Installation of MyPro

Download Virtual Box

https://www.virtualbox.org/wiki/Downloads

Download MyPro.ova: http://sb.nhri.org.tw/MyPro/index.html

Other information at: http://sourceforge.net/projects/sb2nhri/files/MyPro/

Open VirtualBox

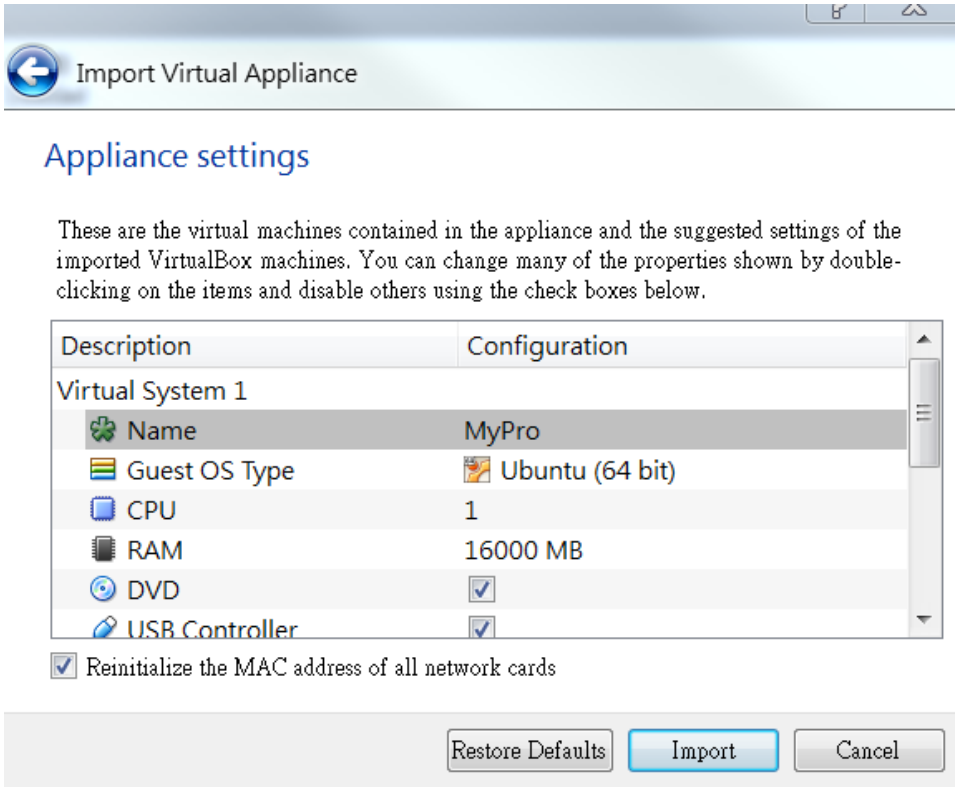
File -> Import Appliance...

Select the file (MyPro.ova) to import

MyPro.ova

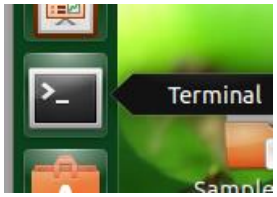
Or, directly double click on MyPro.ova

Please check the box of "Reinitialize the MAC address of all network cards"



Import

Open Terminal



Quick-start guide of MyPro

A. Pre-process

This script is used to trim, pair and sub-sample your raw reads. A total of 100X (paired) reads are generated. This process is strongly recommended, otherwise much computational time is required for genome assembly.

Command:

Preprocess.py -read1 S.gordonii_G9B_TTAGGC_L001_R1_001.fastq -read2 S.gordonii_G9B_TTAGGC_L001_R2_001.fastq -g 2200000

Preprocess.py -h for help.

B. AutoRun

This script is used to perform Assemble, Integrate and Annotate.

Command:

AutoRun.py G9B -read1 50X_R1.fastq -read2 50X_R2.fastq -evaluate -p '--prefix G9B --genus Streptococcus --species gordonii --strain G9B --addgenes --locustag AA01'

AutoRun.py -h, Assemble.py -h, Integrate.py -h and Annotate.py -h for help**.**

C. Post-assemble

This script is to (1) merge your ordered contigs if they are overlapped and (2) fill gaps with the contigs of local assembling.

