Tutorial: QC of sequence data + Assembling reads using Velvet + Annotation using Prokka + Assignment using TAXAassign

(Credit goes to Nick Loman for assembly part of tutorial: http://pathogenomics.bham.ac.uk/blog/author/nick/)

Umer Zeeshan Ijaz

Step 1: Let us have a look inside E. coli sequence data sequenced on the MiSeq instrument (Courtesy of Nick Loman)

```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ head Sample280.fastq
```

Step 2: Convert Sample280.fastq to fasta format and covert nucleotides to lower case with quality less than 30.

```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ seqtk fq2fa -q 30 Sample280.fastq > Sample280.fasta
```

Step 3: Use fastqc to generate QC statistics

```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ fastqc Sample280.fastq
Started analysis of Sample280.fastq
Approx 5% complete for Sample280.fastq
Approx 10% complete for Sample280.fastq
Approx 15% complete for Sample280.fastq
Approx 20% complete for Sample280.fastq
Approx 25% complete for Sample280.fastq
Approx 30% complete for Sample280.fastq
Approx 35% complete for Sample280.fastq
Approx 40% complete for Sample280.fastq
Approx 45% complete for Sample280.fastq
Approx 50% complete for Sample280.fastq
Approx 55% complete for Sample280.fastq
Approx 60% complete for Sample280.fastq
Approx 65% complete for Sample280.fastq
```

```
Approx 70% complete for Sample280.fastq
Approx 75% complete for Sample280.fastq
Approx 80% complete for Sample280.fastq
Approx 85% complete for Sample280.fastq
Approx 90% complete for Sample280.fastq
Approx 95% complete for Sample280.fastq
Approx 100% complete for Sample280.fastq
Analysis complete for Sample280.fastq
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ ls
Sample280.fasta  Sample280.fastq  Sample280_fastqc  Sample280_fastqc.zip
[ngswshop@quince-srv2 ~/workshop/assembly_test]$
```

Step 4: Go to Sample280_fastqc folder and read contents of fastqc_data.txt. The section called “Per base sequence quality” shows an overview of the range of quality scores across all based at each position in the fastq file. A quality score of 30 indicates a 1 in 1000 probability of error and a quality score of 20 indicates a 1 in 100 probability of error (see the wikipedia page on the fastq format at http://en.wikipedia.org/wiki/Fastq).

```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ cd Sample280_fastqc
[ngswshop@quince-srv2 ~/workshop/assembly_test/Sample280_fastqc]$ ls
fastqc_data.txt  fastqc_report.html  Icons  Images  summary.txt
[ngswshop@quince-srv2 ~/workshop/assembly_test/Sample280_fastqc]$ cat fastqc_data.txt
##FastQC    0.10.0
>>Basic Statistics      pass
#Measure    Value
Filename    Sample280.fastq
File type   Conventional base calls
Encoding    Sanger / Illumina 1.9
Total Sequences   1769276
Filtered Sequences      0
Sequence length   150
%GC   49
>>END_MODULE
>>Per base sequence quality   fail
```

#Base	Mean	Median	Lower Quartile	Upper Quartile	10th Percentile	90th Percentile
1	32.2537009488627	34.0	31.0	34.0	31.0	34.0
2	32.53245508332222	34.0	31.0	34.0	31.0	34.0
3	32.62741652517753	34.0	31.0	34.0	31.0	34.0
4	36.0641776636319	37.0	37.0	37.0	35.0	37.0
5	36.0501640218937	37.0	37.0	37.0	35.0	37.0
6	36.08056911414612	37.0	37.0	37.0	35.0	37.0
7	36.09596241626518	37.0	37.0	37.0	35.0	37.0
8	36.09040138452113	37.0	37.0	37.0	35.0	37.0
9	37.79904774608371	39.0	38.0	39.0	35.0	39.0
10-14	38.08336189492199	39.4	38.4	39.4	35.2	39.4
15-19	39.218010191739445	41.0	39.2	41.0	36.0	41.0
20-24	39.03441215502839	41.0	39.0	41.0	36.0	41.0
25-29	38.76067476188	40.0	38.8	41.0	35.2	41.0
30-34	38.402329766525966	40.0	38.0	41.0	34.2	41.0
35-39	37.93910345248565	40.0	38.0	41.0	33.2	41.0
40-44	37.48144472654351	40.0	37.0	41.0	32.6	41.0
45-49	37.256771583404735	39.6	36.0	41.0	32.8	41.0
50-54	36.53242230155159	38.6	35.0	40.8	31.6	41.0
55-59	35.6022392210147	36.6	35.0	39.8	30.8	41.0
60-64	34.6650026338457	35.2	34.4	38.6	29.8	40.8
65-69	33.81172581327051	35.0	34.0	36.6	29.0	39.4
70-74	33.12859542547348	35.0	33.6	35.6	28.6	38.0
75-79	32.5744321406044	35.0	33.0	35.0	27.0	36.6
80-84	32.11298056380124	35.0	33.0	35.0	26.2	35.8
85-89	31.718267924280887	35.0	33.0	35.0	25.0	35.0
90-94	31.366530942600253	35.0	32.4	35.0	24.2	35.0
95-99	30.99419943524922	34.4	31.8	35.0	23.2	35.0
100-104	30.63280494394317	34.0	31.0	35.0	20.4	35.0
105-109	30.237300907263762	34.0	31.0	35.0	18.6	35.0
110-114	29.790075262423727	34.0	30.0	35.0	16.6	35.0
115-119	29.28891184868839	34.0	29.2	35.0	9.8	35.0
120-124	28.728337579891434	33.2	29.0	35.0	4.6	35.0
125-129	28.10175619858066	33.0	27.0	35.0	2.0	35.0
130-134	27.366094153766852	33.0	25.8	34.8	2.0	35.0
135-139	26.497391249301973	32.0	24.4	34.0	2.0	35.0
140-144	25.456898414944867	31.4	21.6	34.0	2.0	35.0

```
145-149     23.9764131769153  31.0  10.6  34.0  2.0   35.0
150   22.703088155833232      31.0  2.0   34.0  2.0   35.0
>>END_MODULE
>>Per sequence quality scores pass
#Quality    Count
2     3474.0
3     1253.0
4     1712.0
5     2111.0
6     3074.0
7     3913.0
8     4568.0
9     4841.0
10    5105.0
11    5468.0
12    5540.0
13    5884.0
14    6644.0
15    6868.0
16    7703.0
17    8376.0
18    9365.0
19    10140.0
20    11490.0
21    12772.0
22    14387.0
23    16127.0
24    18391.0
25    20984.0
26    23721.0
27    27454.0
28    32308.0
29    39433.0
30    48352.0
31    61766.0
32    82651.0
33    120945.0
```

```
34    199824.0
35    367618.0
36    432627.0
37    139109.0
38    3278.0
>>END_MODULE
>>Per base sequence content   pass
#Base G     A     T     C
1     21.97735118771746 26.435898073562292      27.60954198214411 23.977208756576136
2     23.645531243623793      27.78152382362699 28.20604847301688 20.36689645973233
3     23.37978924712707 27.20220022201171 26.87924326108532 22.538767269775885
4     24.094092724933816      26.07671160406856 26.017817457536303      23.811378213461325
5     24.720224543824706      25.504387105234006      25.73380297929775 24.041585371643542
6     24.842986622776774      25.23162016553664 24.86282524603284 25.06256796565375
7     25.417345852201688      25.41796757543764 25.327930746813955      23.83675582554672
8     24.474248223567155      25.533382016146717      25.599341199451075      24.393028560835052
9     24.25986674775445 25.201947011093807      26.009961136645725      24.528225104506024
10-14 24.1889447158927  25.788978447219517      25.77059803924228 24.251478797645497
15-19 24.244927921880763      25.701558627297466      25.73376391276245 24.319749538059323
20-24 24.40530396203278 25.51094415244765 25.66991334748905 24.41383853803052
25-29 24.566321775400805      25.449385244534955      25.507160400268447      24.47713257979579
30-34 24.600194000849843      25.360688184159052      25.431316159325768      24.607801655665334
35-39 24.653108775461458      25.336056260873825      25.346659490834707      24.664175472830014
40-44 24.711403932080096      25.332032007285253      25.31360635180952 24.64295770882513
45-49 24.802022622013467      25.319050061273934      25.254831405398328      24.624095911314274
50-54 24.741107329254053      25.25776225847329 25.319369727333257      24.681760684939405
55-59 24.80591649342139 25.323476279009093      25.227040709917713      24.643566517651802
60-64 24.761559070714256      25.323871923707042      25.189228380928068      24.72534062465063
65-69 24.78213800080056 25.28760293414961 25.231240457180398      24.699018607869426
70-74 24.762431536286616      25.31153165642735 25.25974724777342 24.66628955951262
75-79 24.794122082298717      25.268051070610387      25.219036017698777      24.71879082939212
80-84 24.782144237877873      25.266271921792583      25.24033994180565 24.71124389852389
85-89 24.832340477868357      25.251367762119838      25.212639277285327      24.703652482726486
90-94 24.78070229976869 25.283121758351673      25.199345616883374      24.736830324996266
95-99 24.824172614245974      25.2513854573999  25.1802925919629  24.744149336391228
100-104     24.812619345515184      25.27470664120629 25.186238916664895      24.726435096613628
105-109     24.806590806092736      25.257849722268727      25.170760900148437      24.764798571490097
```

```
110-114     24.79303061287356 25.264017028944348      25.20998200905143 24.732970349130664
115-119     24.848729435783348      25.255938512808513      25.16196457862079 24.73336747278735
120-124     24.826221558723375      25.241164571603502      25.154492962609005      24.778120907064118
125-129     24.816928513888428      25.23722026615878 25.133218433917804      24.812632786034996
130-134     24.832372801668196      25.280384771866142      25.090038185754437      24.79720424071122
135-139     24.861439447105298      25.185542494465263      25.08270449579649 24.870313562632944
140-144     24.816853999141756      25.16228899378563 25.121083687129968      24.899773319942646
145-149     24.829062669710776      25.149413312243414      25.11245847243682 24.90906554560899
150   24.82181030555364 25.105726827426462      25.048412204058145      25.024050662961756
>>END_MODULE
>>Per base GC content   warn
#Base %GC
1     45.9545599442936
2     44.012427703356124
3     45.91855651690296
4     47.90547093839514
5     48.76180991546825
6     49.905554588430526
7     49.254101677748416
8     48.86727678440221
9     48.78809185226047
10-14 48.4404235135382
15-19 48.564677459940086
20-24 48.8191425000633
25-29 49.043454355196594
30-34 49.20799565651517
35-39 49.31728424829147
40-44 49.354361640905225
45-49 49.42611853332774
50-54 49.422868014193455
55-59 49.44948301107319
60-64 49.48689969536489
65-69 49.481156608669984
70-74 49.42872109579923
75-79 49.51291291169084
80-84 49.49338813640176
85-89 49.53599296059484
```

```
90-94 49.517532624764954
95-99 49.5683219506372
100-104     49.539054442128815
105-109     49.57138937758284
110-114     49.526000962004225
115-119     49.5820969085707
120-124     49.6043424657875
125-129     49.629561299923424
130-134     49.62957704237942
135-139     49.73175300973824
140-144     49.7166273190844
145-149     49.73812821531977
150   49.845860968515396
>>END_MODULE
>>Per sequence GC content     warn
#GC Content Count
0     0.0
1     0.0
2     0.5
3     0.5
4     2.0
5     3.0
6     3.0
7     3.5
8     3.0
9     3.0
10    4.5
11    8.0
12    17.0
13    49.0
14    91.0
15    94.0
16    122.5
17    153.5
18    180.5
19    355.5
20    570.5
```

