0.0	putative short chain dehydrogenase / reductase family protein
YE2684	31121	17641	32241	10491	31001	16921	0.0.2	conserved hypothetical protein [unresolved tandem repeat]
YE2686	31141	17621	32221	10511	31021	16901	4.1.1	putative OmpA family protein
YE2693	31211	17551	32151	10581	31091	16831	4.1.0	outer membrane usher protein
YE2705	31331	17411	32031	10701	31221	16701	6.1.1	putative luxR-family regulatory protein
-	31341	17410	32021	10711	31222	16691	1.5.0	xanthosine permease
YE2706	-	-	-	-	-	-	7.0.0	transcription regulator/permease fusion protein

YE2708	31391	17391	32001	10731	31241	16661	3.2.3	alpha-ribazole-5'-phosphate phosphatase
YE2710	31411	17371	31981	10751	31261	16641	3.2.3	cobinamide kinase and guanylyltransferase
YE2719	31501	17271	31881	10851	31361	16541	0.0.2	conserved hypothetical protein
YE2724	31571	17211	31821	10911	31421	16481	4.1.1	putative membrane protein
YE2762	31941	16831	31451	11281	31811	16111	1.5.0	putative ABC transport system ATP-binding protein
-	-	-	-	-	31971	15691	5.1.4	transposase
-	-	-	-	-	-	15951	5.1.4	putative transposase
-	-	-	-	-	32011	15941	5.1.4	IS1400 transposase B
YE2810	32391	16251	10721	11731	32471	15541	3.5.2	formate dehydrogenase H
YE2812	32411	16231	10710	11750	32482	15512	1.5.3	putative formate transporter
YE2829	32581	16071	10531	11921	32661	15341	3.2.9	molybdopterin biosynthesis protein
YE2832	32601	16041	10511	11951	32711	15301	0.0.0	hypothetical protein
YE2852	32801	15841	10311	12151	32931	15091	7.0.0	putative peptidase
YE2859	32871	15761	10241	12221	33021	15021	6.1.1	LysR family transcription regulatory protein
YE2866	32941	15691	10171	12291	33111	14951	4.1.1	putative fatty acid desaturase
YE2867	32951	15681	10161	12301	33121	14931	0.0.2	conserved hypothetical protein
YE2883	33121	15521	10001	12461	33291	14771	3.5.1	D-lactate dehydrogenase
YE2885	33141	15491	09971	12491	33331	14750	0.0.2	conserved hypothetical protein
YE2900	33271	15361	09841	12621	33471	14621	0.0.2	conserved hypothetical protein
-	-	-	-	-	-	14361	5.1.2	putative phage protein
-	-	-	-	-	-	14311	5.1.2	putative bacteriophage protein
YE2922	-	15131	09611	12851	33731	13941	7.0.0	putative protease
YE2955A	33811	14811	09291	13171	34121	13621	0.0.2	conserved hypothetical protein
-	33861	14761	09241	13221	34171	13561	5.1.5	putative cytotoxin RtxA
-	33881	14741	09221	13241	34191	13541	0.0.2	putative uncharacterized protein RtxH
YE2966	33961	14661	09141	13321	34281	13461	5.1.5	putative hemolysin
YE3007	34461	14251	08731	13731	34711	13031	3.2.7	lipoate-protein ligase B
YE3014	34531	14181	08661	13801	34791	12962	6.1.1	putative LyxR-family regulatory protein
YE3021	34601	14111	08591	13871	34871	12911	7.0.0	putative phenylalanine and histidine ammonia-lyase
YE3025	24651	27571	39931	13911	-	12861	6.1.1	putative two-component system sensor protein
-	-	27551	39921	13921	-	12841	0.0.0	hypothetical protein
-	-	-	-	-	34901	-	5.1.2	putative phage integrase
-	-	27502	39891	13951	34941	12811	0.0.2	conserved hypothetical protein
-	-	-	39892	13961	34951	12801	0.0.2	conserved hypothetical protein
-	-	27501	23351	13991	35001	-	5.1.2	putative phage primase
YE3026	34661	-	-	-	-	-	4.1.0	putative exported protein
YE3028	34681	-	-	-	-	-	4.1.0	putative exported protein
YE3037	-	-	-	-	-	-	0.0.0	hypothetical protein
YE3038	-	-	-	-	-	-	5.1.4	integrase
-	-	27471	23331	14041	35061	12791	5.1.2	putative integrase
-	-	-	-	-	-	12501	5.1.2	lysozyme
-	-	-	-	-	-	12281	2.2.3	putative adenine-specific modification methylase
YE3066	35381	27281	YE5630-23141	14231	35261	12081	0.0.2	UDP-sugar hydrolase
YE3097	35611	27111	22911	14461	35541	11871	1.5.2	putative potassium efflux system
-	-	26491	22291	15081	36311	11231	0.0.0	hypothetical protein
-	-	26441	22231	15141	36371	11171	5.1.4	putative transposase for IS1667
YE3168	36291	26442	32811	15151	36391	11161	0.0.2	conserved hypothetical protein
YE3180	36381	26351	32721	15241	36491	11071	5.1.4	phosphate binding protein
YE3190	36481	26251	32621	15341	36601	10961	0.0.1	conserved hypothetical protein
YE3196	36541	26191	32561	15401	36681	10901	4.1.4	putative cyclic beta 1-2 glucan synthetase
YE3199	36571	26161	32561	15431	36711	10851	1.5.3	PTS system, alpha-glucoside-specific IIBC component
YE3208	36661	26071	32441	15521	36821	10751	0.0.2	conserved hypothetical protein
-	-	25861	-	-	-	-	5.1.4	putative transposase for IS1667
-	-	25621	-	-	-	-	5.1.2	putative transposase for IS1667

-	-	25601	35981	40831	-	09751	5.1.4	transposase for IS1668
YE3218	36761	25971	32341	15621	36921	10641	3.5.3	Na+-translocating NADH-guinone reductase subunit C
YE3225	36831	25901	32271	15691	37001	10561	5.1.5	putative accessory processing protein
YE3226	36841	25891	32261	15701	37021	10551	5.1.5	putative adhesin
-	-	-	-	-	-	10131	5.1.2	putative phage replication protein
-	-	-	-	-	-	10121	5.1.2	putative DNA replication protein
-	-	-	-	-	-	10111	5.1.2	putative phage replication protein
-	36931	25790	39021	15791	37121	10001	3.4.2	putative amino acid hydrolase
-	36941	25789	39011	15801	37131	09991	3.7.1	putative purine catabolism protein
YE3285	06081	34221	04451	30331	26331	33001	2.2.10	methionine aminopeptidase
-	-	24571	31431	30121	38571	-	5.1.4	putative transposase for IS1667
-	-	24691	30741	31661	38431	08751	0.0.0	hypothetical protein
-	-	24651	37081	31701	38471	08711	3.3.0	putative oxidoreductase
-	-	24632	37100	31720	38511	08701A	5.1.2	putative integrase
YE3321	05760	24441	-	-	25941 *	33332	0.0.1	conserved hypothetical protein
YE3326	05721	24401	31271	29961	25901	33371	4.1.1	putative membrane protein
YE3328	-	-	-	-	25882	-	5.1.4	putative transposase for IS1667
YE3338	05621	24291	31171	29861	25791	33491	1.1.1	putative methyl-accepting chemotaxis protein
YE3341	05591	24253	31131	29821	25741	33531	1.5.5	prepilin peptidase
YE3342	05581	24252	31121	29811	25731	33541	4.1.0	putative exported protein
YE3344	05551	24241	YE6503-31101	29781	25711	33561	1.5.5	general secretion pathway protein K
YE3345	05531	24221	31081	29761	25681	33581	1.5.5	general secretion pathway protein I
YE3347	05511	24201	31061	29741	25661	33601	1.5.5	general secretion pathway protein
YE3349	05491	24181	31041	29721	25631	33631	1.5.5	general secretion pathway protein D
YE3354	05421	24131	30991	29671	25581	33681	7.0.0	putative oxidoreductase
YE3358	05381 *	24081	30941	29631	-	-	1.5.3	putative sugar transporter
-	-	24041	30901	29591	25551	-	5.1.4	putative transposase for IS1667
YE3367	05271	24001	30871	29551	25521	33741	0.0.1	conserved hypothetical protein
-	-	-	-	-	-	33801	5.1.4	tnpA-like protein
-	-	-	-	-	-	33861	5.1.4	tnpA-like protein
YE3376	04861	23831	30691	29471	25431	33901	2.2.10	peptide chain release factor 2 - ALL ORTHOLOGS WITH FS!!! Shown to be functional
YE3383	04801	23751	13611	29411	25371	33961	6.1.1	putative histidine kinase sensor
-	-	-	-	-	25321	34011	5.1.4	putative transposase for IS1667
YE3455	-	-	-	-	-	-	5.1.4	transposase for IS3-like insertion sequence
YE3456	-	-	-	-	-	-	5.1.4	transposase for IS3-like insertion sequence
YE3516	-	34811	05031	32141	38971	08271	5.1.2	putative integrase
YE3521	-	-	-	-	-	-	0.0.2	conserved hypothetical protein
YE3522	-	-	-	-	-	-	7.2.1	Putative exported protein
YE3524	-	-	-	-	-	-	5.1.4	putative transposase
YE3524A	-	-	-	-	-	-	0.0.0	hypothetical protein
YE3527	03881	34921	03881	32221	39061	08201	3.3.20	ornithine decarboxylase, inducible
YE3528	03871	34931	05121	32231	39071	08181	4.1.1	putative membrane protein
YE3532	03831	34971	05161	32271	39121	08151	0.0.0	hypothetical protein
YE3580	-	-	-	-	-	-	5.1.4	putative IS285-like transposase
YE3587	03751	35058	05241	32351	39221	08061	7.0.0	putative O-methyltransferase
-	03741	35059	05251	32361	39241	08051 *	4.1.1	hypothetical membrane protein
-	03721	35062	05270	32382	39280	08040	3.3.0	putative FMN-binding protein
YE3588	03711	35061	05271	32381	39281	08031	1.5.0	putative transporter protein
YE3590	03691	35081	05291	32401	39311	08001	7.0.0	putative hydrolase
-	03641	35130	05341	32451	39360	07951	0.0.1	conserved hypothetical protein
YE3595	03621	35131	05351	32461	39361	07941	1.4.3	putative integral membrane efflux protein
-	-	39380	38529	-	03463	38722	5.1.4	putative transposse for IS1669
-	-	39381	38530	-	03462	38721	5.1.4	transposase for IS1328
YE3603	03531	35251	05451	32561	39461	07841	2.2.9	hydrogenase isoenzymes nickel incorporation protein

YE3610A	-	-	-	-	-	-	1.1.1	putative flagellar biogenesis protein
YE3611	-	-	-	-	-	-	6.1.1	putative transmembrane transcriptional activator
YE3612	-	-	-	-	-	-	5.1.4	transposase for IS1330
YE3617	03401	35751	05981	33081	-	07261	1.5.3	putative membrane transport protein
YE3622	03351	35801	06031	33131	-	07211	1.6.3	enterochelin esterase
YE3624	03341	-	-	-	-	-	1.6.3	gerrichrysobactin receptor precursor
YE3626	03331	-	-	-	-	-	4.1.0	putative exported protein, precursor
-	03441	35729	05941	33041	-	07301	1.1.1	putative flagellar biosynthesis protein
-	03451	35350	05541	32651	39521	07751	1.1.1	flagellar motor protein motB
-	-	35381	05581	32691	39561	07711	1.1.1	flagellar hook-length control protein fliK
-	-	35431	05631	32741	39641	07651	6.0.0	putative regulatory protein
-	-	35591	05801	39821	-	07451	5.1.4	putative transposase for IS1667
-	-	35641	05851	32951	-	07391	1.1.1	putative flagellar M-ring protein
-	-	-	-	-	-	07181	5.1.4	putative transposase for IS1667
-	03962	-	-	-	-	-	2.2.3	putative exinuclease
-	03971	-	-	-	-	-	5.1.4	putative transposase subunit
-	03991	-	-	-	-	-	5.1.4	putative transposase
-	04121	-	-	-	-	-	2.2.9	putative peptidase
-	05041	-	-	-	-	-	3.5.3	nickel-dependent hydrogenase, b-type cytochrome subunit
-	-	35831	06051	33151	39831	07161	5.1.4	putative transposase
-		35850	06060	33160	39851	07142	6.0.0	putative transcriptional regulatory protein
-	-	35880	06092	33192	39901	07103	5.1.4	putative transposase
YE3630	03291	35921	06141	33231	40010	07071	1.5.2	putative cation efflux system protein
YE3636	03241	35981	06191	33281	40051	07021	1.5.0	putative secretion system protein
YE3671	02911	36351	06531	33621	40401	06681	0.0.2	conserved hypothetical protein
-	-	36481	06671	33761	40541	06541	5.1.4	putative transposase for IS1667
-	-	-	-	-	40841	-	5.1.2	putative phage terminase
-	-	-	-	33869	-	-	5.1.2	putative phage terminase, ATPase subunit
-	-	36611	06851	40775	-	-	5.1.2	putative phage protein
-	-	36621	06861	33891	-	-	5.1.2	putative phage replication protein
YE3690	02721	36761	07011	34031	41061	06461	1.5.3	photransferase system IIBC component
YE3694	02681	36792	07042	34062	41101	06430	6.0.0	Putative transcriptional regulator
YE3697	02651	36811	07061	34081	41111	06411	4.1.1	putative membrane protein
YE3700	-	36861	07111	34121	41171	06361	4.0.0	hypothetical protein
YE3722	02401	37081	07331	34341	41421	06141	3.3.4	probable NADP-dependent alcohol dehydrogenase
-	-	-	-	-	-	06041	5.1.4	transposase
-	-	-	-	-	YE1293-41551	06021	0.0.0	hypothetical protein
YE3734	02291	37231	07491	34501	41671	05931	0.0.2	conserved hypothetical protein
YE3735	02281	37241	07501	34511	41681	05921	3.1.7	glutamate synthase [NADPH] large chain precursor
YE3753	02101	37421	07681	34701	41891	05731	1.5.2	putative sodium/calcium exchanger protein
YE3774	01911	37611	07871	34891	42101	05541	1.5.3	PTS system, trehalose-specific IIBC component
YE3776	01891	37631	07891	34911	42131	05501	1.5.2	Mg2+ transporter ATPase
YE3778	01871	37661	07921	34941	42181	05471	4.1.0	putative exported protein
YE3779	01861	37671	07931	34951	42191	05451	0.0.2	conserved hypothetical protein
-	-	37771	08031	-	42291	05341	5.1.4	putative transposase for IS1667
-	01721	-	-	-	-	05301	5.1.2	hypothetical phage-related protein
-	01711	-	-	-	-	05281	5.1.2	hypothetical phage-related protein
-	01581	-	-	-	-	05151	5.1.2	putative tail sheath protein
-	01571	-	-	-	-	05141	5.1.2	hypothetical phage-related protein
-	01501	-	-	-	-	05061	5.1.2	putative portal vertex protein
-	-	37861	08111	-	-	-	5.1.2	putative phage tail sheath protein
-	-	37921	08171	-	-	-	5.1.2	putative phage tail protein
-	-	38151	08501	35141	42391	04881	6.3.6	putative lysR-family transcriptional regulator
-	-	38171	08521	35161	42411	04861	1.4.1	putative insecticidal toxin complex protein

-	-	38181	08531	35171	42421	04841	5.1.5	putative insecticidal toxin complex protein
-	-	38241	26241	-	-	-	1.4.1	putative insecticidal toxin complex protein
YE3818	01101	38481	26471	35441	42711	04551	6.1.1	possible regulator invovled in cell signalling
YE3825	-	-	-	-	-	-	5.1.4	transposase for insertion sequence element IS1328
YE3828	01041 *	38541 *	26512	35500	42772	04502 *	7.0.0	hypothetical protein
YE3837	00971	38641	26581	35561	42851	04431 *	4.1.1	putative inner membrane protein
-	-	-	-	35761	-	-	5.1.2	putative phage tail protein
-	-	-	-	36081	-	-	5.1.2	putative phage protein
YE3860	-	38901	26831	36341	43121	04171	4.1.0	putative exported protein
-	00061	39021	26961	36471	43241	04021	5.1.2	hemolysin
-	-	-	-	36541	43302	03951	5.1.4	putative transposase
-	-	00001	39951	19531	03480	32671	5.1.4	transposase
-	30761	19588	13981	25091	20771	38081	5.1.5	putative type III secretion protein
-	30841	19580	13901	25011	20691	38161	5.1.5	putative pathogenicity island protein
-	30861	19578	13881	24991	20681	38181	5.1.5	putative pathogenicity island protein
-	30891	19575	13851	24961	20651	38202	5.1.5	putative type III secretion apparatus
-	31051	19482	13691	24801	20491	38381	6.1.1	putative sensor protein
-	-	27711	39571	19401	20291	03831	5.1.4	transposase for IS1668
YE3956	37891	28341	27501	18781	19651	03261 *	4.1.1	putative membrane protein
YE3964	37971	28421	27421	18701	19561	03181	3.7.2	cytosine deaminase
YE3967	38001	28541	27391	18671	19521	03151	1.5.4	putative nitrite transporter
YE3970	38031	28481	27361	18641	19491	03111	3.4.3	phosphoglycolate phosphatase
-	-	28811	27031	-	19121	02761	5.1.4	putative transposase for IS1667
YE3989	38361	28880	24151	18271	19071	02711	0.0.2	conserved hypothetical protein
YE4001	38240	28721	27121	18401	19221	02861	1.5.5	putative MFS-family transport protein
YE4002	-	28711	27131	18411	19231	02871	7.0.0	hypothetical protein
YE4026	38581 *	29081	23931	18051	18851	02491	3.4.3	myo-inositol 2-dehydrogenase
YE4041	38751	29231	23771	17891	18681	02331	3.4.2	glutamate mutase
YE4046	38801	29281	23721	17841	18631	02281	1.5.1	putative sodium-dicarboxylic acid transporter family protein, precursor
-	-	-	-	-	-	02271	0.0.2	hypothetical protein
-	-	-	-	-	-	02241	0.0.2	hypothetical protein
-	-	-	-	-	-	02141	5.1.4	Tn7-like transposition protein
YE4052	38861	29341	23661	17781	18571	02061	2.1.4	metalloprotease
-	39081	-	-	-	-		5.1.4	putative integrase
YE4062	39151	29421	23581	17701	18491	01971	4.1.0	putative exported protein
YE4072	39261	29521	23481	17601	18381	01871	4.1.4	probable BcsC protein involved in cellulose synthesis
YE4074	39291 *	29551	23451	17571	18351	08141	4.1.1	cellulose synthase 1 catalytic subunit[UDP-forming]
YE4086	39411	29711	38471	17411	18141	39841	4.1.0	putative exported protein, precursor
-	-	29751	38431	17371	18091	39891	5.1.3	transposase
-	39521	29882	38921	17231	17911	-	4.1.5	putative fimbrial usher protein
YE4096	39511	29921	38272	17222	17902	-	0.0.0	hypothetical protein
YE4097	-	29911	38273	17223	17903	-	0.0.0	hypothetical protein
YE4101	39561	29851	38331	17271	17971	-	4.1.3	Putative outer membrane porin
YE4106	39611	29801	38381	17321	18021	39941	7.0.0	Hypothetical ROK family protein
-	-	29891	38281	-	-	-	5.1.4	putative transposase for IS1667
YE4113	-	30041	38202	13991	35001	01631	5.1.2	putative phage primase
YE4114	-	30061	38204	-	-	01611	7.2.1	conserved hypothetical protein
YE4115	-	30091	38205	-	-	01601	7.2.1	conserved hypothetical protein
-	-	-	-	17181	17861 *	-	5.1.4	transposase subunit istA
YE4118	39741	30111	21141	17131	17791	01581	6.1.1	putative AraC-family transcriptional regulator
YE4123	39791	30161	21091	17081	17741	01531	3.4.3	xylulose kinase
YE4125	39821	30191	21061	17051	17711	01501	6.1.1	Putative transcriptional regulator
YE4135	39901	30291	20981	16971	17631	01421	3.5.3	formate dehydrogenase-o, major subunit - ALL ORTHOS WITH READ-THROUGH OF STOP CODON

YE4142	39961	30351	20921	16911	17561	01351	7.0.0	hypothetical protein
YE4145	39991	30381	20891	16881	17531	01321	4.1.1	Putative inner membrane protein
YE4167	40211	30651	20631	16621	17301	40661	7.0.0	pyridoxal-phosphate dependent enzyme
YE4174	40291	30730	20561	16551	17212	40722	0.0.0	hypothetical protein
YE4185	40401	30831	20451	16441	17101	40811	6.1.1	Probable two-component sensor kinase
-	40981	32651	17481	36711	01891	40041	5.1.1	putative Colicin

Table 9.8: Number of total CDSs, pseudogenes, and partial genes per group for the *Y. enterocolitica* reference genomes.

group	description	5303	8081	21202	5603	14902	1203	3094
1	Unknown function, no known homologues	857	662	662	663	686	664	722
	pseudo	7	3	10	11	11	28	36
	partial	8	9	13	12	14	19	14
2	Cell processes	95	84	118	118	118	102	117
	pseudo	2	0	1	1	1	1	5
	partial	7	1	0	1	1	2	0
3	Protection responses	69	54	60	59	57	55	56
	pseudo	3	0	0	0	0	0	7
	partial	3	0	3	1	2	2	1
4	Transport/binding proteins	533	497	483	483	483	477	472
	pseudo	4	3	6	5	5	14	23
	partial	1	8	10	9	10	14	17
5	Adaptation	52	56	49	49	49	45	49
	pseudo	0	1	0	0	0	0	3
	partial	1	0	3	3	1	2	1
6	Cell division	25	24	24	24	24	24	24
	pseudo	0	0	0	0	0	0	0
	partial	0	0	0	0	0	0	0
7	Macromolecule metabolism	87	86	89	89	89	88	86
	pseudo	0	1	0	0	0	1	2
	partial	0	1	3	3	4	3	3
8	Macromolecule synthesis, modification	255	229	233	233	235	232	232
	pseudo	2	0	2	1	2	2	4
	partial	5	0	2	2	1	5	1
9	Metabolism of small molecules	108	100	100	101	101	100	99
	pseudo	0	0	1	1	1	2	2
	partial	1	1	0	0	0	1	0
10	Biosynthesis of cofactors, carriers	129	128	128	128	128	128	128
	pseudo	1	0	1	0	0	1	2
	partial	0	1	2	2	3	2	1
11	Central intermediary metabolism	170	158	159	158	161	160	160
	pseudo	1	0	0	0	1	4	4
	partial	0	2	5	4	5	4	9
12	Degradation of small molecules	124	123	121	120	120	120	118
	pseudo	2	0	1	1	1	2	3
	partial	0	2	1	1	1	2	4
13	Energy metabolism, carbon	159	127	143	143	142	142	142
	pseudo	1	0	0	0	1	0	6
	partial	1	1	2	2	3	3	1
14	Fatty acid biosynthesis	28	28	23	24	24	24	24
	pseudo	0	0	0	0	1	1	1
	partial	0	0	1	1	1	1	1

15	Nucleotide biosynthesis	27	26	27	27	28	27	27
	pseudo	0	0	1	1	0	0	0
	partial	0	0	0	0	0	0	1
16	Cell envelope	648	674	635	636	630	632	624
	pseudo	8	5	4	1	4	10	26
	partial	4	2	10	10	8	15	19
17	Ribosome constituents	67	67	67	67	67	67	67
	pseudo	0	0	0	0	0	0	0
	partial	0	0	0	0	0	0	0
18	Extrachromosomal	370	340	265	260	300	278	380
	pseudo	6	11	27	24	17	25	43
	partial	6	20	26	25	29	37	37
19	Regulation	284	280	270	271	272	266	267
	pseudo	2	0	3	5	3	7	15
	partial	0	3	10	8	11	11	13
20	Not classified (included putative assignments)	278	310	264	264	264	262	258
	pseudo	2	3	4	4	1	11	11
	partial	3	6	5	2	3	3	6

Table 9.9: Signal values for *Y. enterocolitica* biotypes.