(Optional) To use r2cat.jar for ordering your contigs against a related reference genome. You can Export contigs order (FASTA) and Export unmatched contigs (FASTA) separately.

Command:

Postassemble.py -o cisa.ordered.fa -u unmatched.fa -read1 ../50X_R1.fastq -read2 ../50X_R2.fastq

No reference genome:

Postassemble.py -u cisa.ctg.fa -read1 ../50X_R1.fastq -read2 ../50X_R2.fastq

Postassemble.py -h for help.

Validation of MyPro on three bacterial species

Dataset: *E. coli* MG1655

Paired reads of *E. coli* MG1655 are available at Illumina website. Mate1 and Mate 2 were downloaded separately.

Mate1:

ftp://webdata:webdata@ussd-ftp.illumina.com/Data/SequencingRuns/MG1655/MiSeq_Ecoli_MG1655_110721_PF_R1.fastq.gz

Mate2:

ftp://webdata:webdata@ussd-ftp.illumina.com/Data/SequencingRuns/MG1655/MiSeq_Ecoli_MG1655_110721_PF_R2.fastq.gz

Preprocess.py -read1 MiSeq_Ecoli_MG1655_110721_PF_R1.fastq -read2 MiSeq_Ecoli_MG1655_110721_PF_R2.fastq -g 4650000

```
manager@bl8vbox[Data] GetReadInfo.py 4650000 50X_R1.fastq
4650000
50X_R1.fastq
seqs amount:1550000
seqLen Mean:150.000000
seqLen Std:0.000000
total:232.50 Mb
depth: 50.00X
```

This process took about 40 min.

AutoRun.py MG1655 -read1 50X_R1.fastq -read2 50X_R2.fastq

This process took about 4 hr for running in a VirtualBox with 16GB RAM @ Dell Precisions Workstations T1600 Computer Workstation (Quad Core Xeon E3-1245, 3.30 GHz with 32GB RAM)

cisa.ctg.fa Alignment:99.34%

whole:4625288

N50: 88512

Number of contigs: 104

Length of the longest contig: 315566

Use **r2cat** to align cisa.ctg.fa against a reference (Ecoli DH10B, NC_010473), then export the ordered assembly.