21	978.0
22	1548.5
23	2112.0
24	2873.0
25	3994.5
26	5111.5
27	6476.0
28	8182.5
29	9934.0
30	11398.0
31	12658.0
32	14125.5
33	15440.0
34	16873.0
35	19448.0
36	22716.5
37	24661.0
38	25284.0
39	27862.0
40	30914.0
41	35005.0
42	40963.0
43	45456.5
44	51173.0
45	59144.5
46	65750.0
47	71533.5
48	79197.5
49	84858.5
50	86144.5
51	86112.0
52	86459.0
53	86907.5
54	84526.5
55	80203.5
56	76130.0
57	70203.5

58	62057.0
59	54197.5
60	44959.0
61	34282.5
62	26934.5
63	23545.5
64	20548.0
65	15726.0
66	10860.0
67	7028.5
68	4715.5
69	3398.5
70	2297.5
71	1499.0
72	1000.0
73	643.0
74	414.0
75	292.0
76	212.5
77	141.0
78	82.0
79	52.5
80	41.5
81	34.0
82	28.0
83	28.0
84	24.0
85	19.0
86	17.0
87	23.0
88	24.0
89	21.5
90	20.5
91	18.0
92	18.5
93	18.5
94	9.0

```
95    3.0
96    2.5
97    1.0
98    1.5
99    2.0
100   7.0
>>END_MODULE
>>Per base N content    pass
#Base N-Count
1     0.0
2     1.695608825304814E-4
3     0.0
4     0.0
5     0.0
6     0.0
7     0.0
8     0.0
9     0.0
10-14 4.521623534146171E-5
15-19 6.782435301219256E-5
20-24 1.3564870602438513E-4
25-29 1.9216900020121224E-4
30-34 3.052095885548665E-4
35-39 2.1477711787194312E-4
40-44 2.1477711787194312E-4
45-49 4.6346641224998247E-4
50-54 5.652029417682713E-4
55-59 6.669394712865601E-4
60-64 6.669394712865601E-4
65-69 7.799800596402143E-4
70-74 9.947571775121575E-4
75-79 0.0013338789425731202
80-84 0.0018086494136584684
85-89 0.0020121224726950457
90-94 0.002102554943377969
95-99 0.0023060280024145468
100-104     0.002441676708438932
```

```
105-109     0.0027016700616523366
110-114     0.002995575591371838
115-119     0.0035042582389632824
120-124     0.0039112043570364375
125-129     0.004318150475109593
130-134     0.00464596818133519
135-139     0.004838137181536402
140-144     0.004883353416877864
145-149     0.00506421835824371
150   0.0053129076526217504
>>END_MODULE
>>Sequence Length Distribution      pass
#Length     Count
150   1769276.0
>>END_MODULE
>>Sequence Duplication Levels fail
#Total Duplicate Percentage   52.74701757772342
#Duplication Level      Relative count
1     100.0
2     24.06891039872125
3     18.096083829144835
4     13.47482461593109
5     8.53121392416304
6     4.952490897788829
7     2.569043601811562
8     1.4030725512831899
9     0.7192966876831542
10++  3.7891839090666903
>>END_MODULE
>>Overrepresented sequences   pass
>>END_MODULE
>>Kmer Content    warn
#Sequence   Count Obs/Exp Overall   Obs/Exp Max Max Obs/Exp Position
TTTTT 990190      3.6758697   3.8426633   1
AAAAA 964645      3.5600014   4.20533     1
>>END_MODULE
[ngswshop@quince-srv2 ~/workshop/assembly_test/Sample280_fastqc]$ cd ..
```

Step 5: We will now assemble the reads using Velvet. Velvet requires an index file to be built before the assembly takes place. Longer k-mers result in a more stringent assembly with fewer overlaps, at the expense of coverage. There is no precise value of k for any given project. However, general rule of thumb is that k must be less than the read length and it should be an odd number. We will pick a value of k between 21 and 99 as a starting point and run velveth to build the hash index. You can use the command velveth my_assembly_directory value_of_k -shortPaired -fastq Sample280.fastq

```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ velveth Sample280_assembly 35 -shortPaired -fastq Sample280.fastq
[0.000000] Reading FastQ file Sample280.fastq;
[8.035066] 1769276 sequences found
[8.035084] Done
[8.035212] Reading read set file Sample280_assembly/Sequences;
[8.382253] 1769276 sequences found
[10.168766] Done
[10.168796] 1769276 sequences in total.
[10.169018] Writing into roadmap file Sample280_assembly/Roadmaps...
[11.092935] Inputting sequences...
[11.092972] Inputting sequence 0 / 1769276
[38.464658] Inputting sequence 1000000 / 1769276
[60.450223]  === Sequences loaded in 49.357296 s
[60.450303] Done inputting sequences
[60.450310] Destroying splay table
[60.504584] Splay table destroyed
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ ls
Sample280_assembly  Sample280.fasta  Sample280.fastq  Sample280_fastqc  Sample280_fastqc.zip
```

Step 6: Have a look inside Sample280_assembly directory, which contains three files, namely, Log, Roadmaps, and Sequences. Log file contains what commands you have used to get this assembly result and is useful for reproducing results later on

```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ cd Sample280_assembly
[ngswshop@quince-srv2 ~/workshop/assembly_test/Sample280_assembly]$ ls
Log  Roadmaps  Sequences
[ngswshop@quince-srv2 ~/workshop/assembly_test/Sample280_assembly]$ cat Log
Sun Jun 16 11:40:19 2013
 velveth Sample280_assembly 35 -shortPaired -fastq Sample280.fastq
Version 1.2.08
Copyright 2007, 2008 Daniel Zerbino (zerbino@ebi.ac.uk)
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
Compilation settings:
CATEGORIES = 2
MAXKMERLENGTH = 99

[ngswshop@quince-srv2 ~/workshop/assembly_test/Sample280_assembly]$ cd ..
```