		BT 1A O:5			BT 1B O:8			BT 2 O:9			BT 3 O:9		
	well	5303_1	5303_2	5303_3	8081_1	8081_2	8081_3	21202_1	21202_2	21202_3	5603_1	5603_2	5603_3
C	PM1-A01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C	PM1-A02	241.176	200.562	193.850	184.028	175.074	163.372	196.353	194.261	144.022	208.628	210.233	151.099
C	PM1-A03	258.504	185.932	195.264	172.622	122.717	94.532	246.721	214.416	107.101	239.961	216.084	188.005
C	PM1-A04	-0.799	-2.058	-6.778	-1.818	-8.464	-1.624	-0.813	-2.779	-2.804	-3.782	2.097	-1.625
C	PM1-A05	199.845	220.725	191.599	161.857	157.349	139.443	156.592	66.034	159.289	11.783	174.145	6.780
C	PM1-A06	272.304	261.607	255.479	243.979	255.362	249.279	271.713	260.684	223.811	250.451	268.886	198.951
C	PM1-A07	248.851	261.086	206.379	211.217	197.992	192.501	180.938	11.867	183.176	9.071	205.122	1.705
C	PM1-A08	52.307	6.146	5.325	17.457	66.507	47.261	105.617	67.653	57.139	7.689	65.739	8.313
C	PM1-A09	44.171	33.714	18.509	76.626	76.807	71.609	3.842	2.163	4.772	-1.895	1.209	-6.080
C	PM1-A10	165.200	126.650	133.330	135.545	101.422	107.236	173.805	147.395	86.961	179.257	166.201	126.221
C	PM1-A11	178.542	88.459	94.772	105.843	77.858	79.176	168.592	152.493	124.851	167.374	170.997	93.203
C	PM1-A12	8.103	7.603	7.604	-0.067	3.974	6.517	52.500	12.192	7.700	1.417	6.555	10.053
C	PM1-B01	271.741	269.005	275.807	2.857	3.175	0.120	157.834	0.103	-0.739	2.117	5.300	-0.816
C	PM1-B02	236.516	211.220	209.955	156.905	138.324	131.226	212.780	189.246	103.321	177.779	177.983	142.509
C	PM1-B03	210.745	147.296	146.320	161.575	133.720	120.101	112.500	113.183	82.274	94.466	113.720	77.020
C	PM1-B04	4.589	-7.330	-3.214	-0.122	-4.076	-2.812	3.050	0.183	-0.099	-1.275	-2.601	-4.904
C	PM1-B05	292.233	290.274	288.182	216.776	237.013	219.259	282.005	274.286	227.104	250.759	266.679	213.230
C	PM1-B06	301.589	292.374	294.604	225.409	248.554	224.046	285.282	284.932	240.789	276.405	283.451	216.333
C	PM1-B07	208.853	213.496	210.010	96.216	116.664	93.525	207.004	177.822	105.737	165.067	195.411	111.458
C	PM1-B08	242.786	190.076	195.857	204.086	183.771	171.134	194.387	184.447	152.320	204.467	200.436	143.457
C	PM1-B09	8.624	-1.492	-0.086	4.503	-6.278	-2.543	85.530	0.992	0.138	3.533	1.651	1.825
C	PM1-B10	72.570	90.757	78.026	-5.771	-2.329	0.403	8.195	13.230	0.097	1.938	9.842	6.596
C	PM1-B11	224.943	187.486	194.759	137.768	108.242	111.591	171.203	168.376	82.026	197.683	150.679	136.724
C	PM1-B12	68.588	76.179	55.338	62.912	82.192	90.726	137.732	133.289	137.988	13.807	124.067	18.720
C	PM1-C01	260.664	249.316	240.900	220.117	212.755	199.447	254.759	263.343	227.571	240.076	261.934	194.120
C	PM1-C02	48.946	118.237	117.133	0.284	1.005	-2.307	0.104	1.767	-3.214	-1.354	-2.133	-0.103
C	PM1-C03	202.671	216.174	201.872	177.388	159.433	137.899	190.914	179.005	136.212	4.307	247.880	3.836
C	PM1-C04	218.421	171.712	168.029	174.987	144.312	139.836	154.822	148.120	100.616	183.468	187.746	97.167
C	PM1-C05	49.307	45.971	23.326	8.905	4.489	10.839	12.801	8.346	5.841	4.739	6.933	8.642
C	PM1-C06	2.054	-6.151	-1.367	-2.237	-8.070	-3.624	3.470	-2.738	-2.525	-2.828	-1.841	-3.892
C	PM1-C07	120.701	58.826	64.508	79.913	52.370	61.076	186.318	159.091	76.389	220.229	144.541	151.151
C	PM1-C08	193.379	160.957	161.138	113.049	102.834	114.853	139.628	124.705	82.374	137.882	151.916	87.482
C	PM1-C09	176.970	99.007	115.584	85.746	53.350	59.462	224.408	196.166	71.659	213.571	174.557	162.949
C	PM1-C10	208.180	155.657	162.913	173.616	162.661	161.334	205.725	179.722	160.012	189.782	199.507	156.158
C	PM1-C11	2.659	-7.974	-6.992	0.022	-5.943	-4.357	84.626	-1.849	-0.380	-2.224	0.624	-2.063
C	PM1-C12	42.786	19.500	20.496	11.458	0.350	6.138	1.484	11.196	13.107	13.707	7.188	11.307
C	PM1-D01	192.583	187.875	169.132	154.275	129.886	128.361	9.817	8.287	148.975	2.663	103.457	0.205
C	PM1-D02	-1.546	-8.368	-6.075	-6.984	-9.787	-9.424	-3.036	-10.766	-3.447	-6.758	-10.609	-13.279
C	PM1-D03	0.904	-4.408	-7.175	-3.245	-8.914	-5.817	1.612	-3.395	-0.245	-3.789	-4.225	-8.375
C	PM1-D04	3.843	0.251	4.491	3.103	-1.051	-1.624	4.889	2.650	-1.608	0.939	4.592	0.822
C	PM1-D05	106.316	4.674	118.725	-0.328	-1.014	-0.870	4.445	3.950	1.025	-0.942	1.217	0.578
C	PM1-D06	0.005	-7.545	-4.566	-6.112	-7.522	-6.874	1.879	-1.220	-5.036	-3.964	-9.222	-4.228
C	PM1-D07	0.272	-5.789	-7.878	-6.605	-8.149	-11.078	-4.718	3.197	-8.232	-3.330	-1.299	-5.497
C	PM1-D08	2.124	-7.179	1.314	-1.957	-2.129	-3.954	3.883	1.143	-2.733	-3.876	-0.992	-1.145
C	PM1-D09	203.461	116.243	138.660	-2.500	-6.442	-2.845	1.789	0.557	-0.953	0.829	1.193	1.407
C	PM1-D10	90.712	122.411	80.599	0.134	-3.201	-1.783	3.872	2.464	1.668	0.686	0.113	-4.101
C	PM1-D11	111.366	68.858	79.520	79.405	64.463	71.200	207.958	143.382	60.322	174.809	86.155	147.410
C	PM1-D12	158.414	150.678	153.528	39.620	20.712	21.962	144.024	124.174	41.292	83.617	136.028	31.461
C	PM1-E01	56.732	62.997	20.966	5.676	8.482	6.624	6.528	53.154	106.855	3.545	89.401	5.584
C	PM1-E02	-1.070	-2.562	-2.862	-5.188	-2.211	-8.500	-0.930	0.225	-2.675	-8.321	-8.043	-1.347
C	PM1-E03	258.734	254.450	250.682	207.187	195.982	190.928	253.762	235.328	193.595	223.129	248.471	185.716
C	PM1-E04	246.358	235.766	233.250	187.483	196.397	183.096	258.391	244.387	199.704	220.537	254.901	135.305
C	PM1-E05	14.482	5.634	7.464	27.249	4.743	28.817	7.625	8.578	4.539	3.130	11.492	1.829
C	PM1-E06	18.229	20.754	-5.682	-3.392	-4.878	-2.182	8.228	4.997	-2.041	1.947	0.084	-2.855
C	PM1-E07	2.842	-8.454	-0.593	-0.887	-5.676	-5.564	0.609	0.745	-2.766	-1.509	-2.004	-3.861
C	PM1-E08	208.886	132.305	150.013	182.275	157.529	155.555	193.567	190.892	159.505	192.682	176.257	138.218
C	PM1-E09	2.849	0.976	-0.789	-1.914	-8.813	-7.399	2.257	2.705	-3.564	-0.155	2.017	-4.870
C	PM1-E10	169.264	98.225	110.663	143.363	127.538	128.174	191.792	159.907	130.422	190.428	169.634	139.701
C	PM1-E11	224.818	224.801	231.818	121.275	131.200	133.020	200.750	191.437	121.782	181.137	194.845	139.207
C	PM1-E12	233.875	231.226	224.045	126.228	136.720	109.512	236.863	208.037	143.195	191.362	212.401	136.118
C	PM1-F01	215.928	194.618	192.146	92.704	83.592	97.579	191.304	16.622	38.392	20.128	178.766	13.864
C	PM1-F02	-2.395	-6.904	-2.045	-3.589	-3.134	-5.724	1.896	-3.982	-1.100	-6.184	-9.836	-8.259
C	PM1-F03	2.954	76.620	1.467	215.187	220.780	202.538	242.091	218.986	197.404	220.105	240.278	191.329
C	PM1-F04	0.409	-9.493	-6.205	-2.196	-8.167	-6.082	3.170	-4.192	0.536	-3.789	-7.757	-8.076
C	PM1-F05	246.484	260.901	239.008	177.589	170.076	110.815	42.109	21.083	155.938	3.654	27.105	2.759

C	PM1-F06	131.317	121.233	121.243	52.467	28.437	-0.378	3.900	73.403	52.876	1.183	105.816	0.237
C	PM1-F07	-3.200	-8.238	-6.449	6.495	-3.780	-9.079	2.588	0.925	-1.112	0.822	3.929	-6.934
C	PM1-F08	3.441	-1.024	1.270	1.654	1.214	-2.553	-2.139	2.661	0.442	0.826	6.375	0.474
C	PM1-F09	-1.070	-2.696	-9.886	-4.407	-8.366	-5.534	-4.295	-4.047	-4.329	-6.207	-6.197	-7.439
C	PM1-F10	-4.728	-9.846	-5.589	-4.451	-5.372	-7.003	-9.161	-2.942	-6.586	-4.695	-2.911	-7.832
C	PM1-F11	200.979	151.621	164.954	147.455	134.616	129.005	183.676	171.095	118.855	171.130	179.690	127.735
C	PM1-F12	222.025	215.132	227.580	79.664	95.016	81.447	186.547	179.491	117.062	180.424	183.278	127.030
C	PM1-G01	161.858	140.783	132.726	94.113	127.093	105.617	130.611	115.337	84.855	99.676	149.724	56.808
C	PM1-G02	-1.700	-3.057	-2.516	-5.451	-2.922	-4.451	0.816	-0.632	-4.159	-1.671	-1.441	-6.296
C	PM1-G03	275.480	269.484	265.546	184.238	212.125	201.120	262.463	249.742	224.174	243.280	260.899	198.788
C	PM1-G04	2.021	-2.509	-1.804	20.045	1.109	27.195	3.620	3.926	45.403	0.988	-0.307	-3.179
C	PM1-G05	188.217	187.704	188.710	91.826	79.745	79.626	177.689	161.100	135.155	140.301	161.792	131.428
C	PM1-G06	153.349	163.438	139.703	117.095	107.917	101.040	142.672	130.771	105.701	125.588	152.766	107.493
C	PM1-G07	7.843	0.762	4.750	4.421	3.950	5.205	2.149	6.100	2.020	2.967	3.789	1.620
C	PM1-G08	238.850	231.224	239.439	213.174	216.525	205.396	229.687	231.812	184.976	220.921	234.486	197.307
C	PM1-G09	73.941	38.395	29.055	43.491	32.000	15.042	14.271	9.174	17.396	10.647	24.113	6.082
C	PM1-G10	213.525	193.962	189.749	140.216	121.899	132.945	201.913	192.164	141.628	170.509	189.950	118.196
C	PM1-G11	1.841	2.491	1.612	1.055	2.312	0.463	-0.338	2.732	2.426	-1.970	-4.178	0.343
C	PM1-G12	232.237	240.071	228.122	183.451	173.880	158.120	188.845	185.821	166.749	41.950	238.621	10.413
C	PM1-H01	177.061	159.684	152.147	137.889	160.524	147.657	227.857	196.172	188.199	195.671	212.283	172.714
C	PM1-H02	2.654	1.342	2.979	-0.462	2.875	-0.821	94.688	9.586	2.499	0.325	5.324	1.738
C	PM1-H03	-2.028	0.428	0.317	0.296	-0.296	-5.263	131.853	9.445	4.001	-0.218	-1.724	-5.779
C	PM1-H04	2.080	7.314	3.737	-2.126	4.441	-0.700	5.333	3.417	4.158	-1.245	3.561	1.271
C	PM1-H05	11.886	6.453	7.900	2.979	6.096	2.575	0.499	5.136	4.757	-0.254	0.043	2.741
C	PM1-H06	31.442	2.022	4.642	3.908	-0.300	-3.692	-2.587	-7.930	-1.130	-2.300	-1.575	-4.601
C	PM1-H07	14.878	-2.236	4.293	11.078	3.796	1.161	-0.780	4.087	2.743	10.392	7.054	1.964
C	PM1-H08	239.428	240.089	241.874	143.795	175.682	145.436	236.333	241.818	199.788	184.316	251.008	143.707
C	PM1-H09	4.289	1.017	4.814	4.296	5.084	1.207	2.129	5.292	4.228	2.378	12.146	5.634
C	PM1-H10	253.942	250.847	260.285	203.979	218.941	216.903	245.039	247.797	239.274	246.408	249.561	222.618
C	PM1-H11	-1.345	-1.134	2.116	-3.558	3.141	-3.250	-2.478	2.271	3.711	-1.153	8.046	4.313
C	PM1-H12	-1.774	2.711	1.474	-4.511	0.049	-4.057	-3.461	3.107	2.446	-3.024	5.055	2.803
C	PM2-AA01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C	PM2-AA02	-4.030	-8.268	-8.393	-2.882	-2.705	-4.283	-7.175	-6.688	-5.455	3.453	-7.305	-4.884
C	PM2-AA03	-3.608	-2.251	-4.446	-0.516	-1.305	-3.113	-2.345	-5.796	-3.608	2.643	-7.445	-5.066
C	PM2-AA04	-3.767	-3.163	-4.151	-1.987	-1.609	-2.780	-3.651	-5.612	-4.267	1.900	-6.633	-6.026
C	PM2-AA05	-4.263	-5.033	-10.083	-2.792	-2.012	-5.028	-5.645	-9.436	-5.022	-1.255	-8.380	-6.847
C	PM2-AA06	84.113	68.784	48.717	37.413	41.446	34.183	40.901	7.401	32.172	70.026	42.251	3.570
C	PM2-AA07	-4.289	-5.196	-9.259	-3.574	-4.401	-3.951	-8.046	-8.087	-6.908	-4.720	-10.857	-5.825
C	PM2-AA08	1.304	-0.576	-1.991	-2.793	-1.336	-1.055	-1.186	-5.608	-1.743	5.083	-9.297	-6.266
C	PM2-AA09	-4.476	-7.682	-5.775	-2.143	-1.324	-2.184	-3.971	-10.668	-3.920	-3.263	-10.550	-8.114
C	PM2-AA10	-7.659	-9.803	-7.070	-3.145	-2.168	-7.501	-4.359	-9.909	-6.668	-1.842	-12.332	-6.864
C	PM2-AA11	-3.193	-1.345	-3.772	-3.033	-2.666	-1.943	-2.200	-8.655	-2.237	2.183	-9.703	-7.782
C	PM2-AA12	176.596	157.680	129.262	94.907	113.217	96.012	201.807	160.011	104.447	30.587	175.884	87.991
C	PM2-AB01	171.041	174.616	191.495	-4.203	-2.136	-2.876	167.903	151.432	143.399	168.296	144.942	130.468
C	PM2-AB02	241.600	246.405	243.889	185.030	206.486	182.668	236.112	219.287	210.128	223.542	227.770	187.729
C	PM2-AB03	-7.157	-9.030	-11.143	-7.830	-9.433	-10.793	-10.172	-13.784	-12.589	-2.733	-13.689	-13.642
C	PM2-AB04	4.792	0.738	-6.534	-4.184	-2.839	-4.112	-5.214	-11.182	-6.447	-3.258	-11.154	-9.388
C	PM2-AB05	-8.080	-7.347	-10.511	-6.057	-5.197	-7.920	-9.391	-15.375	-10.505	-0.772	-12.218	-11.311
C	PM2-AB06	259.628	252.371	254.628	-6.722	-5.959	-6.609	-6.737	-11.676	-9.305	-1.430	-14.500	-8.783
C	PM2-AB07	-5.503	-7.434	-7.579	-6.195	-8.121	-8.017	-5.096	-15.211	-9.616	-2.157	-14.117	-9.991
C	PM2-AB08	194.016	190.438	199.961	135.978	136.563	157.158	38.659	7.476	2.832	5.762	-2.807	-3.088
C	PM2-AB09	132.229	114.605	125.655	109.992	112.372	89.586	174.718	85.866	57.663	119.925	154.543	105.616
C	PM2-AB10	-13.439	-13.664	-14.936	-9.788	-9.662	-13.896	-9.982	-16.588	-12.204	-6.874	-19.558	-13.437
C	PM2-AB11	-9.664	-9.147	-11.433	-6.283	-8.514	-7.996	-10.168	-13.951	-8.095	-5.346	-15.672	-10.946
C	PM2-AB12	200.166	201.497	194.883	12.466	8.143	15.305	9.742	1.541	2.114	28.141	-1.271	7.220
C	PM2-AC01	125.987	126.261	124.750	124.547	112.450	115.466	204.958	187.463	177.218	200.283	201.737	144.384
C	PM2-AC02	-4.653	-4.514	-9.197	-5.543	-7.084	-8.786	-6.961	-11.955	-8.366	-1.418	-10.049	-6.687
C	PM2-AC03	-5.147	-7.275	-9.809	-5.455	-9.629	-8.562	-7.293	-13.876	-8.754	-1.842	-14.384	-14.755
C	PM2-AC04	-5.974	-7.903	-10.957	-5.982	-6.704	-6.787	-7.454	-13.529	-8.193	-1.530	-13.589	-11.932
C	PM2-AC05	-5.801	-7.455	-7.409	-4.341	-9.880	-6.042	-4.732	-10.637	-6.553	-0.317	-11.316	-8.891
C	PM2-AC06	-7.238	-6.963	-9.838	-9.421	-9.534	-9.334	-10.822	-12.462	-10.276	0.533	-15.599	-12.032
C	PM2-AC07	90.151	106.462	84.850	-3.091	-2.899	-3.491	1.386	-3.509	1.346	1.676	-8.014	-7.403
C	PM2-AC08	-4.339	-4.687	-7.137	-5.962	-8.884	-8.376	-7.846	-11.571	-10.967	-1.276	-14.122	-9.346
C	PM2-AC09	-8.334	-9.972	-13.882	-8.978	-11.393	-12.009	-8.051	-15.539	-13.253	-4.134	-17.767	-14.687
C	PM2-AC10	-11.954	-11.379	-13.968	-10.649	-9.363	-12.447	-9.524	-16.803	-11.912	-5.758	-18.675	-12.253
C	PM2-AC11	-7.538	-1.525	-9.247	-5.934	-4.711	-7.187	-8.063	-10.299	-9.339	-2.484	-14.432	-8.888
C	PM2-AC12	-6.859	-4.755	-7.418	-5.124	-4.287	-3.736	-3.076	-9.955	-3.486	0.766	-15.296	-5.970
C	PM2-AD01	55.949	83.774	42.850	-3.132	-3.822	-5.403	-2.888	-9.970	-5.195	1.878	-10.579	-8.983
C	PM2-AD02	213.626	214.393	219.499	-6.159	-4.207	-6.387	-0.704	-11.020	-8.101	1.251	-10.379	-10.518

C	PM2-AD03	-5.505	-4.546	-8.216	-5.712	-4.816	-4.668	-6.347	-9.621	-8.433	-0.586	-8.632	-8.663
C	PM2-AD04	161.620	165.739	170.645	115.897	118.236	101.391	158.566	123.836	115.480	123.340	122.493	87.286
C	PM2-AD05	-8.207	-7.603	-10.432	-3.793	-6.464	-8.033	-10.317	-10.003	-8.150	-0.543	-12.987	-9.108
C	PM2-AD06	-4.207	-2.096	-4.841	-5.286	-4.674	-5.692	-5.001	-7.150	-7.542	2.045	-6.842	-5.709
C	PM2-AD07	-2.518	-9.987	-11.263	-3.389	-8.512	-5.842	-6.450	-12.678	-4.530	0.796	-10.567	-10.184
C	PM2-AD08	-6.091	-9.770	-10.933	-6.392	-6.566	-8.053	-6.587	-10.253	-8.013	-2.668	-11.937	-9.139
C	PM2-AD09	-7.525	-9.499	-12.866	-7.411	-10.496	-8.075	-8.029	-13.195	-12.338	-2.143	-16.672	-10.850
C	PM2-AD10	-7.543	-7.386	-9.667	-8.164	-8.504	-8.737	-7.147	-13.467	-8.958	-3.688	-15.250	-8.561
C	PM2-AD11	-9.529	-9.611	-10.613	-8.559	-5.564	-12.454	-10.811	-13.391	-10.799	-2.824	-15.403	-9.547
C	PM2-AD12	-4.093	-5.176	-8.763	0.651	-0.030	-2.900	-5.272	-12.261	-5.061	1.558	-8.526	-3.911
C	PM2-AE01	-11.563	-9.364	-22.391	-16.491	-12.083	-16.639	-17.833	-13.649	-14.696	-7.538	-17.779	-15.443
C	PM2-AE02	-2.020	-3.533	-5.378	-2.847	-6.058	-3.762	-3.334	-4.884	-14.187	4.717	-7.607	-8.282
C	PM2-AE03	-4.034	-1.757	-5.912	-5.422	-4.222	-8.836	-4.453	-7.746	-8.583	-0.728	-7.443	-9.709
C	PM2-AE04	-7.028	-6.933	-7.630	-6.632	-8.457	-8.964	-6.286	-8.186	-12.241	-1.426	-6.611	-13.826
C	PM2-AE05	220.189	229.026	234.767	185.068	204.612	194.866	225.476	213.812	215.143	214.036	223.161	189.596
C	PM2-AE06	-9.579	-11.017	-13.242	-7.479	-9.079	-11.520	-12.512	-16.439	-14.907	-3.596	-15.983	-13.283
C	PM2-AE07	-6.876	-5.720	-8.776	-6.800	-5.180	-8.687	-6.542	-12.263	-12.714	-2.046	-12.095	-9.033
C	PM2-AE08	-6.062	-1.829	-7.134	-6.342	-6.059	-9.526	-7.312	-12.621	-9.291	0.146	-11.842	-8.583
C	PM2-AE09	-7.030	-3.997	-7.653	-5.163	-8.172	-9.491	-7.336	-9.322	-8.471	-1.100	-9.546	-5.892
C	PM2-AE10	-8.593	-8.347	-10.707	-6.536	-3.962	-11.989	-7.511	-14.126	-11.153	-3.075	-15.587	-11.412
C	PM2-AE11	-11.699	-11.404	-23.445	-10.225	-16.689	-12.711	-14.475	-18.568	-9.662	-2.405	-20.230	-14.899
C	PM2-AE12	199.638	188.380	194.180	103.034	117.379	121.055	218.175	183.068	146.968	160.259	172.609	158.288
C	PM2-AF01	0.747	2.909	-2.516	-1.758	-1.559	-2.109	1.447	-2.943	-5.621	6.214	-2.361	-1.928
C	PM2-AF02	-1.324	-0.821	-5.959	-0.271	3.647	-2.654	-5.013	-5.030	-8.099	4.297	-4.163	-6.041
C	PM2-AF03	1.145	-0.011	-4.375	-3.746	-7.757	-7.078	108.522	-10.716	-5.703	3.008	-7.654	-6.871
C	PM2-AF04	-4.718	-2.234	-7.734	-5.770	-3.525	-5.458	-8.084	-8.254	-6.099	0.045	-6.855	-8.493
C	PM2-AF05	-4.589	-4.247	-6.649	-4.117	-4.868	-1.747	-6.670	-12.146	-7.012	-2.553	-7.416	-11.616
C	PM2-AF06	-3.520	-4.504	-8.741	-5.788	-5.034	-9.357	-6.464	-10.513	-9.463	-1.563	-8.754	-8.384
C	PM2-AF07	-5.664	-1.713	-8.516	-4.887	-4.457	-5.514	-8.468	-9.147	-9.930	1.105	-10.087	-7.034
C	PM2-AF08	-4.896	-6.557	-12.197	-9.246	-7.825	-10.705	-4.153	-16.007	-13.738	-0.650	-12.251	-14.064
C	PM2-AF09	-3.726	-8.339	-8.814	-7.218	-9.099	-9.483	-6.438	-11.661	-11.274	-1.263	-14.416	-10.642
C	PM2-AF10	31.155	47.866	21.967	23.347	27.474	17.200	14.225	-5.979	14.142	-4.200	2.921	-5.551
C	PM2-AF11	-9.280	-4.542	-8.229	-7.463	-5.750	-6.884	-11.221	-8.778	-8.817	-1.884	-14.501	-6.461
C	PM2-AF12	-0.637	1.991	-1.604	1.459	0.514	2.730	-1.767	-2.575	-5.083	5.107	-8.880	1.113
C	PM2-AG01	-0.966	-2.857	-4.370	-4.868	2.539	-1.888	-6.162	-3.495	-4.129	1.000	-3.041	-3.682
C	PM2-AG02	-0.184	-0.889	-2.307	-3.012	-4.832	-5.764	3.787	-2.391	-6.068	64.716	-0.839	-7.200
C	PM2-AG03	-7.745	-0.621	-9.089	-3.564	-6.804	-5.474	-5.679	-18.117	-9.921	2.064	-7.576	-8.886
C	PM2-AG04	-4.195	-4.301	-7.151	-3.871	-3.905	-7.786	-5.724	-11.159	-7.267	0.939	-11.407	-10.841
C	PM2-AG05	-4.632	-5.700	-7.241	-8.745	-6.564	-6.789	-6.045	-11.500	-11.407	1.626	-10.196	-12.030
C	PM2-AG06	205.472	251.191	215.575	44.820	65.678	56.500	-0.357	97.570	146.709	3.475	125.930	-11.974
C	PM2-AG07	-8.400	-4.196	-10.653	-8.158	-7.582	-9.734	-11.211	-14.387	-12.005	-1.139	-15.964	-9.230
C	PM2-AG08	-5.279	-5.243	-8.133	-4.446	-3.751	-6.925	-4.204	-11.449	-5.729	3.784	-11.082	-9.189
C	PM2-AG09	-5.829	-10.149	-11.936	-6.526	-5.800	-7.284	-11.211	-13.625	-9.472	-2.554	-12.037	-10.839
C	PM2-AG10	-14.782	-20.379	-17.521	-5.808	-8.333	-6.861	-14.683	-26.279	-8.146	-5.379	-18.267	-8.813
C	PM2-AG11	-7.305	-6.612	-7.445	-4.937	-5.409	-5.784	-8.086	-11.124	-6.442	-3.441	-14.904	-6.971
C	PM2-AG12	-7.074	-3.374	-3.016	-1.553	0.566	-1.141	-4.717	-8.384	-1.407	4.593	-9.880	0.629
C	PM2-AH01	0.795	1.282	-4.974	-0.818	-3.562	-5.417	-0.346	-1.492	-3.022	2.364	-5.525	-5.033
C	PM2-AH02	1.804	2.949	-1.278	-0.276	-4.580	-2.386	-1.520	-8.658	1.914	5.626	-4.850	-3.833
C	PM2-AH03	5.808	4.659	0.195	3.561	1.683	-0.817	-1.041	-1.561	-1.332	9.271	-4.812	-4.282
C	PM2-AH04	0.984	-2.830	-4.289	-3.751	-5.297	-3.504	-5.164	-5.891	0.204	5.743	-4.303	-6.354
C	PM2-AH05	-4.429	-0.883	-4.559	-6.897	0.408	-3.684	-5.843	-8.516	-4.272	-0.607	-7.324	-8.992
C	PM2-AH06	-8.143	-7.067	-13.016	-2.745	-7.493	-9.758	-12.634	-14.839	-12.799	0.549	-11.455	-16.837
C	PM2-AH07	-1.136	0.466	-0.203	-0.530	2.284	-2.536	-0.289	-1.825	-0.082	4.929	-4.296	-4.939
C	PM2-AH08	2.080	1.636	-9.158	-1.609	12.557	-5.976	-0.087	-6.730	1.292	2.129	-5.295	-6.171
C	PM2-AH09	-11.184	-18.337	-9.995	-10.637	-13.282	-17.012	-14.700	8.251	-15.466	-5.150	-25.574	2.845
C	PM2-AH10	-7.128	-4.532	-8.916	-2.209	-5.472	-4.426	-5.564	-10.024	-2.438	1.603	-11.038	-6.082
C	PM2-AH11	1.699	-0.083	-2.720	2.505	1.814	-0.647	-2.943	-12.826	-3.097	4.401	-12.154	-0.145
C	PM2-AH12	2.757	4.589	3.855	2.971	4.554	-0.678	3.167	-3.274	0.358	5.043	-4.941	-3.696
N	PM3-BA01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
N	PM3-BA02	229.267	231.834	221.362	115.321	128.717	109.933	196.847	208.853	174.132	191.216	201.893	148.771
N	PM3-BA03	8.424	10.351	24.266	7.496	2.279	8.916	7.370	2.399	2.964	2.120	3.914	-0.346
N	PM3-BA04	12.258	5.925	5.892	1.825	-0.983	2.457	12.053	5.572	3.714	5.537	3.536	-0.825
N	PM3-BA05	244.296	250.391	234.711	105.335	106.484	84.716	184.004	213.971	149.730	160.884	205.879	111.811
N	PM3-BA06	3.149	1.408	0.761	-2.791	-7.055	-5.487	20.011	2.176	-2.200	0.599	-1.709	-3.276
N	PM3-BA07	234.904	234.387	228.783	116.053	140.055	112.147	192.301	209.433	188.767	172.126	205.260	148.236
N	PM3-BA08	226.804	225.266	211.930	121.407	129.401	108.368	146.808	151.687	138.900	111.926	162.287	96.443
N	PM3-BA09	256.011	261.001	261.284	217.472	224.754	222.343	229.305	228.316	219.867	185.650	246.305	192.324
N	PM3-BA10	267.053	275.892	262.553	222.835	238.580	229.550	253.667	243.185	223.863	230.828	251.049	213.784
N	PM3-BA11	270.136	261.661	268.571	20.355	284.349	280.904	67.601	276.488	274.117	270.684	277.157	260.636