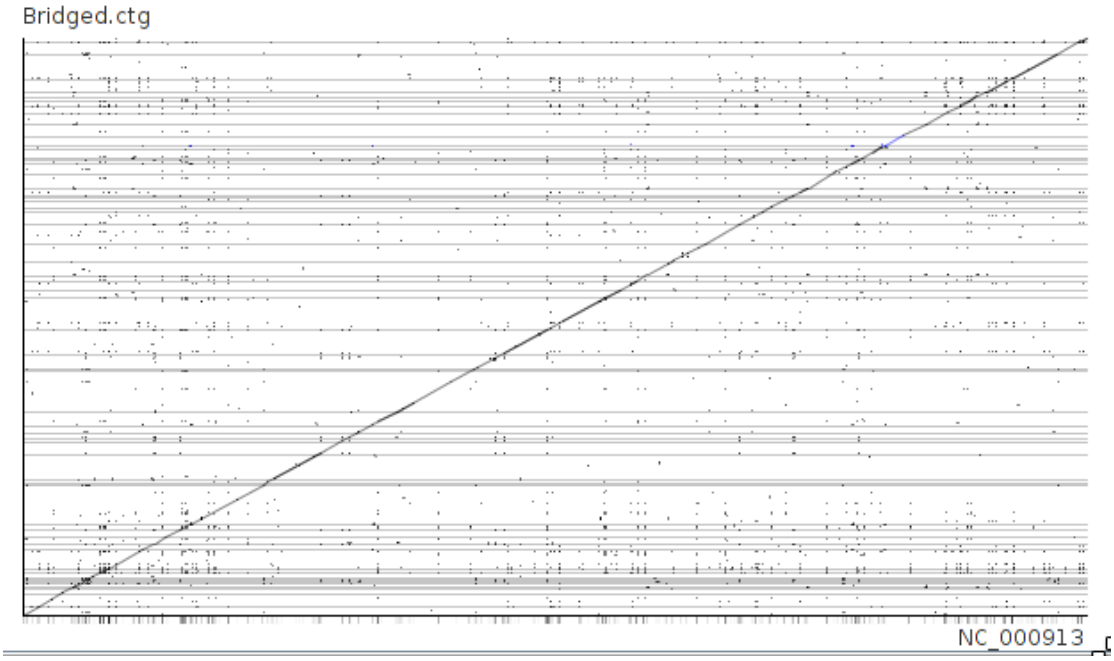
Click on r2cat.far located on Desktop

File --> Match new

Query: cisa.ctg.fa
Target: NC_010473.fna
Start Matching
Continue
Options --> Sort queries
File --> Export contigs order (FASTA)

Postassemble.py -o cisa.ctg.ordered.fa -read1 ../ 50X_R1.fastq -read2 ../50X_R2.fastq

Alignment:99.32%
N50.py Bridged.ctg.fa
whole:4608751
N50: 105185
Number of contigs: 71
Length of the longest contig: 335840
Dot plot against the reference genome (NC_000913) by r2cat:



Quast 2.3 Evaluation:

quast.py -o quast -R NC_000913.fna -G NC_000913.gff raw.abyss.ctg.fa raw.edena.ctg.fa raw.velvet.ctg.fa raw.soap.ctg.fa raw.spades.ctg.fa cisa.ctg.fa Bridged.ctg.fa

NC_000913.fna and NC_000913.gff can be downloaded from here.

QUAST report

13 November 2014, Thursday, 02:43:14

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.)

Extended report

Genome: 4 641 652 bp, G+C content: 50.79 %
4516 genes

Statistics without reference	raw.abyss.ctg	raw.edena.ctg	raw.velvet.ctg	raw.soap.ctg	raw.spades.ctg	cisa.ctg	Bridged.ctg
# contigs	362	363	318	526	131	104	71
Largest contig	85 610	68 181	85 370	58 371	224 038	315 566	335 840
Total length	4 576 900	4 525 060	4 576 844	4 555 491	4 567 001	4 625 288	4 608 751
N50	21 569	21 492	27 298	15 329	77 529	88 512	105 185
Misassemblies							
# misassemblies	1	2	6	0	0	1	4
Misassembled contigs length	37 134	47 106	224 058	0	0	222 869	447 995
Mismatches							
# mismatches per 100 kbp	2.260	0.24	2.75	0.35	3.4	4.59	4.940
# indels per 100 kbp	2.220	0.02	0.570	0.09	0.46	0.39	0.43
# N's per 100 kbp	64.59	0	483.65	0	0	0	0
Genome statistics							
Genome fraction (%)	98.014	97.449	97.871	97.767	98.08	99.066	99.087
Duplication ratio	1.006	1	1.007	1.003	1	1.006	1.003
# genes	4167 + 316 part	4139 + 320 part	4056 + 378 part	4033 + 380 part	4361 + 60 part	4420 + 76 part	4434 + 62 part
NGA50	21 419	21 163	26 259	14 936	77 529	88 512	104 044

To copy Bridged.ctg.fa to the folder of Assemble

cp Bridged.ctg.fa Assemble/

Annotate.py MG1655 -i Bridged.ctg.fa -p ' --genus Escherichia --species coli --strain MG1655 --prefix post'

organism: Escherichia coli MG1655

contigs: 71

bases: 4608751

rRNA: 8

tmRNA: 1

tRNA: 74

CDS: 4309

repeat_region: 2

Detailed log: post.log

Walltime used: 8.00 minutes

Because we have updated the databases for Prokka, the number of hypothetical proteins was reduced from 668 to 657 (Swiss-Prot update) and to 414 (High quality genus database). Meanwhile, the running time was minor increased from 13.13 min to 13.73 min (Swiss-Prot update) but greatly decreased to 8 min (using genus db). If RefSeq bacterial genomes were used as genus database, the number of hypothetical proteins was reduced to 369 but the running time increased to more than 20 min.