Step 7: We will now run velvetg to create contigs: velvetg my_assembly_directory

```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ velvetg Sample280_assembly
[0.000000] Reading roadmap file Sample280_assembly/Roadmaps
[3.959825] 1769276 roadmaps read
[3.960255] Creating insertion markers
[4.631896] Ordering insertion markers
[7.241368] Counting preNodes
[7.651526] 1844350 preNodes counted, creating them now
[12.824422] Sequence 1000000 / 1769276
[16.044911] Adjusting marker info...
[16.466661] Connecting preNodes
[19.609723] Connecting 1000000 / 1769276
[22.795455] Cleaning up memory
```

```
[22.796235] Done creating preGraph
[22.796248] Concatenation...
[23.437325] Renumbering preNodes
[23.437360] Initial preNode count 1844350
[23.502439] Destroyed 1449365 preNodes
[23.502470] Concatenation over!
[23.502472] Clipping short tips off preGraph
[23.605406] Concatenation...
[23.785744] Renumbering preNodes
[23.785778] Initial preNode count 394985
[23.830982] Destroyed 77078 preNodes
[23.831017] Concatenation over!
[23.831019] 44263 tips cut off
[23.831021] 317907 nodes left
[23.831305] Writing into pregraph file Sample280_assembly/PreGraph...
[24.900370] Reading read set file Sample280_assembly/Sequences;
[25.243913] 1769276 sequences found
[27.034404] Done
[27.905219] Reading pre-graph file Sample280_assembly/PreGraph
[27.906468] Graph has 317907 nodes and 1769276 sequences
[28.437759] Scanning pre-graph file Sample280_assembly/PreGraph for k-mers
[28.634110] 9108783 kmers found
[29.540252] Sorting kmer occurence table ...
[37.151443] Sorting done.
[37.151475] Computing acceleration table...
[37.320944] Computing offsets...
[37.439130] Ghost Threading through reads 0 / 1769276
[90.497873] Ghost Threading through reads 1000000 / 1769276
[131.544984]  === Ghost-Threaded in 94.105854 s
[131.545021] Threading through reads 0 / 1769276
[186.600692] Threading through reads 1000000 / 1769276
[229.770908]  === Threaded in 98.225887 s
[229.782483] Correcting graph with cutoff 0.200000
[229.793932] Determining eligible starting points
[230.175998] Done listing starting nodes
[230.176030] Initializing todo lists
[230.254062] Done with initilization
```

```
[230.254095] Activating arc lookup table
[230.400269] Done activating arc lookup table
[230.517399] 10000 / 317907 nodes visited
[230.671419] 20000 / 317907 nodes visited
[230.869498] 30000 / 317907 nodes visited
[231.060229] 40000 / 317907 nodes visited
[231.218727] 50000 / 317907 nodes visited
[231.399032] 60000 / 317907 nodes visited
[231.585998] 70000 / 317907 nodes visited
[231.776716] 80000 / 317907 nodes visited
[231.910510] 90000 / 317907 nodes visited
[232.086366] 100000 / 317907 nodes visited
[232.274794] 110000 / 317907 nodes visited
[232.492884] 120000 / 317907 nodes visited
[232.717592] 130000 / 317907 nodes visited
[232.901985] 140000 / 317907 nodes visited
[233.055681] 150000 / 317907 nodes visited
[233.247460] 160000 / 317907 nodes visited
[233.373794] 170000 / 317907 nodes visited
[233.482261] 180000 / 317907 nodes visited
[233.579969] 190000 / 317907 nodes visited
[233.690244] 200000 / 317907 nodes visited
[233.807521] 210000 / 317907 nodes visited
[233.928592] 220000 / 317907 nodes visited
[234.063637] 230000 / 317907 nodes visited
[234.207627] 240000 / 317907 nodes visited
[234.350775] 250000 / 317907 nodes visited
[234.505313] 260000 / 317907 nodes visited
[234.664713] 270000 / 317907 nodes visited
[234.837260] 280000 / 317907 nodes visited
[235.019604] 290000 / 317907 nodes visited
[235.225146] 300000 / 317907 nodes visited
[235.373969] 310000 / 317907 nodes visited
[235.479852] 320000 / 317907 nodes visited
[235.516433] 330000 / 317907 nodes visited
[235.530404] 340000 / 317907 nodes visited
```

```
[235.531268] Concatenation...
[235.548231] Renumbering nodes
[235.548258] Initial node count 317907
[235.553804] Removed 225109 null nodes
[235.553839] Concatenation over!
[235.553842] Clipping short tips off graph, drastic
[235.677319] Concatenation...
[235.721235] Renumbering nodes
[235.721265] Initial node count 92798
[235.726243] Removed 28092 null nodes
[235.726260] Concatenation over!
[235.726262] 64706 nodes left
[235.726478] Writing into graph file Sample280_assembly/Graph...
[236.417174] WARNING: NO COVERAGE CUTOFF PROVIDED
[236.417194] Velvet will probably leave behind many detectable errors
[236.417197] See manual for instructions on how to set the coverage cutoff parameter
[236.417200] Removing contigs with coverage < -1.000000...
[236.421267] Concatenation...
[236.428782] Renumbering nodes
[236.428802] Initial node count 64706
[236.428892] Removed 0 null nodes
[236.428895] Concatenation over!
[236.432183] Concatenation...
[236.439032] Renumbering nodes
[236.439048] Initial node count 64706
[236.439138] Removed 0 null nodes
[236.439140] Concatenation over!
[236.439153] Clipping short tips off graph, drastic
[236.441575] Concatenation...
[236.448333] Renumbering nodes
[236.448351] Initial node count 64706
[236.448438] Removed 0 null nodes
[236.448440] Concatenation over!
[236.448442] 64706 nodes left
[236.448444] WARNING: NO EXPECTED COVERAGE PROVIDED
[236.448446] Velvet will be unable to resolve any repeats
[236.448448] See manual for instructions on how to set the expected coverage parameter
```

```
[236.448451] Concatenation...
[236.455218] Renumbering nodes
[236.455234] Initial node count 64706
[236.455319] Removed 0 null nodes
[236.455322] Concatenation over!
[236.455323] Removing reference contigs with coverage < -1.000000...
[236.459503] Concatenation...
[236.466314] Renumbering nodes
[236.466329] Initial node count 64706
[236.466417] Removed 0 null nodes
[236.466420] Concatenation over!
[236.476331] Writing contigs into Sample280_assembly/contigs.fa...
[236.971091] Writing into stats file Sample280_assembly/stats.txt...
[237.135531] Writing into graph file Sample280_assembly/LastGraph...
Final graph has 64706 nodes and n50 of 1081, max 10938, total 6016815, using 0/1769276 reads
[ngswshop@quince-srv2 ~/workshop/assembly_test]$
```

Step 8: By looking again at the Sample280_assembly, we find the contigs as contigs.fa, a textual representation of the contig graph as Graph, and a file containing statistics on each contig as stats.txt. Additionally, we find the N50 score of assembly as 1081

```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ ls Sample280_assembly/
contigs.fa  Graph  LastGraph  Log  PreGraph  Roadmaps  Sequences  stats.txt
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ cat Sample280_assembly/Log
Sun Jun 16 11:40:19 2013
 velveth Sample280_assembly 35 -shortPaired -fastq Sample280.fastq
Version 1.2.08
Copyright 2007, 2008 Daniel Zerbino (zerbino@ebi.ac.uk)
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
Compilation settings:
CATEGORIES = 2
MAXKMERLENGTH = 99
```

```
Sun Jun 16 12:10:25 2013
 velvetg Sample280_assembly
Version 1.2.08
Copyright 2007, 2008 Daniel Zerbino (zerbino@ebi.ac.uk)
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
Compilation settings:
CATEGORIES = 2
MAXKMERLENGTH = 99

Final graph has 64706 nodes and n50 of 1081, max 10938, total 6016815, using 0/1769276 reads
```

Step 9: We will now re-run velvetg with an optimized assembly. We need to set a new parameter, exp_cov . This parameter tells Velvet the mean coverage for non-repetitive parts of the genome you are assembling. With this information it can decide whether ambiguous parts in the network graph can be confidently traversed, or whether the sequence needs to be broken into contigs. You can find a very good explanation of exp_cov at http://www.homolog.us/blogs/blog/2012/06/08/an-explanation-of-velvet-parameter-exp_cov/

```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ velvetg Sample280_assembly -cov_cutoff auto -exp_cov auto
[0.000001] Reading read set file Sample280_assembly/Sequences;
[0.759457] 1769276 sequences found
[2.771274] Done
[3.674340] Reading pre-graph file Sample280_assembly/PreGraph
[3.674471] Graph has 317907 nodes and 1769276 sequences
[4.245224] Scanning pre-graph file Sample280_assembly/PreGraph for k-mers
[4.437157] 9108783 kmers found
[5.412444] Sorting kmer occurence table ...
[13.708229] Sorting done.
[13.708279] Computing acceleration table...
[13.940194] Computing offsets...
[14.056067] Ghost Threading through reads 0 / 1769276
[75.296141] Ghost Threading through reads 1000000 / 1769276
[122.827166]  === Ghost-Threaded in 108.771099 s
```

```
[122.827221] Threading through reads 0 / 1769276
[190.911470] Threading through reads 1000000 / 1769276
[243.875455]  === Threaded in 121.048235 s
[246.657643] Correcting graph with cutoff 0.200000
[246.677992] Determining eligible starting points
[247.062764] Done listing starting nodes
[247.062805] Initializing todo lists
[247.138506] Done with initilization
[247.138554] Activating arc lookup table
[247.292377] Done activating arc lookup table
[247.932368] 10000 / 317907 nodes visited
[248.708499] 20000 / 317907 nodes visited
[249.641791] 30000 / 317907 nodes visited
[250.278464] 40000 / 317907 nodes visited
[250.551107] 50000 / 317907 nodes visited
[250.859518] 60000 / 317907 nodes visited
[251.168683] 70000 / 317907 nodes visited
[251.488976] 80000 / 317907 nodes visited
[251.697516] 90000 / 317907 nodes visited
[251.963795] 100000 / 317907 nodes visited
[252.241659] 110000 / 317907 nodes visited
[252.577524] 120000 / 317907 nodes visited
[252.934726] 130000 / 317907 nodes visited
[253.221640] 140000 / 317907 nodes visited
[253.448758] 150000 / 317907 nodes visited
[253.742837] 160000 / 317907 nodes visited
[253.916664] 170000 / 317907 nodes visited
[254.078956] 180000 / 317907 nodes visited
[254.211596] 190000 / 317907 nodes visited
[254.351891] 200000 / 317907 nodes visited
[254.486830] 210000 / 317907 nodes visited
[254.616745] 220000 / 317907 nodes visited
[254.750382] 230000 / 317907 nodes visited
[254.892818] 240000 / 317907 nodes visited
[255.045215] 250000 / 317907 nodes visited
[255.212691] 260000 / 317907 nodes visited
[255.387274] 270000 / 317907 nodes visited
```

```
[255.576947] 280000 / 317907 nodes visited
[255.777861] 290000 / 317907 nodes visited
[256.002368] 300000 / 317907 nodes visited
[256.175985] 310000 / 317907 nodes visited
[256.290106] 320000 / 317907 nodes visited
[256.328767] 330000 / 317907 nodes visited
[256.343706] 340000 / 317907 nodes visited
[256.344934] Concatenation...
[256.361220] Renumbering nodes
[256.361250] Initial node count 317907
[256.366309] Removed 225109 null nodes
[256.366345] Concatenation over!
[256.366348] Clipping short tips off graph, drastic
[256.489077] Concatenation...
[256.570586] Renumbering nodes
[256.570633] Initial node count 92798
[256.575001] Removed 28092 null nodes
[256.575013] Concatenation over!
[256.575016] 64706 nodes left
[256.575435] Writing into graph file Sample280_assembly/Graph2...
[263.719138] Measuring median coverage depth...
[263.749931] Median coverage depth = 15.722191
[263.750026] Removing contigs with coverage < 7.861095...
[263.786854] Concatenation...
[264.859348] Renumbering nodes
[264.859392] Initial node count 64706
[264.859917] Removed 59621 null nodes
[264.859922] Concatenation over!
[264.860093] Concatenation...
[264.860476] Renumbering nodes
[264.860479] Initial node count 5085
[264.860487] Removed 0 null nodes
[264.860489] Concatenation over!
[264.860509] Clipping short tips off graph, drastic
[264.862329] Concatenation...
[264.894920] Renumbering nodes
[264.894971] Initial node count 5085
```