N	PM3-BA12	220.151	224.312	216.001	188.060	201.047	203.780	203.634	171.829	191.896	127.210	185.313	118.129
N	PM3-BB01	247.001	234.417	245.291	201.087	200.367	207.368	225.962	226.722	210.932	210.764	230.967	183.603
N	PM3-BB02	194.325	190.395	186.763	80.071	70.300	54.984	149.954	127.783	103.188	95.389	150.404	58.524
N	PM3-BB03	255.350	250.421	246.432	182.412	192.971	177.216	238.730	209.846	199.643	180.150	220.628	82.688
N	PM3-BB04	45.514	58.003	64.130	10.711	17.546	8.629	22.497	7.738	11.263	6.421	19.138	2.489
N	PM3-BB05	128.980	123.318	121.695	10.591	7.833	4.374	80.357	91.239	22.712	57.530	107.879	21.626
N	PM3-BB06	11.339	13.343	9.158	-5.103	-8.811	-8.245	2.537	0.547	-3.847	-3.833	-1.158	-6.568
N	PM3-BB07	130.651	136.630	118.395	76.354	85.739	62.229	46.575	61.645	46.800	6.530	78.866	-3.437
N	PM3-BB08	120.064	122.704	116.425	46.000	41.232	37.087	38.405	22.132	28.841	0.405	46.705	0.467
N	PM3-BB09	150.611	164.278	151.203	131.385	143.512	124.303	166.813	163.458	168.936	107.668	165.854	109.512
N	PM3-BB10	224.437	229.861	219.447	154.407	160.149	139.991	215.267	213.295	187.872	201.732	213.437	182.480
N	PM3-BB11	97.901	136.407	119.453	55.741	63.638	64.411	51.079	88.371	66.897	3.878	100.145	6.243
N	PM3-BB12	43.597	140.845	24.279	68.905	73.186	64.014	-2.175	0.396	0.264	-4.887	1.843	2.695
N	PM3-BC01	-17.313	-17.409	-19.088	-9.376	-26.143	-22.383	-17.351	-17.978	-22.497	-12.412	-24.442	-27.728
N	PM3-BC02	40.155	54.782	49.100	12.422	25.695	6.908	7.486	22.938	20.978	3.551	30.462	0.508
N	PM3-BC03	198.241	203.822	194.645	128.864	142.212	110.413	174.841	144.185	162.242	77.267	157.504	76.188
N	PM3-BC04	127.761	125.671	116.353	38.403	50.868	33.389	8.600	3.349	3.564	0.508	2.343	-4.612
N	PM3-BC05	-4.891	-5.751	-5.907	-9.603	-8.493	-9.645	-6.263	-3.663	-6.655	-8.372	-5.845	-7.895
N	PM3-BC06	-13.608	-11.493	-10.295	-9.600	-11.611	-8.489	-9.047	-7.179	-8.471	-10.907	-9.779	-10.047
N	PM3-BC07	-5.291	-4.983	-3.087	-4.801	-7.100	-5.905	-4.317	-3.671	-4.363	-6.432	-7.459	-7.109
N	PM3-BC08	203.561	198.287	221.396	-6.066	-7.372	-5.151	-7.645	-0.862	-4.763	-3.296	-3.816	-3.225
N	PM3-BC09	-0.357	0.721	3.404	-1.954	-1.361	1.964	-1.930	-2.093	0.903	-5.287	-4.729	-5.350
N	PM3-BC10	9.113	13.707	11.197	-1.171	0.947	1.753	0.638	-0.751	-2.282	-5.572	-1.266	-3.901
N	PM3-BC11	-0.767	2.971	2.868	-1.655	-2.359	3.721	2.247	2.570	14.825	-1.147	1.838	-0.832
N	PM3-BC12	81.571	119.576	76.972	10.205	18.908	19.463	30.482	26.599	37.162	25.486	42.246	26.609
N	PM3-BD01	-7.775	-6.717	-3.678	-0.425	-2.254	0.684	-1.424	-3.412	0.900	-4.934	-6.217	-4.112
N	PM3-BD02	-10.703	-7.797	-6.271	-5.738	-6.111	-5.857	-5.604	-5.208	-8.967	-5.508	-7.275	-6.922
N	PM3-BD03	1.182	7.349	6.000	3.533	9.284	8.643	2.095	18.333	56.093	-0.918	47.792	-1.950
N	PM3-BD04	-5.926	-6.120	-4.321	-3.071	-4.726	-1.813	-1.264	-4.433	-1.283	-4.264	-6.662	-5.707
N	PM3-BD05	-11.747	-11.254	-11.592	-10.667	-11.726	-11.750	-10.253	-9.671	-10.455	-11.349	-10.221	-8.900
N	PM3-BD06	-9.137	-7.779	-7.288	-7.057	-8.970	-7.979	-6.257	-5.495	-7.391	-8.989	-7.088	-8.562
N	PM3-BD07	-4.601	-3.617	-1.645	-5.005	-5.379	-4.132	-3.779	-3.537	-0.886	-7.003	-5.412	-5.297
N	PM3-BD08	-5.186	-4.142	-1.695	-5.876	-6.457	-4.255	-5.447	-4.072	-3.491	-5.645	-5.575	-5.779
N	PM3-BD09	-2.278	-4.745	-2.016	-4.678	-5.828	-2.764	-3.978	-3.593	-2.912	-5.967	-4.143	-5.484
N	PM3-BD10	-8.537	-7.222	-3.212	-6.338	-3.266	-0.230	-0.708	-5.443	-0.513	-5.075	-6.880	-5.432
N	PM3-BD11	-9.946	-6.201	-4.570	-8.111	-4.770	-2.075	-2.288	-3.638	-3.882	-6.214	-4.396	-4.909
N	PM3-BD12	-5.732	-1.786	-0.239	-2.672	0.451	6.918	9.729	-1.143	3.675	-3.275	-0.171	0.541
N	PM3-BE01	-7.384	-7.892	-4.851	-2.037	-3.053	-0.661	-0.700	-2.325	-2.151	-4.583	-5.661	-3.812
N	PM3-BE02	-9.261	-6.420	-5.120	-4.307	-5.408	-4.937	-3.442	-4.718	-6.211	-6.758	-6.726	-6.012
N	PM3-BE03	-7.103	-4.355	-1.647	-3.050	-4.541	-4.667	-2.733	-2.867	-5.392	-5.293	-5.051	-5.007
N	PM3-BE04	-4.862	-6.355	-2.283	-3.353	-4.345	-2.151	-2.454	-2.313	-1.971	-5.293	-5.728	-5.370
N	PM3-BE05	-5.520	-3.775	-0.463	-2.576	-2.003	2.496	2.659	-2.946	3.167	-2.124	-4.186	-4.659
N	PM3-BE06	1.675	37.995	39.243	11.839	36.476	24.729	3.767	11.588	37.279	6.913	33.416	2.233
N	PM3-BE07	186.483	192.587	181.737	-3.036	-4.543	-2.645	40.607	131.799	140.792	41.811	132.867	71.121
N	PM3-BE08	212.559	178.724	204.001	145.191	135.601	126.074	219.958	184.834	149.532	188.432	192.200	167.297
N	PM3-BE09	-6.151	-7.783	-2.649	-3.637	-3.803	-2.163	-1.945	-3.700	-1.233	-4.743	-6.200	-4.462
N	PM3-BE10	-11.628	-7.789	-6.816	-6.978	-6.341	-2.271	-2.345	-6.478	-0.571	-7.374	-8.108	-6.634
N	PM3-BE11	81.864	83.641	85.046	80.614	78.634	40.987	203.524	171.637	20.425	186.867	170.093	164.775
N	PM3-BE12	116.980	144.559	138.961	-4.529	-2.926	2.225	185.283	158.942	48.558	201.555	169.284	153.362
N	PM3-BF01	160.178	167.407	169.228	143.738	160.400	147.682	196.414	163.137	127.899	161.417	182.276	168.263
N	PM3-BF02	-11.925	-17.022	-16.800	-5.253	-20.212	-12.445	-3.557	-12.166	-11.339	-10.976	-13.228	-7.322
N	PM3-BF03	232.649	232.334	235.349	186.666	201.917	192.072	227.030	225.495	203.201	214.182	232.667	196.938
N	PM3-BF04	262.188	260.393	264.663	202.088	213.946	209.576	264.713	250.355	227.530	238.472	243.296	206.818
N	PM3-BF05	26.566	33.299	39.358	34.780	50.878	31.096	1.286	-2.830	1.633	-4.076	-5.678	-5.007
N	PM3-BF06	-7.503	-8.686	9.868	-6.517	-2.225	-3.579	3.357	-6.187	-8.962	2.392	-8.833	-8.180
N	PM3-BF07	-6.900	-13.882	-9.741	-23.397	-18.687	-21.324	-3.661	-1.582	-12.287	-19.661	-25.003	-21.537
N	PM3-BF08	-3.900	-0.695	5.301	9.241	14.512	11.442	1.782	-4.239	5.325	-2.733	-4.762	-3.562
N	PM3-BF09	11.936	12.153	15.879	26.180	42.683	36.103	-1.263	-3.388	0.026	-4.208	-5.030	-3.793
N	PM3-BF10	-0.122	-0.566	5.614	6.318	9.287	9.263	-0.326	-5.600	-0.036	-4.937	-5.709	-6.242
N	PM3-BF11	5.093	2.966	8.161	15.518	28.693	20.609	0.368	-4.811	0.013	-5.049	-5.749	-4.486
N	PM3-BF12	-8.408	-5.809	-3.268	-4.909	-3.417	0.001	-3.170	-4.788	-5.038	-5.459	-3.250	-5.347
N	PM3-BG01	16.086	25.880	49.124	38.329	51.676	42.301	25.004	7.289	2.021	0.430	3.384	3.450
N	PM3-BG02	30.061	66.441	75.593	34.841	46.445	27.287	53.353	3.900	-0.383	4.387	5.739	0.937
N	PM3-BG03	42.891	44.738	51.359	8.074	13.466	7.738	25.576	-18.108	13.259	11.483	14.188	-5.447
N	PM3-BG04	2.222	20.730	28.971	-10.817	-0.539	-1.534	-8.213	0.061	-0.651	0.136	-0.425	-0.182
N	PM3-BG05	18.064	31.039	34.395	-0.397	0.836	2.204	5.691	4.541	5.709	2.046	2.949	2.697
N	PM3-BG06	6.542	30.336	32.846	3.536	8.576	14.163	12.234	3.039	12.796	3.772	2.559	2.832
N	PM3-BG07	135.647	151.072	120.501	29.226	50.822	24.483	59.413	101.007	91.701	102.792	129.683	89.440
N	PM3-BG08	78.613	102.480	113.499	46.841	72.826	51.311	103.407	28.116	47.032	18.720	70.445	8.092

N	PM3-BG09	71.400	68.758	99.978	45.924	63.795	51.714	97.172	11.578	36.839	8.188	48.457	5.622
N	PM3-BG10	-22.226	63.738	60.813	-6.516	19.501	9.339	-6.179	48.709	36.400	39.433	67.392	37.872
N	PM3-BG11	-1.888	-3.324	2.089	-2.545	0.804	3.943	1.142	-3.618	1.401	-5.189	-4.324	-2.832
N	PM3-BG12	90.707	96.747	73.607	12.020	28.282	8.639	37.821	48.083	27.049	27.480	48.387	17.537
N	PM3-BH01	255.661	258.078	255.558	235.605	226.807	232.179	244.353	229.755	223.279	230.489	240.943	208.203
N	PM3-BH02	255.439	257.039	254.549	233.830	232.986	237.488	242.246	239.364	213.350	219.700	235.766	186.961
N	PM3-BH03	267.121	261.945	261.837	246.993	235.889	242.178	246.612	232.163	214.690	216.989	239.764	193.236
N	PM3-BH04	238.042	234.471	234.516	174.434	191.588	156.057	220.538	215.355	195.737	203.397	215.971	172.603
N	PM3-BH05	233.564	241.458	240.551	85.534	116.218	72.513	212.901	209.875	177.832	197.066	219.099	156.809
N	PM3-BH06	191.786	190.587	195.028	35.559	71.217	40.875	113.051	166.722	81.730	162.934	172.310	169.217
N	PM3-BH07	226.434	232.549	225.700	119.276	147.786	131.947	205.360	204.549	188.443	193.350	204.126	185.433
N	PM3-BH08	253.689	257.287	262.670	223.283	226.626	235.434	252.067	232.072	224.951	231.160	245.947	207.417
N	PM3-BH09	250.718	249.621	262.222	205.005	215.826	230.007	249.824	228.407	215.096	233.074	238.574	188.368
N	PM3-BH10	212.459	218.068	225.055	176.050	197.607	202.538	229.459	196.834	200.032	184.880	205.318	157.480
N	PM3-BH11	189.479	199.892	191.497	157.458	185.117	162.108	183.849	171.058	167.941	161.821	183.779	155.312
N	PM3-BH12	175.339	177.297	165.995	96.542	135.014	111.375	169.728	151.255	133.986	142.395	154.121	127.382
P	PM4-AA01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
P	PM4-AA02	169.771	151.013	142.095	71.012	57.632	57.047	86.080	117.215	86.551	141.657	76.458	89.299
P	PM4-AA03	176.870	127.400	132.272	29.771	33.168	31.166	73.180	105.661	66.651	113.853	69.074	82.437
P	PM4-AA04	175.101	154.185	149.204	63.163	46.413	44.607	91.442	109.372	82.078	148.013	91.097	101.859
P	PM4-AA05	158.262	138.125	112.416	42.437	41.912	15.708	83.071	109.250	81.516	143.863	85.343	105.230
P	PM4-AA06	0.638	70.146	1.183	6.379	4.858	-2.658	-11.484	23.678	4.437	1.488	8.740	1.933
P	PM4-AA07	40.462	86.457	36.201	23.367	9.446	17.993	18.233	25.055	-22.188	17.407	14.821	0.364
P	PM4-AA08	188.903	168.974	167.263	76.759	64.079	44.000	117.312	134.178	112.571	177.226	113.259	154.004
P	PM4-AA09	193.157	167.993	171.367	144.795	133.583	152.942	130.762	140.592	128.822	162.033	122.666	147.880
P	PM4-AA10	195.662	29.278	-55.947	-23.376	136.788	164.628	140.730	136.779	-19.400	173.558	117.591	141.986
P	PM4-AA11	175.901	153.693	-60.689	129.051	128.957	131.025	140.155	132.547	111.179	158.225	102.305	130.915
P	PM4-AA12	172.274	146.246	146.992	119.078	116.240	118.988	122.474	119.912	90.091	153.130	87.872	115.950
P	PM4-AB01	141.986	128.941	118.842	43.516	37.083	51.007	77.862	95.665	82.309	134.307	69.147	102.616
P	PM4-AB02	184.989	142.654	133.267	76.991	74.055	67.804	81.586	105.225	96.943	140.713	84.343	107.279
P	PM4-AB03	188.987	166.732	173.688	138.464	131.296	128.671	126.796	124.280	129.088	167.309	109.017	127.395
P	PM4-AB04	170.038	150.083	142.899	130.403	133.130	134.828	95.325	112.021	113.489	141.854	89.724	111.642
P	PM4-AB05	167.928	159.449	153.458	73.358	71.840	61.616	85.775	116.138	100.399	146.970	98.975	121.478
P	PM4-AB06	170.325	164.107	151.928	67.507	59.868	51.362	94.905	112.997	96.466	141.162	94.592	127.075
P	PM4-AB07	162.272	145.925	151.158	65.154	48.643	50.351	86.253	113.568	95.054	137.089	88.653	127.661
P	PM4-AB08	198.228	189.118	192.080	159.629	159.962	172.578	135.628	145.157	159.768	179.276	124.240	167.051
P	PM4-AB09	196.692	166.224	170.084	128.051	123.826	127.026	138.317	124.355	115.926	163.716	106.897	137.104
P	PM4-AB10	180.408	156.299	161.774	119.849	125.109	127.612	123.328	124.166	112.800	152.247	104.388	143.429
P	PM4-AB11	161.580	135.026	143.292	108.357	109.013	105.204	108.120	103.115	96.183	131.778	78.107	124.751
P	PM4-AB12	161.814	138.074	138.863	112.299	108.580	106.041	118.305	104.037	89.239	134.297	76.142	113.987
P	PM4-AC01	160.741	142.263	141.432	58.232	33.918	49.026	82.557	102.486	75.562	134.337	81.924	94.009
P	PM4-AC02	156.783	143.017	129.967	57.558	63.716	47.128	81.637	102.265	84.804	131.351	79.984	109.904
P	PM4-AC03	198.455	174.930	184.367	147.111	145.250	156.414	133.442	136.953	133.899	178.766	112.178	153.429
P	PM4-AC04	-33.116	189.057	195.974	159.387	158.829	171.301	127.550	148.721	144.012	182.755	127.993	142.053
P	PM4-AC05	-43.782	-62.450	-56.796	-34.496	-46.871	-32.782	-84.195	-85.870	-72.649	-35.949	-103.054	-35.004
P	PM4-AC06	218.637	199.409	202.796	173.054	168.605	179.604	149.925	154.718	156.493	195.863	136.996	190.561
P	PM4-AC07	187.546	165.962	190.197	103.111	98.966	106.574	114.816	127.965	117.974	160.814	105.142	151.175
P	PM4-AC08	214.082	175.826	174.657	148.200	129.345	121.271	147.395	145.429	132.691	179.304	128.709	161.190
P	PM4-AC09	208.482	188.193	-5.517	166.092	161.887	183.034	155.633	146.313	150.155	181.101	110.670	2.980
P	PM4-AC10	-53.953	-54.075	-57.214	141.713	143.376	152.004	126.382	134.313	-65.659	163.803	91.042	156.136
P	PM4-AC11	165.318	149.397	152.339	139.524	134.409	148.308	117.845	116.595	109.275	149.249	86.668	131.592
P	PM4-AC12	178.737	156.151	164.099	150.443	131.666	146.554	130.474	129.004	93.358	155.030	92.125	121.603
P	PM4-AD01	182.311	166.707	176.084	144.950	124.405	139.428	-79.309	127.046	110.591	158.296	109.255	140.924
P	PM4-AD02	193.318	174.988	181.758	133.583	137.830	145.586	100.679	126.918	119.109	161.987	116.592	149.142
P	PM4-AD03	159.467	141.251	132.180	69.466	64.032	53.834	90.822	117.449	98.814	147.057	93.883	109.967
P	PM4-AD04	163.013	151.439	143.476	82.986	78.745	51.950	84.724	118.924	99.703	143.480	98.851	140.520
P	PM4-AD05	177.836	165.179	160.246	65.221	59.940	57.130	59.512	104.263	71.818	138.396	89.550	113.157
P	PM4-AD06	186.453	169.229	158.424	97.176	89.058	78.863	99.436	126.465	119.688	157.897	99.649	156.949
P	PM4-AD07	183.764	164.859	155.439	92.853	86.266	89.108	108.470	124.170	110.225	156.328	103.722	163.807
P	PM4-AD08	211.716	190.347	183.880	158.443	-22.224	56.339	-76.228	147.507	-49.147	179.711	118.471	109.384
P	PM4-AD09	189.434	188.307	184.501	161.854	121.743	159.838	142.974	152.863	133.287	184.996	119.042	165.596
P	PM4-AD10	-37.411	166.842	177.183	112.257	140.337	145.661	-74.699	132.857	117.509	165.820	107.418	153.436
P	PM4-AD11	181.116	155.251	168.392	136.345	134.315	135.828	133.771	122.141	101.447	160.849	92.324	136.040
P	PM4-AD12	185.525	159.697	169.699	128.207	123.812	126.901	132.422	138.618	94.808	166.251	95.586	120.030
P	PM4-AE01	180.497	160.776	155.392	67.754	48.826	35.917	94.944	118.567	90.532	-32.691	92.179	117.408
P	PM4-AE02	168.238	160.741	-19.067	-26.250	-17.109	32.379	85.883	105.346	85.963	135.932	84.947	124.679
P	PM4-AE03	171.230	140.921	139.332	48.786	57.805	57.101	91.703	116.116	101.034	143.842	84.668	126.688
P	PM4-AE04	167.168	139.897	133.380	39.575	51.088	45.583	64.055	103.515	75.759	127.967	71.693	118.020
P	PM4-AE05	173.305	162.484	156.238	68.704	65.532	60.511	91.601	120.453	101.617	145.297	94.225	136.413

P	PM4-AE06	138.951	120.834	130.096	-4.822	-35.978	0.428	7.940	58.238	25.535	77.525	47.200	80.609
P	PM4-AE07	168.742	165.884	150.872	90.645	88.654	76.009	34.104	115.022	94.479	99.647	84.915	86.212
P	PM4-AE08	30.241	34.609	35.971	51.280	10.839	-0.745	6.144	23.840	4.236	3.661	14.820	36.014
P	PM4-AE09	177.717	166.597	146.934	118.805	120.271	106.979	110.065	126.492	105.137	157.559	93.601	141.396
P	PM4-AE10	183.034	170.041	170.250	127.850	125.738	136.170	131.757	133.518	124.004	158.386	99.096	146.483
P	PM4-AE11	116.292	78.814	57.211	22.391	14.614	11.996	16.763	44.879	19.976	41.212	21.965	40.862
P	PM4-AE12	168.721	150.424	145.276	113.659	102.649	91.151	118.732	113.813	108.214	145.083	80.379	111.383
S	PM4-AF01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
S	PM4-AF02	201.892	212.070	156.468	104.555	111.920	91.537	169.459	194.714	169.449	176.132	179.867	145.058
S	PM4-AF03	218.338	215.709	164.388	115.062	135.220	112.782	177.713	209.803	187.345	195.114	191.328	157.213
S	PM4-AF04	209.433	221.001	173.284	111.133	124.529	100.472	182.434	198.475	171.189	175.259	195.680	155.063
S	PM4-AF05	202.439	207.549	137.380	101.279	117.343	91.891	176.859	183.150	133.268	137.291	186.479	89.280
S	PM4-AF06	211.268	223.201	164.865	124.850	131.055	112.843	183.842	207.013	180.955	185.279	197.820	151.692
S	PM4-AF07	190.209	207.488	145.786	135.701	121.097	117.346	186.112	204.976	167.728	181.643	203.397	146.304
S	PM4-AF08	197.688	215.026	160.245	95.645	110.691	92.008	180.011	204.813	179.272	182.222	198.225	166.037
S	PM4-AF09	173.325	203.530	138.392	131.867	144.567	122.972	174.776	197.854	178.003	177.799	193.640	164.405
S	PM4-AF10	-5.416	-2.020	-44.941	-5.078	-0.555	1.072	1.125	4.768	2.299	-2.279	-2.358	-0.866
S	PM4-AF11	4.867	8.259	-38.721	0.526	3.571	5.322	7.733	11.811	6.853	-0.889	12.001	-1.497
S	PM4-AF12	174.599	183.018	133.028	78.801	90.520	74.347	142.538	177.578	146.255	154.154	158.408	114.725
S	PM4-AG01	-0.034	0.492	-42.401	0.017	-0.601	-0.325	-0.575	2.382	-1.709	0.400	0.436	-0.267
S	PM4-AG02	144.496	157.539	72.561	14.726	12.536	11.037	36.678	36.295	22.370	12.555	36.772	9.262
S	PM4-AG03	213.179	210.196	163.778	118.122	124.858	97.582	183.704	206.429	171.084	190.037	194.343	161.086
S	PM4-AG04	219.522	218.209	166.280	130.904	138.810	110.220	184.036	211.461	186.771	190.900	198.570	167.780
S	PM4-AG05	224.939	108.299	9.030	43.261	52.968	45.713	192.778	34.947	35.642	7.742	37.176	-1.299
S	PM4-AG06	48.518	66.325	3.809	0.704	-1.939	1.946	65.617	104.633	23.911	78.160	99.061	83.520
S	PM4-AG07	210.822	214.196	163.763	126.507	133.126	115.307	187.084	206.296	154.639	182.407	193.699	147.103
S	PM4-AG08	195.808	196.134	134.475	80.905	91.124	87.404	155.655	188.838	117.692	149.512	182.253	142.038
S	PM4-AG09	204.400	211.664	171.507	135.837	144.058	128.353	194.475	204.662	178.942	183.460	194.346	144.490
S	PM4-AG10	197.667	214.575	164.150	113.778	133.697	110.825	182.353	198.679	149.717	172.388	179.675	136.188
S	PM4-AG11	185.571	196.441	154.888	110.997	120.989	106.947	174.549	200.316	159.314	163.437	173.263	147.916
S	PM4-AG12	-7.349	-6.607	-51.654	-4.972	-4.693	-3.176	-2.874	1.001	-0.884	-1.253	-3.311	-3.582
S	PM4-AH01	179.675	193.851	133.603	65.882	68.696	63.163	141.809	183.928	99.768	163.578	190.909	119.515
S	PM4-AH02	42.480	85.532	173.187	30.596	42.113	32.645	27.528	31.136	23.400	34.925	47.359	22.454
S	PM4-AH03	41.471	3.414	-43.459	4.632	0.849	1.695	47.811	3.807	-2.747	5.138	2.845	3.904
S	PM4-AH04	19.962	27.491	-29.754	19.770	21.975	17.605	17.588	26.612	20.558	7.986	39.953	15.009
S	PM4-AH05	34.705	32.113	-19.372	10.793	25.893	11.689	24.625	22.158	15.096	13.433	22.653	12.095
S	PM4-AH06	135.885	134.778	59.678	14.305	23.572	8.033	56.043	97.163	31.521	81.101	79.876	71.770
S	PM4-AH07	103.414	116.996	48.015	20.811	31.099	23.537	58.300	89.329	40.255	79.308	60.666	69.120
S	PM4-AH08	13.593	11.553	-39.808	6.603	5.074	5.697	2.370	6.950	7.678	5.704	6.118	3.904
S	PM4-AH09	-10.230	2.228	-43.397	2.555	2.522	1.196	1.872	7.516	3.567	5.137	2.405	1.093
S	PM4-AH10	-1.095	-1.324	-43.926	-5.497	2.688	-0.959	2.433	3.025	2.030	2.649	-3.055	0.689
S	PM4-AH11	-4.654	-4.782	-47.017	-4.726	-1.034	-1.187	-2.678	-0.396	0.213	-0.016	-3.630	-3.916
S	PM4-AH12	-6.896	-5.213	-44.375	-5.551	1.204	-2.603	-0.788	-0.555	1.022	-2.078	-5.601	-4.636