Dataset: *S. aureus*

Strain: *S. aureus* MW2

Sequencing reads were downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR857/SRR857914

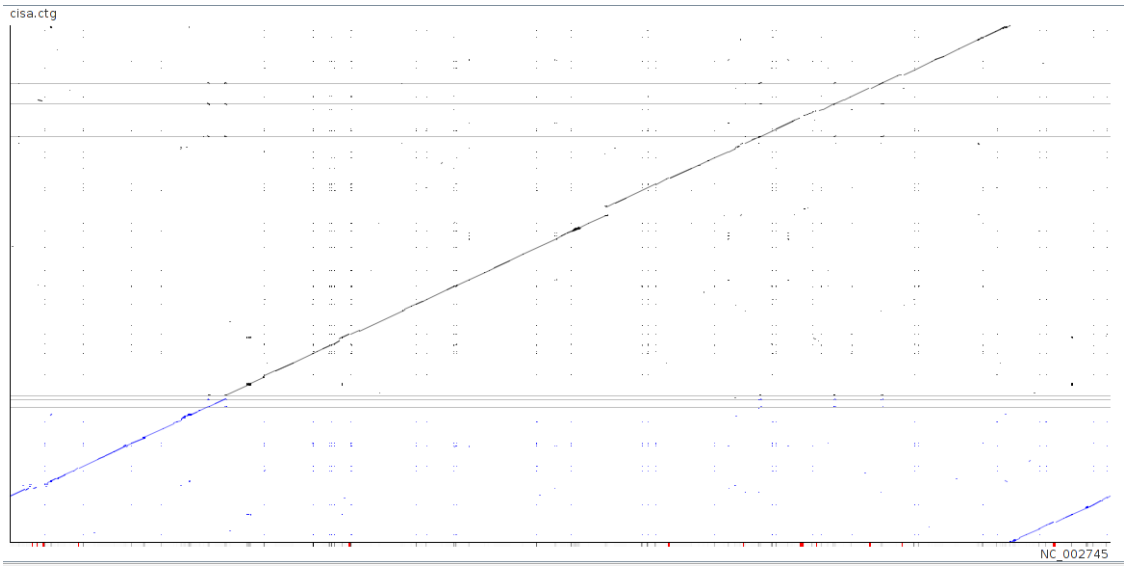
Preprocess.py -read1 SRR857914_1.fastq -read2 SRR857914_2.fastq -g 2800000
This process took about 20 min.

AutoRun.py MW2 -read1 50X_R1.fastq -read2 50X_R2.fastq
2014-11-21,02:27 ==> 2014-11-21,04:34
Alignment %:

- raw.soap.ctg.fa: 94.97
- raw.abyss.ctg.fa: 95.17
- raw.spades.ctg.fa: 99.31
- raw.velvet.ctg.fa: 99.46
- raw.edena.ctg.fa: 99.59

cisa.ctg.fa Alignment:93.86%
whole:2768641
N50: 1419734
Number of contigs: 5
Length of the longest contig: 1419734

To increase genome size for assembly integration:
Integrate.py MW2 -i abyss.ctg.fa,edena.ctg.fa,velvet.ctg.fa,spades.ctg.fa -evaluate **-gs 3000000**
cisa.ctg.fa Alignment:99.51%
whole:2830305
N50: 1419734
Number of contigs: 7
Length of the longest contig: 1419734



Postassemble.py -o cisa.ordered.fa -read1 ../../50X_R1.fastq -read2 ../../50X_R2.fastq -m 423 -s 71

N50.py Bridged.ctg.fa
whole:2830305
N50: 1419734
Number of contigs: 7
Length of the longest contig: 1419734
Alignment:99.51%
Identical assembly was obtained (to cisa.ctg.fa) after post-assembly!

Quast 2.3 Evaluation

quast.py -o quast -R NC_003923.fna -G NC_003923.gff raw.abyss.ctg.fa raw.edena.ctg.fa raw.velvet.ctg.fa raw.soap.ctg.fa raw.spades.ctg.fa cisa.ctg.fa

QUAST report

21 November 2014, Friday, 06:01:59

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.)