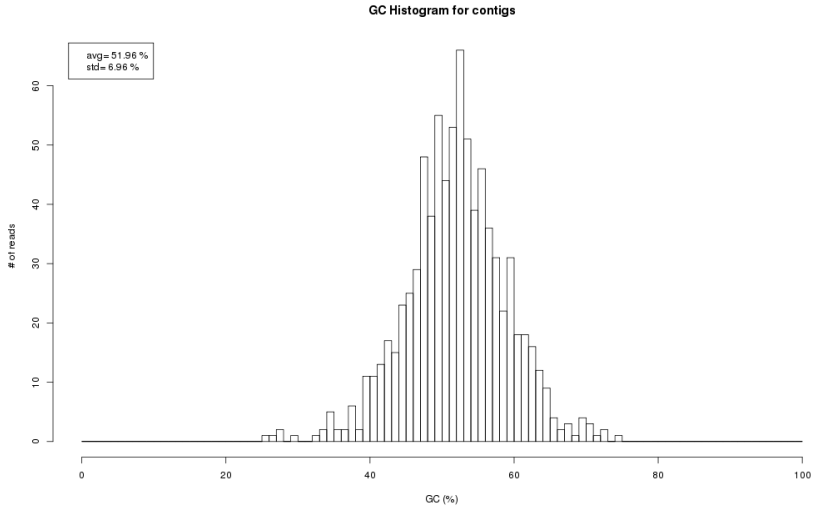
```
[264.895111] Removed 1620 null nodes
[264.895115] Concatenation over!
[264.895118] 3465 nodes left
[264.895122] Read coherency...
[264.895217] Identifying unique nodes
[264.895348] Done, 981 unique nodes counted
[264.895351] Trimming read tips
[264.895531] Renumbering nodes
[264.895535] Initial node count 3465
[264.895541] Removed 0 null nodes
[264.895543] Confronted to 0 multiple hits and 0 null over 0
[264.895546] Read coherency over!
[264.910946] Starting pebble resolution...
[264.923031] Computing read to node mapping array sizes
[264.977072] Computing read to node mappings
[265.575256] Estimating library insert lengths...
[265.713206] Paired-end library 1 has length: 444, sample standard deviation: 33
[265.713261] Done
[265.713322] Computing direct node to node mappings
[265.836511] Scaffolding node 0
[265.953769]  === Nodes Scaffolded in 0.240444 s
[265.980185] Preparing to correct graph with cutoff 0.200000
[266.187894] Cleaning memory
[266.187943] Deactivating local correction settings
[266.188002] Pebble done.
[266.188006] Starting pebble resolution...
[266.205387] Computing read to node mapping array sizes
[266.303886] Computing read to node mappings
[267.020103] Estimating library insert lengths...
[267.153167] Paired-end library 1 has length: 444, sample standard deviation: 34
[267.153213] Done
[267.153246] Computing direct node to node mappings
[267.271372] Scaffolding node 0
[267.383610]  === Nodes Scaffolded in 0.230361 s
[267.392252] Preparing to correct graph with cutoff 0.200000
[267.399091] Cleaning memory
[267.399099] Deactivating local correction settings
```

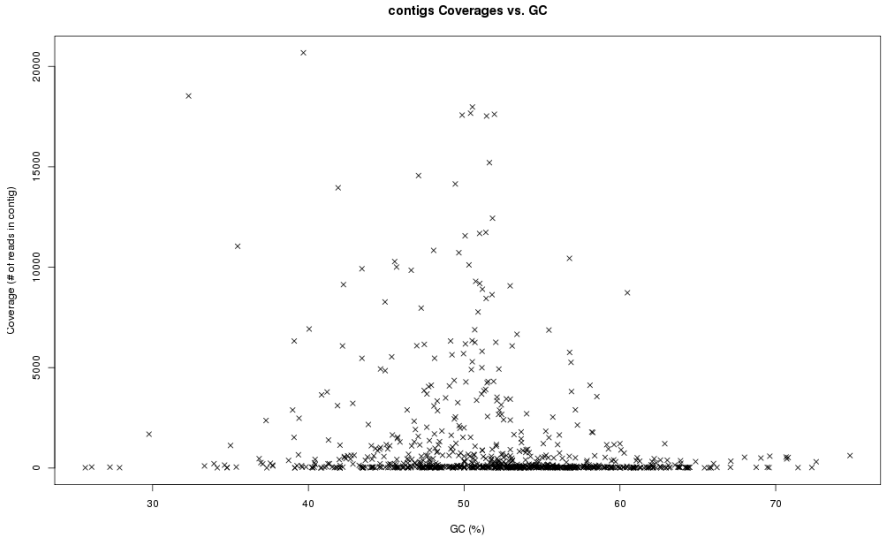
```
[267.399151] Pebble done.
[267.399156] Concatenation...
[267.421692] Renumbering nodes
[267.421713] Initial node count 3465
[267.421810] Removed 1408 null nodes
[267.421836] Concatenation over!
[267.421840] Removing reference contigs with coverage < 7.861095...
[267.421951] Concatenation...
[267.422071] Renumbering nodes
[267.422075] Initial node count 2057
[267.422080] Removed 0 null nodes
[267.422103] Concatenation over!
[267.434047] Writing contigs into Sample280_assembly/contigs.fa...
[268.304976] Writing into stats file Sample280_assembly/stats.txt...
[268.335988] Writing into graph file Sample280_assembly/LastGraph...
[270.215872] Estimated Coverage = 15.722191
[270.215986] Estimated Coverage cutoff = 7.861095
Final graph has 2057 nodes and n50 of 113713, max 258956, total 5303438, using 1691137/1769276 reads
[ngswshop@quince-srv2 ~/workshop/assembly_test]$
```

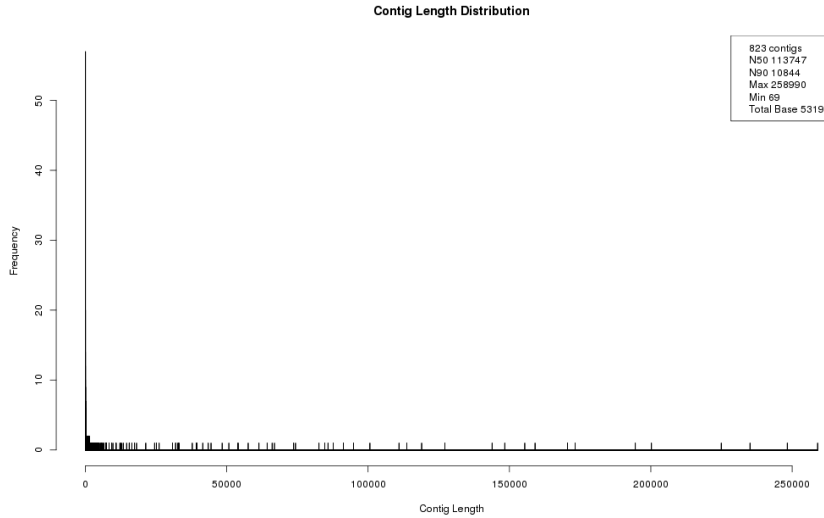
Step 10: We will now use some perl scripts (http://biostuff2010.googlecode.com/svn/trunk/perlscripts/) to look at the assemblies in detail. We will use the script contigs_stats.pl to generate five types of graphs (GC Histogram of contigs, contigs Coverges vs GC, Contig Length Distribution, Contigs Mean Coverage Vs. Contigs Length, and Contigs Coverage Vs. Length). These graphs are shown below.

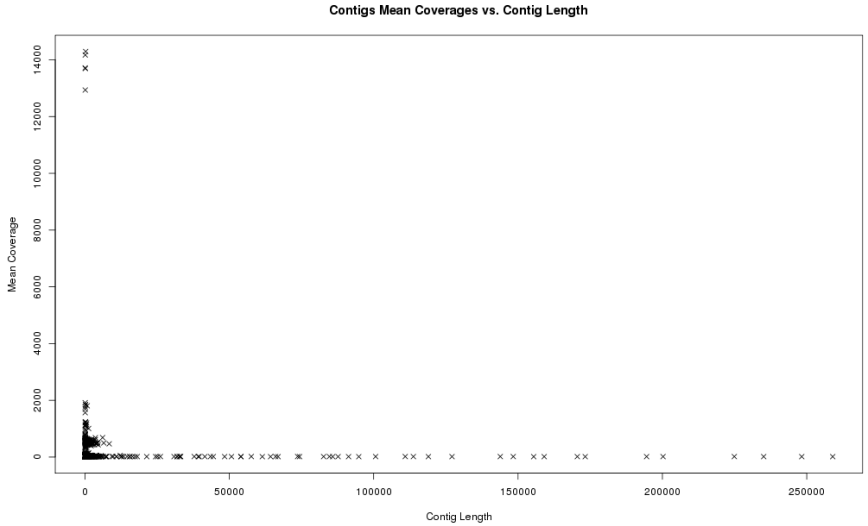
```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ perl /home/opt/perl_scripts/contigs_stats.pl -t Velvet
Sample280_assembly/contigs.fa -plot
Expected_coverage:      15.722191
Assembled_reads:  95.58%
Singleton:  78139
Contigs_number:   823
N50:  113747
N90:  10844
Max:  258990
```