		BT 3 O:5,27			BT 4 O:3			BT 5 O:2a,2b,3		
	well	14902_1	14902_2	14902_3	1203_1	1203_2	1203_3	3094_1	3094_2	3094_3
C	PM1-A01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C	PM1-A02	191.763	180.437	171.599	248.688	227.921	230.820	91.213	51.400	50.911
C	PM1-A03	81.895	72.968	40.004	170.624	115.474	104.292	90.553	90.889	90.832
C	PM1-A04	-0.063	-7.708	-2.250	-8.832	-4.942	-2.005	1.587	0.000	0.000
C	PM1-A05	20.217	12.604	7.271	28.301	8.432	8.711	168.197	114.390	104.534
C	PM1-A06	245.749	225.503	224.664	281.925	264.450	272.850	145.934	113.079	118.176
C	PM1-A07	7.511	7.789	5.374	9.392	8.008	7.446	144.818	99.782	86.321
C	PM1-A08	8.266	6.093	3.584	0.722	4.741	1.008	1.608	1.029	0.000
C	PM1-A09	2.768	2.078	-1.257	-5.929	-4.904	-4.841	9.787	3.229	0.000
C	PM1-A10	93.118	67.850	64.383	143.111	101.662	114.182	0.000	0.503	0.000
C	PM1-A11	61.709	46.438	35.436	118.388	74.233	81.093	109.083	94.889	97.292
C	PM1-A12	6.258	11.263	10.836	58.889	2.692	3.304	11.729	11.086	2.489
C	PM1-B01	3.999	0.854	5.150	165.604	-4.074	0.993	1.133	0.755	0.000
C	PM1-B02	123.761	104.984	87.593	160.212	118.886	118.501	4.171	3.229	1.011
C	PM1-B03	98.879	95.393	77.364	174.705	123.021	148.220	84.584	62.903	62.876
C	PM1-B04	2.295	-1.912	-1.045	-3.232	-4.492	-4.009	0.000	0.000	0.000
C	PM1-B05	242.643	179.788	166.942	269.642	244.167	257.245	112.932	98.113	81.784
C	PM1-B06	271.707	218.561	211.862	200.468	86.075	106.887	48.537	39.111	40.268
C	PM1-B07	170.484	137.267	168.772	185.067	179.174	186.757	75.682	59.976	47.505
C	PM1-B08	105.146	123.625	98.808	-4.704	-4.675	-5.711	94.442	57.650	56.916
C	PM1-B09	2.995	0.984	1.384	-0.843	-3.472	-3.261	0.000	0.000	0.000
C	PM1-B10	5.375	5.286	5.074	-2.005	4.679	2.153	0.000	0.000	0.000
C	PM1-B11	89.390	65.630	50.475	134.130	102.093	97.986	97.645	79.311	64.687

C	PM1-B12	9.561	13.346	14.801	131.707	115.463	136.234	20.599	27.186	19.089
C	PM1-C01	253.249	185.212	181.246	241.478	211.797	231.790	11.459	10.528	9.221
C	PM1-C02	3.663	-0.979	-1.137	-4.732	0.812	-3.695	0.000	0.000	0.000
C	PM1-C03	8.724	5.075	17.366	4.263	4.267	0.446	89.150	63.755	58.103
C	PM1-C04	79.347	84.125	75.199	195.251	180.771	159.059	60.682	44.988	44.924
C	PM1-C05	11.851	5.917	9.478	5.526	5.616	2.543	4.066	3.818	2.542
C	PM1-C06	1.905	-5.372	-0.743	-7.239	-4.267	-7.503	0.000	0.000	0.000
C	PM1-C07	52.005	46.637	40.762	95.757	64.314	60.955	49.705	45.432	37.979
C	PM1-C08	53.343	74.926	84.972	122.792	105.537	110.513	38.045	39.634	37.129
C	PM1-C09	56.530	47.100	37.664	106.904	65.732	72.895	79.187	64.426	46.916
C	PM1-C10	150.180	137.159	133.837	193.632	175.118	177.161	5.705	7.132	1.005
C	PM1-C11	-0.555	-0.557	1.904	-2.095	-6.679	-7.245	0.000	0.000	0.000
C	PM1-C12	20.587	13.414	17.336	13.913	9.059	13.264	37.383	32.325	19.753
C	PM1-D01	7.078	0.947	1.313	7.934	4.874	0.096	44.333	27.970	20.343
C	PM1-D02	-3.987	-4.386	-4.439	-4.745	-9.737	-14.016	0.000	0.505	0.000
C	PM1-D03	2.045	-3.978	-0.814	83.716	-6.945	-6.757	0.000	0.000	0.000
C	PM1-D04	1.763	-0.793	-0.658	-0.700	-0.217	-3.774	0.000	0.000	0.000
C	PM1-D05	3.911	2.286	-1.276	-1.184	-1.512	-2.211	0.000	0.000	0.000
C	PM1-D06	-0.224	-2.022	-5.957	-5.621	-5.817	-8.896	0.000	0.000	0.000
C	PM1-D07	-1.567	-0.293	-6.532	-8.026	-3.874	-1.258	0.000	0.000	0.000
C	PM1-D08	0.692	0.862	-1.080	56.884	-2.555	-5.179	0.000	0.000	0.000
C	PM1-D09	-1.400	-0.647	5.061	-2.600	-6.170	-7.938	3.221	3.082	0.000
C	PM1-D10	1.430	1.137	-0.413	-0.701	-0.517	-1.724	4.550	4.189	0.000
C	PM1-D11	128.849	130.186	111.221	94.272	60.288	57.475	56.661	49.461	52.184
C	PM1-D12	59.651	69.145	69.774	185.062	135.718	112.276	47.820	44.580	33.305
C	PM1-E01	6.163	0.766	2.441	7.822	6.613	8.792	2.854	3.737	1.824
C	PM1-E02	-3.547	0.383	0.564	-7.488	-0.646	-6.125	0.000	0.000	0.000
C	PM1-E03	226.509	185.351	184.972	250.266	239.974	244.562	113.237	89.368	80.787
C	PM1-E04	211.538	88.720	170.958	253.454	215.241	225.707	0.000	0.000	0.000
C	PM1-E05	1.741	5.470	3.514	-0.063	3.795	-4.521	0.000	0.000	0.000
C	PM1-E06	2.789	-2.789	0.863	-3.587	-0.291	-3.036	0.000	0.000	0.000
C	PM1-E07	-2.405	-3.322	-3.113	-6.170	-5.805	-5.762	0.000	0.000	0.000
C	PM1-E08	118.975	118.812	96.018	222.768	161.568	156.816	0.000	0.000	0.000
C	PM1-E09	1.725	-6.416	0.650	-0.262	-2.886	-5.497	0.000	0.000	0.000
C	PM1-E10	128.600	112.471	111.188	187.186	156.047	157.166	18.311	16.582	9.984
C	PM1-E11	167.987	169.829	173.762	208.043	207.017	199.067	88.595	56.468	59.000
C	PM1-E12	222.786	225.061	223.034	233.712	216.914	224.371	129.980	104.309	103.136
C	PM1-F01	25.300	8.307	12.270	170.401	143.313	141.668	14.878	16.639	10.679
C	PM1-F02	-1.395	-5.887	-3.546	-2.425	-3.374	-4.305	0.000	0.000	0.000
C	PM1-F03	192.125	193.014	190.626	130.459	114.850	142.458	10.689	9.405	5.703
C	PM1-F04	-1.612	-4.264	-2.216	-5.332	-5.116	-6.800	0.000	0.000	0.000
C	PM1-F05	4.237	3.566	-0.257	-0.922	0.418	-1.621	118.750	86.074	69.061
C	PM1-F06	0.757	-1.434	-2.724	-1.183	-1.978	-4.699	57.521	39.037	30.329
C	PM1-F07	-2.343	-2.636	-6.655	3.929	-1.887	-6.203	0.000	0.000	0.000
C	PM1-F08	-2.249	2.320	-2.366	-5.293	0.433	-3.234	0.000	0.000	0.000
C	PM1-F09	-2.616	-4.170	-2.872	-12.432	-6.354	-8.075	0.000	0.000	0.000
C	PM1-F10	-6.754	-1.841	-5.242	-6.496	-3.686	-8.358	0.000	0.000	0.000
C	PM1-F11	132.738	128.247	116.138	143.032	129.767	123.051	103.766	85.433	65.676
C	PM1-F12	155.280	167.291	174.488	215.782	197.564	190.429	124.090	96.182	80.770
C	PM1-G01	80.182	68.333	73.670	110.776	99.334	102.862	14.550	8.568	6.732
C	PM1-G02	2.880	-1.682	-0.083	173.288	-1.803	-5.489	0.000	0.000	0.000
C	PM1-G03	226.791	212.900	210.161	240.321	239.567	244.382	128.732	113.868	109.997
C	PM1-G04	2.516	0.714	0.609	5.700	5.158	-5.716	8.539	3.729	2.529
C	PM1-G05	53.320	91.346	47.561	133.976	122.392	109.162	19.274	6.776	4.308
C	PM1-G06	99.834	110.420	107.338	123.251	121.372	112.237	15.374	16.292	15.132
C	PM1-G07	-2.426	5.191	2.013	-3.382	2.450	-1.917	6.147	0.000	0.000
C	PM1-G08	195.690	178.717	190.683	223.388	223.957	204.190	76.547	59.058	47.976
C	PM1-G09	7.770	8.895	3.563	2.318	7.922	2.874	46.918	42.382	34.382
C	PM1-G10	172.963	131.386	128.232	193.228	155.701	164.082	54.642	53.453	46.845
C	PM1-G11	2.230	4.339	4.364	-6.271	2.871	-1.318	3.539	4.063	0.886
C	PM1-G12	11.195	34.547	28.892	67.567	24.937	35.451	138.122	114.908	102.386
C	PM1-H01	163.753	158.768	139.293	189.801	187.383	167.980	99.171	77.916	55.614
C	PM1-H02	7.212	4.150	3.687	3.543	3.413	0.304	2.047	2.521	-0.433
C	PM1-H03	5.667	-0.068	2.559	0.297	-0.363	-13.236	6.368	5.867	0.725
C	PM1-H04	5.363	3.788	2.989	86.639	4.043	4.789	1.784	2.066	0.264
C	PM1-H05	6.525	4.117	4.553	0.849	7.128	3.076	4.033	3.643	0.550
C	PM1-H06	-1.438	-0.059	-0.636	-3.933	-3.380	-4.476	-4.297	-2.057	-3.705
C	PM1-H07	8.370	5.979	6.658	1.179	6.946	2.345	12.262	8.592	3.671
C	PM1-H08	112.350	107.814	117.218	177.047	187.363	187.343	66.111	53.405	54.839

C	PM1-H09	3.014	5.463	4.312	-0.712	5.962	0.864	4.201	4.479	0.000
C	PM1-H10	211.090	206.628	216.346	247.755	227.932	238.630	114.328	107.280	93.709
C	PM1-H11	-0.393	3.639	6.571	0.267	2.796	0.628	6.208	9.012	4.737
C	PM1-H12	1.888	4.182	2.980	-11.988	1.109	-2.572	7.188	8.050	3.971
C	PM2-AA01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C	PM2-AA02	-3.401	-3.495	-3.949	-2.245	-5.422	-6.005	0.000	0.505	1.587
C	PM2-AA03	-1.651	-3.647	-0.601	-1.045	-5.276	-4.689	0.000	0.000	0.000
C	PM2-AA04	-4.695	-3.612	-2.047	-3.034	-5.671	-5.358	0.000	0.000	1.133
C	PM2-AA05	-6.349	-5.021	-4.132	-3.517	-7.967	-8.039	0.000	0.000	0.000
C	PM2-AA06	21.405	10.274	10.242	41.074	2.139	6.449	9.282	9.839	9.711
C	PM2-AA07	-5.158	-4.880	-5.987	-7.575	-6.578	-7.316	0.000	0.000	0.000
C	PM2-AA08	-4.880	-2.962	-2.038	-1.954	-5.128	-5.529	0.000	0.000	1.016
C	PM2-AA09	-6.929	-3.568	-3.791	-6.580	-12.062	-7.664	0.000	0.000	0.000
C	PM2-AA10	-7.805	-10.150	-3.984	-5.250	-14.182	-7.075	0.505	0.000	3.450
C	PM2-AA11	-4.687	-3.580	-1.605	-2.991	-5.338	-4.789	1.133	0.388	1.589
C	PM2-AA12	132.471	87.495	86.693	155.650	159.637	120.025	49.370	40.067	45.863
C	PM2-AB01	107.791	93.739	92.567	167.209	175.200	182.253	0.000	0.000	0.000
C	PM2-AB02	211.784	174.055	170.901	219.074	204.268	204.975	74.774	61.150	47.189
C	PM2-AB03	-9.742	-8.866	-9.437	-8.014	-9.705	-13.420	0.000	0.000	0.000
C	PM2-AB04	-5.612	-4.237	-4.404	-6.283	-8.046	-8.341	0.000	0.000	0.000
C	PM2-AB05	-9.288	-9.450	-6.341	-4.711	-10.172	-10.325	0.000	0.000	0.000
C	PM2-AB06	-8.836	-7.530	-6.534	-8.614	-7.328	-9.464	0.000	0.000	0.000
C	PM2-AB07	-10.742	-8.457	-8.420	-6.886	-12.647	-12.270	0.000	0.000	0.000
C	PM2-AB08	8.389	6.632	8.354	43.276	50.824	38.955	0.000	0.000	0.000
C	PM2-AB09	-12.309	-13.247	-8.107	-6.346	-14.607	-9.239	0.000	0.000	0.000
C	PM2-AB10	-14.818	-8.693	-10.100	-12.292	-11.918	-13.155	0.000	0.000	0.000
C	PM2-AB11	-8.742	-8.241	-7.670	-9.512	-9.975	-12.546	0.000	0.000	0.000
C	PM2-AB12	3.978	7.336	6.305	4.667	3.280	-3.074	6.245	3.329	7.786
C	PM2-AC01	157.390	126.680	112.536	187.963	168.518	186.634	0.000	0.000	0.000
C	PM2-AC02	-8.675	-7.271	-7.211	-5.434	-9.132	-8.729	0.000	0.000	0.000
C	PM2-AC03	-8.997	-9.333	-8.108	-9.376	-10.551	-11.291	0.000	0.000	0.000
C	PM2-AC04	-8.254	-9.680	-7.057	-9.633	-9.984	-10.953	0.000	0.000	0.000
C	PM2-AC05	-9.649	-6.974	-4.905	-7.907	-11.547	-10.830	0.000	0.000	0.000
C	PM2-AC06	-11.146	-10.317	-8.943	-8.657	-11.853	-12.629	0.000	0.000	0.000
C	PM2-AC07	-7.911	-5.579	-2.529	-2.228	-2.068	-4.336	9.282	0.000	0.000
C	PM2-AC08	-9.586	-6.495	-6.464	-7.329	-9.759	-13.446	0.000	0.000	0.000
C	PM2-AC09	-13.243	-9.454	-7.524	-8.386	-12.786	-13.533	0.000	0.000	0.000
C	PM2-AC10	-13.070	-8.795	-8.122	-12.212	-12.143	-11.661	0.000	0.000	0.000
C	PM2-AC11	-9.925	-5.689	-6.047	-8.068	-5.825	-10.774	0.000	0.000	0.000
C	PM2-AC12	-4.484	-3.225	-2.404	-3.620	-6.696	-7.422	6.432	3.771	11.821
C	PM2-AD01	-3.488	-5.403	-3.253	-3.064	-7.739	-8.161	1.011	2.953	1.058
C	PM2-AD02	-6.778	-7.499	-7.034	-5.813	-8.757	-9.245	0.000	0.000	0.000
C	PM2-AD03	-7.037	-6.776	-6.680	-5.717	-5.142	-9.399	0.000	0.000	0.000
C	PM2-AD04	124.192	115.782	102.041	122.018	109.134	118.765	33.453	18.574	24.732
C	PM2-AD05	-12.451	-10.746	-8.336	-9.254	-11.811	-11.913	0.000	0.000	0.000
C	PM2-AD06	-7.716	-6.099	-5.393	-3.937	-6.995	-8.768	0.000	0.000	0.000
C	PM2-AD07	-7.475	-10.650	-5.384	5.325	-11.333	-10.108	0.000	0.000	0.000
C	PM2-AD08	-10.453	-9.224	-8.055	-6.726	-8.179	-10.174	0.000	0.000	0.000
C	PM2-AD09	-11.441	-7.143	-7.139	-5.763	-10.342	-11.729	0.000	0.000	0.000
C	PM2-AD10	-10.453	-6.396	-8.175	-7.353	-9.061	-9.688	0.000	0.000	0.000
C	PM2-AD11	-15.142	-8.063	-8.255	-6.704	-9.095	-13.974	0.000	0.000	0.000
C	PM2-AD12	-8.709	-2.389	-1.392	-2.476	-6.961	-8.876	1.008	1.912	5.149
C	PM2-AE01	-18.336	-12.741	-12.047	-13.666	-19.766	-23.768	1.654	-0.066	2.626
C	PM2-AE02	-1.395	-4.921	-1.661	-3.028	-10.761	-14.150	0.000	0.000	0.000
C	PM2-AE03	-7.782	-4.861	-7.863	-4.353	-7.459	-7.533	0.000	0.000	0.000
C	PM2-AE04	-12.400	-9.116	-8.593	-4.541	-8.371	-11.375	0.000	0.000	0.000
C	PM2-AE05	72.595	57.013	48.062	190.467	186.547	196.622	62.211	68.892	61.211
C	PM2-AE06	-13.645	-10.937	-12.384	-11.234	-14.042	-14.866	0.000	0.000	0.000
C	PM2-AE07	-9.614	-7.001	-7.407	-3.382	-6.551	-9.357	0.000	0.000	0.000
C	PM2-AE08	-10.997	-5.861	-7.789	-4.646	-3.839	-9.825	0.000	0.000	0.000
C	PM2-AE09	-7.501	-5.084	-4.461	-4.384	-4.126	-8.211	0.000	0.000	0.000
C	PM2-AE10	-10.401	-6.067	-6.611	-6.580	-9.179	-9.396	0.000	0.000	0.000
C	PM2-AE11	-13.258	-13.859	-7.850	-10.741	-16.114	-16.692	0.000	0.000	0.000
C	PM2-AE12	169.597	154.522	169.114	2.922	-2.008	-2.664	14.591	11.309	17.016
C	PM2-AF01	-1.107	-1.055	1.333	1.162	-1.096	-3.501	0.000	0.000	0.000
C	PM2-AF02	-5.537	-4.695	-3.988	-2.483	-4.321	-6.712	0.000	0.000	0.000
C	PM2-AF03	-9.358	-3.734	-5.780	-2.403	-6.842	-7.021	0.000	0.000	0.000
C	PM2-AF04	-8.905	-4.761	-8.253	-3.488	-5.979	-7.807	0.000	0.000	0.000
C	PM2-AF05	-11.291	-7.301	-6.587	-4.614	-9.864	-7.511	0.000	0.000	0.000

C	PM2-AF06	-12.445	-7.555	-8.961	-2.216	-7.014	-9.766	0.000	0.000	0.000
C	PM2-AF07	-8.600	-7.375	-8.872	-4.588	-6.168	-8.258	0.000	0.000	0.000
C	PM2-AF08	-9.321	-6.671	-7.776	-3.051	-8.541	-14.036	0.000	0.000	0.000
C	PM2-AF09	-4.701	-4.532	-8.279	-0.045	-11.426	-13.988	0.000	0.000	0.000
C	PM2-AF10	-4.986	-2.183	-2.807	0.454	-7.196	-5.830	0.000	0.000	0.000
C	PM2-AF11	-9.464	-4.497	-4.059	-11.337	-10.657	-11.784	0.000	0.000	0.000
C	PM2-AF12	-8.038	0.334	-3.363	1.117	-1.909	-4.018	0.000	0.000	0.000
C	PM2-AG01	-7.250	-3.616	-1.017	-2.841	-4.300	-3.988	0.000	0.000	0.000
C	PM2-AG02	0.154	5.016	3.546	8.959	6.718	3.436	0.000	0.000	0.000
C	PM2-AG03	-7.746	-5.096	-6.375	-4.038	-5.068	-7.704	0.000	0.000	0.000
C	PM2-AG04	-9.297	-6.229	-6.816	-3.295	-5.995	-8.670	0.000	0.000	0.000
C	PM2-AG05	-9.225	-6.407	-8.457	-8.039	-8.630	-11.424	0.000	0.000	0.000
C	PM2-AG06	-5.651	-6.463	-7.053	22.449	6.637	26.289	0.000	0.000	0.000
C	PM2-AG07	-13.030	-8.218	-7.916	-5.839	-11.261	-11.184	0.000	0.000	0.000
C	PM2-AG08	-7.241	-5.795	-4.946	-3.964	-9.937	-8.643	0.000	0.000	0.000
C	PM2-AG09	-9.829	-7.805	-7.474	-8.197	-12.108	-11.658	0.000	0.000	0.000
C	PM2-AG10	-11.628	-7.264	-8.749	-6.543	-9.071	-13.870	0.000	0.000	0.000
C	PM2-AG11	-11.170	-3.580	-5.736	-3.251	-6.786	-12.018	0.000	0.000	0.000
C	PM2-AG12	-2.655	0.261	0.041	4.116	-7.255	-1.754	0.503	0.000	0.000
C	PM2-AH01	-1.488	-4.400	-1.599	0.276	-2.261	-5.932	0.000	0.000	0.000
C	PM2-AH02	-5.697	-2.988	0.195	-2.428	-4.038	-3.717	0.000	0.000	0.000
C	PM2-AH03	-0.549	-0.529	0.472	1.330	-0.821	-5.771	0.000	0.000	0.000
C	PM2-AH04	-3.408	-5.280	-3.200	-4.197	-6.333	-5.346	0.000	0.000	0.000
C	PM2-AH05	-6.542	-2.582	-3.850	-3.638	-6.675	-6.436	0.000	0.000	0.000
C	PM2-AH06	-14.933	-10.257	-14.750	-10.664	-12.059	-15.617	0.000	0.000	0.000
C	PM2-AH07	-5.742	-1.449	-0.253	-0.251	-3.014	-4.989	0.000	0.000	0.000
C	PM2-AH08	-6.383	-5.489	-1.837	0.776	-0.592	-2.321	0.000	0.000	0.000
C	PM2-AH09	-17.825	-13.089	-10.892	-13.637	-17.711	-14.899	-2.234	-7.828	-6.912
C	PM2-AH10	-11.055	-3.274	-2.751	-5.909	-8.213	-7.055	0.000	0.000	0.000
C	PM2-AH11	-6.712	-6.599	-0.951	-0.199	-2.264	-3.770	0.000	0.000	0.000
C	PM2-AH12	-5.405	0.393	-1.618	-0.070	-1.004	-2.303	1.008	0.000	1.032
N	PM3-BA01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
N	PM3-BA02	3.214	2.604	-1.039	76.260	87.670	81.564	2.520	0.505	
N	PM3-BA03	-5.457	-8.203	-7.330	7.789	1.591	1.884	0.000	0.000	0.000
N	PM3-BA04	2.530	3.439	-0.318	1.942	1.717	5.099	0.000	1.008	0.000
N	PM3-BA05	-0.071	-0.225	-0.525	88.812	83.833	73.916	0.000	0.000	0.000
N	PM3-BA06	-1.897	-4.399	-7.591	33.075	27.350	30.286	0.000	0.000	0.000
N	PM3-BA07	15.411	37.454	21.617	74.943	77.445	69.446	0.000	0.000	0.000
N	PM3-BA08	10.700	33.520	9.050	65.604	75.524	72.216	0.000	0.000	0.000
N	PM3-BA09	148.580	125.410	116.953	209.146	194.426	219.753	16.308	13.895	11.200
N	PM3-BA10	191.047	142.234	136.449	227.370	207.401	220.829	31.018	20.958	20.134
N	PM3-BA11	83.897	271.534	265.458	15.557	123.183	130.895	1.076	0.758	0.000
N	PM3-BA12	11.368	42.712	19.737	196.725	209.700	203.838	19.203	15.112	22.011
N	PM3-BB01	10.925	35.750	79.674	187.287	180.404	193.786	11.239	12.171	14.524
N	PM3-BB02	-1.095	-0.424	-1.295	148.264	136.638	135.682	0.503	0.000	0.000
N	PM3-BB03	94.876	106.493	87.307	203.116	197.520	206.133	0.000	0.000	0.000
N	PM3-BB04	-1.618	1.558	-3.307	34.605	25.351	30.018	0.000	0.000	0.000
N	PM3-BB05	-4.847	-1.789	-4.838	58.867	53.204	58.507	0.000	0.000	0.000
N	PM3-BB06	-3.016	-4.441	-7.951	22.833	8.100	23.128	0.000	0.000	0.000
N	PM3-BB07	2.384	-3.970	7.764	25.342	27.482	46.647	0.000	0.000	0.000
N	PM3-BB08	-4.236	-2.304	-5.475	14.843	20.714	29.301	0.000	0.000	1.005
N	PM3-BB09	20.412	61.350	20.705	89.749	82.811	89.843	0.000	0.000	0.000
N	PM3-BB10	77.283	98.171	117.205	163.483	152.263	161.625	0.000	0.000	0.000
N	PM3-BB11	0.625	12.568	2.988	38.166	57.950	61.709	0.000	0.000	0.000
N	PM3-BB12	0.068	2.339	2.800	-3.641	-3.426	-1.883	6.814	3.305	1.599
N	PM3-BC01	-15.729	-14.439	-28.813	-15.361	-23.208	-25.857	-12.232	-13.026	-6.820
N	PM3-BC02	-4.268	-2.864	-4.042	-3.126	2.766	1.886	0.000	0.000	0.000
N	PM3-BC03	-3.080	-1.279	-3.187	35.074	46.653	37.657	0.000	0.000	0.000
N	PM3-BC04	-2.708	-1.157	-3.295	-1.374	-1.817	1.480	0.000	0.000	0.000
N	PM3-BC05	-6.593	-7.292	-7.832	-11.658	-8.863	-9.382	0.000	0.000	0.000
N	PM3-BC06	-10.034	-11.051	-13.059	-14.216	-9.666	-10.625	0.000	0.000	0.000
N	PM3-BC07	-7.943	-6.467	-8.421	-9.397	-7.671	-5.824	0.000	0.000	0.000
N	PM3-BC08	-2.543	-1.795	-4.361	-12.089	-9.733	-7.034	0.000	0.000	0.000
N	PM3-BC09	-8.339	-3.707	-6.432	-8.233	-5.587	-3.364	0.000	0.000	0.000
N	PM3-BC10	-3.584	-2.914	-3.487	3.564	1.945	16.597	0.758	0.000	0.000
N	PM3-BC11	-3.011	1.034	-1.730	-5.668	-1.480	0.746	0.000	0.000	0.000
N	PM3-BC12	1.958	8.863	3.361	8.393	13.297	18.930	7.529	2.267	1.562
N	PM3-BD01	-4.080	-2.211	-4.338	-2.545	-1.904	-1.497	0.434	1.237	2.347
N	PM3-BD02	-6.672	-7.247	-6.875	-7.234	-6.357	-6.872	0.886	0.000	0.000

