Bio 346 Bioinformatics Lab Guide 4:

Gene Annotation and Misc Gene Detection

James B. Herrick, Ph.D.
James Madison University

I'm going to list links and very general instructions for annotating and detecting genes. The instructions won't be as detailed as you might like… sometimes because I may have done little with a particular tool yet. You don't have to use all of these, or explore them in any order. Go where your interests take you. This is where the fun really begins!

Annotation of Your Entire Genome

"Annotation" is the process of where genes (open reading frames or ORFs) might be in your genome and then attempting to assign a function to each of them based upon their similarity to other genes of known function (in various databases). Many genes won't be identified (yet!). These are based on your assemblies. The main methods for microbial genomes are *Prokka*, *RAST*, and PGAP. The latter is a proprietary annotator from NCBI so we won't be using it until we upload our assemblies to NCBI.

Prokka

Prokka can be accessed using Galaxy (not GalaxyTrakr). You'll have to sign up separately for Galaxy itself, and upload your files just as you have in GalaxyTrakr; the interface is the same.

Look for the Prokka tool under "NGS Assembly" (or just do a search under 'Tools'). The default parameters should be sufficient, although you might consider excluding somewhat larger contigs than the 200 bp minimum in the default settings). Prokka gives various output files, such as table and text files for searching.

Visualization

There are a number of "genome viewers" of various flavors that will allow you to visualize your annotated contigs. Artemis is an oldie but goodie. It has an ancient interface but is still being maintained by the Sanger Wellcome Institute. Open Artemis and upload your .gff file output by Prokka.

RAST

RAST is a web-based annotation server found on the PATRIC website. PATRIC has a number of other programs you might be interested in; however, the interface takes some getting used to.

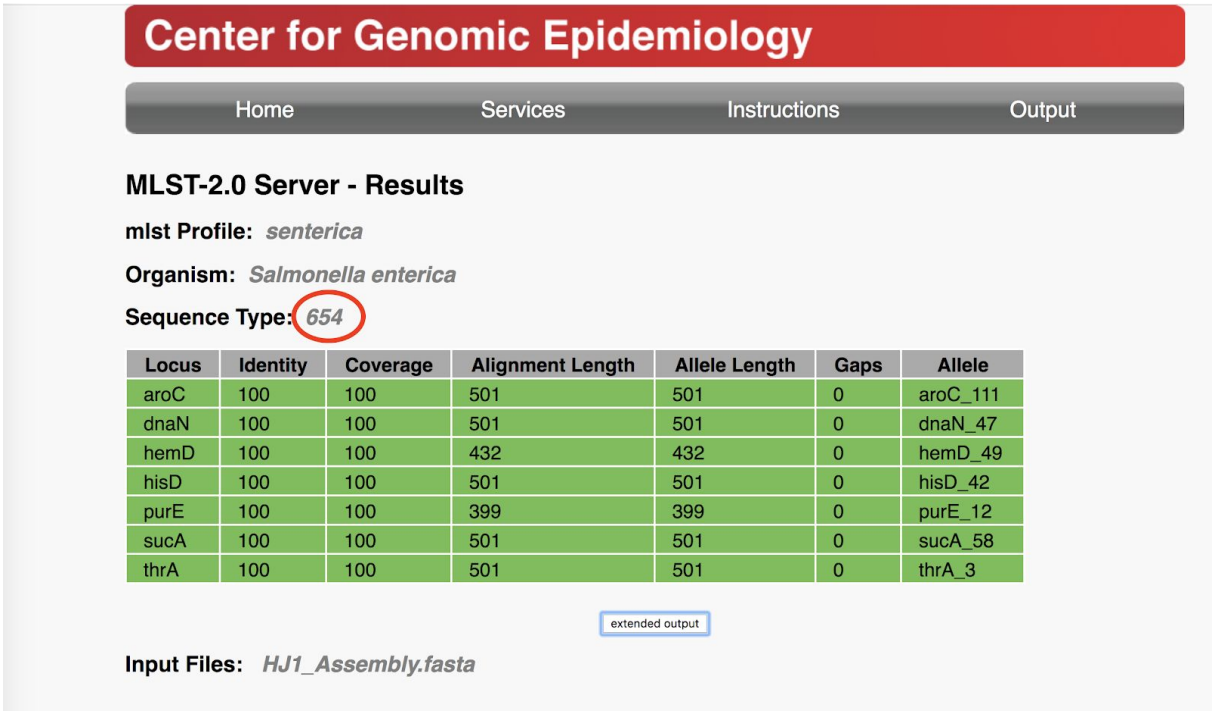
Center for Genomic Epidemiology Programs

For all of these CGE programs, it pays to read the instructions, the explanation of the output files and, if possible, the paper upon which the analysis is based. These are normally at the top of the 'home' page for each analysis.

