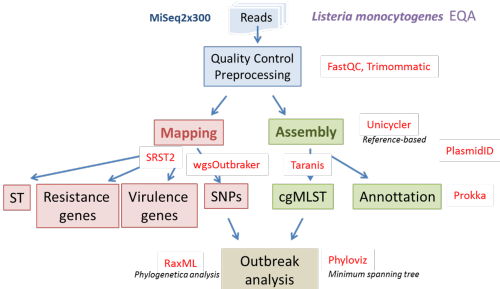


Bacterial WGS training : Exercise 5

Title	Chromosome, plasmid, resistance and virulence annotation
Training dataset:	
Questions:	<ul style="list-style-type: none">How many genes there are in my sample?Are there virulence and/or antibiotic resistance genes?Where are the genes located?Which plasmids are present in the sample?How do I visualize the results?
Objectives:	<ul style="list-style-type: none">Annotate virulence and ABR genesDetermine gene variantsDetermine plasmidomeLocate annotated genesResults interpretation
Time estimation:	1 h
Key points:	<ul style="list-style-type: none">Comparing annotation using mapping vs assemblyPlasmid, virulence and resistance determination

<<<<<< HEAD



1ac4d30d8ca8e4a68f1238b9bba5705ab11fea69

- Introduction
- Exercise
 - Mapping based annotation
 - Assembly based annotation

Introduction

In this exercise we are going to determine the genomic content of a multidrug-resistant (MDR) *K. pneumoniae* isolate. First we will use [srst2](#) to assess the resistome and later, we will use [plasmidID](#) to infer biological and positional information to sequences and see where the genes, detected with mapping strategy, are located.

Training dataset description

The sample we are going to analyse is an *in silico* dataset obtained with [wgsim](#) using a sample of *Klebsiella pneumoniae subsp. pneumoniae* [HS11286](#) available at [ncbi](#).

Exercise

Mapping based annotation

To execute [srst2](#), which maps the reads against a antibiotic resistance genes database (ARGannot), lets execute this command:

```
cd
cd Documents/wgs
nextflow run BU-ISCIIB/bacterial_wgs_training \
-profile singularity \
--reads 'training_dataset/plasmidid_test/KPN_TEST_R{1,2}.fastq.gz' \
--fasta training_dataset/listeria_NC_021827.1_NoPhages.fna \
--gtf training_dataset/listeria_NC_021827.1_NoPhages.gff \
--srst2_resistance training_dataset/ARGannot.r1.fasta \
--srst2_virulence training_dataset/EcOH.fasta \
--step mapAnnotation
```

Results should look like that

Sample	DB	gene	allele	coverage	depth	diffs	uncertainty	divergence	length	maxMAF	clusterid	seqid	annotation
KPN_TEST_R	ARGannot.r1	RmtB_AgLy	RmtB_1580	100.0	12.09	1snp		0.132	756	0.125	309	1580	no;no;RmtB;AGly;AB263754;2843-3598;756
KPN_TEST_R	ARGannot.r1	TEM-1D_Bla	TEM-117_968	100.0	33.386	2snp		0.262	764	0.382	205	968	no;no;TEM-117;Bla;AY130282;1-764;764
KPN_TEST_R	ARGannot.r1	KPC-1_Bla	KPC-14_809	100.0	5.412	1indel		0.0	876	0.333	184	809	no;no;KPC-14;Bla;JX524191;396-1271;876
KPN_TEST_R	ARGannot.r1	Amph_Bla	Amph_634	100.0	11.373	14snp		1.206	1161	0.143	86	634	no;no;Amph;Bla;CP003785;4208384-4209544;1161
KPN_TEST_R	ARGannot.r1	CTX-M-9_Bla	CTX-M-14_102	100.0	26.676	1snp		0.114	876	0.412	190	102	no;yes;CTX-M-14;Bla;AF252622;1741-2616;876
KPN_TEST_R	ARGannot.r1	StrA_AgLy	StrA_1501	100.0	12.502	2snp		0.249	804	0.167	263	1501	no;no;StrA;AGly;AJ627643;3725-4528;804