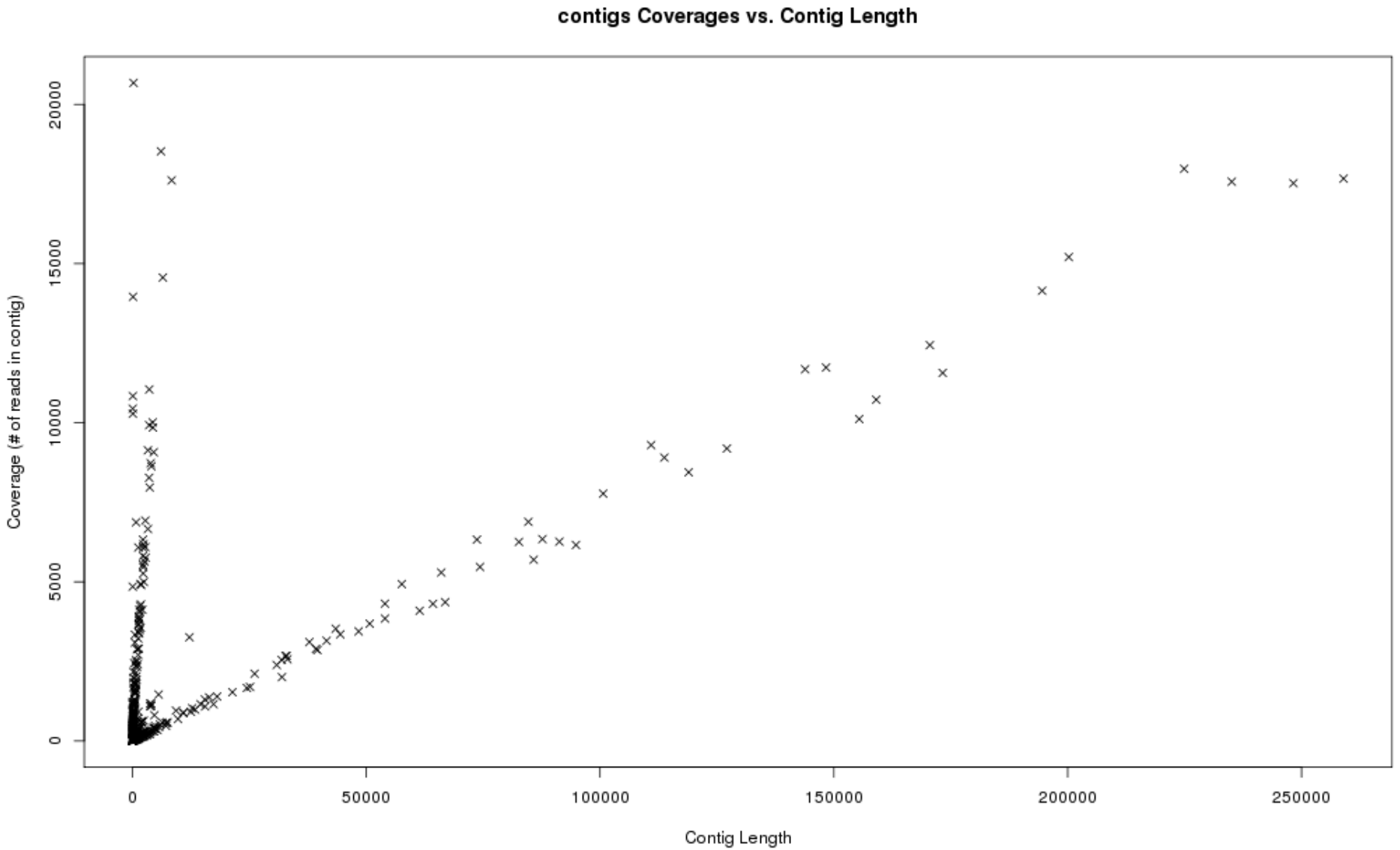
```
Min:  69
Total_bases:      5319373
Top10_bases:      1864839
Top20_bases:      3070009
Top40_bases:      4290477
Top100_bases:     4985788
>100kb_bases:     2883857
>50kb_bases:      4033831
>25kb_bases:      4573212
>10kb_bases:      4789218
>5kb_bases: 4887546
>3kb_bases: 5022637
>2kb_bases: 5102190
>1kb_bases: 5187404
>100kb_reads:     211958
>50kb_reads:      296184
>25kb_reads:      337178
>10kb_reads:      355801
>5kb_reads: 413670
>3kb_reads: 523877
>2kb_reads: 603627
>1kb_reads: 696166
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ ls
contigs_gc_hist.png           contigs_len_hist.png     Sample280_assembly  Sample280_fastqc
contigs_gc_vs_depth.png       contigs_len_vs_cov.png   Sample280.fasta     Sample280_fastqc.zip
contigs_len_gc_depth_cov.txt  contigslen_vs_depth.png  Sample280.fastq
```


GC Histogram for contigs
avg= 51.96 %
std= 6.96 %
# of reads
GC (%)


contigs Coverages vs. GC
Coverage (# of reads in contig)
GC (%)


Contig Length Distribution
823 contigs
N50 113747
N90 10844
Max 258990
Min 69
Total Base 5319373
Frequency
Contig Length


Contigs Mean Coverages vs. Contig Length
Mean Coverage
Contig Length


contigs Coverages vs. Contig Length
Coverage (# of reads in contig)
0
5000
10000
15000
20000
0
50000
100000
150000
200000
250000
Contig Length