MLST

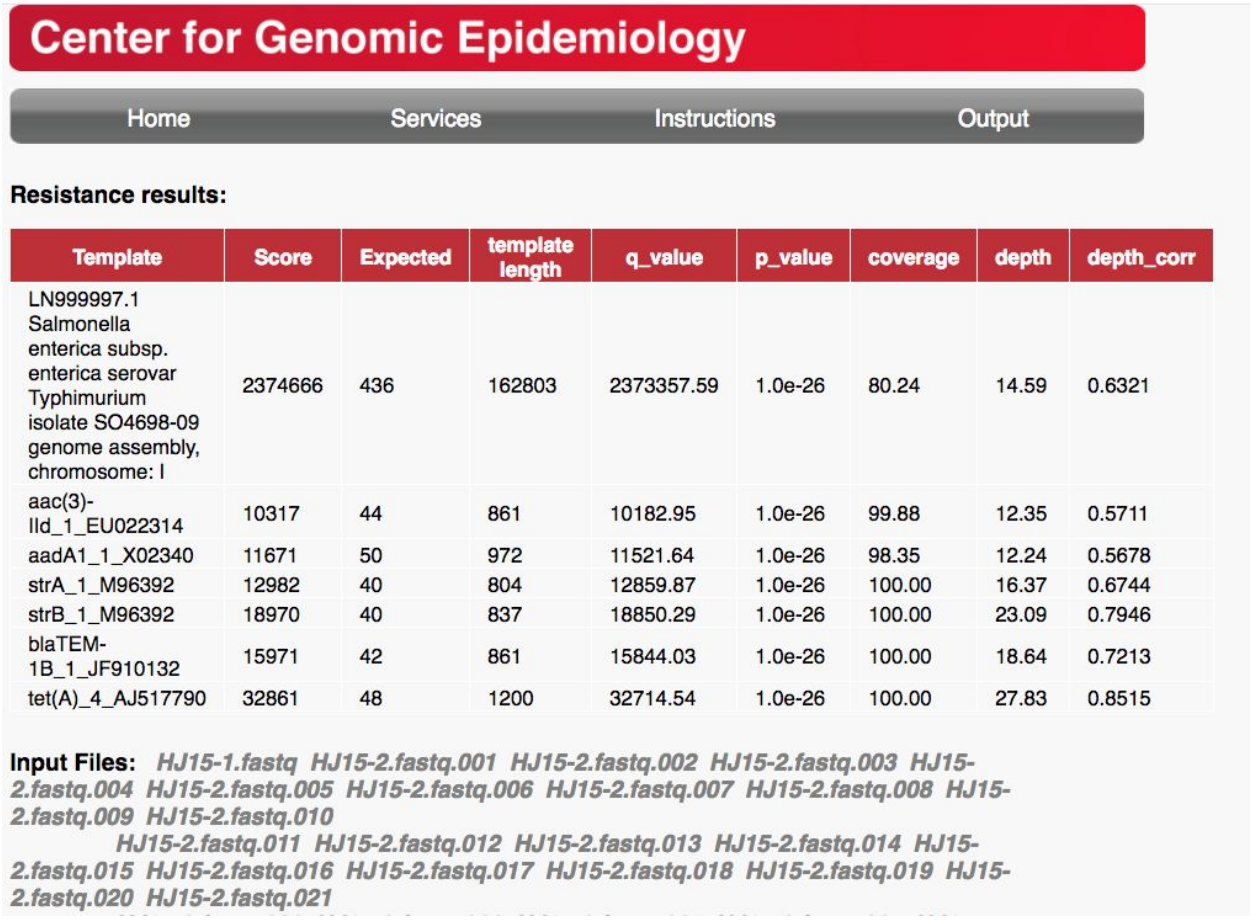
Although not really concerned with gene annotation oedetection *per se*, MLST is a method to type organisms that, in the case of *Salmonella,* resolves at a somewhat finer level than serotyping. MLST is organism-specific (i.e. there is a different scheme, or set of genes used for each organism). This is the classic, 7-gene MLST.

