Final Project:

Yersinia Pestis Genome Assembly and Annotation

Emma Hoover

Introduction

- Strain was isolated from *Citellophilus tesquorum* fleas found at the entrance of gopher burrows in Kabardino-Balker Republic, Russia (2000)
- Known to cause Central-Caucasian High Mountain Plague
- Studied in order to identify possible molecular targets to prevent and treat the plague

Kislichkina AA, Mazurina EM, Platonov ME, Skyrabin YP, Sizova AA, Solomentsev VI, Galkina EV, Trunyakova AS, Gapel'chenkova TV, Dentovskaya SV, Bogun AG, Anisimov AP. 2022. Complete Genome Assembly of Yersinia pestis subsp. pestis bv. Medievalis SCPM-O-B6530, a Proline-Dependent Strain Isolated from the Central-Caucasian High-Mountain Plague Focus in Kabardino-Balker Republic (Russia). Microbiol Resour Announc 11:e01115-21

ABySS

What it does:

- Software that assembles genomes (use QUAST to check quality and completeness of assembly)

Code:

- mamba activate genomeassembly
- abyss-pe name=ypestis k=96 B=2G in='ypestisreads1.fastq ypestisreads2.fastq'
  - pe=paired end reads
  - name=prefix name of output files
  - k=k-mer
  - B=memory amount
  - in=input

ABySS Output

Report

	ypestis-8_fa	ypestis-8_fa_broken
# contigs (>= 0 bp)	1114	-
# contigs (>= 1000 bp)	397	397
Total length (>= 0 bp)	4669048	-
Total length (>= 1000 bp)	4518129	4515405
# contigs	449	453
Largest contig	56670	56670
Total length	4552839	4552639
Reference length	4658411	4658411
GC (%)	47.58	47.58
Reference GC (%)	47.63	47.63
N50	17523	17523
NG50	16980	16980
N90	5751	5751
NG90	4849	4849
auN	19845.1	19843.0
auNG	19395.4	19392.5
L50	84	84
LG50	87	87
L90	256	256
LG90	275	275
# misassemblies	29	29
# misassembled contigs	27	27
Misassembled contigs length	473856	473856
# local misassemblies	13	13
# scaffold gap ext. mis.	0	-
# scaffold gap loc. mis.	0	-
# unaligned mis. contigs	0	0
# unaligned contigs	12 + 16 part	13 + 15 part
Unaligned length	137929	137891
Genome fraction (%)	95.635	95.629
Duplication ratio	1.009	1.009
# N's per 100 kbp	4.39	0.00
# mismatches per 100 kbp	9.36	9.36
# indels per 100 kbp	4.92	4.74
# genomic features	7775 + 521 part	7773 + 521 part
Complete BUSCO (%)	97.97	97.97
Partial BUSCO (%)	0.00	0.00
# predicted rRNA genes	3 + 1 part	2 + 2 part
Largest alignment	55685	55685
Total aligned length	4411191	4411043
NA50	16553	16553
NGA50	16189	16189
NA90	4282	4282
NGA90	3375	3375
auNA	18519.8	18519.1
auNGA	18100.1	18098.6
LA50	89	89
LGA50	93	93
LA90	284	284
LGA90	308	308

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

ABySS Output

Misassemblies report

	ypestis-8_fa	ypestis-8_fa_broken
# misassemblies	29	29
# contig misassemblies	29	29
# c. relocations	28	28
# c. translocations	0	0
# c. inversions	1	1
# scaffold misassemblies	0	0
# s. relocations	0	0
# s. translocations	0	0
# s. inversions	0	0
# misassembled contigs	27	27
Misassembled contigs length	473856	473856
# local misassemblies	13	13
# scaffold gap ext. mis.	0	-
# scaffold gap loc. mis.	0	-
# unaligned mis. contigs	0	0
# mismatches	413	413
# indels	217	209
# indels (<= 5 bp)	92	86
# indels (> 5 bp)	125	123
Indels length	4567	4466