Step 11: Next, from the same website, we will use contig_size_select.pl to extract contigs above 200000 bp length and annotate them using prokka (http://www.vicbioinformatics.com/prokka-manual.html). Prokka is a contraction of "prokaryotic annotation". It is specifically designed for Bacteria, Archaea and Viruses. It can't handle multi-exon gene models; you can use MAKER 2 for that purpose. Prokka pipeline uses Aragorn to find transfer RNA features (tRNA), RNAmmer to find ribosomal RNA features (rRNA), Prodigal to find protein-coding features (CDS), SignalP to find signal peptide features in CDS, BLAST+ to find similarity searching against protein sequence libraries, HMMER3 for similarity searching against protein family profiles, and Infernal for similarity searching against ncRNA family profiles.

```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ perl /home/opt/perl_scripts/contig_size_select.pl  -low 200000 -
high 290000 Sample280_assembly/contigs.fa > contigs_filtered.fa
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ mkdir annotation
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ mv contigs_filtered.fa annotation/.
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ cd annotation
[ngswshop@quince-srv2 ~/workshop/assembly_test/annotation]$ prokka contigs_filtered.fa
[13:14:33] This is prokka 1.5.2
[13:14:33] Written by Torsten Seemann <torsten.seemann@monash.edu>
[13:14:33] Victorian Bioinformatics Consortium - http://www.vicbioinformatics.com
[13:14:33] Local time is Sun Jun 16 13:14:33 2013
[13:14:33] You are ngswshop
[13:14:33] Annotating as >>> Bacteria <<<
[13:14:33] Creating new output folder: PROKKA_06162013
[13:14:33] Running: mkdir -p PROKKA_06162013
[13:14:33] Using filename prefix: PROKKA_06162013.XXX
[13:14:33] Writing log to: PROKKA_06162013/PROKKA_06162013.log
[13:14:33] Command: contigs_filtered.fa
[13:14:33] Need 'less' - using /usr/bin/less
[13:14:33] Need 'grep' - using /bin/grep
[13:14:33] Need 'egrep' - using /bin/egrep
[13:14:33] Need 'sed' - using /bin/sed
[13:14:33] Need 'find' - using /bin/find
[13:14:33] Need 'tbl2asn' - using /home/opt/tbl2asn_linux/tbl2asn
[13:14:33] Need 'makeblastdb' - using /home/opt/ncbi-blast-2.2.28+/bin/makeblastdb
```

```
[13:14:33] Need 'blastp' - using /home/opt/ncbi-blast-2.2.28+/bin/blastp
[13:14:33] Need 'aragorn' - using /home/opt/aragorn1.2.36/aragorn
[13:14:33] Need 'prodigal' - using /home/opt/prodigal.v2_60/prodigal
[13:14:33] Need 'rnammer' - using /home/opt/rnammer-1.2/rnammer
[13:14:33] Need 'parallel' - using /usr/local/bin/parallel
[13:14:33] Need 'hmmscan' - using /usr/local/bin/hmmscan
[13:14:33] Need 'cmscan' - using /usr/local/bin/cmscan
[13:14:33] Determined aragorn version is 1.2.36
[13:14:33] Determined prodigal version is 2.60
[13:14:33] Determined rnammer version is 1.2
[13:14:33] Determined infernal version is 1.1
[13:14:34] Determined signalp version is 4.1
[13:14:34] Using genetic code table 11.
[13:14:34] Loading and checking input file: contigs_filtered.fa
[13:14:34] Wrote 5 contigs
[13:14:34] Predicting tRNAs and tmRNAs
[13:14:34] Running: aragorn -gc11 -w PROKKA_06162013/PROKKA_06162013.fna
[13:14:35] 1 tRNA-Met [60618,60694] 35 (cat)
[13:14:35] 2 tRNA-Met [60727,60803] 35 (cat)
[13:14:35] 3 tRNA-Gly c[119146,119219] 33 (ccc)
[13:14:35] 4 tRNA-Phe [234679,234754] 34 (gaa)
[13:14:35] 1 tRNA-Ala c[48349,48424] 34 (ggc)
[13:14:35] 2 tRNA-Ala c[48464,48539] 34 (ggc)
[13:14:35] 3 tRNA-Val [51239,51314] 34 (tac)
[13:14:35] 4 tRNA-Val [51359,51434] 34 (tac)
[13:14:35] 5 tRNA-Lys [51439,51514] 34 (ttt)
[13:14:35] 1 tRNA-Pro [204891,204967] 35 (ggg)
[13:14:35] Found 10 tRNAs
[13:14:35] Predicting Ribosomal RNAs
[13:14:35] Running: rnammer -S bac -multi -xml PROKKA_06162013/rnammer.xml PROKKA_06162013/PROKKA_06162013.fna
[13:14:38] Deleting temporary file: PROKKA_06162013/rnammer.xml
[13:14:38] Found 0 rRNAs
[13:14:38] Disabling ncRNA search, can't find /home/opt/prokka-1.5.2/bin/../db/cm/Bacteria file.
[13:14:38] Total of 10 RNA features
[13:14:38] Predicting coding sequences
[13:14:38] Contigs total 1167462 bp, so using single mode
[13:14:38] Running: prodigal -i PROKKA_06162013/PROKKA_06162013.fna -c -m -g 11 -p single -f sco -q
```

```
[13:14:42] Found 1043 CDS
[13:14:42] Connecting features back to sequences
[13:14:42] Option --gram not specified, will NOT check for signal peptides.
[13:14:42] Not using genus-specific database. Try --usegenus to enable it.
[13:14:42] Annotating CDS, please be patient.
[13:14:42] Will use all available CPUs for similarity searching.
[13:14:43] blast 1043 (of 1043) proteins against /home/opt/prokka-1.5.2/bin/../db/kingdom/Bacteria/sprot
[13:14:43] Running: nice parallel blastp -query {} -db /home/opt/prokka-1.5.2/bin/../db/kingdom/Bacteria/sprot -
evalue 1e-06 -num_threads 1 -out {}.out -num_descriptions 1 -num_alignments 1 2>/dev/null ::: PROKKA_06162013/*.seq
[13:14:48] Modify product: Uncharacterized peptidase SA1530 => putative peptidase
[13:14:48] Modify product: Uncharacterized FAD-linked oxidoreductase Rv2280 => putative FAD-linked oxidoreductase
[13:14:48] Modify product: Uncharacterized ABC transporter ATP-binding protein YheS => putative ABC transporter ATP-
binding protein YheS
[13:14:48] Modify product: Uncharacterized ABC transporter ATP-binding protein YbhF => putative ABC transporter ATP-
binding protein YbhF
[13:14:48] Modify product: Uncharacterized sufE-like protein ygdK => putative sufE-like protein ygdK
[13:14:48] Modify product: Probable 3-phenylpropionic acid transporter => putative 3-phenylpropionic acid
transporter
[13:14:49] Modify product: Probable GTPase ArgK => putative GTPase ArgK
[13:14:49] Modify product: Probable sensor-like histidine kinase YehU => putative sensor-like histidine kinase YehU
[13:14:49] Modify product: Probable copper-binding protein pcoE precursor => putative copper-binding protein pcoE
precursor
[13:14:49] Modify product: Probable adenosine monophosphate-protein transferase fic => putative adenosine
monophosphate-protein transferase fic
[13:14:49] Modify product: Uncharacterized ferredoxin-like protein yfhL => putative ferredoxin-like protein yfhL
[13:14:49] Modify product: Uncharacterized metalloprotease yggG => putative metalloprotease yggG
[13:14:49] Modify product: DnaA-homolog protein hda => hypothetical protein
[13:14:49] Modify product: Uncharacterized metalloprotease yggG => putative metalloprotease yggG
[13:14:49] Modify product: Uncharacterized HTH-type transcriptional regulator yegW => putative HTH-type
transcriptional regulator yegW
[13:14:50] Modify product: Phosphotriesterase homology protein => hypothetical protein
[13:14:50] Modify product: Probable transcriptional regulatory protein YehT => putative transcriptional regulatory
protein YehT
[13:14:50] Modify product: Probable deferrochelatase/peroxidase YfeX => putative deferrochelatase/peroxidase YfeX
[13:14:50] Modify product: Probable endopeptidase Spr precursor => putative endopeptidase Spr precursor
[13:14:50] Modify product: Putative membrane protein igaA homolog => hypothetical protein
```

```
[13:14:50] Modify product: Probable poly(glycerol-phosphate) alpha-glucosyltransferase => putative poly(glycerol-
phosphate) alpha-glucosyltransferase
[13:14:50] Modify product: Probable HTH-type transcriptional regulator ygaV => putative HTH-type transcriptional
regulator ygaV
[13:14:51] Modify product: Uncharacterized lipoprotein ygdR precursor => putative lipoprotein ygdR precursor
[13:14:51] Modify product: Probable oxalyl-CoA decarboxylase => putative oxalyl-CoA decarboxylase
[13:14:51] Modify product: Uncharacterized oxidoreductase yhhX => putative oxidoreductase yhhX
[13:14:51] Modify product: PGL/p-HBAD biosynthesis glycosyltransferase Rv2957/MT3031 => PGL/p-HBAD biosynthesis
glycosyltransferase/MT3031
[13:14:51] Modify product: Probable Fe(2+)-trafficking protein => putative Fe(2+)-trafficking protein
[13:14:51] Modify product: Probable tRNA-dihydrouridine synthase => putative tRNA-dihydrouridine synthase
[13:14:51] Modify product: Uncharacterized GTP-binding protein YjiA => putative GTP-binding protein YjiA
[13:14:51] Modify product: Uncharacterized lipoprotein yehR precursor => putative lipoprotein yehR precursor
[13:14:51] Modify product: Uncharacterized HTH-type transcriptional regulator ypdC => putative HTH-type
transcriptional regulator ypdC
[13:14:51] Modify product: Uncharacterized lipoprotein ygdR precursor => putative lipoprotein ygdR precursor
[13:14:51] Modify product: Uncharacterized protease yhbU precursor => putative protease yhbU precursor