Output example (HJ1)



KmerResistance

Maps the co-occurrence of k-mers between the WGS data and a database of resistance genes.



Center for Genomic Epidemiology

Home | Services | Instructions | Output

Resistance results:

Template	Score	Expected	template length	q_value	p_value	coverage	depth	depth_corr
LN999997.1 Salmonella enterica subsp. enterica serovar Typhimurium isolate SO4698-09 genome assembly, chromosome: I	2374666	436	162803	2373357.59	1.0e-26	80.24	14.59	0.6321
aac(3)-IId_1_EU022314	10317	44	861	10182.95	1.0e-26	99.88	12.35	0.5711
aadA1_1_X02340	11671	50	972	11521.64	1.0e-26	98.35	12.24	0.5678
strA_1_M96392	12982	40	804	12859.87	1.0e-26	100.00	16.37	0.6744
strB_1_M96392	18970	40	837	18850.29	1.0e-26	100.00	23.09	0.7946
blaTEM-1B_1_JF910132	15971	42	861	15844.03	1.0e-26	100.00	18.64	0.7213
tet(A)_4_AJ517790	32861	48	1200	32714.54	1.0e-26	100.00	27.83	0.8515

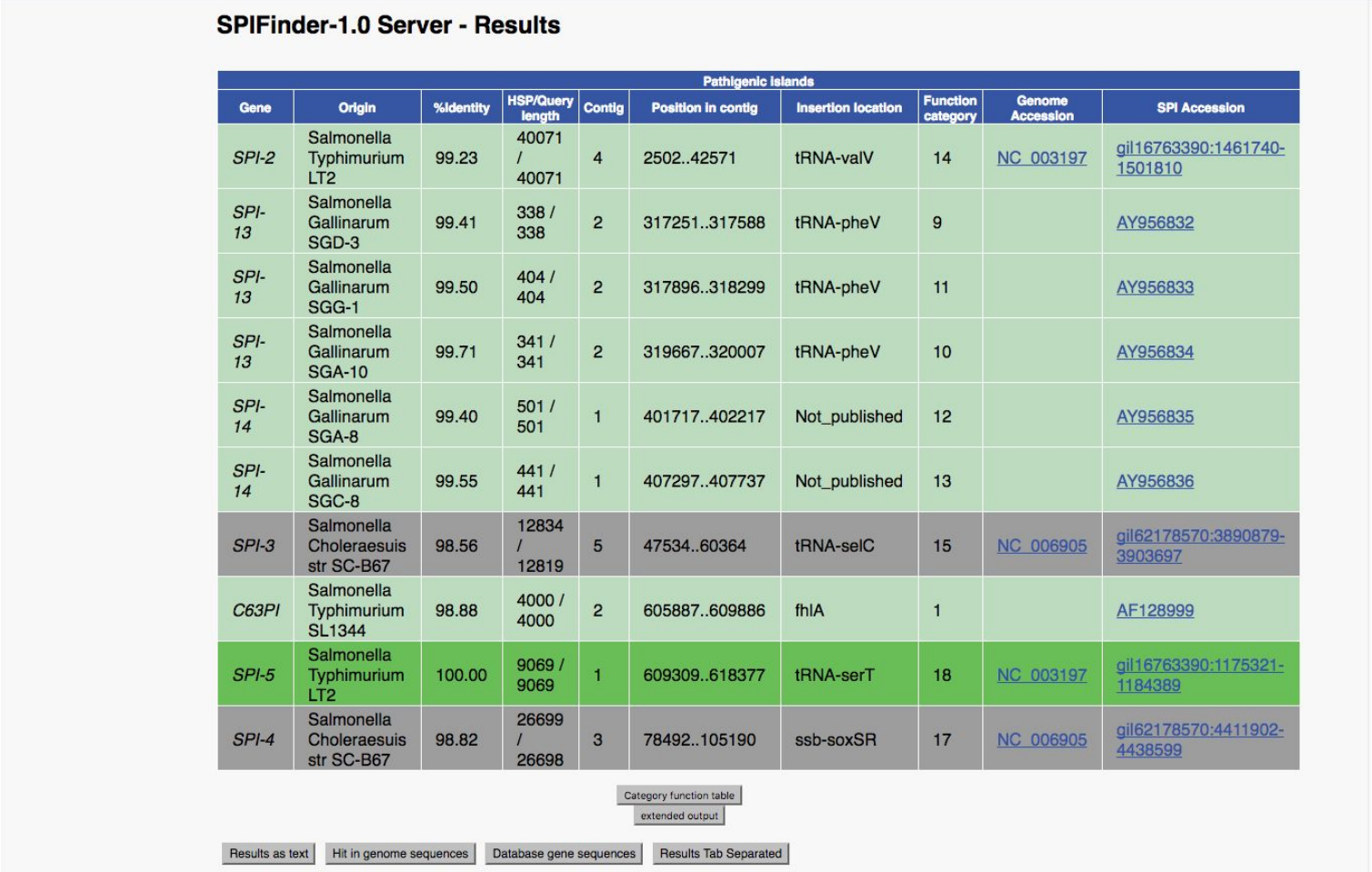
Input Files: *HJ15-1.fastq HJ15-2.fastq.001 HJ15-2.fastq.002 HJ15-2.fastq.003 HJ15-2.fastq.004 HJ15-2.fastq.005 HJ15-2.fastq.006 HJ15-2.fastq.007 HJ15-2.fastq.008 HJ15-2.fastq.009 HJ15-2.fastq.010 HJ15-2.fastq.011 HJ15-2.fastq.012 HJ15-2.fastq.013 HJ15-2.fastq.014 HJ15-2.fastq.015 HJ15-2.fastq.016 HJ15-2.fastq.017 HJ15-2.fastq.018 HJ15-2.fastq.019 HJ15-2.fastq.020 HJ15-2.fastq.021*