Extended report worst.....best

Genome: 2 820 462 bp, G+C content: 32.83 %
2703 genes

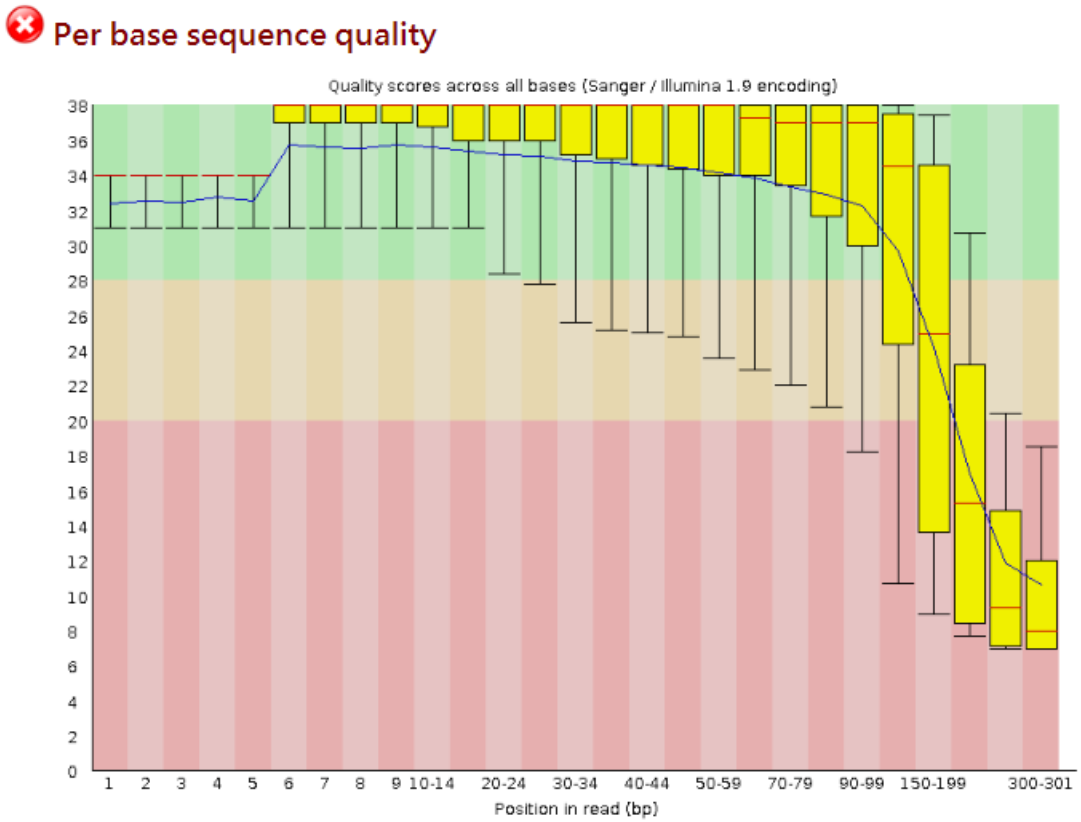
Statistics without reference	raw.abyss.ctg	raw.edena.ctg	raw.velvet.ctg	raw.soap.ctg	raw.spades.ctg	cisa.ctg
# contigs	18	18	8	78	43	7
Largest contig	940 710	539 806	1 413 768	138 695	410 165	1 419 734
Total length	2 826 349	2 816 795	2 815 477	2 783 529	2 808 157	2 830 305
N50	383 680	268 019	1 413 768	69 354	147 491	1 419 734
Misassemblies						
# misassemblies	0	0	2	0	2	0
Misassembled contigs length	0	0	1 413 768	0	347 121	0
Mismatches						
# mismatches per 100 kbp	4.83	1	1.93	0.04	6.1	4.020
# indels per 100 kbp	0.11	0	0.68	0	0.43	0
# N's per 100 kbp	1.8	0	79.84	0	0	0
Genome statistics						
Genome fraction (%)	99.752	99.126	99.011	98.62	98.748	99.606
Duplication ratio	1.005	1	1.001	1.001	1	1
# genes	2697 + 3 part	2685 + 10 part	2674 + 10 part	2632 + 41 part	2640 + 15 part	2694 + 1 part
NGA50	383 680	268 019	1 054 961	69 354	147 491	1 419 520