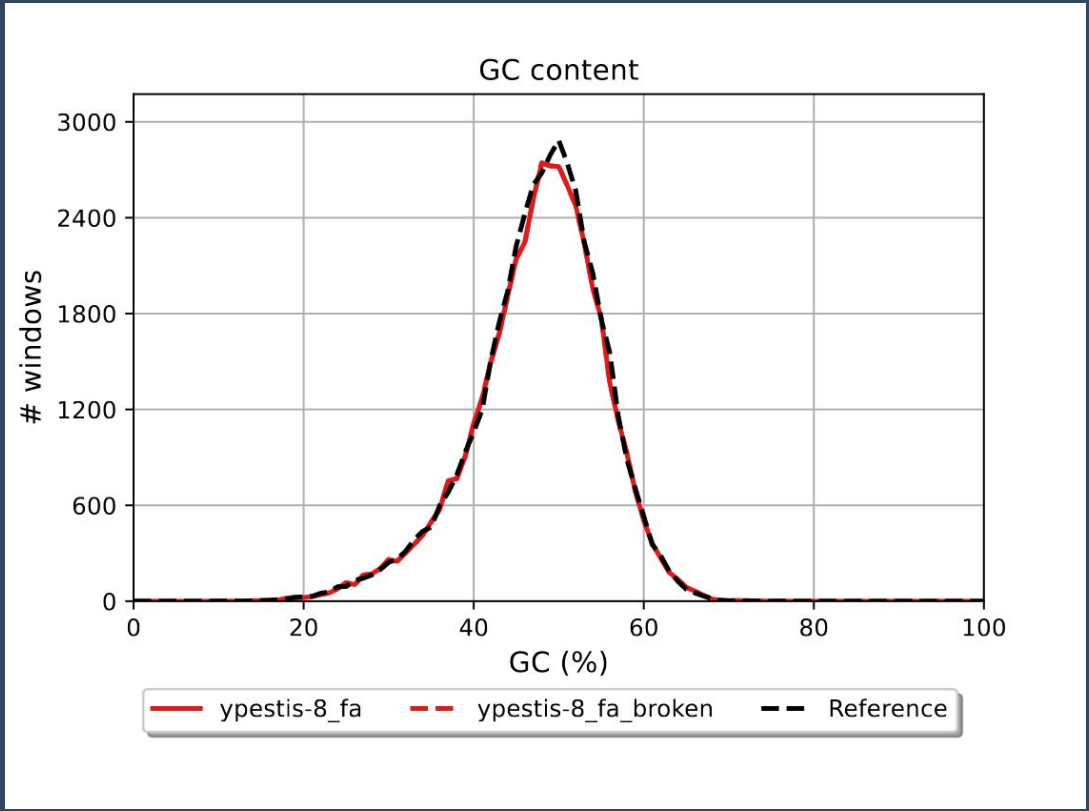
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Unaligned report

	ypestis-8_fa	ypestis-8_fa_broken
# fully unaligned contigs	12	13
Fully unaligned length	31748	32709
# partially unaligned contigs	16	15
Partially unaligned length	106181	105182
# N's	200	0

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

ABySS Output



Prokka Annotation

What it does:

- Software that annotates genome and identifies predicted coding sequences

Code:

- mamba activate genomeassembly
- prokka --outdir ProkkaAnnotation --prefix ypestis ABySSOutput/ypestis-8.fa
  - 'prokka'=calls the program
  - '--outdir ProkkaAnnotation'=creates an output directory that is called 'ProkkaAnnotation'
  - '--prefix ypestis'=gives the files a prefix name 'ypestis'
  - 'ABySSOutput/ypestis-8.fa'=tells the program which file we want it to annotate in the ABySSOutput folder