VirulenceFinder

Identification of acquired virulence genes. Only available so far for *E. coli, S. aureus, Listeria,* and *Enterococcus*. Choose your assembly (fasta) as the input file.

SPIFinder

Identifies “Salmonella Pathogenicity Islands”. Choose your assembly (fasta) as the input file.



SPIFinder-1.0 Server - Results

Pathigenic islands									
Gene	Origin	%Identity	HSP/Query length	Contig	Position in contig	Insertion location	Function category	Genome Accession	SPI Accession
SPI-2	Salmonella Typhimurium LT2	99.23	40071 / 40071	4	2502..42571	tRNA-valV	14	NC_003197	gi|16763390:1461740-1501810
SPI-13	Salmonella Gallinarum SGD-3	99.41	338 / 338	2	317251..317588	tRNA-pheV	9		AY956832
SPI-13	Salmonella Gallinarum SGG-1	99.50	404 / 404	2	317896..318299	tRNA-pheV	11		AY956833
SPI-13	Salmonella Gallinarum SGA-10	99.71	341 / 341	2	319667..320007	tRNA-pheV	10		AY956834
SPI-14	Salmonella Gallinarum SGA-8	99.40	501 / 501	1	401717..402217	Not_published	12		AY956835
SPI-14	Salmonella Gallinarum SGC-8	99.55	441 / 441	1	407297..407737	Not_published	13		AY956836
SPI-3	Salmonella Choleraesuis str SC-B67	98.56	12834 / 12819	5	47534..60364	tRNA-selC	15	NC_006905	gi|62178570:3890879-3903697
C63PI	Salmonella Typhimurium SL1344	98.88	4000 / 4000	2	605887..609886	fhlA	1		AF128999
SPI-5	Salmonella Typhimurium LT2	100.00	9069 / 9069	1	609309..618377	tRNA-serT	18	NC_003197	gi|16763390:1175321-1184389
SPI-4	Salmonella Choleraesuis str SC-B67	98.82	26699 / 26698	3	78492..105190	ssb-soxSR	17	NC_006905	gi|62178570:4411902-4438599

Category function table
extended output

Results as text | Hit in genome sequences | Database gene sequences | Results Tab Separated