N	PM3-BD03	-3.613	-1.547	-4.371	-9.957	-8.313	-8.633	0.000	0.000	0.000
N	PM3-BD04	-4.568	-5.293	-7.949	-7.571	-6.009	-4.474	0.000	0.000	0.000
N	PM3-BD05	-11.296	-11.462	-11.975	-14.658	-12.736	-12.739	0.000	0.000	0.000
N	PM3-BD06	-9.511	-9.033	-10.709	-12.454	-9.605	-9.154	1.513	0.000	0.000
N	PM3-BD07	-8.458	-7.054	-9.493	-10.143	-7.612	-6.447	0.000	0.000	0.000
N	PM3-BD08	-9.853	-7.849	-6.695	-10.863	-8.987	-6.964	0.000	0.000	0.000
N	PM3-BD09	-7.850	-3.089	-6.793	-11.132	-7.736	-5.645	0.000	0.000	0.000
N	PM3-BD10	-9.307	-3.087	-7.976	-11.038	-7.424	-4.759	0.758	0.503	0.000
N	PM3-BD11	-8.101	-3.403	-6.370	-13.437	-7.655	-6.349	0.000	0.000	0.000
N	PM3-BD12	2.995	2.779	-0.982	-4.717	-4.334	-1.764	8.805	2.158	2.393
N	PM3-BE01	-3.468	-2.420	-4.675	-4.009	-1.891	-1.562	2.829	1.024	0.633
N	PM3-BE02	-6.299	-5.457	-7.583	-8.721	-5.293	-5.434	0.000	0.000	0.000
N	PM3-BE03	-5.150	-2.789	-5.787	-6.986	-3.086	-4.836	0.000	0.000	0.000
N	PM3-BE04	-4.854	-2.105	-5.942	-6.776	-3.999	-3.546	0.000	0.000	0.000
N	PM3-BE05	-4.220	-1.414	-6.761	-6.764	-5.418	-3.521	0.000	0.000	0.000
N	PM3-BE06	-1.534	-0.288	-2.661	-1.709	34.759	43.171	0.000	0.000	0.000
N	PM3-BE07	-5.426	-2.189	-5.801	0.662	-4.083	7.063	0.000	0.000	0.000
N	PM3-BE08	22.945	15.083	10.788	155.688	99.862	100.149	0.000	0.000	0.000
N	PM3-BE09	-7.007	-3.413	-6.042	-10.258	-8.001	-5.791	0.000	0.000	0.000
N	PM3-BE10	-9.367	-4.667	-8.862	-12.254	-11.495	-8.972	1.438	1.391	0.000
N	PM3-BE11	13.258	22.836	9.243	9.976	31.792	13.436	2.108	0.000	0.000
N	PM3-BE12	12.239	15.037	33.403	33.438	40.214	31.538	4.200	0.616	1.279
N	PM3-BF01	44.629	54.203	71.850	159.137	158.530	154.349	18.609	20.430	17.250
N	PM3-BF02	-7.253	-14.005	-9.921	-10.283	-12.161	-13.059	1.162	1.939	0.503
N	PM3-BF03	97.724	81.629	83.532	203.700	207.282	212.499	5.797	0.000	0.000
N	PM3-BF04	24.972	14.761	37.278	210.474	215.101	218.675	0.000	0.000	0.000
N	PM3-BF05	-4.788	-1.004	-4.905	-6.391	-4.237	-2.533	0.000	0.000	0.000
N	PM3-BF06	-10.803	8.553	6.995	1.028	0.116	1.882	11.197	31.908	0.378
N	PM3-BF07	-16.862	-8.600	-10.432	-22.251	-6.025	-20.605	-0.171	0.899	0.000
N	PM3-BF08	-6.151	0.224	-6.876	-6.730	-7.832	-1.639	1.013	0.000	0.000
N	PM3-BF09	-8.155	-3.370	-6.763	-6.429	-5.075	-3.324	0.000	0.000	0.000
N	PM3-BF10	-7.387	-2.097	-8.179	-9.653	-7.870	-5.207	3.595	1.047	0.000
N	PM3-BF11	-7.457	-0.550	-5.945	-10.679	-6.996	-4.549	1.839	0.766	0.000
N	PM3-BF12	-9.507	-1.280	-3.824	-11.228	-4.921	-5.945	4.764	1.176	1.707
N	PM3-BG01	1.192	7.437	-2.416	21.658	39.843	31.586	-1.587	-4.286	-4.729
N	PM3-BG02	-2.442	-0.576	-3.722	32.434	35.316	34.878	0.000	0.000	0.000
N	PM3-BG03	-8.174	-6.747	-1.717	15.557	7.479	15.376	1.178	0.787	1.409
N	PM3-BG04	-11.886	-1.545	-4.433	-16.462	14.554	17.436	0.195	0.255	0.000
N	PM3-BG05	-2.821	0.911	-4.191	-0.170	4.296	14.932	0.000	0.000	0.000
N	PM3-BG06	-2.718	6.017	-2.470	10.184	13.863	25.847	0.000	0.000	0.000
N	PM3-BG07	-5.309	-1.114	-6.567	-4.579	1.678	0.596	0.000	0.000	0.000
N	PM3-BG08	-3.342	5.966	-2.559	39.961	40.471	46.245	0.000	0.000	0.000
N	PM3-BG09	-4.504	7.833	-1.789	37.272	34.617	41.559	0.380	0.000	0.000
N	PM3-BG10	-8.921	-8.099	-13.216	-11.597	32.679	27.226	0.592	0.680	1.021
N	PM3-BG11	-5.450	2.341	-2.439	-4.182	-3.504	0.193	5.537	3.105	1.289
N	PM3-BG12	-3.346	4.241	-2.529	3.375	5.257	2.526	2.638	0.859	-1.709
N	PM3-BH01	153.988	127.255	140.574	220.393	212.479	241.221	5.342	4.114	4.378
N	PM3-BH02	157.000	169.583	155.809	230.737	222.796	219.862	0.072	-0.976	-1.099
N	PM3-BH03	82.588	95.342	58.428	222.287	212.357	211.597	5.917	2.599	0.318
N	PM3-BH04	13.321	7.857	5.157	154.760	151.651	151.280	1.247	1.001	1.142
N	PM3-BH05	1.066	4.636	2.371	198.837	195.163	207.513	2.721	0.295	0.480
N	PM3-BH06	-1.453	6.314	0.199	51.559	52.491	35.212	-0.074	0.576	0.800
N	PM3-BH07	1.809	11.108	3.595	78.410	100.365	98.976	4.061	1.774	0.000
N	PM3-BH08	66.543	116.976	112.787	223.613	203.236	226.164	27.674	29.111	27.413
N	PM3-BH09	154.589	171.649	168.205	215.274	199.122	227.526	7.078	10.775	9.161
N	PM3-BH10	3.176	67.501	84.997	192.207	178.104	191.779	13.649	15.672	14.242
N	PM3-BH11	148.101	193.208	154.484	117.395	126.054	134.553	4.678	3.480	1.266
N	PM3-BH12	160.236	169.295	134.714	65.328	58.795	64.205	9.454	1.563	1.325
P	PM4-AA01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
P	PM4-AA02	1.833	-1.451	-1.796	28.876	49.566	28.472	0.000	0.503	
P	PM4-AA03	-8.100	-3.983	-3.333	48.263	43.112	21.932	5.416	-3.028	-2.383
P	PM4-AA04	-4.141	-6.245	-5.141	72.668	51.743	27.496	0.000	0.000	-2.092
P	PM4-AA05	-2.966	-2.068	-6.054	50.533	47.061	28.000	0.000	0.000	-3.100
P	PM4-AA06	-7.416	-5.483	-5.399	35.366	16.886	3.724	0.000	0.000	-3.100
P	PM4-AA07	-5.192	-5.779	-5.528	53.737	31.143	13.263	0.000	0.000	-3.100
P	PM4-AA08	79.878	11.118	76.608	142.714	128.746	94.054	0.000	0.000	-3.100
P	PM4-AA09	2.961	64.064	-10.420	150.254	138.342	137.957	0.000	1.529	-3.100
P	PM4-AA10	128.295	-13.441	9.062	171.991	138.251	-44.620	0.000	0.000	-0.508
P	PM4-AA11	78.501	-10.454	-11.696	173.700	134.224	127.139	5.316	8.507	4.268

P	PM4-AA12	116.257	72.225	108.387	152.426	121.161	117.635	4.745	4.475	4.814
P	PM4-AB01	153.253	139.814	117.189	42.171	17.172	3.707	0.000	0.000	-3.100
P	PM4-AB02	155.691	147.236	131.634	30.845	44.633	17.063	0.000	0.000	-3.100
P	PM4-AB03	71.584	39.787	69.897	185.683	153.341	150.558	0.000	0.000	-3.100
P	PM4-AB04	-7.550	-7.524	-7.539	145.514	128.504	112.104	0.000	0.000	-3.100
P	PM4-AB05	-7.787	-6.634	-6.333	76.424	54.380	35.154	0.000	0.000	-3.100
P	PM4-AB06	-7.332	-9.682	-6.376	144.939	47.901	65.557	0.000	0.000	-3.100
P	PM4-AB07	-10.401	-8.329	-9.013	59.984	47.036	25.662	0.000	0.000	-3.100
P	PM4-AB08	45.968	54.401	116.072	178.476	161.093	155.083	0.000	0.000	-3.100
P	PM4-AB09	111.197	9.807	72.122	176.326	138.757	126.225	0.000	0.000	-3.100
P	PM4-AB10	44.603	34.545	88.418	168.266	133.847	118.422	0.000	0.000	-3.100
P	PM4-AB11	36.779	21.987	50.168	142.538	117.476	104.697	0.000	0.000	-3.100
P	PM4-AB12	66.301	46.818	81.196	149.766	119.530	111.557	1.011	3.299	2.301
P	PM4-AC01	0.504	-7.028	-0.293	63.353	48.209	30.768	0.000	0.000	-3.100
P	PM4-AC02	-5.004	-7.618	-7.399	84.089	64.949	42.555	0.000	0.000	-1.022
P	PM4-AC03	77.170	93.295	11.189	200.322	170.576	155.272	0.000	0.000	-3.100
P	PM4-AC04	-14.457	72.184	-7.866	160.359	154.608	127.343	0.000	0.000	-3.100
P	PM4-AC05	-16.728	-15.180	-16.109	-15.154	-40.957	-50.413	0.000	0.000	-3.100
P	PM4-AC06	118.199	111.345	102.521	193.542	163.750	138.772	0.000	0.000	-3.100
P	PM4-AC07	-8.305	-9.149	-7.532	87.203	56.226	33.462	0.000	0.000	-3.100
P	PM4-AC08	37.425	31.496	56.811	182.985	144.930	133.354	0.000	0.000	-3.100
P	PM4-AC09	56.242	15.011	-14.243	180.695	156.453	-47.107	0.000	0.000	-3.100
P	PM4-AC10	18.399	4.338	-14.387	143.420	106.076	99.685	0.000	0.000	-3.100
P	PM4-AC11	28.211	-5.083	26.008	142.201	122.008	113.175	0.000	0.000	-3.100
P	PM4-AC12	23.734	4.524	41.229	149.410	116.479	119.822	0.503	5.136	4.154
P	PM4-AD01	20.547	80.864	97.903	165.300	152.080	144.512	0.000	0.000	-3.100
P	PM4-AD02	32.228	43.492	31.458	101.432	68.212	4.583	0.000	0.000	-3.100
P	PM4-AD03	64.347	18.309	6.949	82.603	50.101	36.943	0.000	0.000	-3.100
P	PM4-AD04	30.804	8.059	14.355	26.907	35.105	13.761	0.000	0.000	-3.100
P	PM4-AD05	-9.213	-5.439	-6.112	18.674	36.553	24.184	0.000	0.000	-3.100
P	PM4-AD06	-4.666	-6.372	-4.432	73.634	49.218	30.346	0.000	0.000	-3.100
P	PM4-AD07	-5.658	-5.112	-5.041	90.768	54.966	30.393	0.000	0.000	-3.100
P	PM4-AD08	-0.257	-4.158	51.413	174.974	146.055	92.897	0.000	0.000	-3.100
P	PM4-AD09	-6.779	13.499	40.021	120.787	136.029	122.782	0.000	0.000	-3.100
P	PM4-AD10	-12.914	14.726	-13.289	110.863	127.747	116.975	0.000	0.000	-3.100
P	PM4-AD11	-4.024	-2.499	19.368	155.201	122.416	119.433	0.000	0.000	-3.100
P	PM4-AD12	17.738	3.399	37.078	131.062	111.778	100.167	0.000	3.966	1.014
P	PM4-AE01	-1.616	-1.066	-8.682	80.907	55.839	35.766	0.000	0.000	-3.100
P	PM4-AE02	-5.011	-5.459	-9.833	80.351	49.846	-47.463	0.000	0.000	-3.100
P	PM4-AE03	-10.816	-4.861	-4.325	82.941	47.021	25.553	0.000	0.000	-3.100
P	PM4-AE04	-7.638	-7.134	-8.208	34.713	21.211	5.137	0.000	0.000	-3.100
P	PM4-AE05	-6.763	-6.218	-7.332	60.664	33.322	15.992	0.000	0.000	-3.100
P	PM4-AE06	-8.330	-6.582	-5.336	-1.916	-28.666	-40.380	0.000	0.000	-3.100
P	PM4-AE07	-7.392	-7.391	-6.471	70.475	40.850	20.487	0.000	0.000	-3.100
P	PM4-AE08	-5.188	-8.522	-5.733	41.046	19.512	-2.293	0.000	0.000	-3.100
P	PM4-AE09	-4.916	-4.358	-4.967	145.704	72.922	53.525	0.000	0.000	-3.100
P	PM4-AE10	47.220	9.176	54.762	176.179	140.717	130.710	0.000	0.000	-3.100
P	PM4-AE11	-9.992	3.508	3.767	1.838	-13.674	-26.737	0.000	1.412	-1.558
P	PM4-AE12	45.934	27.799	48.125	160.713	128.400	130.667	0.000	0.000	-2.597
S	PM4-AF01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
S	PM4-AF02	3.441	-0.101	-1.853	79.425	79.026	74.303	0.000	0.000	
S	PM4-AF03	192.607	160.508	146.010	96.636	87.126	84.274	0.000	0.000	0.000
S	PM4-AF04	179.574	164.603	136.870	104.876	82.816	78.193	0.000	0.000	0.000
S	PM4-AF05	166.420	124.759	98.622	79.207	75.039	74.111	0.000	0.000	0.000
S	PM4-AF06	180.543	157.907	139.333	73.172	95.418	81.084	0.000	0.000	0.000
S	PM4-AF07	175.554	150.195	135.151	69.974	92.067	88.722	0.000	0.000	0.000
S	PM4-AF08	176.199	166.482	145.654	107.866	103.375	87.049	0.000	0.000	0.000
S	PM4-AF09	163.796	163.922	147.695	19.230	76.391	62.539	0.000	0.000	0.000
S	PM4-AF10	-3.180	-4.018	-7.874	-1.634	2.141	2.003	0.000	0.000	0.000
S	PM4-AF11	4.030	8.634	0.575	-0.862	16.767	3.093	0.000	0.000	0.000
S	PM4-AF12	162.564	140.708	123.125	18.422	52.647	53.226	1.203	1.225	2.588
S	PM4-AG01	-0.668	-1.526	-1.095	105.749	1.839	0.651	1.805	1.611	0.795
S	PM4-AG02	57.934	58.067	40.364	6.128	15.039	3.049	0.000	0.000	0.000
S	PM4-AG03	181.575	154.942	134.641	91.770	86.261	76.926	0.000	0.000	0.000
S	PM4-AG04	204.475	176.053	157.489	121.817	97.236	83.705	0.000	0.000	0.000
S	PM4-AG05	165.355	25.330	13.507	61.733	3.982	-5.797	0.000	0.000	0.000
S	PM4-AG06	25.380	34.092	1.207	12.441	38.053	13.342	0.000	0.000	0.000
S	PM4-AG07	176.850	146.238	129.578	113.667	118.142	114.450	0.000	0.000	0.000
S	PM4-AG08	138.288	125.513	105.285	67.259	89.895	69.974	0.000	0.000	0.000

S	PM4-AG09	180.024	144.409	139.475	141.645	135.838	141.692	0.000	0.000	0.000
S	PM4-AG10	172.629	142.817	119.516	86.661	113.526	112.212	0.000	0.000	0.000
S	PM4-AG11	159.791	137.612	123.789	93.912	103.847	103.250	0.000	0.000	0.000
S	PM4-AG12	-0.757	-1.761	-0.771	-2.941	0.717	1.164	3.995	4.063	5.359
S	PM4-AH01	151.968	110.154	105.784	58.324	68.024	59.249	2.803	1.788	2.216
S	PM4-AH02	27.053	32.964	24.988	20.661	32.745	12.634	0.000	0.000	1.005
S	PM4-AH03	40.650	2.057	1.991	6.557	5.139	-4.189	0.000	0.000	0.000
S	PM4-AH04	11.120	9.391	7.351	1.943	15.316	0.011	0.000	0.000	0.000
S	PM4-AH05	32.188	24.051	14.475	3.807	28.011	1.645	0.000	0.000	0.000
S	PM4-AH06	83.996	81.445	53.807	-1.164	35.829	2.042	0.000	0.000	0.000
S	PM4-AH07	86.434	80.762	58.399	17.934	43.989	13.975	0.000	0.000	0.000
S	PM4-AH08	2.342	2.816	2.924	5.313	2.999	3.430	0.000	0.000	0.000
S	PM4-AH09	0.362	1.220	0.653	-1.712	4.525	2.905	0.000	0.000	0.000
S	PM4-AH10	3.457	1.822	1.267	-6.551	-0.382	1.779	0.000	0.000	0.000
S	PM4-AH11	-0.238	-1.630	-1.691	-5.391	-3.683	-2.222	0.000	0.000	0.503
S	PM4-AH12	-0.028	-2.549	0.458	-5.645	-0.620	1.580	3.436	4.892	8.084

Table 9.10: Positive signal values and significant changes compared to BT 1B strain 8081 for *Y. enterocolitica* biotypes. p – p value; Λ / V - significantly higher/lower activity.

	Nutrient sources		BT 1A 5303	p	BT 1B 8081	BT 2 21202	p	BT 3 5603	p	BT 3 14902	p	BT 4 1203	p	BT 5 3094	p
C	L-Arabinose	PM1-A02	+		+	+		+		+		+	0.030 Λ	+	0.000 V
C	N-Acetyl-D-Glucosamine	PM1-A03	+	0.045 Λ	+	+		+		+		+		+	
C	D-Galactose	PM1-A06	+		+	+		+		+		+		+	0.000 V
C	D-Mannose	PM1-A11	+		+	+		+		+		+		+	
C	Glycerol	PM1-B03	+		+	+		+		+	0.050 V	+		+	0.004 V
C	D-Glucuronic Acid	PM1-B05	+	0.009 Λ	+	+		+		+		+		+	0.000 V
C	D-Gluconic Acid	PM1-B06	+		+	+		+		+		+	0.012 V	+	0.000 V
C	D,L-a-Glycerol Phosphate	PM1-B07	+	0.001 Λ	+	+		+		+		+	0.018 Λ	+	
C	D-Mannitol	PM1-B11	+	0.006 Λ	+	+		+		+		+		+	
C	D-Ribose	PM1-C04	+		+	+		+		+	0.014 V	+		+	0.001 V
C	D-Fructose	PM1-C07	+		+	+	0.044 Λ	+	0.005 Λ	+		+		+	
C	a-D-Glucose	PM1-C09	+		+	+	0.033 Λ	+	0.010 Λ	+		+		+	
C	Sucrose	PM1-D11	+		+	+		+		+		+		+	
C	D-Glucose-1-Phosphate	PM1-E03	+	0.010 Λ	+	+		+		+		+		+	0.000 V
C	2'-Deoxyadenosine	PM1-E11	+	0.000 Λ	+	+		+		+		+	0.005 Λ	+	0.007 V
C	Adenosine	PM1-E12	+	0.000 Λ	+	+	0.018 Λ	+		+	0.001 Λ	+	0.012 Λ	+	
C	D-Cellobiose	PM1-F11	+		+	+		+		+		+		+	0.018 V
C	Inosine	PM1-F12	+	0.000 Λ	+	+	0.006 Λ	+	0.004 Λ	+	0.002 Λ	+	0.000 Λ	+	
C	L-Serine	PM1-G03	+	0.000 Λ	+	+	0.018 Λ	+		+		+	0.037 Λ	+	0.000 V
C	N-Acetyl-D-Mannosamine	PM1-G08	+		+	+		+		+		+		+	0.000 V
C	Methylpyruvate	PM1-G10	+	0.007 Λ	+	+		+		+		+		+	0.002 V
C	glycyl-L-proline	PM1-H01	+		+	+	0.008 Λ	+	0.023 Λ	+		+		+	0.000 V
C	Pyruvic Acid	PM1-H08	+	0.001 Λ	+	+	0.015 Λ	+		+		+		+	0.000 V
C	D-Galacturonic Acid	PM1-H10	+	0.000 Λ	+	+	0.007 Λ	+	0.015 Λ	+		+	0.029 Λ	+	0.000 V
C	Pectin	PM2-AA12	+		+	+		+		+		+		+	
C	N-Acetyl-Neuraminic Acid	PM2-AB02	+	0.002 Λ	+	+		+		+		+		+	0.000 V
C	D-Glucosamine	PM2-AE05	+	0.003 Λ	+	+		+		+	0.000 V	+		+	0.000 V
C	D-Trehalose	PM1-A10	+		+	+		+		+		+			0.000 V
C	D-Sorbitol	PM1-B02	+	0.008 Λ	+	+		+		+		+			0.000 V
C	D-Glucose-6-Phosphate	PM1-C01	+		+	+		+		+		+			0.000 V
C	Acetic Acid	PM1-C08	+	0.005 Λ	+	+		+		+		+			0.001 V
C	Maltose	PM1-C10	+		+	+		+		+		+			0.000 V
C	Uridine	PM1-D12	+	0.001 Λ	+	+		+		+		+	0.005 Λ		
C	D-Fructose-6-Phosphate	PM1-E04	+		+	+		+		+		+			0.000 V
C	b-Methyl-D-Glucoside	PM1-E08	+		+	+		+		+		+			0.000 V
C	Maltotriose	PM1-E10	+		+	+		+		+		+			0.000 V
C	Glycyl-L-glutamic acid	PM1-G01	+		+	+		+		+		+			0.000 V
C	L-Alanine	PM1-G05	+	0.000 Λ	+	+	0.000 Λ	+	0.002 Λ	+		+	0.041 Λ		0.000 V
C	L-alanyl-glycine	PM1-G06	+	0.003 Λ	+	+		+		+		+			0.000 V
C	Gentiobiose	PM2-AC01	+		+	+	0.001 Λ	+	0.003 Λ	+		+	0.004 Λ		0.000 V
C	L-Sorbose	PM2-AD04	+	0.001 Λ	+	+		+		+		+			0.000 V
C	5-Keto-D-Gluconic Acid	PM2-AE12	+	0.000 Λ	+	+	0.001 Λ	+	0.006 Λ	+	0.003 Λ		0.000 V		0.000 V

C	2-Deoxy-D-Ribose	PM2-AB09	+			+			+			+		0.000	V		0.001	V		0.000	V
C	D-Xylose	PM1-B08	+			+			+			+		0.003	V		0.000	V	+	0.000	V
C	Succinic Acid	PM1-A05	+			+			+					0.005	V		0.013	V	+		
C	D,L-Malic Acid	PM1-C03	+			+			+					0.016	V		0.027	V	+		
C	Fumaric Acid	PM1-F05	+	0.007	Λ	+			+	0.045	V		0.001	V		0.000	V		0.001	V	+		
C	L-Aspartic Acid	PM1-A07	+			+								0.004	V		0.011	V	+		
C	L-Malic Acid	PM1-G12	+			+			+					0.007	V	+	0.035	V	+		
C	L-Glutamic Acid	PM1-B12	+			+			+					0.016	V	+				0.020	V
C	Glycyl-L-aspartic acid	PM1-F01	+	0.043	Λ	+										+					
C	Arbutin	PM2-AB08	+	0.000	Λ	+		0.000	V		0.000	V		0.000	V	+	0.000	V		0.000	V
C	N-Acetyl-D-galactosamine	PM2-AB01	+	0.000	Λ		+	0.000	Λ	+	0.000	Λ	+	0.000	Λ	+	0.000	Λ			
C	D-Alanine	PM1-A09	+	0.000	V	+		0.000	V		0.000	V		0.000	V		0.000	V		0.000	V
C	L-Asparagine	PM1-D01	+			+					0.027	V		0.003	V		0.007	V		0.009	V
C	Dextrin	PM2-AA06	+			+															
C	succinamic acid	PM2-AF10	+			+					0.006	V		0.002	V		0.005	V		0.004	V
C	L-Histidine	PM2-AG06	+	0.001	Λ	+															
C	D-Serine	PM1-B01	+	0.000	Λ																
C	Formic acid	PM1-B10	+	0.000	Λ																
C	D-galactonic acid-g-lactone	PM1-C02	+	0.000	Λ																
C	tween 20	PM1-C05	+	0.000	Λ																
C	thymidine	PM1-C12	+	0.006	Λ															0.002	V
C	a-D-lactose	PM1-D09	+	0.000	Λ																
C	Lactulose	PM1-D10	+	0.000	Λ																
C	L-glutamine	PM1-E01	+																		
C	Bromo succinic acid	PM1-F06	+	0.003	Λ																
C	mono methyl succinate	PM1-G09	+																		
C	D-Arabitol	PM2-AB06	+	0.000	Λ															0.045	Λ
C	3-0-b-D-Galactopyranosyl-D-arabinose	PM2-AB12	+	0.000	Λ																
C	b-Methyl-D-galactoside	PM2-AC07	+	0.000	Λ																
C	D-Raffinose	PM2-AD01	+	0.000	Λ																
C	Salicin	PM2-AD02	+	0.000	Λ																
C	Glycogen	PM1-A08					+													0.040	V
C	m-Inositol	PM1-F03		0.000	V	+	+			+			+			+	0.003	V		0.000	V
N	L-Asparagine	PM3-BA09	+	0.019	Λ	+	+			+			+	0.000	V	+				0.000	V
N	L-Aspartic Acid	PM3-BA10	+	0.023	Λ	+	+			+			+	0.000	V	+				0.000	V
N	L-Histidine	PM3-BB03	+	0.027	Λ	+	+			+			+	0.006	V	+				0.000	V
N	L-Serine	PM3-BB10	+	0.000	Λ	+	+	0.001	Λ	+	0.003	Λ	+	0.000	V	+				0.000	V
N	N-Acetyl-D-Mannosamine	PM3-BF01	+			+	+			+			+	0.000	V	+				0.000	V
N	Adenosine	PM3-BF03	+	0.001	Λ	+	+			+			+	0.000	V	+				0.000	V
N	Ala-Asp	PM3-BH01	+	0.030		+	+			+			+	0.000	V	+				0.000	V
N	Ala-Gln	PM3-BH02	+			+	+			+			+	0.000	V	+				0.000	V
N	Ala-Glu	PM3-BH03	+			+	+			+			+	0.000	V	+				0.000	V
N	Gly-Asn	PM3-BH08	+			+	+			+			+	0.000	V	+				0.000	V
N	Gly-Gln	PM3-BH09	+	0.020	Λ	+	+			+			+	0.003	V	+				0.000	V
N	Gly-Met	PM3-BH11	+			+	+			+			+			+	0.012	V		0.000	V
N	Met-Ala	PM3-BH12	+	0.000	Λ	+	+	0.023	Λ	+			+	0.007	Λ	+	0.004	V		0.000	V
N	Ammonia	PM3-BA02	+	0.000	Λ	+	+	0.000	Λ	+	0.001	Λ		0.000	V	+	0.033	V		0.000	V