Sample	DB	gene	allele	coverage	depth	diffs	uncertainty	divergence	length	maxMAF	clusterid	seqid	annotation
KPN_TEST_R	ARGannot.r1	StrB_AGly	StrB_1614	100.0	9.545	1snp		0.119	837	0.167	227	1614	no;no;StrB;AGly;KR091911;169145-169981;837
KPN_TEST_R	ARGannot.r1	AadA_AGly	AadA2_1605	100.0	9.306	2snp		0.256	780	0.167	229	1605	yes;no;AadA2;AGly;X68227;166-945;780
KPN_TEST_R	ARGannot.r1	SHV-OKP-LEN_Bla	SHV-11_1287	100.0	9.401			0.0	861	0.143	164	1287	yes;no;SHV-11;Bla;HM751098;1-861;861
KPN_TEST_R	ARGannot.r1	TetRG_Tet	TetRG_605	96.209	6.48	10snp24holes	edge0.0	1.642	633	0.5	373	605	no;no;TetRG;Tet;S52438;113-745;633
KPN_TEST_R	ARGannot.r1	DfrA_Tmt	DfrA12_1089	99.799	8.389	1indel		0.0	498	0.143	418	1089	yes;no;DfrA12;Tmt;Z21672;310-807;498
KPN_TEST_R	ARGannot.r1	TetG_Tet	TetG_632	100.0	9.963			0.0	1176	0.25	80	632	no;no;TetG;Tet;NC_010410;3672607-3671432;1176
KPN_TEST_R	ARGannot.r1	SulII_Sul	SulII_1219	100.0	11.094	1snp		0.123	816	0.2	256	1219	no;no;SulII;Sul;KR091911;167466-168281;816

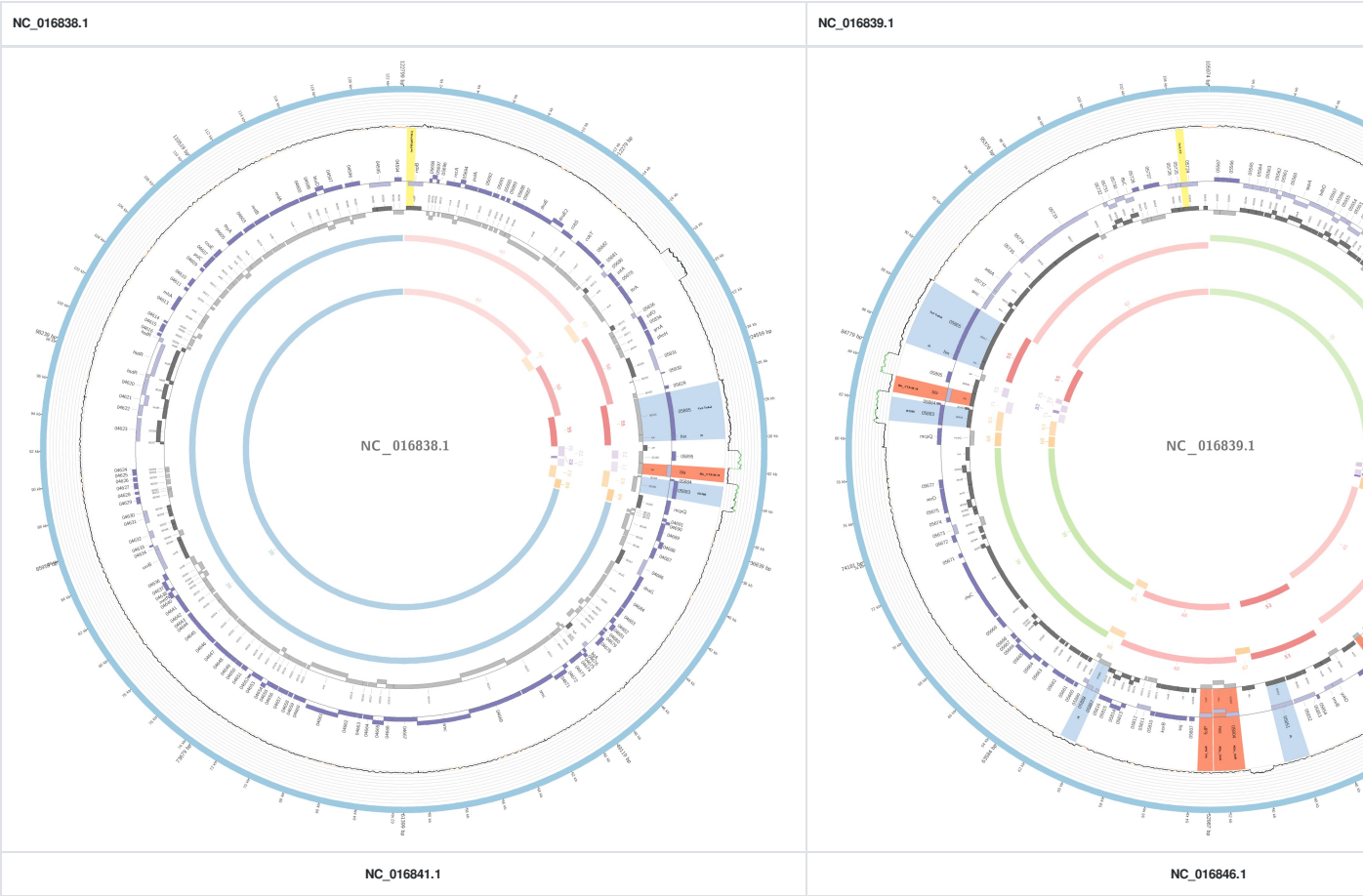
This table is a full report of all the ARG found with all mapping stats.