Pathogenicity Island Functional Categories

Function table	
Category	**Function**
1	Iron uptake system, sit operon
2	P4-like integrase; located within a high-pathogenicity-island (HPI) region
3	Type VI secretion system effector
4	Multi drug resistance (ampicillin, gentamicin, streptomycin, spectinomycin, sulfathiazole, tetracycline, and nalidixic acid)
5	Type III secretion system, invasion into epithelial cells, apoptosis (InvA, OrgA, SptP, SipA, SipB, SipC, SipD, SopE, prgH)
6	Phage 46 and the sefA-R chaperone-usher fimbrial operon
7	sopB, mppA, icdA, envF, msgA, envE, pagD, pagC
8	msgA, narP
9	Acetyl-coA dehydrolase (gacD)
10	LysR family transcriptional regulation (gtrB)
11	Trancriptional regulation (gtrA)
12	Electron transfer favoprotein beta subunit (gpiA)
13	Trancriptional regulation (gpiB)
14	Type III secretion system, required for systemic infection and intracellular pathogenesis by facilitating replication of intracellular bacteria within membrane-bound Salmonella-containing vacuoles
15	Invasion, survival in monocytes, Mg2+ uptake (MgtC, B, MarT, MisL)
16	Invasion, survival in monocytes, Mg2+ uptake (sugR, rhuMMgtC, B, MarT, MisL), putative fimbrial-like protein(yadC/K./L/M), probable pilin chaperone(ecpD1/D2)
17	Type I secretion system, putative toxin secretion, apoptosis, required for intracellular survival in macrophages; large secretion protein(siiE) and Type I secretion system components(siiC,D,F), genes weakly similar to RTX-like toxins
18	Effector proteins for SPI-1 and SPI-2 (SopB, SigD, PipB)
19	safA-D and tcsA-R chaperone-usher fimbrialoperons6
20	Vi exopolysaccharide, SopE prophage and a type IVB pilus operon
21	Two bacteriocin pseudogenes, genes conferring immunity to the bacteriocins
22	Type I secretory apparatus, large RTX-like protein
23	Intestinal colonization and persistence determinants (shdA, ratA, ratB, sivI, sivH)

Selected %ID threshold: *95.00 %*

Selected minimum length: *60 %*

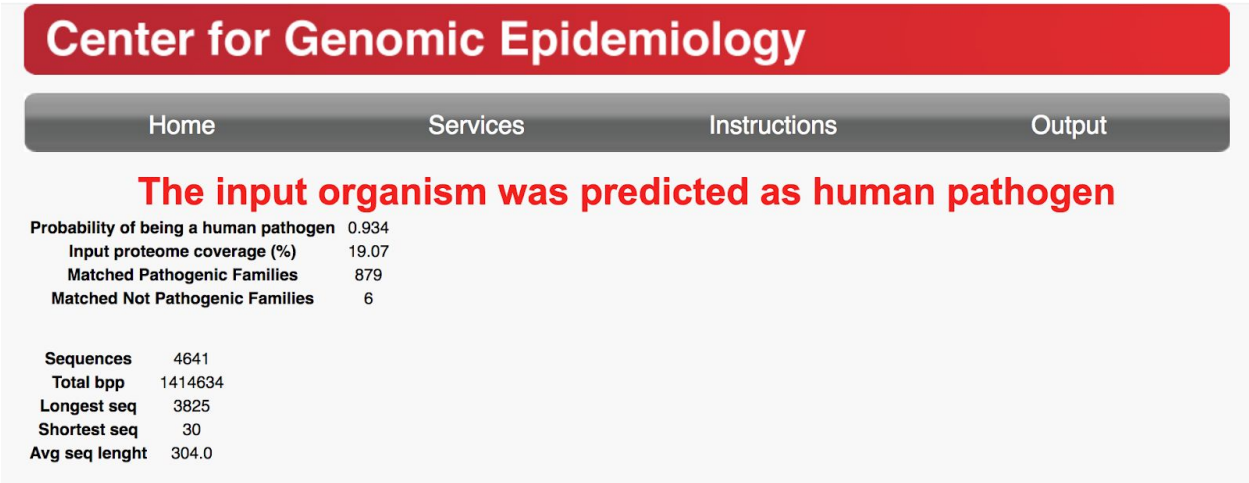
Input Files: *HJ15_assembly_galaxy.fasta*

SeroTypeFinder

Prediction of serotypes in total or partial sequenced isolates of *E. coli*. (Not an annotation tool but listed here because it's potentially useful for typing *E. coli*.). Again, use your assembly (fasta) as the input file.