N	Urea	PM3-BA05	+	0.000	Λ	+	+	0.004	Λ	+	0.023	Λ	0.000	V	+		0.000	V
N	L-Alanine	PM3-BA07	+	0.000	Λ	+	+	0.000	Λ	+	0.003	Λ	0.000	V	+	0.006	V	0.000	V
N	L-Arginine	PM3-BA08	+	0.000	Λ	+	+			+			0.000	V	+	0.012	V	0.000	V
N	L-Glutamic Acid	PM3-BA12	+			+	+			+	0.008	V	0.000	V	+			0.000	V
N	L-Glutamine	PM3-BB01	+	0.027	Λ	+	+			+			0.000	V	+			0.000	V
N	Glycine	PM3-BB02	+	0.000	Λ	+	+	0.018	Λ	+			0.003	V	+	0.006	Λ	0.002	V
N	L-Proline	PM3-BB09	+			+	+			+			0.000	V	+	0.027	V	0.000	V
N	D-Alanine	PM3-BC03	+	0.001	Λ	+	+			+			0.000	V	+	0.001	V	0.000	V
N	D-Glucosamine	PM3-BE08	+	0.005	Λ	+	+	0.045	Λ	+			0.000	V	+			0.000	V
N	Cytidine	PM3-BF04	+	0.000	Λ	+	+	0.007	Λ	+			0.000	V	+			0.000	V
N	Ala-Gly	PM3-BH04	+	0.000	Λ	+	+	0.015	Λ	+			0.000	V	+			0.000	V
N	Ala-His	PM3-BH05	+	0.000	Λ	+	+	0.000	Λ	+	0.000	Λ	0.000	V	+	0.000	Λ	0.000	V
N	Ala-Leu	PM3-BH06	+	0.000	Λ	+	+	0.003	Λ	+	0.000	Λ	0.019	V	+			0.010	V
N	Ala-Thr	PM3-BH07	+	0.000	Λ	+	+	0.000	Λ	+	0.000	Λ	0.000	V	+	0.001	V	0.000	V
N	Gly-Glu	PM3-BH10	+			+	+			+			0.000	V	+			0.000	V
N	L-Threonine	PM3-BB11	+	0.025	Λ	+	+						0.033	V	+			0.013	V
N	L-Methionine	PM3-BB07	+	0.009	Λ	+	+				0.035	V	0.001	V				0.001	V
N	N-Acetyl-D-Glucosamine	PM3-BE11	+			+				+	0.020	Λ							
N	g-Amino-N-Butyric Acid	PM3-BG08	+			+							0.016	V	+			0.010	V
N	e-Amino-N-Caproic Acid	PM3-BG09	+			+							0.018	V	+			0.011	V
N	L-Phenylalanine	PM3-BB08	+	0.000	Λ	+												0.001	V
N	D-Asparagine	PM3-BC04	+	0.000	Λ	+		0.000	V		0.000	V	0.000	V		0.000	V	0.000	V
N	L-Cysteine	PM3-BA11	+				+			+			+					0.026	V
N	D,L-Lactamide	PM3-BE07	+	0.000	Λ		+	0.003	Λ	+	0.010	Λ							
N	D,L-a-Amino-N-Butyric Acid	PM3-BG07	+				+	0.006	Λ	+	0.000	Λ	0.012	V		0.041	V	0.015	V
N	N-Acetyl-D-Galactosamine	PM3-BE12	+				+	0.001	Λ	+	0.000	Λ			+				
N	L-Leucine	PM3-BB05	+	0.000	Λ			0.029	Λ		0.034	Λ			+				
N	Xanthosine	PM3-BG02	+										0.020	V	+			0.018	V
N	D,L-a-amino-caprylic acid	PM3-BG10								+									
N	L-Tryptophan	PM3-BB12				+		0.018	V		0.016	V	0.012	V		0.017	V	0.009	V
N	xanthine	PM3-BG01				+		0.006	V		0.001	V						0.000	V
N	L-Isoleucine	PM3-BB04	+	0.000	Λ											0.028	Λ	0.049	V
N	L-Valine	PM3-BC02	+	0.001	Λ								0.047	V					
N	D-Serine	PM3-BC08	+	0.000	Λ														
N	L-Ornithine	PM3-BC12	+	0.000	Λ														
N	cytosine	PM3-BF05	+					0.000	V		0.000	V				0.000	V	0.000	V
N	Uric Acid	PM3-BG03	+	0.001	Λ														
N	a-Amino-N-Valeric Acid	PM3-BG12	+	0.000	Λ														
P	adenosine-3'.5'-cyclic monphosphate (3',5' cyclic AMP)	PM4-AA12	+			+	+			+			+		+			0.000	V
P	D-glucosamine-6-phosphate	PM4-AC06	+			+	+			+			+	0.002	V	+		0.000	V
P	dithiophosphate	PM4-AB02	+	0.000	Λ	+	+			+			+	0.001	Λ			0.000	V
P	phosphate	PM4-AA02	+	0.000	Λ	+	+			+	0.041	Λ	0.001	V				0.001	V
P	carbamyl phosphate	PM4-AB05	+	0.000	Λ	+	+			+	0.003	Λ	0.000	V				0.000	V
P	6-phospho-gluconic	PM4-AC07	+	0.000	Λ	+	+			+			0.000	V		0.033	V	0.000	V

	acid															
P	D-mannose-6-phosphate	PM4-AD02	+			+			+			0.000	V	0.004	V	0.000	V
P	o-phospho-D-serine	PM4-AD05	+			+			+	0.003	Λ	0.000	V	0.047	V	0.000	V
P	o-phospho-L-serine	PM4-AD06	+			+			+	0.016	Λ	0.000	V			0.000	V
P	o-phospho-L-threonine	PM4-AD07	+	0.000	Λ	+			+	0.016	Λ	0.000	V			0.000	V
P	O-phosphoryl-ethanolamine	PM4-AE05	+	0.000	Λ	+	+	0.029	Λ	+	0.002	Λ	0.000	V			0.000	V
P	thymidine-3'-monophosphate	PM4-AE09	+			+			+			0.000	V			0.000	V
P	adenosine-3'-monophosphate (3'-AMP)	PM4-AA09	+			+			+			0.000	V	+		0.000	V
P	D,L-alpha-glycerol phosphate	PM4-AB03	+	0.010	Λ	+			+			0.000	V	+		0.000	V
P	beta-glycerol phosphate	PM4-AB04	+			+			+			0.000	V	+		0.000	V
P	guanosine-2'-monophosphate	PM4-AB08	+			+			+			0.000	V	+		0.000	V
P	guanosine-3'-monophosphate	PM4-AB09	+			+			+			0.026	V	+		0.000	V
P	guanosine-5'-monophosphate	PM4-AB10	+	0.030	Λ	+			+			0.002	V	+		0.000	V
P	guanosine-2',3'-cyclic monophosphate	PM4-AB11	+	0.018	Λ	+			+			0.000	V	+		0.000	V
P	guanosine-3',5'-cyclic monophosphate	PM4-AB12	+	0.036	Λ	+			+			0.017	V	+		0.000	V
P	D-glucose-1-phosphate	PM4-AC03	+			+			+			0.001	V	+		0.000	V
P	cytidine-2'-monophosphate	PM4-AC08	+	0.005	Λ	+			+			0.000	V	+		0.000	V
P	cytidine-2',3'-cyclic monophosphate	PM4-AC11	+			+			+			0.000	V	+		0.000	V
P	cytidine-3',5'-cyclic monophosphate	PM4-AC12	+			+			+			0.000	V	+		0.000	V
P	uridine-3'-monophosphate	PM4-AD09	+	0.042	Λ	+			+			0.000	V	+		0.000	V
P	uridine-2',3'-cyclic monophosphate	PM4-AD11	+			+			+			0.000	V	+		0.000	V
P	uridine-3',5'-cyclic monophosphate	PM4-AD12	+	0.026	Λ	+			+			0.000	V	+		0.000	V
P	thymidine-5'-monophosphate	PM4-AE10	+	0.021	Λ	+			+			0.000	V	+		0.000	V
P	thymidine-3',5'-cyclic monophosphate	PM4-AE12	+	0.003	Λ	+			+			0.000	V	+		0.000	V
P	thiophosphate	PM4-AB01	+	0.000	Λ		+	0.045	Λ	+	0.006	Λ	+	0.000	Λ	0.013	V
P	pyrophosphate	PM4-AA03	+	0.000	Λ		+	0.015	Λ	+	0.005	Λ	0.047	V			
P	trimetaphosphate	PM4-AA04	+	0.000	Λ		+	0.320	Λ	+	0.003	Λ	0.003	V			0.004	V
P	tripolyphosphate	PM4-AA05	+	0.000	Λ		+	0.007	Λ	+	0.001	Λ	0.050	V			0.047	V
P	D-2-phospho-glyceric acid	PM4-AB06	+	0.000	Λ		+			+	0.021	Λ	0.007	V			0.010	V
P	D-3-phospho-glyceric acid	PM4-AB07	+	0.000	Λ		+	0.011	Λ	+	0.001	Λ	0.000	V			0.001	V
P	phosphoenoly pyruvate	PM4-AC01	+	0.000	Λ		+	0.026	Λ	+	0.003	Λ	0.003	V			0.003	V

P	phospho-glycolic acid	PM4-AC02	+	0.000	Λ		+		+	0.005	Λ	0.000	V			0.001	V
P	cysteamine-S-phosphate	PM4-AD03	+	0.000	Λ		+		+	0.019	Λ					0.003	V
P	phospho-L-arginine	PM4-AD04	+	0.000	Λ		+		+	0.004	Λ	0.002	V		0.015	V	0.000	V
P	phosphocreatine	PM4-AE03	+	0.000	Λ		+	0.028	Λ	+	0.004	Λ	0.003	V			0.005	V
P	phosphoryl choline	PM4-AE04	+	0.000	Λ		+			+	0.003	Λ	0.005	V			0.008	V
P	adenosine-2'-monophosphate	PM4-AA08	+	0.000	Λ		+	0.032	Λ	+	0.003	Λ			+	0.038	Λ	0.010	V
P	O-phospho-D-tyrosine	PM4-AE01	+	0.002	Λ		+												
P	uridine-2'-monophosphate	PM4-AD08	+	0.046	Λ					+					+				
P	D-mannose-1-phosphate	PM4-AD01	+			+				+					+			0.006	V
P	2-aminoethyl phosphonic acid	PM4-AE07	+	0.001	Λ	+				+			0.000	V				0.000	V
P	adenosine-2',3'-cyclic monophosphate	PM4-AA11				+	+			+					+			0.027	V
P	D-glucose-6-phosphate	PM4-AC04				+	+			+			0.015	V	+			0.005	V
P	cytidine-3'-monophosphate	PM4-AC09				+	+											0.024	V
P	cytidine-5'-monophosphate	PM4-AC10		0.000	V	+				+			0.007	V	+			0.004	V
P	uridine-5'-monophosphate	PM4-AD10				+				+					+			0.039	V
P	O-phospho-L-tyrosine	PM4-AE02		0.043	Λ		+			+	0.046	Λ							
P	adenosine-5'-monophosphate	PM4-AA10								+									
P	phosphono acetic acid	PM4-AE06	+	0.000	Λ			0.026	Λ		0.000	Λ							
P	inositol hexaphosphate	PM4-AE11	+	0.000	Λ														
S	thiosulfate	PM4-AF03	+	0.000	Λ	+	+	0.003	Λ	+	0.007	Λ	+	0.026	Λ	+			0.000	V
S	tetrathionate	PM4-AF04	+	0.000	Λ	+	+	0.002	Λ	+	0.003	Λ	+	0.012	Λ	+			0.000	V
S	thiophosphate	PM4-AF05	+	0.012	Λ	+	+			+			+			+			0.001	V
S	dithiophosphate	PM4-AF06	+	0.000	Λ	+	+	0.004	Λ	+	0.012	Λ	+			+			0.000	V
S	L-cysteine	PM4-AF07	+	0.010	Λ	+	+	0.015	Λ	+	0.029	Λ	+			+			0.000	V
S	D-cysteine	PM4-AF08	+	0.000	Λ	+	+	0.000	Λ	+	0.000	Λ	+	0.001	Λ	+			0.000	V
S	cystathionine	PM4-AG03	+	0.000	Λ	+	+	0.002	Λ	+	0.003	Λ	+	0.026	Λ	+			0.000	V
S	lanthionine	PM4-AG04	+	0.001	Λ	+	+	0.006	Λ	+	0.010	Λ	+	0.013	Λ	+			0.000	V
S	L-methionine	PM4-AG07	+	0.002	Λ	+	+	0.018	Λ	+	0.036	Λ	+			+			0.000	V
S	D-methionine	PM4-AG08	+	0.001	Λ	+	+	0.015	Λ	+	0.008	Λ	+			+			0.000	V
S	glycyl-L-methionine	PM4-AG09	+	0.001	Λ	+	+	0.007	Λ	+			+			+			0.000	V
S	N-acetyl-D,L-methionine	PM4-AG10	+	0.002	Λ	+	+	0.021	Λ	+			+			+			0.000	V
S	L-methionine sulfoxide	PM4-AG11	+	0.000	Λ	+	+	0.001	Λ	+	0.006	Λ	+			+			0.000	V
S	L-djenkolic acid	PM4-AH01	+	0.001	Λ	+	+	0.018	Λ	+	0.004	Λ	+	0.049	Λ	+			0.017	V
S	L-cysteinyl-glycine	PM4-AF09	+			+	+	0.029	Λ	+	0.045	Λ	+				0.002	V	0.000	V
S	L-cystein sulfinic acid	PM4-AF12	+	0.000	Λ	+	+	0.003	Λ	+	0.008	Λ	+	0.005	Λ				0.000	V
S	sulfate	PM4-AF02	+	0.000	Λ	+	+	0.000	Λ	+	0.001	Λ		0.000	V	+			0.000	V
S	taurine	PM4-AH06	+	0.001	Λ					+	0.024	Λ		0.024	Λ					
S	S-methyl-L-cystein	PM4-AG02	+	0.000	Λ															
S	hypotaurine	PM4-AH07		0.003	Λ					+			+	0.017	Λ					
S	D,L-ethionine	PM4-AG06								+	0.002	Λ								

References

454 LIFE SCIENCES, 2008. 3K long-tag paired-end sequencing with the genome sequencer FLX system. *Nature Methods*, **5**: i – ii.

ACHTMAN, M., ZURTH, K., MORELLI, G., TORREA, G., GUIYOULE, A., and CARNIEL, E., 1999. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *PNAS*, **96** (24): 14043 – 14048.
Correction to Achtman, et al., 1999, in *PNAS*, **97** (14): 8192.

ACHTMAN, M., MORELLI, G., ZHU, P., WIRTH, T., DIEHL, I., KUSECEK, B., VOGLER, A.J., WAGNER, D.M., ALLENDER, C.J., EASTERDAY, W.R., CHENAL-FRANCISQUE, V., WORSHAM, P., THOMSON, N.R., PARKHILL, J., LINDLER, L.E., CARNIEL, E., and KEIM, P., 2004. Microevolution and history of the plague bacillus, *Yersinia pestis*. *PNAS*, **101** (51): 17837 – 17842.

AKIBA, M., 2007. Salmonella enterica subsp. enterica serovar Dublin str. L-789 plasmid pMAK2 complete sequence. NCBI accession number: AB366441.1.

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W., and LIPMAN, D.J., 1990. Basic local alignment search tool. *J Mol Biol*, **215** (3): 403 – 410.

ANSORGE, W.J., 2009. Next-generation DNA sequencing techniques. *New Biotechnology*, **25** (4): 195 – 203.

ASAI, T., ZAPOROJETS, D., SQUIRES, C., and SQUIRES, C.L., 1999. An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. *PNAS*, **96**: 1971 – 1976.

ASSEFA, S., KEANE, T.M., OTTO, T.D., NEWBOLD, C., and BERRIMAN, M., 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, **25** (15): 1968 – 1969.

AURY, J.M., CRUAUD, C., BARBE, V., ROGIER, O., MANGENOT, S., SAMSON, G., POULAIN, J., ANTHOUARD, V., SCARPELLI, C., ARTIGUENAVE, F., and WINCKER, P., 2008. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics*, **9**: 603.

BACH, S., BUCHRIESER, C., PRENTICE, M., GUIYOULE, A., MSADEK, T., and CARNIEL, E., 1999. The high-pathogenicity island of *Yersinia enterocolitica* Ye8081 undergoes low-frequency deletion but not precise excision, suggesting recent stabilization in the genome. *Inf Immun*, **67** (10): 5091 – 5099.

BADGER, J.L., and MILLER, V.L., 1998. Expression of invasin and motility are coordinately regulated in *Yersinia enterocolitica*. *J Bact*, **180** (4): 793 – 800.

BALADA-LLASAT, J.-M., and MECSAS, J., 2006. *Yersinia* has a tropism for B and T cell zones of lymph nodes that is independent of the type III secretion system. *PLoS Pathogens*, **2** (9): e86.

BARTLING, C., WÖLFEL, R., NIKOLAOU, K., PETRY, T., THIEDE, S., HILDEBRANDT, T., FASSBENDER, M., GÖRITZ, F., BLOTTNER, S., SPITTLER, H., and NEUBAUER, H., 2004. Prevalence of anti-*Yersinia* antibodies in European brown hares in North-Rhine Westphalia, Germany. *Dtsch tierärztl Wschr*, **111**: 259 – 264. [German]

BATZILLA, J., HÖPER, D., ANTONENKA, U., HEESEMANN, J., and RAKIN, A., 2011a. Complete genome sequence of *Yersinia enterocolitica* subsp. *palearctica* serogroup O:3. *J Bact*, **193** (8): 2067.

BATZILLA, J., ANTONENKA, U., HÖPER, D., HEESEMANN, J., and RAKIN, A., 2011b. *Yersinia enterocolitica palearctica* serobiotype O:3/4 – a successful group of emerging zoonotic pathogens. *BMC Genomics*, **12**: 348.

BATZILLA, J., HEESEMANN, J., and RAKIN, A., 2011. The pathogenic potential of *Yersinia enterocolitica* 1A. *Int J Med Microbiol* [doi: 10.1016/j.ijmm.2011.05.002]

BERRIMAN, M., and RUTHERFORD, K., 2003. Viewing and annotating sequence data with Artemis. *Brief Bioinf,* **4**: 124 – 132.

BLEASDALE, B., LOTT, P.J., JAGANNATHAN, A., STEVENS, M.P., BIRTLES, R.J., and WIGLEY, P., 2009. The *Salmonella* pathogenicity island 2-encoded type III secretion system is essential for the survival of *Salmonella enterica* serovar Typhimurium in free-living amoebae. *Appl Env Microbiol*, **75** (6): 1793 – 1795.

BOCHNER, B.R., GADZINSKI, P., and PANOMITROS, E., 2001. Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res*, **11**: 1246 – 1255.

BOCHNER, B.R., 2003. New technologies to assess genotype-phenotype relationships. *Nature Reviews Genetics*, **4**: 309 – 314.

BOCHNER, B.R., 2009. Global phenotypic characterization of bacteria. *FEMS Microbiol Rev*, **33**: 191 – 205.

BOLAND, A., HAVAUX, S., and CORNELIS, G.R., 1998. Heterogeneity of the *Yersinia* YopM protein. *Micro Path*, **25** (6): 343 – 348.

BONFIELD, J.K., SMITH, K., and STADEN, R., 1995. A new DNA sequence assembly program. *Nucleic Acids Res*, **23** (24): 4992 – 4999. (http://staden.sourceforge.net/)

BOTTONE, E.J., 1977. *Yersinia enterocolitica*: a panoramic view of a charismatic microorganism. *Crit Rev Microbiol*, **5** (2): 211 – 214.

BOTTONE, E.J., 1997. *Yersinia enterocolitica*: the charisma continues. *Clin Micro Rev*, **10** (20): 257 – 276.

BOTTONE, E.J., 1999. *Yersinia enterocolitica*: overview and epidemiologic correlates. *Microbes Inf*, **1**: 323 – 330.

BRENDTSEN, J.D., NIELSEN, H., VON HEIJNE, G., and BRUNAK, S., 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340**: 783 – 795.

BRESOLIN, G., MORGAN, J.A., ILGEN, D., SCHERER, S., and FUCHS, T.M., 2006a. Low temperature-induced insecticidal activity of *Yersinia enterocolitica*. *Mol Microbiol*, **59** (2): 503 – 512.

BRESOLIN, G., NEUHAUS, K., SCHERER, S., and FUCHS, T.M., 2006b. Transcriptional analysis of long-term adaptation of *Yersinia enterocolitica* to low-temperature growth. *J Bact*, **188** (8): 2945 – 2958.

BRESOLIN, G., TRČEK, J., SCHERER, S., and FUCHS, T.M., 2008. Presence of a functional flagellar cluster Flag-2 and low-temperature expression of flagellar genes in *Yersinia enterocolitica* W22703. *Microbiology*, **154**: 196 – 206.

BRINKKÖTTER, A., KLÖß, H., ALPERT, C.-A., and LENGELER, J.W., 2000. Pathways for the utilization of N-acetyl-galactosamine and galactosamine in *Escherichia coli*. *Mol Microbiol*, **37** (1): 125 – 135.

BRUBAKER, R.R., 1991. Factors promoting acute and chronic disease caused by *Yersiniae*. *Clin Micro Rev*, **4** (3): 309 – 324.

BRUNETEAU, M., and MINKA, S., 2003. Lipopolysaccharides of bacterial pathogens from the genus *Yersinia*: a mini-review. *Biochimie*, **85**: 145 – 152.

BURROWS, T.W., and GILLETT, W.A., 1966. The nutritional requirements of some Pasteurella species. *J gen Microbiol*, **45**: 333 – 345.

CARNIEL, E., and MOLLARET, H.H., 1990. Yersiniosis. *Comp Immun Microbiol infect Dis*, **13** (2): 51 – 58.

CARNIEL, E., GUILVOUT, I., and PRENTICE, M., 1996. Characterization of a large chromosomal “High-Pathogenicity Island” in biotype 1B *Yersinia enterocolitica*. *JBact*, **178**(23): 6743 – 6751.

CARNIEL, E., 2002. Plasmids and pathogenicity islands of *Yersinia*. *Curr Top Microbiol Immunol*, **264** (1): 89 – 108.

CARNIEL, E., 2003. Evolution of pathogenic *Yersinia*, some lights in the dark. In: M. Skurnik, ed, 2003. *The Genus Yersinia*. New York, USA: Klywer Academic / Plenum Publishers, pp. 3 – 12.

CARVER, T.J., RUTHERFORD, K.M., BERRIMAN, M., RAJANDREAM, M.A., BARRELL, B.G., 2005. ACT: The Artemis Comparison Tool. *Bioinformatics*, **21**: 3422 – 3423.

CDC (CENTERS FOR DISEASE CONTROL AND PREVENTION), 2005. Information on plague [online]. Available at: http://www.cdc.gov/ncidod/dvbid/plague/info.htm [Accessed 23rd August 2011]

CDC (CENTERS FOR DISEASE CONTROL AND PREVENTION), 2009. Plague home page [online]. Available at: http://www.cdc.gov/ncidod/dvbid/plague/index.htm [Accessed 23rd August 2011]

CHAIN, P.S.G., CARNIEL, E., LARIMER, F.W., LAMERDIN, J., STOUTLAND, P.O., REGALA, W.M., GEORGESCU, A.M., VERGES, L.M., LAND, M.L., MOTIN, V.L., BRUBAKER, R.R., FOWLER, J., HINNEBUSCH, J., MARCEAU, M., MEDIGUE, C., SIMONET, M., CHENAL-FRANCISQUE, V., SOUA, B., DACHEUX, D., ELLIOTT, J.M., DERBISE, A., HAUSER, L.J., and GARCIA, E., 2004. Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *PNAS*, **101** (38): 13826 – 13831.

CHAIN, P.S., GRAFHAM, D.V., FULTON, R.S., FITZGERALD, M.G., HOSTETLER, J., MUZNY, D., ALI, J., BIRREN, B., BRUCE, D.C., BUHAY, C., COLE, J.R., DING, Y, DUGAN, S., FIELD, D., GARRITY, G.M., GIBBS, R., GRAVES, T., HAN, C.S., HARRISON, S.H., HIGHLANDER, S., HUGENHOLTZ, P., KHOURI, H.M., KODIRA, C.D., KOLDER, E., KYRPIDES, N.C., LANG, D., LAPIDUS, A., MALFATTI, S.A., MARKOWITZ, V., METHA, T., NELSON, K.E., PARKHILL, J., PITLUCK, S., QIN, X., READ, T.D., SCHMUTZ, J., SOZHAMANNAN, S., STERK, P., STRAUSBERG, R.L., SUTTON, G., THOMSON, N.R., TIEDJE, J.M., WEINSTOCK, G., WOLLAM, A., GENOMICS STANDARDS CONSORTIUM HUMAN MICROBIOME PROJECT JUMPSTART CONSORTIUM, DETTLER, J.C., 2009. Genomics. Genome project standards in a new era of sequencing. *Science* **326** (5950): 236 – 237.

CHAISSON, M.J., and PEVZNER, P.A., 2007. Short read fragment assembly of bacterial genomes. *Genome Research*, **18**: 324 – 330.

CHAISSON, M.J., BRINZA, D., and PEVZNER, P.A., 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Research*, **19**: 336 – 346.

CHAUDHURI, R.R., SEBAIHIA, M., HOBMAN, J.L., WEBBER, M.A., LEYTON, D., GOLDBERG, M.D., CUNNINGHAM, A.F., SCOTT-TUCKER, A., FERGUSON, P.R., THOMAS, C.M., FRANKEL, G., TANG, C.M., DUDLEY, E.G., ROBERTS, I.S., RASKO, D.A., PALLEN, M.J., PARKHILL, J., NATARO, J.P., THOMSON, N.R., and HENDERSON, I.R., 2010. Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative *Escherichia coli* strain 042. *PLoS One*, **5** (1): e8801.