Prokka Annotation Output

	A	B	C	D	E	F	G	H	I	J
1	locus_tag	ftype	length_bp	gene	EC_number	COG	product			
2	OBENKMEN_00001	CDS	399	mqsA		COG1396	Antitoxin MqsA			
3	OBENKMEN_00002	CDS	297	mqsR	3.1.-.-		mRNA interferase toxin MqsR			
4	OBENKMEN_00003	CDS	363				hypothetical protein			
5	OBENKMEN_00004	CDS	261				hypothetical protein			
6	OBENKMEN_00005	CDS	309	xerC_1			Tyrosine recombinase XerC			
7	OBENKMEN_00006	tRNA	76				tRNA-Phe(gaa)			
8	OBENKMEN_00007	CDS	1149	adeP		COG2252	Adenine permease AdeP			
9	OBENKMEN_00008	CDS	621				hypothetical protein			
10	OBENKMEN_00009	CDS	282				hypothetical protein			
11	OBENKMEN_00010	CDS	1200	fabV	1.3.1.9	COG3007	Enoyl-[acyl-carrier-protein] reductase [NADH]			
12	OBENKMEN_00011	CDS	1365	mnmE	3.6.-.-	COG0486	tRNA modification GTPase MnmE			
13	OBENKMEN_00012	CDS	1641	yidC			Membrane protein insertase YidC			
14	OBENKMEN_00013	CDS	258	yidD		COG0759	Putative membrane protein insertion efficiency factor			
15	OBENKMEN_00014	CDS	360	rnpA	3.1.26.5	COG0594	Ribonuclease P protein component			
16	OBENKMEN_00015	CDS	141	rpmH		COG0230	50S ribosomal protein L34			
17	OBENKMEN_00016	CDS	141				hypothetical protein			
18	OBENKMEN_00017	CDS	1389	dnaA		COG0593	Chromosomal replication initiator protein DnaA			
19	OBENKMEN_00018	CDS	1101	dnaN		COG0592	Beta sliding clamp			
20	OBENKMEN_00019	CDS	1086	recF		COG1195	DNA replication and repair protein RecF			
21	OBENKMEN_00020	CDS	2415	gyrB	5.6.2.2	COG0187	DNA gyrase subunit B			
22	OBENKMEN_00021	CDS	810	yidA	3.1.3.23	COG0561	Sugar phosphatase YidA			

RAST

What it does:

- Rapid Annotations using Subsystems Technology
- Annotate genome and identifies protein-encoding genes
- Organizes identified genes into the subsystems and pathways they are present in

RAST Output

Organism Overview for Yersinia pestis (632.984)

Genome	Yersinia pestis (Taxonomy ID: 632)
Domain	Bacteria
Taxonomy	Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Yersiniaceae; Yersinia; Yersinia pseudotuberculosis complex; Yersinia pestis; Yersinia pestis
Neighbors	View closest neighbors
Size	4,669,048
GC Content	47.5
N50	16980
L50	87
Number of Contigs (with PEGs)	1114
Number of Subsystems	352
Number of Coding Sequences	5067
Number of RNAs	79

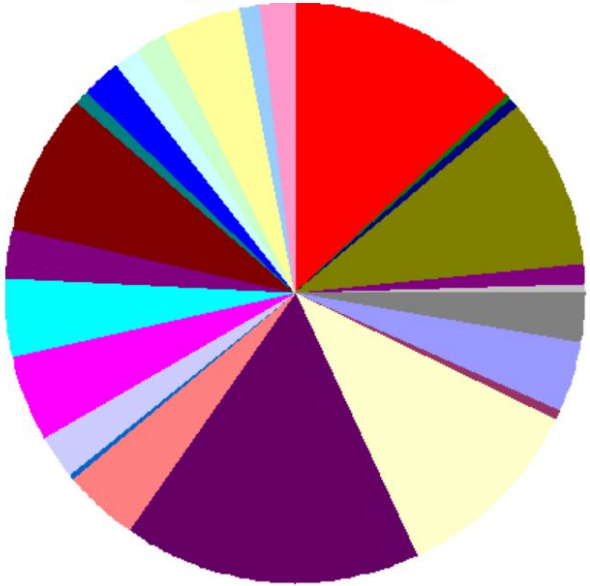
RAST Output

Subsystem Information

Subsystem Statistics | Features in Subsystems

Subsystem Coverage

Subsystem Category Distribution



Subsystem Feature Counts

- Fatty Acids, Lipids, and Isoprenoids (48)
- Nucleosides and Nucleotides (86)
- Stress Response (73)
- Secondary Metabolism (6)
- Amino Acids and Derivatives (314)
- Protein Metabolism (201)
- RNA Metabolism (52)
- Potassium metabolism (12)
- DNA Metabolism (74)
- Nitrogen Metabolism (8)
- Metabolism of Aromatic Compounds (17)
- Membrane Transport (179)
- Phages, Prophages, Transposable elements, Plasmids (9)
- Cell Division and Cell Cycle (6)
- Carbohydrates (244)
- Cell Wall and Capsule (33)
- Motility and Chemotaxis (21)
- Respiration (84)
- Phosphorus Metabolism (28)
- Sulfur Metabolism (25)
- Photosynthesis (0)
- Regulation and Cell signaling (44)
- Miscellaneous (15)
- Cofactors, Vitamins, Prosthetic Groups, Pigments (145)
- Virulence, Disease and Defense (51)
- Iron acquisition and metabolism (85)
- Dormancy and Sporulation (1)

Reactive Oxygen Species, Damage and Protection Mechanisms (Bacteria)