PathogenFinder (HJ1)

Prediction of a bacterium's pathogenicity towards human hosts. Choose *γ-proteobacteria* under 'phylum or class' and *assembled genome* under 'Sequencing Platform'. Then use your assembly (fasta) as the input file.



Center for Genomic Epidemiology

Home | Services | Instructions | Output

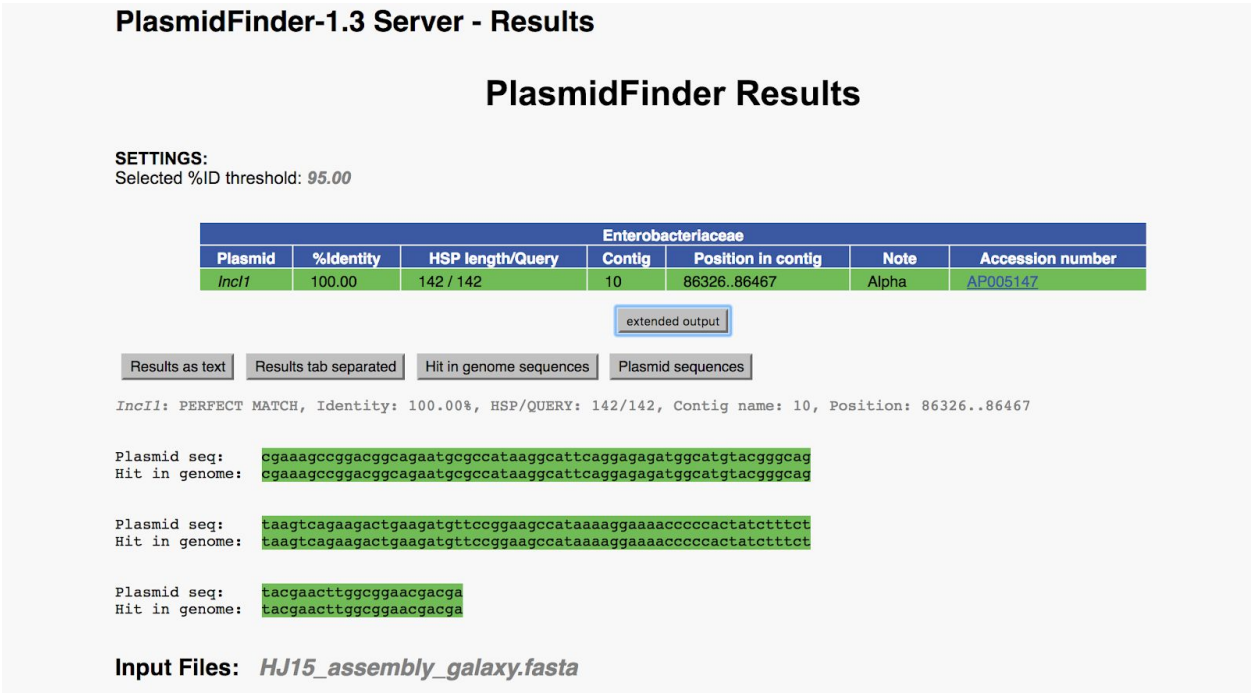
The input organism was predicted as human pathogen

Probability of being a human pathogen	0.934
Input proteome coverage (%)	19.07
Matched Pathogenic Families	879
Matched Not Pathogenic Families	6

Sequences	4641
Total bpp	1414634
Longest seq	3825
Shortest seq	30
Avg seq lenght	304.0

PlasmidFinder

Identifies plasmids in total or partial sequenced isolates of bacteria. It is very accurate, with few false positives (although it does suffer from false negatives). Select the *Enterobacteriace* database. You can experiment with the thresholds. Use your assembly as the input file. "The filename must not contain spaces.