Assembly based annotation

Now, using the contigs assembled using those same reads, we can determine the exact location of those ARG. ARG can be located on the chromosome but motly on plasmids. In that case, we are going to focus on plasmid derived ARG using the annotation feature of plasmidID. To run the analysis lets use this command:

```
cd
cd Documents/wgs
nextflow run BU-ISCIIII/bacterial_wgs_training \
-profile singularity \
--reads 'training_dataset/plasmidid_test/KPN_TEST_R{1,2}.fastq.gz' \
--fasta training_dataset/listeria_NC_021827.1_NoPhages.fna \
--gtf training_dataset/listeria_NC_021827.1_NoPhages.gff \
--plasmidid_database training_dataset/plasmidid_test/plasmids_TEST_database.fasta \
--plasmidid_config training_dataset/plasmidid_test/plasmidid_config.txt \
--step plasmidID
```

##Results should look like that



A circular genome map of the NC_016841.1 sequence. The map displays several concentric rings representing different genomic features. The outermost ring is a thick blue line. Moving inward, there are several thin grey lines. A prominent orange ring is visible, followed by a black ring and a purple ring. A yellow sector is highlighted on the right side of the map. Various labels are placed around the perimeter, including '100 bp', '200 bp', '300 bp', '400 bp', '500 bp', '600 bp', '700 bp', '800 bp', '900 bp', '1000 bp', '1100 bp', '1200 bp', '1300 bp', '1400 bp', '1500 bp', '1600 bp', '1700 bp', '1800 bp', '1900 bp', '2000 bp', '2100 bp', '2200 bp', '2300 bp', '2400 bp', '2500 bp', '2600 bp', '2700 bp', '2800 bp', '2900 bp', '3000 bp', '3100 bp', '3200 bp', '3300 bp', '3400 bp', '3500 bp', '3600 bp', '3700 bp', '3800 bp', '3900 bp', '4000 bp', '4100 bp', '4200 bp', '4300 bp', '4400 bp', '4500 bp', '4600 bp', '4700 bp', '4800 bp', '4900 bp', '5000 bp', '5100 bp', '5200 bp', '5300 bp', '5400 bp', '5500 bp', '5600 bp', '5700 bp', '5800 bp', '5900 bp', '6000 bp', '6100 bp', '6200 bp', '6300 bp', '6400 bp', '6500 bp', '6600 bp', '6700 bp', '6800 bp', '6900 bp', '7000 bp', '7100 bp', '7200 bp', '7300 bp', '7400 bp', '7500 bp', '7600 bp', '7700 bp', '7800 bp', '7900 bp', '8000 bp', '8100 bp', '8200 bp', '8300 bp', '8400 bp', '8500 bp', '8600 bp', '8700 bp', '8800 bp', '8900 bp', '9000 bp', '9100 bp', '9200 bp', '9300 bp', '9400 bp', '9500 bp', '9600 bp', '9700 bp', '9800 bp', '9900 bp', '10000 bp'. The center of the map contains the text 'NC_016841.1'.