Dataset: *R. sphaeroides* 2.4.1

http://systems.illumina.com/systems/miseq/scientific_data.html

The *B. cereus* ATCC 10987 & ***R. sphaeroides*** **ATCC BAA-808** samples were sequenced on the MiSeq System using the new MiSeq Reagent Kit v3 at a 2 x 300 bp read length configuration with dual indexing. The total yield was 17.8 Gb with 70% of bases at or above Q30. The average fragment length for this data set is 475 bp.

fastqc Rhodo_S2_L001_R2_001.fastq



Preprocess.py -read1 Rhodo_S2_L001_R1_001.fastq -read2 Rhodo_S2_L001_R2_001.fastq -g 4600000

This process took about 4 hr! Paired reads were trimmed to 200 bp.

AutoRun.py Miseq300 -read1 50X_R1.fastq -read2 50X_R2.fastq

2014-11-14,00:23 ==> 2014-11-14,04:19

n50:

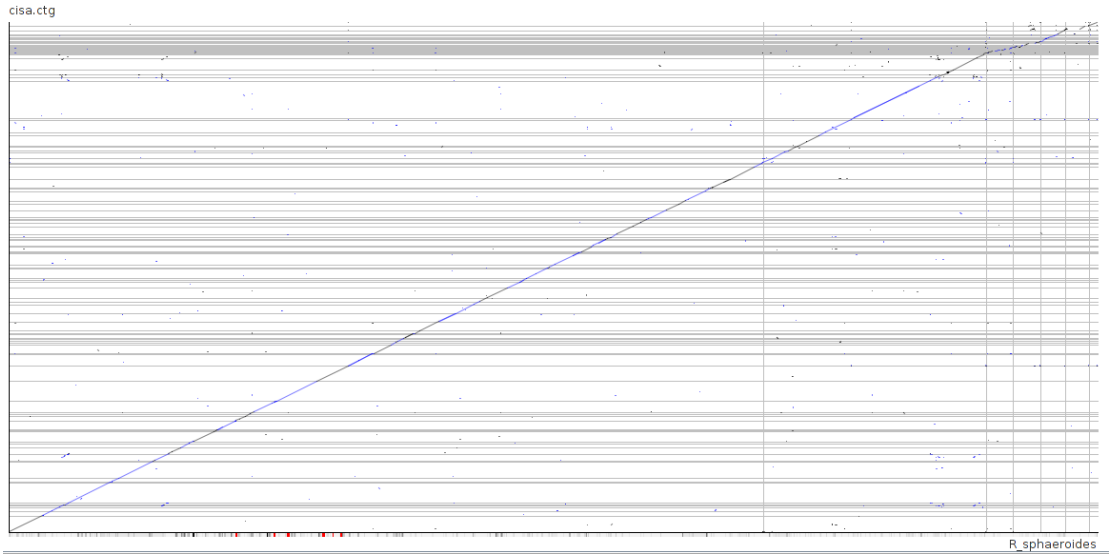
soap.ctg.fa: 19291

velvet.ctg.fa: 20714

edena.ctg.fa: 22179
abyss.ctg.fa: 22737
spades.ctg.fa: 77772
Ctgs:
soap.ctg.fa: 403
velvet.ctg.fa: 373
abyss.ctg.fa: 345
edena.ctg.fa: 293
spades.ctg.fa: 142
The longest ctg's length:
soap.ctg.fa: 69823
abyss.ctg.fa: 97096
edena.ctg.fa: 97244
velvet.ctg.fa: 97422
spades.ctg.fa: 320329
WholeGenome:
edena.ctg.fa: 4014722
abyss.ctg.fa: 4292653
soap.ctg.fa: 4339440
velvet.ctg.fa: 4369665
spades.ctg.fa: 4429967
Alignment %:
raw.edena.ctg.fa: 95.82
raw.abyss.ctg.fa: 97.90
raw.soap.ctg.fa: 98.19
raw.velvet.ctg.fa: 98.21
raw.spades.ctg.fa: 98.24
Keep all assemblies.

cisa.ctg.fa Alignment:98.26%
whole:4401027
N50: 79366
Number of contigs: 130
Length of the longest contig: 320634

r2cat against the reference *R.sphaeroides* 2.4.1
(http://sb.nhri.org.tw/CISA/upload/en/2013/8/R_sphaeroides.fna-01021413.gz).