PlasmidFinder-1.3 Server - Results

PlasmidFinder Results

SETTINGS:
Selected %ID threshold: ***95.00***

Enterobacteriaceae						
Plasmid	**%Identity**	**HSP length/Query**	**Contig**	**Position in contig**	**Note**	**Accession number**
IncI1	100.00	142 / 142	10	86326..86467	Alpha	AP005147

extended output

Results as text | Results tab separated | Hit in genome sequences | Plasmid sequences

IncI1: PERFECT MATCH, Identity: 100.00%, HSP/QUERY: 142/142, Contig name: 10, Position: 86326..86467

Plasmid seq: cgaaagccggacggcagaatgcgccataaggcattcaggagagatggcatgtacgggcag
Hit in genome: cgaaagccggacggcagaatgcgccataaggcattcaggagagatggcatgtacgggcag

Plasmid seq: taagtcagaagactgaagatgttccggaagccataaaaggaaaaccccactatctttct
Hit in genome: taagtcagaagactgaagatgttccggaagccataaaaggaaaaccccactatctttct

Plasmid seq: tacgaacttggcggaacgacga
Hit in genome: tacgaacttggcggaacgacga

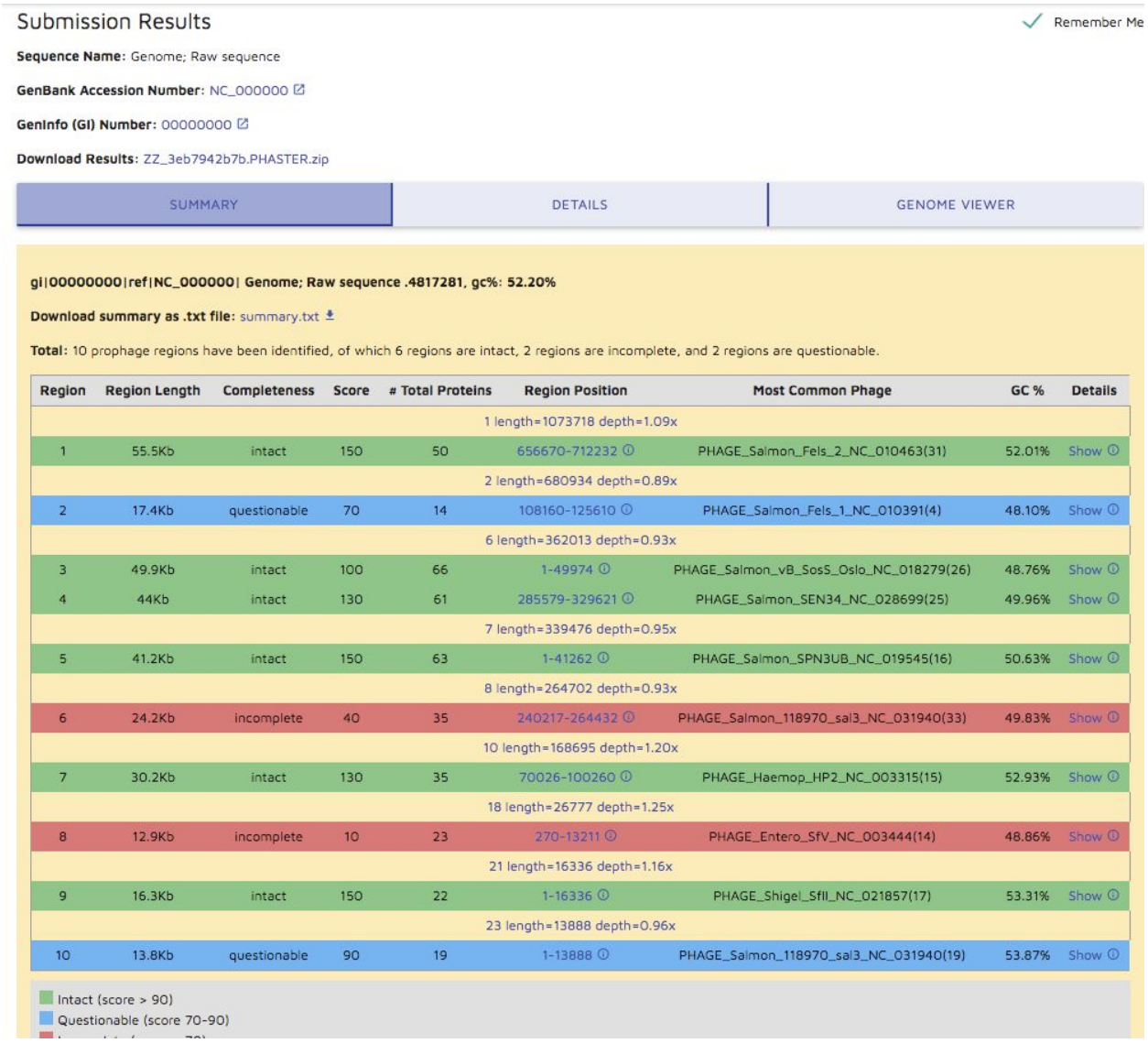
Input Files: ***HJ15_assembly_galaxy.fasta***