A circular genome map of the NC_016846.1 plasmid. The map displays several concentric rings representing different genomic features. The outermost ring shows the plasmid's size in base pairs (bp) from 0 to 10,000. The next ring inward shows the plasmid's map, with various features labeled, including genes (e.g., *oriT*, *repA*, *repB*, *repC*, *repD*, *repE*, *repF*, *repG*, *repH*, *repI*, *repJ*, *repK*, *repL*, *repM*, *repN*, *repO*, *repP*, *repQ*, *repR*, *repS*, *repT*, *repU*, *repV*, *repW*, *repX*, *repY*, *repZ*, *repAA*, *repAB*, *repAC*, *repAD*, *repAE*, *repAF*, *repAG*, *repAH*, *repAI*, *repAJ*, *repAK*, *repAL*, *repAM*, *repAN*, *repAO*, *repAP*, *repAQ*, *repAR*, *repAS*, *repAT*, *repAU*, *repAV*, *repAW*, *repAX*, *repAY*, *repAZ*, *repBA*, *repBB*, *repBC*, *repBD*, *repBE*, *repBF*, *repBG*, *repBH*, *repBI*, *repBJ*, *repBK*, *repBL*, *repBM*, *repBN*, *repBO*, *repBP*, *repBQ*, *repBR*, *repBS*, *repBT*, *repBU*, *repBV*, *repBW*, *repBX*, *repBY*, *repBZ*, *repCA*, *repCB*, *repCC*, *repCD*, *repCE*, *repCF*, *repCG*, *repCH*, *repCI*, *repCJ*, *repCK*, *repCL*, *repCM*, *repCN*, *repCO*, *repCP*, *repCQ*, *repCR*, *repCS*, *repCT*, *repCU*, *repCV*, *repCW*, *repCX*, *repCY*, *repCZ*, *repDA*, *repDB*, *repDC*, *repDD*, *repDE*, *repDF*, *repDG*, *repDH*, *repDI*, *repDJ*, *repDK*, *repDL*, *repDM*, *repDN*, *repDO*, *repDP*, *repDQ*, *repDR*, *repDS*, *repDT*, *repDU*, *repDV*, *repDW*, *repDX*, *repDY*, *repDZ*, *repEA*, *repEB*, *repEC*, *repED*, *repEE*, *repEF*, *repEG*, *repEH*, *repEI*, *repEJ*, *repEK*, *repEL*, *repEM*, *repEN*, *repEO*, *repEP*, *repEQ*, *repER*, *repES*, *repET*, *repEU*, *repEV*, *repEW*, *repEX*, *repEY*, *repEZ*, *repFA*, *repFB*, *repFC*, *repFD*, *repFE*, *repFF*, *repFG*, *repFH*, *repFI*, *repFJ*, *repFK*, *repFL*, *repFM*, *repFN*, *repFO*, *repFP*, *repFQ*, *repFR*, *repFS*, *repFT*, *repFU*, *repFV*, *repFW*, *repFX*, *repFY*, *repFZ*, *repGA*, *repGB*, *repGC*, *repGD*, *repGE*, *repGF*, *repGG*, *repGH*, *repGI*, *repGJ*, *repGK*, *repGL*, *repGM*, *repGN*, *repGO*, *repGP*, *repGQ*, *repGR*, *repGS*, *repGT*, *repGU*, *repGV*, *repGW*, *repGX*, *repGY*, *repGZ*, *repHA*, *repHB*, *repHC*, *repHD*, *repHE*, *repHF*, *repHG*, *repHH*, *repHI*, *repHJ*, *repHK*, *repHL*, *repHM*, *repHN*, *repHO*, *repHP*, *repHQ*, *repHR*, *repHS*, *repHT*, *repHU*, *repHV*, *repHW*, *repHX*