[13:14:51] Cleaned 33 /product names
[13:14:52] hmmer3 202 (of 1043) proteins against /home/opt/prokka-1.5.2/bin/../db/hmm/CLUSTERS.hmm
[13:14:52] Running: nice parallel hmmscan --noali --notextw --acc -E 1e-06 --cpu 1 -o {}.out /home/opt/prokka-
1.5.2/bin/../db/hmm/CLUSTERS.hmm {} 2>/dev/null ::: PROKKA_06162013/*.seq
[13:14:55] hmmer3 125 (of 1043) proteins against /home/opt/prokka-1.5.2/bin/../db/hmm/Cdd.hmm
[13:14:55] Running: nice parallel hmmscan --noali --notextw --acc -E 1e-06 --cpu 1 -o {}.out /home/opt/prokka-
1.5.2/bin/../db/hmm/Cdd.hmm {} 2>/dev/null ::: PROKKA_06162013/*.seq
[13:14:58] Modify product: Uncharacterized protein with protein kinase and helix-hairpin-helix DNA-binding domains
=> putative protein with protein kinase and helix-hairpin-helix DNA-binding domains
[13:14:58] Modify product: Uncharacterized protein encoded in hypervariable junctions of pilus gene clusters =>
putative protein encoded in hypervariable junctions of pilus gene clusters
[13:14:58] Modify product: Uncharacterized conserved protein => hypothetical protein
[13:14:58] Modify product: Uncharacterized protein conserved in bacteria => hypothetical protein
[13:14:58] Modify product: Predicted metalloprotease => putative metalloprotease
[13:14:58] Modify product: Glucoamylase and related glycosyl hydrolases => Glucoamylase hydrolases
[13:14:58] Modify product: Uncharacterized protein conserved in bacteria => hypothetical protein
[13:14:58] Modify product: Uncharacterized protein conserved in bacteria => hypothetical protein
[13:14:59] Modify product: Uncharacterized protein conserved in bacteria => hypothetical protein
```

```
[13:14:59] Modify product: Uncharacterized protein encoded in toxicity protection region of plasmid R478, contains
von Willebrand factor (vWF) domain => putative protein encoded in toxicity protection region of plasmid R478,
contains von Willebrand factor (vWF) domain
[13:14:59] Modify product: Uncharacterized conserved protein => hypothetical protein
[13:14:59] Modify product: Predicted DNA-binding protein with PD1-like DNA-binding motif => putative DNA-binding
protein with PD1-like DNA-binding motif
[13:14:59] Modify product: Predicted secreted protein => putative secreted protein
[13:14:59] Modify product: ABC-type cobalt transport system, permease component CbiQ and related transporters =>
ABC-type cobalt transport system, permease component CbiQ
[13:14:59] Cleaned 14 /product names
[13:14:59] hmmer3 108 (of 1043) proteins against /home/opt/prokka-1.5.2/bin/../db/hmm/TIGRFAMs.hmm
[13:14:59] Running: nice parallel hmmscan --noali --notextw --acc -E 1e-06 --cpu 1 -o {}.out /home/opt/prokka-
1.5.2/bin/../db/hmm/TIGRFAMs.hmm {} 2>/dev/null ::: PROKKA_06162013/*.seq
[13:15:01] Modify product: PRD domain protein, EF_0829/AHA_3910 family => PRD domain protein,/AHA_3910 family
[13:15:01] Modify product: conserved hypothetical protein => hypothetical protein
[13:15:01] Cleaned 2 /product names
[13:15:01] hmmer3 101 (of 1043) proteins against /home/opt/prokka-1.5.2/bin/../db/hmm/Pfam.hmm
[13:15:01] Running: nice parallel hmmscan --noali --notextw --acc -E 1e-06 --cpu 1 -o {}.out /home/opt/prokka-
1.5.2/bin/../db/hmm/Pfam.hmm {} 2>/dev/null ::: PROKKA_06162013/*.seq
[13:15:04] Modify product: ERF superfamily => ERF superfamily protein
[13:15:04] Modify product: Protein of unknown function (DUF1175) => hypothetical protein
[13:15:04] Modify product: Glycine zipper 2TM domain => Glycine zipper 2TM domain protein
[13:15:04] Modify product: Protein of unknown function (DUF1456) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF3816) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF3300) => hypothetical protein
[13:15:04] Modify product: Alanine racemase, N-terminal domain => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF554) => hypothetical protein
[13:15:04] Modify product: WGR domain => WGR domain protein
[13:15:04] Modify product: Protein of unknown function (DUF2545) => hypothetical protein
[13:15:04] Modify product: YGGT family => YGGT family protein
[13:15:04] Modify product: Protein of unknown function (DUF2500) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF2737) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF2684) => hypothetical protein
[13:15:04] Modify product: MG2 domain => MG2 domain protein
[13:15:04] Modify product: Protein of unknown function (DUF1434) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function DUF2620 => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF2856) => hypothetical protein
```

```
[13:15:04] Modify product: Protein of unknown function (DUF2502) => hypothetical protein
[13:15:04] Modify product: Alpha-2-macroglobulin family N-terminal region => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF1176) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF2933) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF2531) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF1202) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF2542) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF550) => hypothetical protein
[13:15:04] Modify product: Putative transcription regulator (DUF1323) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF2574) => hypothetical protein
[13:15:04] Modify product: Domain of unknown function (DUF477) => hypothetical protein
[13:15:04] Modify product: Protein of unknown function (DUF2633) => hypothetical protein
[13:15:04] Modify product: Predicted permease => putative permease
[13:15:04] Cleaned 31 /product names
[13:15:04] Labelling remaining 59 proteins as 'hypothetical protein'
[13:15:04] Possible /pseudo 'Gamma-glutamyltranspeptidase precursor' at gnl|VBC|contig000001 position 107430
[13:15:04] Possible /pseudo 'Gluconate utilization system GNT-I transcriptional repressor' at gnl|VBC|contig000001
position 114100
[13:15:04] Possible /pseudo '1,4-alpha-glucan branching enzyme GlgB' at gnl|VBC|contig000001 position 119913
[13:15:04] Possible /pseudo 'GTP pyrophosphokinase' at gnl|VBC|contig000002 position 25377
[13:15:04] Possible /pseudo 'Glucarate dehydratase-related protein' at gnl|VBC|contig000002 position 33792
[13:15:04] Possible /pseudo 'tRNA pseudouridine synthase C' at gnl|VBC|contig000002 position 36857
[13:15:04] Possible /pseudo 'Surface presentation of antigens protein spaS' at gnl|VBC|contig000002 position 113730
[13:15:04] Possible /pseudo 'Surface presentation of antigens protein spaS' at gnl|VBC|contig000002 position 113894
[13:15:04] Possible /pseudo 'type III secretion system protein SpaO' at gnl|VBC|contig000002 position 116805
[13:15:04] Possible /pseudo 'Methylmalonyl-CoA mutase' at gnl|VBC|contig000002 position 181966
[13:15:04] Possible /pseudo 'Glutathione synthetase' at gnl|VBC|contig000002 position 216691
[13:15:04] Possible /pseudo 'Sulfate transport system permease protein CysW' at gnl|VBC|contig000004 position 70775
[13:15:04] Possible /pseudo 'Multidrug transporter MdtC' at gnl|VBC|contig000005 position 77291
[13:15:04] Possible /pseudo 'D-tagatose-1,6-bisphosphate aldolase subunit gatZ' at gnl|VBC|contig000005 position
93617
[13:15:04] Found 776 unique /gene codes.
[13:15:04] Fixed 44 colliding /gene names.
[13:15:04] Assigned 1053 locus_tags to CDS and RNA features.
[13:15:04] Writing outputs to PROKKA_06162013/
[13:15:05] Generating Genbank and Sequin files
```

```
[13:15:05] Running: tbl2asn -N 1 -y 'Annotated using prokka 1.5.2 from http://www.vicbioinformatics.com' -Z
PROKKA_06162013/PROKKA_06162013.err -M n -V b -i PROKKA_06162013/PROKKA_06162013.fsa -f
PROKKA_06162013/PROKKA_06162013.tbl 2> /dev/null
[13:15:07] Deleting temporary file: PROKKA_06162013/errorsummary.val
[13:15:07] Deleting temporary file: PROKKA_06162013/PROKKA_06162013.dr
[13:15:07] Deleting temporary file: PROKKA_06162013/PROKKA_06162013.fixedproducts
[13:15:07] Deleting temporary file: PROKKA_06162013/PROKKA_06162013.ecn
[13:15:07] Deleting temporary file: PROKKA_06162013/PROKKA_06162013.val
[13:15:07] Output files:
[13:15:07] PROKKA_06162013/PROKKA_06162013.faa
[13:15:07] PROKKA_06162013/PROKKA_06162013.gff
[13:15:07] PROKKA_06162013/PROKKA_06162013.fsa
[13:15:07] PROKKA_06162013/PROKKA_06162013.fna
[13:15:07] PROKKA_06162013/PROKKA_06162013.tbl
[13:15:07] PROKKA_06162013/PROKKA_06162013.err
[13:15:07] PROKKA_06162013/PROKKA_06162013.sqn
[13:15:07] PROKKA_06162013/PROKKA_06162013.ffn
[13:15:07] PROKKA_06162013/PROKKA_06162013.gbk
[13:15:07] PROKKA_06162013/PROKKA_06162013.log
[13:15:07] Walltime used: 0.57 minutes
[13:15:07] Thank you, come again.
[ngswshop@quince-srv2 ~/workshop/assembly_test/annotation]$
```