Postassemble.py -o ordered.cisa.fa

N50.py Bridged.ctg.fa
whole:4391111
N50: 87718
Number of contigs: 120
Length of the longest contig: 320634
Alignment:98.26%

Validation of MyPro on 18 oral streptococcal species

Take G9B as an example:
Illumina HiSeq2500 reads for *Streptococcus gordonii* G9B

AutoRun.py G9B -read1 S.gordonii_G9B_TTAGGC_L001_R1_001.fastq -read2 S.gordonii_G9B_TTAGGC_L001_R2_001.fastq

2014-09-30,13:21 ==> 2014-09-30,23:58
This process took about 10 hr in a linux server!

cisa.ctg.fa Alignment:91.11%
whole:2206208
N50: 556432
Number of contigs: 7
Length of the longest contig: 641667

Use **r2cat** to align cisa.ctg.fa against NC_009785, then export the ordered assembly and the unmatched assembly.

Postassemble.py -o cisa.ordered.fa -u cisa.unmatched.fa
-read1 ../S.gordonii_G9B_TTAGGC_L001_R1_001.fastq
-read2 ../S.gordonii_G9B_TTAGGC_L001_R2_001.fastq

Alignment:91.11%
whole:2203707
N50: 1555780
Number of contigs: 3
Length of the longest contig: 1555780

Please note that there are three contigs in Bridged.ctg.fa.
>ref_3_len:6260
>ref_1_len:1555780
>ref_2_len:641667
The first one is a circular genome, and is likely to be a phage!

Quast 2.3 Evaluation

quast.py -o Quast_G9B raw.abyss.ctg.fa raw.edena.ctg.fa raw.velvet.ctg.fa raw.soap.ctg.fa raw.spades.ctg.fa CISA.fa cisa.ctg.fa Bridged.ctg.fa

QUAST report

24 December 2014, Wednesday, 16:07:26

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.)

Extended report worst.....best

Statistics without reference	raw.abyss.ctg	raw.edena.ctg	raw.velvet.ctg	raw.soap.ctg	raw.spades.ctg	CISA	cisa.ctg	Bridged.ctg
# contigs	15	16	16	80	381	10	7	3
Largest contig	555 590	484 400	898 711	222 575	639 860	556 432	641 667	1 555 780
Total length	2 215 919	2 197 541	2 196 493	2 191 062	2 426 588	2 208 238	2 206 208	2 203 707
N50	343 271	335 948	483 119	78 330	334 420	344 471	556 432	1 555 780
Mismatches								
# N's per 100 kbp	0	0	26.36	0	0.04	0.45	0	0

For 8 genomes, complete reference genome was available for the same oral streptococcal species. The N50 values and number of contigs from these genomes are summarized in Table 1 of the article. For 10 genomes, complete reference genome was not available. The N50 values and number of contigs from these genomes are summarized in Table 2 of the article.

Reference genomes:

S. gordonii G9B:

NC_009785, ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Streptococcus_gordonii_Challis_substr__CH1_uid57667/NC_009785.fna

S. salivarius KB005:

NC_015760, ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Streptococcus_salivarius_CCHSS3_uid70481/NC_015760.fna

S. parasanguinis MGH413:

NC_015678, ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Streptococcus_parasanguinis_ATCC_15912_uid49313/NC_015678.fna

S. cristatus CC5A, CR3:

AEVC00000000.1,

http://www.ncbi.nlm.nih.gov/nuccore?term=GL732518:GL732522[PACC]

S. mitis OT25, SK145, SK137:

NC_013853, ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Streptococcus_mitis_B6_uid46097/NC_013853.fna