CHEN, P.E., COOK, C., STEWART, A.C., NAGARAJAN, N., SOMMER, D.D., POP, M., THOMASON, B., THOMASON, M.P.K., LENTZ, S., NOLAN, N., SOZHAMANNAN, S., SULAKVELIDZE, A., MATECZUN, A., DU, L., ZWICK, M.E., and READ, T.D., 2010. Genomic characterization of the *Yersinia* genus. *Genome Biol*, **11**: R1.

CHENG, L., CONNOR, T.R., AAENENSEN, D.M., SPRATT, B.G., and CORANDER, J., 2011. Bayesian semi-supervised classification of bacterial samples using MLST databases. *BMC Bioinf*, **12**: 302.

CH'NG, S.L., OCTAVIA, S., XIA, Q., DUONG, A., TANAKA, M.M., FUKUSHIMA, H., and LAN, R., 2011. Population structure and evolution of pathogenicity of *Yersinia pseudotuberculosis*. *Appl Env Microbiol*, **77** (3): 768 – 775.

COLE, S.T., EIGLMEIER, K., PARKHILL, J., JAMES, K.D., THOMSON, N.R., WHEELER, P.R., HONORÉ, N., GARNIER, T., CHURCHER, C., HARRIS, D., MUNGALL, K., BASHAM, D., BROWN, D., CHILLINGWORTH, T., CONNOR, R., DAVIES, R.M., DEVLIN, K., DUTHOY, S., FELTWELL, T., FRASER, A., HAMLIN, N., HOLROYD, S., HORNSBY, T., JAGELS, K., LACROIX, C., MACLEAN, J., MOULE, S., MURPHY, L., OLIVER, K., QUAIL, M.A., RAJANDREAM, M.A., RUTHERFORD, K.M., RUTTER, S., SEEGER, K., SIMON, S., SIMMONDS, M., SKELTON, J., SQUARES, R., SQUARES, S., STEVENS, K., TAYLOR, K., WHITEHEAD, S., WOODWARD, J.R., and BARRELL, B.G., 2001. Massive gene decay in the leprosy bacillus. *Nature*, **409** (6823): 1007 – 1011.

COLLYN, F., BILLAULT, A., MULLET, C., SIMONET, M., and MARCEAU, M., 2004. YAPI, a new *Yersinia pseudotuberculosis* pathogenicity island. *Inf Immun*, **72** (8): 4784 – 4790.

CORNELIS, G., LUKE, R.K.J., and RICHMOND, M.H., 1978. Fermentation of raffinose by lactose-fermenting strains of *Yersinia enterocolitica* and by sucrose-fermenting strains of *Escherichia coli*. *J Clin Micro*, **7** (2): 180 – 183.

CORNELIS, G., LAROCHE, Y., BALLIGAND, G., SORY, M.-P., and WAUTERS, G., 1987. *Yersinia enterocolitica*, a primary model for bacterial invasiveness. *Rev Infec Dis*, **9** (1): 64 – 87.

CORNELIS, G.R., 1994. *Yersinia* pathogenicity factors. *Curr Top Microbiol Immunol*, **192**: 243 – 263.

CORNELIS, G.R., BOLAND, A., BOYD, A.P., GEUIJEN, C., IRIARTE, M., NEYT, C., SORY, M.-P., and STAINIER, I., 1998. The virulence plasmid of *Yersinia*, an antihost genome. *Microbiol Mol Biol Rev*, **62** (4): 1315 – 1352.

CORNELIS, G.R., 2002. *Yersinia* type III secretion: send in the effectors. *JCB*, **158** (3): 401 – 408.

CROUCHER, N.J., FOOKES, M.C., PERKINS, T.T., TURNER, D.J., MARGUERAT, S.B., KEANE, T., QUAIL, M.A., HE, M., ASSEFA, S., BÄHLER, J., KINGSLEY, R.A., PARKHILL, J., BENTLEY, S.D., DOUGAN, G., and THOMSON, N.R., 2009. A simple method for directional transcriptome sequencing using Illumina technology. *Nucleic Acids Research*, **37** (22): e148.

CROUCHER, N.J., HARRIS, S.R., FRASER, C., QUAIL, M.A., BURTON, J., VAN DER LINDEN, M., McGEE, L., VON GOTTBERG, A., SONG, J.H., KO, K.S., PICHON, B., BAKER, S., PARRY, C.M., LAMBERTSEN, L.M., SHAHINAS, D., PILLAI, D.R., MITCHELL, T.J., DOUGAN, G., TOMASZ, A., KLUGMAN, K.P., PARKHILL, J., HANAGE, W.P., and BENTLEY, S.D., 2011. Rapid pneumococcal evolution in response to clinical intervention. *Science*, **331**: 430 – 434.

DELCHER, A.L., BRATKE, K.A., POWERS, E.C., and SALZBERG, S.L., 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23** (6): 673 – 679.

DENG, W., BURLAND, V., PLUNKETT III, G., BOUTIN, A., MAYHEW, G.F., LISS, P., PERNA, N.T., ROSE, D.J., MAU, B., ZHOU, S., SCHWARTZ, D.C., FETHERSTON, J.D., LINDLER, L.E., BRUBAKER, R.R., PLANO, G.V., STRALEY, S.C., McDONOUGH, K.A., NILLES, M.L., MATSON, J.S., BLATTNER, F.R., and PERRY, R.D., 2002. Genome sequence of *Yersinia pestis* KIM. *JBact*, **184** (16): 4601 – 4611.

DE PAEPE, M., GABORIAU-ROUTHIAU, V., RAINTEAU, D., RAKOTOBE, S., TADDEI, F., and CERF-BENSUSSAN, N., 2011. Trade-off between bile resistance and nutritional competence drives *Escherichia coli* diversification in the mouse gut. *PLoS Genetics*, **7** (6): e1002107.

DRANCOURT, M., and RAOULT, D., 2002. *rpoB* gene sequence-based identification of *Staphylococcus* species. *J Clin Micro*, **40** (4): 1333 – 1338.

EBI, 2009. Velvet Sequence assembler for very short reads. [online]. Available at: http://www.ebi.ac.uk/~zerbino/velvet/ [Accessed 25 January 2011]

ELLISON, D.W., LAWRENCE, M.B., and MILLER, V.L., 2004. Invasin and beyond: regulation of *Yersinia* virulence by RovA. *Trends Microbiol*, **12** (6): 296 – 300.

ENGLESBERG, E., 1957a. Mutation to rhamnose utilization in *Pasteurella pestis*. *J Bact*, **73** (5): 641 – 648.

ENGLESBERG, E., 1957b. Physiological basis for rhamnose utilization by a mutant of *Pasteurella pestis*. I. Experiments with resting cells; the isolation of lactic aldehyde. *J Bact*, **74** (1): 8 – 11.

EPPINGER, M., ROSOVITZ, M.J., FRICKE, W.F., RASKO, D.A., KOKORINA, G., FAYOLLE, C., LINDLER, L.E., CARNIEL, E., and RAVEL, J., 2007. The complete genome sequence of *Yersinia pseudotuberculosis* IP31758, the causative agent of far east scarlet-like fever. *PLoS Genetics*, **3** (8): e142.

EPPINGER, M., WORSHAM, P.L., NIKOLICH, M.P., RILEY, D.R., SEBASTIAN, Y., MOU, S., ACHTMAN, M., LINDLER, L.E., and RAVEL, J., 2010. Genome sequence of the deep-rooted *Yersinia pestis* strain Angola reveals new insights into the evolution and pangenome of the plague bacterium. *J Bact*, **192** (6): 1685 – 1699.

FANTASIA MAZZOTTI, M., GIRALDI, V., PUGINA, P., MINGRONE, M.G., and BENZI, G., 1985. Survey of some biochemical characteristics of *Yersinia sp.* Strains. *Microbiologica*, **8** (1): 43 – 50.

FANTASIA MAZZOTTI, M., MINGRONE, M.G., and MARTINI, A., 1987. Fermenting ability of *Yersinia spp.* strains on esculin, L-fucose and D-arabitol. *Microbiologica*, **10** (4): 413 – 416.

FARMER III, J.J., CARTER, G.P., MILLER, V.L., FALKOW, S., and WACHSMUTH, I.K., 1992. Pyrazinamidase, CR-MOX agar, salicin fermentation-esculin hydrolysis, and D-xylose fermentation for identifying pathogenic serotypes of *Yersinia enterocolitica*. *J Clin Micro*, **30** (10): 2589 – 2594.

FEARNLEY, C., ON, S.L.W., KOKOTOVIC, B., MANNING, G., CHEASTY, T., and NEWELL, D.G., 2005. Application of fluorescent amplified fragment length polymorphism for comparison of human and animal isolates of *Yersinia enterocolitica*. *Appl Env Microbiol*, **71** (9): 4960 – 4965.

FERNÁNDEZ, L., MÉNDEZ, J., and GUIJARRO, J.A., 2007. Molecular virulence mechanisms of the fish pathogen *Yersinia ruckeri*. *Vet Microbiol*, **125**: 1 – 10.

FINN, R.D., MISTRY, J., TATE, J., COGGILL, P., HEGER, A., POLLINGTON, J.E., GAVIN, O.L., GUNASEKARAN, P., CERIC, G., FORSLUND, K., HOLM, L., SONNHAMMER, E.L.L., EDDY, S.R., and BATEMAN, A., 2010. The Pfam protein families database. *Nucl Acids Res*, **38** (1): D211 – D222.

FODOR, I.K., HOLTZ-MORRIS, A.E., and McCUTCHEN-MALONEY, S.L., 2005. Growth curve models for the analysis of phenotype arrays for a systems biology overview of *Yersinia pestis*. *Joint Statistical Meetings*. Minneapolis, MN, United States, 6 – 10 August 2005. Available from: https://e-reports-ext.llnl.gov/pdf/324837.pdf [accessed 13 July 2011].

FOULTIER, B., TROISFONTAINES, P., MÜLLER, S., OPPERDOES, F.R., and CORNELIS, G.R., 2002. Characterization of the *ysa* pathogenicity locus in the chromosome of *Yersinia enterocolitica* and phylogeny analysis of the type III secretion systems. *J Mol Evol*, **55**: 37 – 51.

FOULTIER, B., and CORNELIS, G.R., 2003. DNA sequence and analysis of the pYVa127/90 virulence plasmid of *Yersinia enterocolitica* strain A127/90. *Res Microbiol*, **154** (8): 553 – 557.

FRASER-LIGGETT, C.M., 2005. Insights on biology and evolution from microbial genome sequencing. *Genome Res*, **15**: 1603 – 1610.

FRECH, G., and SCHWARZ, S., 1998. Tetracycline resistance in *Salmonella enterica* subsp. *enterica* serovar Dublin. *Antimicrob Agents Chemother*, **42** (5): 1288 – 1289.

FUCHS, T.M, BRESOLIN, G., MARCINOWSKI, L., SCHACHTNER, J., and SCHERER, S. , 2008. Insecticidal genes of *Yersinia* spp.: taxonomical distribution, contribution to toxicity towards *Manduca sexta* and *Galleria mellonella*, and evolution. *BMC Microbiology*, **8**: 214.

FUCHS, T.M., BRANDT, K., STARKE, M., and RATTEI, T., 2011. Shotgun sequencing of *Yersinia enterocolitica* strain W22703 (biotype 2, serotype O:9): genomic evidence for oscillation between invertebrates and mammals. *BMC Genomics*, **12**: 168.

FUKUSHIMA, H., GOMYODA, M., ALEKSIC, S., and TSUBOKURA, M., 1993. Differentiation of *Yersinia enterocolitica* serotype O:5,27 strains by phenotypic and molecular techniques. *J Clin Micro*, **31** (6): 1672 – 1674.

FUKUSHIMA, H., MATSUDA, Y., SEKI, R., TSUBOKURA, M., TAKEDA, N., SHUBIN, F.N., PAIK, I.K., and ZHENG, X.B., 2001. Geographical heterogeneity between Far Eastern and Western countries in prevalence of the virulence plasmid, the superantigen *Yersinia pseudotuberculosis*-derived mitogen, and the high-pathogenicity island among *Y. pseudotuberculosis* strains. *J Clin Micro*, **39** (10): 3541 – 3547.

GALTIER, N., GOUY, M., and GAUTIER, C., 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequencing alignment and molecular phylogeny. *Comput Appl Biosci*, **12**: 543 – 548.

GARCIA, J.A., DOMINGUEZ, L., LARSEN, J.L., and PEDERSEN, K., 1988. Ribotyping and plasmid profiling of *Yersinia ruckeri*. *J Appl Microbiol*, **85**: 949 – 955.

GARDNER, J., DAUB, J., TATE, J., MOORE, B.L., OSUCH, I.H., GRIFFITHS-JONES, S., FINN, R.D., NAWROCKI, E.P., KOLBE, D.L., EDDY, S.R., and BATEMAN, A., 2011. Rfam: Wikipedia, clans, and the „decimal“ release. *Nucleic Acids Research* [doi: 10.1093/nar/gkq1129]

GHEBREMEDHIN, B., LAYER, F., KÖNIG, W., and KÖNIG, B., 2008. Genetic classification and distinguishing of *Staphylococcus* species based on different partial *gap*, 16S rRNA, *hsp60*, *rpoB*, *sodA*, and *tuf* gene sequences. *J Clin Micro*, **46** (3): 1019 – 1025.

GORSHKOVA, R.P., KALMYKOVA, E.N., ISAKOV, V.V., and OVODOV, Y.S., 1986. Structural studies on O-specific polysaccharides of lipopolysaccharides from *Yersinia enterocolitica* serovars O:5 and O:5,27. *Eur J Biochem*, **156**: 391 – 397.

GOLDBERG, S.M.D., JOHNSON, J., BUSAM, D., FELDBLYUM, T., FERRIERA, S., FRIEDMAN, R., HALPERN, A., KHOURI, H., KRAVITZ, S.A., LAURO, F.M., LI, K., ROGERS, Y.-H., STRAUSBERG, R., SUTTON, G., TALLON, L., THOMAS, T., VENTER, E., FRAZIER, M., and VENTER, J.C., 2006. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *PNAS*, **103** (30): 11240 – 11245.

GOLUBOV, A., HEESEMANN, J., and RAKIN, A., 2003. Uncovering genomic differences in human pathogenic *Yersinia enterocolitica*. *FEMS Immunol Med Microbiol*, **38**: 107 – 111.

GOUY, M., GUINDON, S., and GASCUEL, O., 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, **27** (2): 221 – 224.

GRANT, T., BENNETT-WOOD, V., and ROBINS-BROWNE, R.M., 1998. Identification of virulence-associated characteristics in clinical isolates of *Yersinia enterocolitica* lacking classical virulence markers. *Inf Immun*, **66**(3): 1113 – 1120.

GRIFFITHS-JONES, S., BATEMAN, A., MARSHALL, M., KHANNA, A., and EDDY, S.R., 2003. Rfam: an RNA family database. *Nucleic Acids Research*, **31** (1): 439 – 441.

GUIYOULE, A., GUINET, F., MARTIN, L., BENOIT, C., DESPLACES, N., and CARNIEL, E., 1998. Phenotypic and genotypic characterization of virulent *Yersinia enterocolitica* strain unable to ferment sucrose. *J Clin Micro*, **36** (9): 2732 – 2734.

GUILVOUT, I., QUILICI, M.L., RABOT, S., LESEL, R., and MAZIGH, D., 1988. *Bam*HI restriction endonuclease analysis of *Yersinia ruckeri* plasmids and their relatedness to the genus *Yersinia* 42- to 47-Megadalton plasmid. *Appl Env Microbiol*, **54** (10): 2594 – 2597.

HALL, N.,, 2007. Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol*, **210** (9): 1518 – 1525.

HALLER, J.C., CARLSON, S., PEDERSON, K.J., and PIERSON, D.E., 2000. A chromosomally encoded type III secretion pathway in *Yersinia enterocolitica* is important in virulence. *Mol Microbiol*, **36**(6): 1436 – 1446.

HAMMERL, J.A., KLEIN, I, LANKA, E., APPEL, B., and HERTWIG, S., 2008. Genetic and functional properties of the self-transmissible *Yersinia enterocolitica* plasmid pYE854, which mobilizes the virulence plasmid pYV. *J Bact*, **190** (3): 991 – 1010.

HARISMENDY, O., NG, P.C., STRAUSBERG, R.L., WANG, X., STOCKWELL, T.B., BEESON, K.Y., SCHORK, N.J., MURRAY, S.S., TOPOL, E.J., LEVY, S., and FRAZER, K.A., 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, **10**: R32.

HARRIS, S.R., FEIL, E.J., HOLDEN, M.T.G., QUAIL, M.A., NICKERSON, E.K., CHANTRATITA, N., GARDETE, S., TAVARES, A., DAY, N., LINDSAY, J.A., EDGEWORTH, J.D., DE LENCASTRE, H., PARKHILL, J., PEACOCK, S.J., and BENTLEY, S.D., 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science*, **327** (5964): 469 – 474.

HENSEL, M., 2000. *Salmonella* pathogenicity island 2. *Mol Microbiol*, **36** (5): 1015 – 1023.

HEESEMANN, J., KELLER, C., MORAWA, R., SCHMIDT, N., SIEMENS, H.J., and LAUFS, R., 1983. Plasmids of human strains of *Yersinia enterocolitica*: molecular relatedness and possible importance for pathogenesis. *J Infect Dis*, **147** (1): 107 – 115.

HERTWIG, S., KLEIN, I., HAMMERL, J.A., and APPEL, B., 2003. Characterization of two conjugative *Yersinia* plasmids mobilizing pYV. In: M. Skurnik, ed, 2003. *The Genus Yersinia*. New York, USA: Klywer Academic / Plenum Publishers, pp. 38 – 38.

HEUEL, H., TURGUT, S., SCHMID, K., and LENGELER, J.W., 1997. Substrate recognition domains as revealed by active hybrids between the D-arabinitol and ribitol transporters from *Klebsiella pneumoniae. J Bact*, **179** (19): 6014 – 6019.

HINNEBUSCH, B.J., RUDOLPH, A.E., CHEREPANOV, P., DIXON, J.E., SCHWAN, T.G., and FORSBERG, Å, 2002. Role of Yersinia murine toxin in survival of *Yersinia pestis* in the midgut of the flea vector. *Science*, **296** (5568): 733 – 735.

HOLTZ-MORRIS, A., 2005. Phenotype array analysis of metabolic differences in *Yersinia pestis*. *The American Society for Microbiology's 105th General Meeting*. Atlanta, GA, United States, 5 – 9 June 2005. Available from: https://e-reports-ext.llnl.gov/pdf/320376.pdf [accessed 13 July 2011].

HOMANN, O.R., CAI, H., BECKER, J.M., and LINDQUIST, S.L., 2005. Harnessing natural diversity to probe metabolic pathways. *PLoS Genetics*, **1** (6): e80.

HOWARD, S.L., GAUNT, M.W., HINDS, J., WITNEY, A.A., STABLER, R., and WREN, B.W., 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.

HULO, N., BAIROCH, A., BULLIARD, V., CERUTTI, L., DE CASTRO, E., LANGENDIJK-GENEVAUX, P.S., PAGNI, M., and SIGRIST, C.J.A., 2006. The PROSITE database. *Nucl Acids Res*, **34**: D227 – D230.

HURST, M.R.H., BECHER, S.A., YOUNG, S.D., NELSON, T.L., and GLARE, T.R., 2011. *Yersinia entomophaga* sp. nov. isolated from New Zealand grass grub *Costelytra zealandica*. *IJSEM*, **61** (4): 844 – 849.

IBRAHIM, A., LIESACK, W., and STACKEBRANDT, E., 1992. Polymerase chain reaction-gene probe detection system specific for pathogenic strains of *Yersinia enterocolitica*. *J Clin Micro*, **30** (8): 1942 – 1947.

IBRAHIM, A., GOEBEL, B.M., LIESACK, W., GRIFFITHS, M., and STACKEBRANDT, E., 1993. The phylogeny of the genus *Yersinia* based on 16S rDNA sequences. *FEMS Microbiology Letters*, **114**: 173 – 178.

IBRAHIM, A., LIESACK, W., STEIGERWALT, A.G., BRENNER, D.J., STACKEBRANDT, E., and ROBINS-BROWNE, R.M., 1997. A cluster of atypical *Yersinia* strains with a distinctive 16S rRNA signature. *FEMS Microbiology Letters*, **146**: 73 – 78.

ILLUMINA TECHNOLOGY, 2011. Available at: http://www.illumina.com/technology/mate_pair_sequencing_assay.ilmn; http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn; http://www.illumina.com/documents/products/datasheets/datasheet_sequencing_multiplex.pdf

IRIARTE, M., and CORNELIS, G.R., 1995. MyfF, an element of the network regulating the synthesis of fibrillae in *Yersinia enterocolitica*. *J Bact*, **177** (3): 738 – 744.

IWOBI, A., HEESEMANN, J., GARCIA, E., IGWE, E., NOELTING, C., and RAKIN, A., 2003. Novel virulence-associated type II secretion system unique to high-pathogenicity *Yersinia enterocolitica*. *Inf Immun*, **71**(4): 1872 – 1879.

JANDA, J.M., and ABBOTT, S.L., 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Micro*, **45** (9): 2761 – 2764.

JOHNSON, D.A., TETU, S.G., PHILLIPPY, K., CHEN, J., REN, Q., and PAULSEN, I.T., 2008. High-throughput phenotypic characterization of *Pseudomonas aeruginosa* membrane transport genes. *PLoS Genetics*, **4** (10): e1000211.

JOURNET, L., AGRAIN, C., BROZ, P., and CORNELIS, G.R., 2003. The needle length of bacterial injectisomes is determined by a molecular ruler. *Science*, **302** (5651): 1757 – 1760.

KADO, C.I., and LIU, S.-T., 1981. Rapid procedure for detection and isolation of large and small plasmids. *J Bact*, **145** (3): 1365 – 1373.

KAPPERUD, G., and NESBAKKEN, T., 1987. Restriction endonuclease analysis of 40- to 50-Mdalton plasmids from *Yersinia enterocolitica* strains of worldwide origin. *Contrib Microbiol Immunol*, **9**: 317 – 323.

KOTETISHVILI, M., KREGER, A., WAUTERS, G., MORRIS, J.G. Jr., SULAKVELIDZE, A., and STINE, O.C., 2005. Multilocus sequence typing for studying genetic relationships among *Yersinia* species. *J Clin Micro*, **43** (6): 2674 – 2684.

KROGH, A., LARSSON, B., VON HEIJNE, G., and SONNHAMMER, E.L.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol*, **305** (3): 567 – 580.

KUO, C.-H., and OCHMAN, H., 2010. The extinction dynamics of bacterial pseudogenes. *PLoS Genetics*, **6** (8): e1001050.

LANGRIDGE, G.C., PHAN, M.-D., TURNER, D.J., PERKINS, T.T., PARTS, L., HAASE, J., CHARLES, I., MASKELL, D.J., PETERS, S.E., DOUGAN, G., WAIN, J., PARKHILL, J., and TURNER, A.K., 2009. Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. *Genome Res*, **19**: 2308 – 2316.

LARKIN, M.A., BLACKSHIELDS, G., BROWN, N.P., CHENNA, R., McGETTIGAN, P.A., McWILLIAM, H., VALENTIN, F., WALLACE, I.M., WILM, A., LOPEZ, R., THOMPSON, J.D., GIBSON, T.J., and HIGGINS, D.G., 2007. Clustal W and Clustal X version 2.0. *Bioinf*, **23** (21): 2947 – 2948.

LAROCHE, Y., VAN BOUCHAUTE, M., and CORNELIS, G., 1984. A restriction map of virulence plasmid pYVE439-80 from a serogroup 9 *Yersinia enterocolitica* strain. *Plasmid*, **12** (1): 67 – 70.

LEPKA, D., KERRINNES, T., SKIEBE, E., HAHN, B., FRUTH, A., and WILHARM, G., 2009. Adding to *Yersinia enterocolitica* gene pool diversity: two cryptic plasmids from a biotype 1A isolate. *J Biomed Biotechnol*, **2009**: 398434.

LESEL, R., LESEL, M., GAVINI, F., and VUILLAUME, A., 1983. Outbreak of enteric redmouth disease in rainbow trout, *Salmo gairdneri* Richardson, in France. *J Fish Dis*, **6**: 385 – 387.

LEWIN, A., STRAUCH, E., HERTWIG, S., HOFFMANN, B., NATTERMANN, H., and APPEL, B., 1996. Comparison of plasmids of strains of *Yersinia enterocolitica* biovar 1A with the virulence plasmid of a pathogenic *Y. enterocolitica* strain. *Zentralblatt Bakt*, **285** (1): 52 – 63.

LI, H., and DURBIN, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**: 1754 – 1760.

MAHAN, M.J., SLAUCH, J.M., and MEKALANOS, J.J., 1993. Selection of bacterial virulence genes that are specifically induced in host tissues. *Science*, **259** (5095): 686 – 688.

MAIDEN, M.C.J., BYGRAVES, J.A., FEIL, E., MORELLI, G., RUSSELL, J.E., URWIN, R., ZHANG, Q., ZHOU, J., ZURTH, K., CAUGANT, D.A., FEAVERS, E.M., ACHTMAN, M., and SPRATT, B.G., 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *PNAS,* **95** (6): 3140 – 3145.

MARCEAU, M., 2004. Transcriptional regulation in *Yersinia*: an update. In: E. Carniel, B.J. Hinnebusch, eds, 2004. *Yersinia – Molecular and cellular biology*. Norfolk, UK: Horizon Bioscience, pp. 109 – 148.

MARDIS, E.R., 2008a. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24** (3): 133 – 141.

MARDIS, E.R., 2008b. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, **9**: 287 – 402.

MARGULIES, M., EGHOLM, M., ALTMAN, W.E., ATTIYA, S., BADER, J.S., BEMBEN, L.A., BERKA, J., BRAVERMAN, M.S., CHEN, Y.J., CHEN, Z., DEWELL, S.B., DU, L., FIERRO, J.M., GOMES, X.V., GODWIN, B.C., HE, W., HELGESEN, S., HO, C.H., IRZYK, G.P., JANDO, S.C., ALENQUER, M.L., JARVIE, T.P., JIRAGE, K.B., KIM, J.B., KNIGHT, J.R., LANZA, J.R., LEAMON, J.H., LEFKOWITZ, S.M., LEI, M., LI, J., LOHMAN, K.L., LU, H., MAKHIJANI, V.B., MCDADE, K.E., MCKENNA, M.P., MYERS, E.W., NICKERSON, E., NOBILE, J.R., PLANT, R., PUC, B.P., RONAN, M.T., ROTH, G.T., SARKIS, G.J., SIMONS, J.F., SIMPSON, J.W., SRINIVASAN, M., TARTARO, K.R., TOMASZ, A., VOGT, K.A., VOLKMER, G.A., WANG, S.H., WANG, Y., WEINER, M.P., YU, P., BEGLEY, R.F., and ROTHBERG, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437** (7057): 376 – 380.

MARTIN, L., LECLERCQ, A., SAVIN, C., and CARNIEL, E., 2009. Characterization of atypical isolates of *Yersinia intermedia* and definition of two new biotypes. *J Clin Micro*, **47** (8): 2377 – 2380.