Step 12: You can look at PROKKA_06162013.gbk inside the folder PROKKA_06162013 to see the annotations in genbank format

```
[ngswshop@quince-srv2 ~/workshop/assembly_test/annotation]$ head -50 PROKKA_06162013/PROKKA_06162013.gbk
LOCUS       contig000001          224906 bp    DNA     linear       16-JUN-2013
DEFINITION  Genus species strain strain.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      Genus species
  ORGANISM  Genus species
            Unclassified.
```

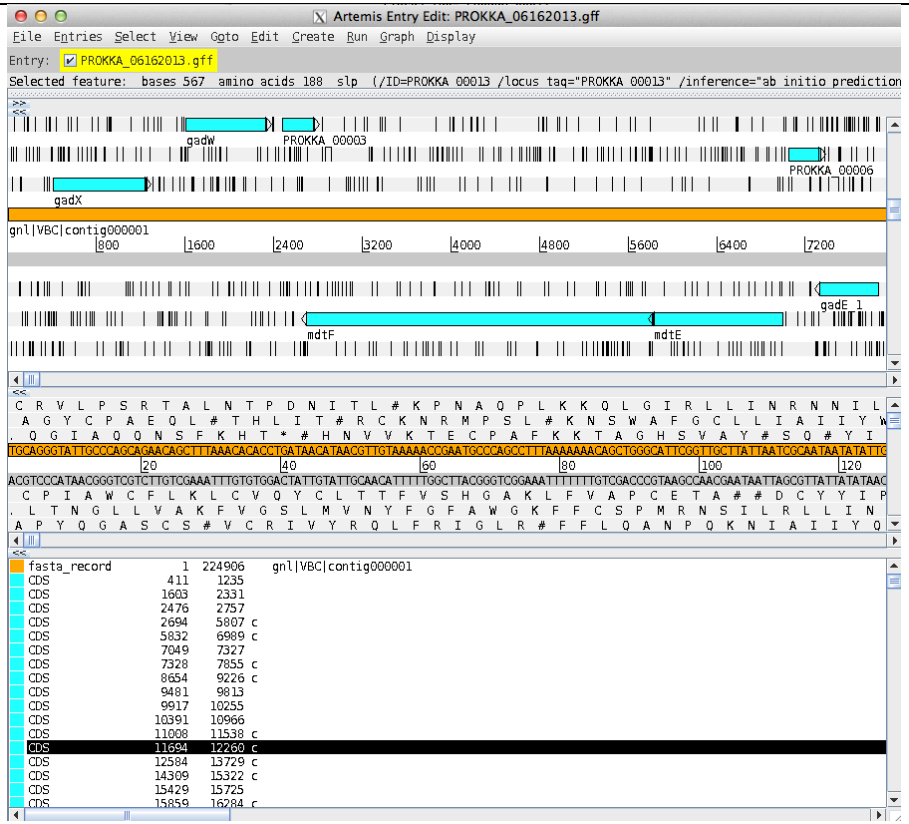
```
COMMENT     Annotated using prokka 1.5.2 from http://www.vicbioinformatics.com.
FEATURES             Location/Qualifiers
     source          1..224906
                     /organism="Genus species"
                     /mol_type="genomic DNA"
                     /strain="strain"
     CDS             411..1235
                     /gene="gadX"
                     /locus_tag="PROKKA_00001"
                     /inference="ab initio prediction:Prodigal:2.60"
                     /inference="similar to AA sequence:UniProtKB:P37639"
                     /codon_start=1
                     /transl_table=11
                     /product="HTH-type transcriptional regulator gadX"
                     /protein_id="VBC:PROKKA_00001"
                     /translation="MQSLHGNCLIAYARHKYILTMVNGEYRYFNGGDLVFADASQIRV
                     DKCVENFVLVSRDTLSLFLPMLKEEALNLHAHKKISSLLVHHCSRDIPVFQEVAQLSQ
                     NKNLRYAEMLRKRALIFALLSVFLEDEHFIPLLLNVLQPNMRTRVCTVINNNIAHEWT
                     LARIASELLMSPSLLKKKLREEETSYSQLLTECRMQRALQLIVIHGFSIKRVAVSCGY
                     HSVSYFIYVFRNYYGMTPTEYQERSAQGLPNRDSAASIVAQGNFYGTDRSAEGIRL"
     CDS             1603..2331
                     /gene="gadW"
                     /locus_tag="PROKKA_00002"
                     /inference="ab initio prediction:Prodigal:2.60"
                     /inference="similar to AA sequence:UniProtKB:P63201"
                     /codon_start=1
                     /transl_table=11
                     /product="HTH-type transcriptional regulator gadW"
                     /protein_id="VBC:PROKKA_00002"
                     /translation="MTHVCSVILIRRSFDIYHEQHKISLHNESIVLLEKNLADDFAFC
                     SPDTRRLDIDELTVCHYLQNIRQLPRNLGLHSKDRLLINQSPPMPLVTAIFDSFNESG
                     VNSPILSNMLYLSCLSMFSHKKELIPLLFNSISTVSGKVERLISFDIAKRWYLRDIAE
                     RMYTSESLIKKKLQDENTCFSKILLASRMSMARRLLELRQIPLHTIAEKCGYSSTSYF
                     INTFRQYYGVTPHQFAQHSPGTFS"
     CDS             2476..2757
                     /locus_tag="PROKKA_00003"
                     /inference="ab initio prediction:Prodigal:2.60"
```

```
                     /codon_start=1
                     /transl_table=11
                     /product="hypothetical protein"
                     /protein_id="VBC:PROKKA_00003"
                     /translation="MFGIIKLTIHTITGMWVSIVLFKLMTNGWSGFYFQCCVLSLVFL
                     TVSWLLSGEWLAGKSKAEPSRSTLLSFTRYAFLKRAKRCSTTTKKTGTK"
     CDS             complement(2694..5807)
                     /gene="mdtF"
                     /locus_tag="PROKKA_00004"
                     /inference="ab initio prediction:Prodigal:2.60"
                     /inference="similar to AA sequence:UniProtKB:P37637"
                     /codon_start=1
                     /transl_table=11
                     /product="Multidrug resistance protein MdtF"
                     /protein_id="VBC:PROKKA_00004"
                     /translation="MANYFIDRPVFAWVLAIIMMLAGGLAIMNLPVAQYPQIAPPTIT
                     VSATYPGADAQTVEDSVTQVIEQNMNGLDGLMYMSSTSDAAGNASITLTFETGTSPDI
                     AQVQVQNKLQLAMPSLPEAVQQQGISVDKSSSNILMVAAFISDNGSLNQYDIADYVAS
                     NIKDPLSRTAGVGSVQLFGSEYAMRIWLDPQKLNKYNLVPSDVISQIKVQNNQISGGQ
                     LGGMPQAADQQLNASIIVQTRLQTPEEFGKILLKVQQDGSQVLLRDVARVELGAEDYS
                     TVARYNGKPAAGIAIKLATGANALDTSRAVKEELNRLSAYFPASLKTVYPYDTTPFIE
                     ISIQEVFKTLVEAIILVFLVMYLFLQNFRATIIPTIAVPVVILGTFAILSAVGFTINT
                     LTMFGMVLAIGLLVDDAIVVVENVERVIAEDKLPPKEATHKSMGQIQRALVGIAVVLS
                     AVFMPMAFMSGATGEIYRQFSITLISSMLLSVFVAMSLTPALCATILKAAPEGGHKPN
                     ALFARFNTLFEKSTQHYTDSTRSLLRCTGRYMVVYLLICAGMAVLFLRTPTSFLPEED
                     QGVFMTTAQLPSGATMVNTTKVLQQVTDYYLTKEKDNVQSVFTVGGFGFSGQGQNNGL
                     AFISLKPWSERVGEENSVTAIIQRAMIALSSINKAVVFPFNLPAVAELGTASGFDMEL
                     LDNGNLGHEKLTQARNELLSLAAQSPDQVTGVRPNGLEDTPMFKVNVNAAKAEAMGVA
                     LSDINQTISTAFGSSYVNDFLNQGRVKKVYVQAGTPFRMLPDNINQWYVRNASGTMAP
                     LSAYSSTEWTYGSPRLERYNGIPSMEILGEAAAGKSTGDAMKFMADLVAKLPAGVGYS
                     WTGLSYQEALSSNQAPALYAISLVVVFLALAALYESWSIPFSVMLVVPLGVVGALLAT
                     DLRGLSNDVYFQVGLLTTIGLSAKNAILIVEFAVEMMQKEGKTPIEAIIEAARMRLRP
                     ILMTSLAFILGVLPLVISHGAGSGAQNAVGTGVMGGMFAATVLAIYFVPVFFVVVEHL
                     FARFKKA"
     CDS             complement(5832..6989)
                     /gene="mdtE"
                     /locus_tag="PROKKA_00005"
```

```
                     /inference="ab initio prediction:Prodigal:2.60"
                     /inference="similar to AA sequence:UniProtKB:P37636"
                     /codon_start=1
                     /transl_table=11
                     /product="Multidrug resistance protein MdtE precursor"
                     /protein_id="VBC:PROKKA_00005"
                     /translation="MNRRRKLLIPLLFCGAMLTACDDKSAENAAAMTPEVGVVTLSPG
                     SVNVLSELPGRTVPYEVAEIRPQVGGIIIKRNFIEGDKVNQGDSLYQIDPAPLQAELN
                     SAKGSLAKALSTASNARITFNRQASLLKTNYVSRQDYDTARTQLNEAEANVTVAKAAV
                     EQATINLQYANVTSPITGVSGKSSVTVGALVTANQADSLVTVQRLDPIYVDLTQSVQD
                     FLRMKEEVASGQIKQVQGSTPVQLNLENGKRYSQTGTLKFSDPTVDETTGSVTLRAIF
                     PNPNGDLLPGMYVTALVDEGSRQNVLLVPQEGVTHNAQGKATALILDKDDVVQLREIE
                     ASKAIGDQWVVTSGLQAGDRVIVSGLQRIRPGIKARAISSSQENASTESKQ"
     CDS             7049..7327
                     /locus_tag="PROKKA_00006"
                     /inference="ab initio prediction:Prodigal:2.60"
                     /codon_start=1
                     /transl_table=11
                     /product="hypothetical protein"
                     /protein_id="VBC:PROKKA_00006"
                     /translation="MRDLQTSGIVGLSASKVGYRYSDRHGEVNDKKNVAVLPTVPASI
                     RATKGSTGGYTYQGNKVVLASPLVITGGNEISICIPRHASHQFQLLMS"
     CDS             complement(7328..7855)
                     /gene="gadE_1"
                     /locus_tag="PROKKA_00007"
                     /inference="ab initio prediction:Prodigal:2.60"
                     /inference="similar to AA sequence:UniProtKB:P63204"
                     /codon_start=1
                     /transl_table=11
                     /product="Transcriptional regulator gadE"
                     /protein_id="VBC:PROKKA_00007"
                     /translation="MIFLMTKDSFLLQGFWQLKDNHEMIKINSLSEIKKVGNKPFKVI
                     IDTYHNHILDEEAIKFLEKLDAERIIVLAPYHISKLKAKAPIYFVSRKESIKNLLEIT
                     YGKHLPHKNSQLCFSHNQFKIMQLILKNKNESNITSTLNISQQTLKIQKFNIMYKLKL
                     RRMSDIVTLGITSYF"
[ngswshop@quince-srv2 ~/workshop/assembly_test/annotation]$
```

Step 13: In the same folder, we can find PROKKA_06162013.gff file which can then be loaded to Artemis (http://www.sanger.ac.uk/resources/software/artemis/) to visualize the annotations. For understanding the GFF format, visit http://www.sequenceontology.org/gff3.shtml

```
[ngswshop@quince-srv2 ~/workshop/assembly_test/annotation]$ art PROKKA_06162013/PROKKA_06162013.gff
starting Artemis with flags: -mx500m -ms20m -noverify -Dartemis.environment=UNIX -DproxySet=true -
Dhttp.proxyHost=wwwcache.gla.ac.uk -Dhttp.proxyPort=8080
```



Step 14: To be sure that we have ECOLI, we will use TAXAassign (http://userweb.eng.gla.ac.uk/umer.ijaz/TAXAassign_tutorial.pdf) to find which organsims these contigs correspond to.

```
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ mkdir assignment
[ngswshop@quince-srv2 ~/workshop/assembly_test]$ cd assignment
[ngswshop@quince-srv2 ~/workshop/assembly_test/assignment]$ cp ../annotation/contigs_filtered.fa .
[ngswshop@quince-srv2 ~/workshop/assembly_test/assignment]$ /home/opt/TAXAassign_v0.2/TAXAassign.sh -r 10 -l species
-f contigs_filtered.fa
TAXAassign v0.2. Copyright (c) 2013 Computational Microbial Genomics Group, University of Glasgow, UK
[2013-06-16 13:51:44] Using  /home/opt/ncbi-blast-2.2.28+/bin/blastn
[2013-06-16 13:51:44] Using  /home/opt/TAXAassign_v0.2/scripts/blast_concat_taxon.py
[2013-06-16 13:51:44] STEP 1: Blast against NCBI's nt database with minimum percent ident of 97, maximum of 10
reference sequences, and evalue of 0.0001 in blastn.
[2013-06-16 13:52:15] blastn took 31 seconds for contigs_filtered.fa.
[2013-06-16 13:52:15] contigs_filtered_B.out generated successfully!
[2013-06-16 13:52:15] STEP 2: Filter blastn hits with minimum query coverage of 97.
[2013-06-16 13:52:15] contigs_filtered_BF.out generated successfully!
[2013-06-16 13:52:15] STEP 3: Annotate blastn hits with NCBI's taxonomy data at species level
[2013-06-16 13:52:17] contigs_filtered_BFT.out generated successfully!
[2013-06-16 13:52:17] STEP 4: Generate taxonomic assignments table from blastn hits.
[2013-06-16 13:52:18] contigs_filtered_ASSIGNMENTS.csv generated successfully!
[2013-06-16 13:52:18] SUMMARY: Reads assigned at species level are 5/5.
[ngswshop@quince-srv2 ~/workshop/assembly_test/assignment]$ cat contigs_filtered_ASSIGNMENTS.csv
NODE_1732_length_258956_cov_14.891905,Escherichia coli
NODE_1068_length_235041_cov_16.361355,Escherichia coli
NODE_1273_length_248226_cov_15.383513,Escherichia coli
NODE_1136_length_200197_cov_16.600334,Escherichia coli
NODE_934_length_224872_cov_17.441233,Escherichia coli
```