Other Programs

PHASTER

Rapid identification and annotation of **prophage** sequences within bacterial genomes and plasmids

- There is a short tutorial video on the home page that I recommend you watch. The *Help* tab at the top of the page also has very good instructions on using the site.
- Upload your fasta assembly file. Check “My FASTA file consists of metagenomic contigs”
- This will take a little while (minutes to hours), depending upon how many are ahead of you in the queue.
- (Be sure and check out the ‘Genome Viewer’ output.)



Submission Results

Remember Me

Sequence Name: Genome; Raw sequence

GenBank Accession Number: NC_000000

GenInfo (GI) Number: 00000000

Download Results: ZZ_3eb7942b7b.PHASTER.zip

SUMMARY | DETAILS | GENOME VIEWER

gi|00000000|ref|NC_000000| Genome; Raw sequence .4817281, gc%: 52.20%

Download summary as .txt file: summary.txt

Total: 10 prophage regions have been identified, of which 6 regions are intact, 2 regions are incomplete, and 2 regions are questionable.

Region	Region Length	Completeness	Score	# Total Proteins	Region Position	Most Common Phage	GC %	Details
				1 length=1073718 depth=1.09x				
1	55.5Kb	intact	150	50	656670-712232	PHAGE_Salmon_Fels_2_NC_010463(31)	52.01%	Show
				2 length=680934 depth=0.89x				
2	17.4Kb	questionable	70	14	108160-125610	PHAGE_Salmon_Fels_1_NC_010391(4)	48.10%	Show
				6 length=362013 depth=0.93x				
3	49.9Kb	intact	100	66	1-49974	PHAGE_Salmon_vB_SosS_Oslo_NC_018279(26)	48.76%	Show
4	44Kb	intact	130	61	285579-329621	PHAGE_Salmon_SEN34_NC_028699(25)	49.96%	Show
				7 length=339476 depth=0.95x				
5	41.2Kb	intact	150	63	1-41262	PHAGE_Salmon_SPN3UB_NC_019545(16)	50.63%	Show
				8 length=264702 depth=0.93x				
6	24.2Kb	incomplete	40	35	240217-264432	PHAGE_Salmon_118970_sal3_NC_031940(33)	49.83%	Show
				10 length=168695 depth=1.20x				
7	30.2Kb	intact	130	35	70026-100260	PHAGE_Haemop_HP2_NC_003315(15)	52.93%	Show
				18 length=26777 depth=1.25x				
8	12.9Kb	incomplete	10	23	270-13211	PHAGE_Entero_SfV_NC_003444(14)	48.86%	Show
				21 length=16336 depth=1.16x				
9	16.3Kb	intact	150	22	1-16336	PHAGE_Shigel_SfII_NC_021857(17)	53.31%	Show
				23 length=13888 depth=0.96x				
10	13.8Kb	questionable	90	19	1-13888	PHAGE_Salmon_118970_sal3_NC_031940(19)	53.87%	Show

Intact (score > 90)
Questionable (score 70-90)

INTEGRALL

identification of **integrons**, which are are genetic systems that allow bacteria to capture and express antibiotic resistance gene cassettes.

- For this site, you actually have to cut and paste your assembled fasta file into the "BLAST" page. Open your file in a text viewer (such as TextEdit on the Mac, or the free viewer SublimeText), if possible. Converting it to text in Word might work. Select the entire file and copy it.
- Click on 'BLAST/Search' on the INTEGRALL site and paste your file. Then hit 'search'. **Note:** cutting and pasting these huge files will tax your computer's clipboard and CPU. Be patient. I found that Safari worked better than Chrome on my Mac.
- The results will appear in the same window. If nothing appears, apprently no integrons were found. (I know, it's kind of silly to me, too.)

Virulence Factors of Pathogenic Bacteria (website)

This is a database, with useful information on the virulence factors -- including pathogenicity islands -- of *Salmonella* and other bacterial pathogens. It will help you understand some of your results obtained above. (Note: last time I checked, this site wouldn't load for me. It might be down or even non-functional now.)