, *repHY*, *repHZ*, *repIA*, *repIB*, *repIC*, *repID*, *repIE*, *repIF*, *repIG*, *repIH*, *repII*, *repIJ*, *repIK*, *repIL*, *repIM*, *repIN*, *repIO*, *repIP*, *repIQ*, *repIR*, *repIS*, *repIT*, *repIU*, *repIV*, *repIW*, *repIX*, *repIY*, *repIZ*, *repJA*, *repJB*, *repJC*, *repJD*, *repJE*, *repJF*, *repJG*, *repJH*, *repJI*, *repJJ*, *repJK*, *repJL*, *repJM*, *repJN*, *repJO*, *repJP*, *repJQ*, *repJR*, *repJS*, *repJT*, *repJU*, *repJV*, *repJW*, *repJX*, *repJY*, *repJZ*, *repKA*, *repKB*, *repKC*, *repKD*, *repKE*, *repKF*, *repKG*, *repKH*, *repKI*, *repKJ*, *repKK*, *repKL*, *repKM*, *repKN*, *repKO*, *repKP*, *repKQ*, *repKR*, *repKS*, *repKT*, *repKU*, *repKV*, *repKW*, *repKX*, *repKY*, *repKZ*, *repLA*, *repLB*, *repLC*, *repLD*, *repLE*, *repLF*, *repLG*, *repLH*, *repLI*, *repLJ*, *repLK*, *repLL*, *repLM*, *repLN*, *repLO*, *repLP*, *repLQ*, *repLR*, *repLS*, *repLT*, *repLU*, *repLV*, *repLW*, *repLX*, *repLY*, *repLZ*, *repMA*, *repMB*, *repMC*, *repMD*, *repME*, *repMF*, *repMG*, *repMH*, *repMI*, *repMJ*, *repMK*, *repML*, *repMM*, *repMN*, *repMO*, *repMP*, *repMQ*, *repMR*, *repMS*, *repMT*, *repMU*, *repMV*, *repMW*, *repMX*, *repMY*, *repMZ*, *repNA*, *repNB*, *repNC*, *repND*, *repNE*, *repNF*, *repNG*, *repNH*, *repNI*, *repNJ*, *repNK*, *repNL*, *repNM*, *repNN*, *repNO*, *repNP*, *repNQ*, *repNR*, *repNS*, *repNT*, *repNU*, *repNV*, *repNW*, *repNX*, *repNY*, *repNZ*, *repOA*, *repOB*, *repOC*, *repOD*, *repOE*, *repOF*, *repOG*, *repOH*, *repOI*, *repOJ*, *repOK*, *repOL*, *repOM*, *repON*, *repOO*, *repOP*, *repOQ*, *repOR*, *repOS*, *repOT*, *repOU*, *repOV*, *repOW*, *repOX*, *repOY*, *repOZ*, *repPA*, *repPB*, *repPC*, *repPD*, *repPE*, *repPF*, *repPG*, *repPH*, *repPI*, *repPJ*, *repPK*, *repPL*, *repPM*, *repPN*, *repPO*, *repPP*, *repPQ*, *repPR*, *repPS*, *repPT*, *repPU*, *repPV*, *repPW*, *repPX*, *repPY*, *repPZ*, *repQA*, *repQB*, *repQC*, *repQD*, *repQE*, *repQF*, *repQG*, *repQH*, *repQI*, *repQJ*, *repQK*, *repQL*, *repQM*, *repQN*, *repQO*, *repQP*, *repQQ*, *repQR*, *repQS*, *repQT*, *repQU*, *repQV*, *repQW*, *repQX*, *repQY*, *repQZ*, *repRA*, *repRB*, *repRC*, *repRD*, *repRE*, *repRF*, *repRG*, *repRH*, *repRI*, *repRJ*, *repRK*, *repRL*, *repRM*, *repRN*, *repRO*, *repRP*, *repRQ*, *repRR*, *repRS*, *repRT*, *repRU*, *repRV*,

Are all the genes located with *srst2* bound to plasmids?