McNALLY, A., CHEASTY, T., FEARNLEY, C., DALZIEL, R.W., PAIBA, G.A., MANNING, G., and NEWELL, D.G., 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. *Lett Appl Microbiol*, **39**: 103 – 108.

McNALLY, A., DALTON, T., LA RAGIONE, R.M., STAPLETON, K., MANNING, G., and NEWELL, D.G., 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.

McNALLY, A., LA RAGIONE, R.M., BEST, A., MANNING, G., and NEWELL, D.G., 2007. An aflagellate mutant *Yersinia enterocolitica* biotype 1A strain displays altered invasion of epithelial cells, persistence in macrophages, and cytokine secretion profiles *in vitro*. *Microbiology*, **153**: 1339 – 1349.

McQUISTON, J.R., FIELDS, P.I., TAUXE, R.V., and LOGSDON JR, J.M., 2008. Do *Salmonella* carry spare tyres? *Trends in Microbiology*, **16** (4): 142 – 148.

MERHEJ, V., ADÉKAMBI, T., PAGNIER, I., RAOULT, D., and DRANCOURT, M., 2008. *Yersinia massiliensis* sp. nov., isolated from fresh water. *IJSEM*, **58** (4): 779 – 784.

METZKER, M.L., 2010. Sequencing technologies – the next generation. *Nature Reviews Genetics*, **11**: 31 – 46.

MILLER, V.L., and FALKOW, S., 1988. Evidence for two genetic loci in *Yersinia enterocolitica* that can promote invasion of epithelial cells. *Inf Immun*, **56**(5): 1242 – 1248.

MILLER, V.L., FARMER III, J.J., HILL, W.E., and FALKOW, S., 1989. The *ail* locus is found uniquely in *Yersinia enterocolitica* serotypes commonly associated with disease. *Inf Immun*, **57**(1): 121 – 131.

MILLER, J.M., and RHODEN, D.L., 1991. Preliminary evaluation of Biolog, a carbon source utilization method for bacterial identification. *J Clin Micro*, **29** (6): 1143 – 1147.

MORELLI, G., SONG, Y., MAZZONI, C.J., EPPINGER, M., ROUMAGNAC, P., WAGNER, D.M., FELDKAMP, M., KUSECEK, B., VOGLER, A.J., LI, Y., CUI, Y., THOMSON, M.R., JOMBART, T., LEBLOIS, R., LICHTNER, P., RAHALISON, L., PETERSEN, J.M., BALLOUX, F., KEIM, P., WIRTH, T., RAVEL, J., YANG, R., CARNIEL, E., and ACHTMAN, M., 2010. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genetics*, **42** (12): 1140 – 1143.

MOTIN, V.L., GEORGESCU, A.M., FITCH, J.P., GU, P.P., NELSON, D.O., MABERY, S.L., GARNHAM, J.B., SOKHANSANJ, B.A., OTT, L.L., COLEMAN, M.A., ELLIOTT, J.M., KEGELMEYER, L.M., WYROBEK, A.J., SLEZAK, T.R., BRUBAKER, R.R., and GARCIA, E., 2004. Temporal global changes in gene expression during temperature transition in *Yersinia pestis*. *J Bact*, **186** (18): 6298 – 6305.

MÜLLER-LOENNIES, S., RUND, S., ERVELÄ, E., SKURNIK, M., and HOLST, O., 1999. The structure of the carbohydrate backbone of the core-lipid A region of the

lipopolysaccharide from a clinical isolate of *Yersinia enterocolitica* O:9. *Eur J Biochem*, **261**: 19 – 24.

MURROS-KONTIAINEN, A., FREDRIKSSON-AHOMAA, M., KORKEALA, H., JOHANSSON, P., RAHKILA, R., and BJÖRKROTH, J., 2010a. *Yersinia nurmii* sp. nov. *IJSEM* [doi:10.1099/ijs.0.024836-0].

MURROS-KONTIAINEN, A., JOHANSSON, P., NISKANEN, T., FREDRIKSSON-AHOMAA, M., KORKEALA, H., and BJÖRKROTH, J., 2010b. *Yersinia pekkanenii* sp. nov. *IJSEM* [doi:10.0199/ijs.0019984-0]

NAKAJIMA, H., INOUE, M., MORI, T., ITOH, K.-I., ARAKAWA, E., and WATANABE, H., 1992. Detection and identification of *Yersinia pseudotuberculosis* and pathogenic *Yersinia enterocolitica* by an improved polymerase chain reaction method. *J Clin Micro*, **30** (9): 2484 – 2486.

NAVID, A., and ALMAAS, E., 2009. Genome-scale reconstruction of the metabolic network in *Yersinia pestis*, strai 91001. *Mol BioSyst*, **5**: 368 – 375.

NELSON, K.M., YOUNG, G.M., and MILLER, V.L., 2001. Identification of a locus involved in systemic dissemination of *Yersinia enterocolitica*. *Inf Immun*, **69**(10): 6201 – 6208.

NEUBAUER, H., SAUER, T., BECKER, H., ALEKSIC, S., and MEYER, H., 1998. Comparison of systems for identification and differentiation of species within the genus *Yersinia*. *J Clin Micro*, **36** (11): 3366 – 3368.

NEUBAUER, H., HENSEL, A., ALEKSIC, S., and MEYER, H., 2000. Identification of *Yersinia enterocolitica* within the genus *Yersinia*. *System Appl Microbiol*, **23**: 58 – 62.

NEUBAUER, H., ALEKSIC, S., HENSEL, A., FINKE, E.-J., and MEYER, H., 2000. *Yersinia enterocolitica* 16S rRNA gene types belong to the same genospecies but form three homology groups. *Int J Med Microbiol*, **290** (1): 61 – 64.

NEUBAUER, H.K.J., and SPRAGUE, L.D., 2003. Epidemiology and diagnostics of *Yersinia*-infections. In: M. Skurnik, ed, 2003. *The Genus Yersinia*. New York, USA: Klywer Academic / Plenum Publishers, pp. 431 – 438.

NEYT, C., IRIARTE, M., THI, V.H., and CORNELIS, G.R., 1997. Virulence and arsenic resistance in Yersiniae. *J Bact*, **179** (3): 612 – 619.

NIELSEN, H., ENGELBRECHT, J., BRUNAK, S., and VON HEIJNE, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering,***10**: 1 – 6.

NING, Z., COX, A.J., and MULLIKIN, J.C., 2001. SSAHA: a fast search method for large DNA databases. *Genome Res*, **11**: 1725 – 1729.

OBERHETTINGER, P., SCHÜTZ, M., RADDATZ, G., KELLER, H., and AUTENRIETH, I.B., 2011. The sequence of the pYV virulence plasmid from

Yersinia enterocolitica strain WA-314 biogroup 1B serotype O:8. *Plasmid*, **65** (1): 20 – 24.

OERTELT, C., LINDNER, B., SKURNIK, M., and HOLST, O., 2001. Isolation and structural characterization of an R-form lipopolysaccharide from *Yersinia enterocolitica* serotype O:8. *Eur J Biochem*, **268**: 554 – 564.

OTTO, T.D., SANDERS, M., BERRIMAN, M., and NEWBOLD, C., 2010. Iterative correction of reference nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26** (14): 1704 – 1707.

PARKHILL, J., WREN, B.W., THOMSON, N.R., TITBALL, R.W., HOLDEN, M.T.G., PRENTICE, M.B., SEBAIHA, M., JAMES, K.D., CURCHER, C., MUNGALL, K.L., BAKER, S., BASHAM, D., BENTELY, S.D., BROOKS, K., CERDEÑO-TÁRRAGA, A.M., CHILLINGWORTH, T., CRONIN, A., DAVIES, R.M., DAVIS, P., DOUGAN, G., FELTWELL, T., HAMLIN, N., HOLROYD, S., JAGELS, K., KARLYSHEV, A.V., LEATHER, S., MOULE, S., OYSTON, P.C.F., QUAIL, M., RUTHERFORD, K., SIMMONDS, M., SKELTON, J., STEVENS, K., WHITEHEAD, S., and BARRELL, B.G., 2001a. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, **413** (6855): 523 – 527.

PARKHILL, J., DOUGAN, G., JAMES, K.D., THOMSON, N.R., PICKARD, D., WAIN, J., CHURCHER, C., MUNGALL, K.L., BENTLEY, S.D., HOLDEN, M.T., SEBAIHIA, M., BAKER, S., BASHAM, D., BROOKS, K., CHILLINGWORTH, T., CONNERTON, P., CRONIN, A., DAVIS, P., DAVIES, R.M., DOWD, L., WHITE, N., FARRAR, J., FELTWELL, T., HAMLIN, N., HAQUE, A., HIEN, T.T., HOLROYD, S., JAGELS, K., KROGH, A., LARSEN, T.S., LEATHER, S., MOULE, S., O'GAORA, P., PARRY, C., QUAIL, M., RUTHERFORD, K., SIMMONDS, M., SKELTON, J., STEVENS, K., WHITEHEAD, S., and BARRELL, B.G., 2001b. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature*, **413** (6858): 848 – 852.

PAYNE, S.H., HUANG, S.-T., and PIEPER, R., 2010. A proteogenomic update to Yersinia: enhancing genome annotation. *BMC Genomics*, **11**: 460.

PEDERSEN, K.B., WINBLAD, S., and BITSCH, V., 1979. Studies on the interaction between different O-serotypes of *Yersinia enterocolitica* and HeLa cells. *Act path microbiol scand Sect B*, **87**: 141 – 145.

PEPE, J.C., and MILLER, V.L., 1993. *Yersinia enterocolitica* invasin: a primary role in the initiation of infection. *PNAS*, **90** (14): 6473 – 6477.

PERRY, R.D., and FETHERSTON, J.D., 1997. *Yersinia pestis* – etiologic agent of plague. *Clin Microbiol Rev*, **10** (1): 35 – 66.

PIERSON, D.E., and FALKOW, S., 1990. Non-pathogenic isolates of *Yersinia enterocolitica* do not contain functional *inv*-homologous sequences. *Inf Immun*, **58** (4): 1059 – 1064.

PIERSON, D.E., and FALKOW, S., 1993. The *ail* gene of *Yersinia enterocolitica* has a role in the ability of the organism to survive serum killing. *Inf Immun*, **61** (5): 1846 – 1852.

POP, M., and SALZBERG, S.L., 2008. Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, **24** (3): 142 – 149.

PORTNOY, D.A., MOSELEY, S.L, and FALKOW, S., 1981. Characterization of plasmids and plasmid-associated determinants of *Yersinia enterocolitica* pathogenesis. *Inf Immun*, **31** (2): 775 – 782.

PORTNOY, D.A., and FALKOW, S., 1981. Virulence-associated plasmids form *Yersinia enterocolitica* and *Y. pestis*. *J Bact*, **148** (3): 877 – 883.

PORTNOY, D.A., WOLF-WATZ, H., BOLIN, I., BEEDER, A.B., and FALKOW, S., 1984. Characterization of common virulence plasmids in *Yersinia* species and their role in the expression of outer membrane proteins. *Inf Immun*, **43** (1): 108 – 114.

PRIOR, J.L., HITCHEN, P.G., WILLIAMSON, E.D., REASON, A.J., MORRIS, H.R., DELL, A., WREN, B.W., and TITBALL, R.W., 2001. Characterization of the lipopolysaccharide of *Yersinia pestis*. *Microbial Pathogenesis*, **30**: 49 – 57.

PUJOL, C., and BLISKA, J.B., 2003. The ability to replicate in macrophages is conserved between *Yersinia pestis* and *Yersinia pseudotuberculosis*. *Inf Immun*, **71** (10): 5892 – 5899.

REEVES, P.R., HOBBS, M., VALVANO, M.A., SKURNIK, M., WHITFIELD, C., COPLIN, D., KIDO, N., KLENA, J., MASKELL, D., RAETZ, C.R.H., and RICK, P.D., 1996a. Bacterial polysaccharide synthesis and gene nomenclature. *Trends in Microbiology*, **4** (12): 495 – 503.

REEVES, P.R., HOBBS, M., VALVANO, M.A., SKURNIK, M., WHITFIELS, C., COPLIN, D., KIDO, N., KLENA, J., MASKELL, D., RAETZ, C.R.H., and RICK, P.D., 1996b. Application of a new nomenclature for bacterial surface polysaccharide genes. Available at: http://sydney.edu.au/science/molecular_bioscience/BPGD/big_paper.pdf [Accessed 21 June 2011]

REIZER, J., RAMSEIER, T.M., REIZER, A., CHARBIT, A., and SAIER JR., M.H., 1996. Novel phosphotransferase genes revealed by bacterial genome sequencing: a gene cluster encoding a putative N-acetylgalactosamine metabolic pathway in *E. coli*. *Microbiology*, **142**: 231 – 250.

REVELL, P.A., and MILLER, V.L., 2001. *Yersinia* virulence: more than a plasmid. *FEMS Micro Lett*, **205** (2): 159 – 164.

ROBINS-BROWNE, R.M., MILIOTIS, M.D., CIANCIOSI, S., MILLER, V.L., FALKOW, S., and MORRIS, G. JR., 1989. Evaluation of DNA colony hybridization and other techniques for detection of virulence in *Yersinia* species. *J Clin Micro*, **27** (4): 644 – 650.

ROCHE, 2011. 454 Sequencing – Products & Solutions. [online] Available at: http://www.454.com/products-solutions/analysis-tools/gs-de-novo-assembler.asp [Accessed 25 January 2011]

ROGGENKAMP, A., GEIGER, A.M., LEITRITZ, L., KESSLER, A., and HEESEMANN, J., 1997. Passive immunity to infection with *Yersinia* spp. Mediated by anti-recombinant V antigen is dependent on polymorphism of V antigen. *Inf Immun*, **65** (2): 446 – 451.

ROHMER, L., HOCQUET, D., and MILLER, S.I., 2011. Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends Microbiol*, **19** (7): 341 – 348.

RYCKAERT, J., BOSSIER, P., D'HERDE, K., DIEZ-FRAILE, A., SORGELOOS, P., HAESEBROUCK, F., and PASMANS, F., 2010. Persistance of *Yersinia ruckeri* in trout macrophages. *Fish Shellfish Immunol*, **29**: 648 – 655.

SALZBERG, S.L., DELCHER, A.L., KASIF, S., and WHITE, O., 1998. Microbial gene identification using interpolated Markov models. *Nucl Acids Res*, **26** (2): 544 – 548.

SAMUEL, G., and REEVES, P., 2003. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. *Carbohydrate Research*, **338**: 2503 – 2519.

SANGER, F., NICKLEN, S., and COULSON, A.R., 1977. DNA sequencing with chain terminating inhibitors. *PNAS*, **74** (12): 5463 – 5467.

SANGER, F., 1988. Sequences, sequences, and sequences. *Ann Rev Biochem*, **57**: 1 – 28.

SCHIEMANN, D.A., and DEVENISH, J.A., 1982. Relationship of HeLa cell infectivity to biochemical, serological, and virulence characteristics of *Yersinia enterocolitica*. *Inf Immun*, **35** (2): 497 – 506.

SCHOULS, L.M., SCHOT, C.S., and JACOBS, J.A., 2003. Horizontal transfer of segments of the 16S rRNA genes between species of the *Streptococcus anginosus* group. *J Bact*, **185** (24): 7241 – 7246.

SCHUBERT, S., FISCHER, D., and HEESEMANN, J., 1999. Ferric enterochelin transport in *Yersinia enterocolitica*: Molecular and evolutionary aspects. *JBact*, **181** (20): 6387 – 6395.

SHEA, J.E., HENSEL, M., GLEESON, C., and HOLDEN, D.W., 1996. Identification of a virulence locus encoding a second type III secretion system in *Salmonella typhimurium*. *PNAS*, **93**: 2593 – 2597.

SHEHEE, M.W., and SOBSEY, M.D., 2004. Development of a L-rhamnose and D-arabitol supplemented MacConkey agar to identify pathogenic *Yersinia enterocolitica* among environmental *Yersinias* in swine production wastes. *J Microbiol Methods*, **57**: 289 – 292.

SHENDURE, J., and JI, H., 2008. Next-generation DNA sequencing. *Nature Biotechnology*, **26** (10): 1135 – 1145.

SIGRIST, C.J.A., CERUTTI, L., HULO, N., GATTIKER, A., FALQUET, L., PAGNI, M., BAIROCH, A., and BUCHER, P., 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, **3**: 265 – 274.

SIRISENA, D.M., and SKURNIK, M., 2003. Mutations in the genes for synthesis of the outer core region of the lipopolysaccharide of *Yersinia enterocolitica* O:3. *J Appl Microbiol*, **94**: 686 – 692.

SKURNIK, M., VENHO, R., BENGOECHEA, J.A., and MORIYÓN, I., 1999. The lipopolysaccharide outer core of *Yersinia enterocolitica* serotype O:3 is required for virulence and plays a role in outer membrane integrity. *Mol Microbiol*, **31** (5): 1443 – 1462.

SKURNIK, M., and BENGOECHEA, J.A., 2003. The biosynthesis and biological role of lipopolysaccharide O-antigens of pathogenic *Yersiniae*. *Carbohydrate Res*, **338**: 2521 – 2529.

SKURNIK, M., BIEDZKA-SAREK, M., LÜBECK, P.S., BLOM, T., BENGOECHEA, J.A., PÉREZ-GUTIÉRREZ, C., AHRENS, P., and HOORFAR, J., 2007. Characterization and biological role of the O-polysaccharide gene cluster of *Yersinia enterocolitica* serotype O:9. *J Bact*, **189** (20): 7244 – 7253.

SMALLA, K., WACHTENDORF, U., HEUER, H., LIU, W.-T., and FORNEY, L., 1998. Analysis of BIOLOG GN substrate utilization patterns by microbial communities. *Appl Env Microbiol*, **64** (4): 1220 – 1225.

SMYTH, G.K., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, **3** (1): Article 3.

SNELLINGS, N.J., POPEK, M., and LINDLER, L.E., 2001. Complete DNA Sequence of *Yersinia enterocolitica* Serotype O:8 Low-Calcium-Response Plasmid Reveals a New Virulence Plasmid-Associated Replicon. *Inf Immun*, **69** (7): 4627 – 4638.

SONG, H., HWANG, J., YI, H., ULIRCH, R.L., YU, Y., NIERMAN, W.C., and KIM, H.S., 2010. The early stage of bacterial genome-reductive evolution in the host. *PLoS Pathogens,* **6** (5): e1000922.

SOUZA, R.A., PITONDO-SILVA, A., FALCÃO, D.P., and FALCÃO, J.P., 2010. Evaluation of four molecular typing methodologies as tools for determining taxonomy relations and for identifying species among *Yersinia* isolates. *J Micro Methods*, **82**: 141 – 150.

SOUZA, R.A., FALCÃO, D.P., and FALCÃO, J.P., 2011. Emended description of the species *Yersinia massiliensis*. *IJSEM*, **61** (5): 1094 – 1097.

SPRAGUE, L.D., and NEUBAUER, H., 2005. *Yersinia aleksiciae* sp. nov. *IJSEM*, **55**: 831 – 835.

SPRAGUE, L.D., SCHOLZ, H.C., AMANN, S., BUSSE, H.-J., and NEUBAUER, H., 2008. *Yersinia similis* sp. nov. *IJSEM*, **58** (4): 952 – 958.

STAINIER, I., BLEVES, S., JOSENHANS, C., KARMANI, L, KERBOURCH, C., LAMBERMONT, I., TÖTEMEYER, S., BOYD, A., and CORNELIS, G.R., 2000. YscP, a *Yersinia* protein required for Yop secretion that is surface exposed, and released in low Ca^{2+}. *Mol Microbiol*, **37** (5): 1005 – 1018.

STAMATAKIS, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22** (21): 2688 – 2690.

STOCK, I., HENRICHFREISE, B., and WIEDEMANN, B., 2002. Natural antibiotic susceptibility and biochemical profiles of *Yersinia enterocolitica*-like strains: *Y. bercovieri*, *Y. mollaretii*, *Y. aldova* and '*Y. ruckeri*'. *J Med Microbiol*, **51**: 56 – 69.

STRALEY, S.C., and PERRY, R.D., 1995. Environmental modulation of gene expression and pathogenesis in *Yersinia*. *Trends in Microbiology*, **3** (8): 310 – 317.

SULAKVELIDZE, A., 2000. Yersiniae other than *Y. enterocolitica*, *Y. pseudotuberculosis*, and *Y. pestis*: the ignored species. *Microbes Inf*, **2**: 497 – 513.

SWAMINATHAN, B., HARMON, M.C., and MEHLMAN, I.J., 1982. A review *Yersinia enterocolitica*. *J Appl Bacteriol*, **52**: 151 – 183.

TENNANT, S.M., GRANT, T.H., and ROBINS-BROWNE, R.M., 2003. Pathogenicity of *Yersinia enterocolitica* biotype 1A. *FEMS Immunol Med Microbiol*, **38**: 127 – 137.

TENNANT, S.M., JOE, A., and ROBINS-BROWNE, R.M., 2004. LPS genes of a serotype O:5 strain of *Yersinia enterocolitica* biotype 1A. NCBI Accession number: AY653208.

TENNANT, S.M., SKINNER, N.A., JOE, A., and ROBINS-BROWNE, R.M., 2005. Homologues of insecticidal toxin complex genes in *Yersinia enterocolitica* biotype 1A and their contribution to virulence. *Inf Immun*, **73** (10): 6860 – 6867.

THOMSON, N.R., HOWARD, S., WREN, B.W, HOLDEN, M.T., CROSSMAN, L., CHALLIS, G.L., CHURCHER, C., MUNGALL, K., BROOKS, K., CHILLINGWORTH, T., FELTWELL, T., ABDELLAH, Z., HAUSER, H., JAGELS, K., MADDISON, M., MOULE, S., SANDERS, M., WHITEHEAD, S., QUAIL, M.A., DOUGAN, G., PARKHILL, J., and PRENTICE, M.B., 2006. The complete genome sequence and comparative genome analysis of the high pathogenicity *Yersinia enterocolitica* strain 8081. *PLoS Genet*, **2**: e206.

THOMSON, N.R., CLAYTON, D.J., WINDHORST, D., VERNIKOS, G., DAVIDSON, S., CHURCHER, C., QUAIL, M.A., STEVENS, M., JONES, M.A., WATSON, M., BARRON, A., LAYTON, A., PICKARD, D., KINGSLEY, R.A., BIGNELL, A., CLARK, L., HARRIS, B., ORMOND, D., ABDELLAH, Z., BROOKS, K., CHEREVACH, I., CHILLINGWORTH, T., WOODWARD, J., NORBERCZAK, H., LORD, A., ARROWSMITH, C., JAGELS, K., MOULE, S., MUNGALL, K., SANDERS, M., WHITEHEAD, S., CHABALGOITY, J.A., MASKELL, D., HUMPHREY, T., ROBERTS, M., BARROW, P.A., DOUGAN, G., and PARKHILL, J., 2008. Comparative genome analysis of *Salmonella* Enteritidis PT4 and

Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res*, **18** (10): 1624 – 1637.

TOBBACK, E., DECOSTERE, A., HERMANS, K., RYCKAERT, J., DUCHATEAU, L., HAESEBROUCK, F., and CHIERS, K., 2009. Route of entry and tissue distribution of *Yersinia ruckeri* in experimentally infected rainbow trout *Oncorhynchus mykiss*. *Dis Aquatic Org*, **84**: 219 – 228.

TOMPKINS, D.S., HUDSON, M.J., SMITH, H.R., EGLIN, R.P., WHEELER, J.G., BRETT, M.M., OWEN, R.J., BRAZIER, J.S., CUMBERLAND, P., KING, V., and COOK, P.E., 1999. A study of infectious intestinal disease in England: microbiological findings in cases and controls. *Commun Dis Pub Health*, **2** (2): 108 – 113.

TORANZO, A.E., BARJA, J.L, COLWELL, R.R., and HETRICK, F.M., 1983. Characterization of plasmids in bacterial fish pathogens. *Inf Immun*, **39** (1): 184 – 192.

VIRDI, J.S., and SACHDEVA, P., 2005. Molecular heterogeneity in *Yersinia enterocolitica* and '*Y. enterocolitica*-like' species – Implications for epidemiology, typing and taxonomy. *FEMS Immunol Med Microbiol*, **45**: 1 – 10.

UNIPROT, 2011. UniProt Protein knowledgebase BLAST server. [online] Available at: http://www.uniprot.org/ [Accessed 25 January 2011]

WAGNER, S., SORG, I., DEGIACOMI, M., JOURNET, L., PERARO, M.D., and CORNELIS, G.R., 2009. The helical content of the YscP molecular ruler determines the length of the *Yersinia* injectisome. *Mol Microbiol*, **71** (3): 692 – 701.

WANG, X., LI, Y., JING, H., REN, Y., ZHOU, Z., WANG, S., KAN, B., XU, J., and WANG, L., 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 Micro*, **49**(4): 1251 – 1259.

WAUTERS, G., KANDOLO, K., and JANSSENS, M., 1987. Revised biogrouping scheme of *Yersinia enterocolitica*. *Contrib Microbiol Immunol*, **9**: 14 – 21.

WELLCOME TRUST SANGER INSTITUTE, 2011. Scientific resources – Software – ACT: Artemis Comparison Tool – User Manual. [online] Available at: ftp://ftp.sanger.ac.uk/pub/pathogens/software/arte-mis/extra/big_blast.pl [Accessed 25 January 2011].

WREN, B.W., 2003. The *Yersiniae* – a model genus to study the rapid evolution of bacterial pathogens. *Nat Rev Micro*, **1** (1): 55 – 64.

WUTHE, H.H., and ALEKSIC, S., 1997. *Yersinia enterocolitica* serovar 2a, wb, 3:b,c biovar 5 in hares and sheep. *Berl Munch Tierarztl Wochenschr*, **110** (5): 176 – 177. [German]

YOUNG, G.M., and MILLER, V.L., 1997. Identification of novel chromosomal loci affecting *Yersinia enterocolitica* pathogenesis. *Mol Microbiol*, **25**(2): 319 – 238.

ZERBINO, D.R., and BIRNEY, E., 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*, **18**: 821 – 829.

ZHANG, L., TOIVANEN, P., and SKURNIK, M., 1996. The gene cluster directing O-antigen biosynthesis in *Yersinia enterocolitica* serotype O:8: identification of the genes for mannose and galactose biosynthesis and the gene for the O-antigen polymerase. *Microbiology*, **142**: 277 – 288.

ZHANG, L., RADZIEJEWSKA-LEBRECHT, J., KRAJEWSKA-PIETRASIK, D., TOIVANEN, P., and SKURNIK, M., 1997. Molecular and chemical characterization of the lipopolysaccharide O-antigen and its role in the virulence of *Yersinia enterocolitica* serotype O:8. *Mol Microbiol*, **23** (1): 63 – 76.

ZHOU, L., LEI, X.-H., BOCHNER, B.R., and WANNER, B.L., 2003. Phenotype microarray analysis of *Escherichia coli* K-12 mutants with deletions of all two-component systems. *J Bact*, **185** (16): 4956 – 4972.