SAG MiSeq Sequencing 1

Tuesday, December 09, 2014   3:51 PM

2/2/2015

Extract all files

```
        [pranjan6@bacphile2 Mon Feb 02 15:46:18 1.RawFiles.20150202]$ gzip -d *
```

2/3/2015

Scrapped previous analysis (till quality processing by TrimGalore) since the name of the species was flipped. Restarted analysis from scratch.

Extract all files

```
        [pranjan6@bacphile2 Tue Feb 03 17:05:34 1.RawFiles.20150202]$ gzip -d *.gz
```

Renamed files

```
        [pranjan6@bacphile2 Tue Feb 03 17:06:22 1.RawFiles.20150202]$ mv C04_S1_L001_R1_001.fastq sar406_R1.fq
        [pranjan6@bacphile2 Tue Feb 03 17:06:24 1.RawFiles.20150202]$ mv C04_S1_L001_R2_001.fastq sar406_R2.fq
        [pranjan6@bacphile2 Tue Feb 03 17:07:26 1.RawFiles.20150202]$ mv C10_S2_L001_R1_001.fastq sar11_R1.fq
        [pranjan6@bacphile2 Tue Feb 03 17:07:35 1.RawFiles.20150202]$ mv C10_S2_L001_R2_001.fastq sar11_R2.fq
```

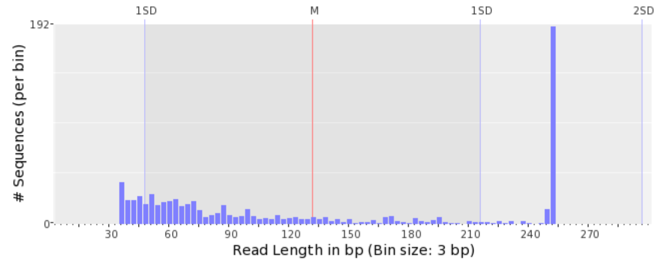
Running PrinSeq to generate initial report

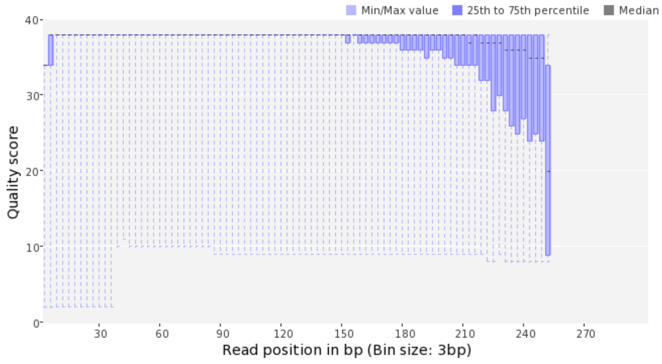
```
        [pranjan6@bacphile2 Tue Feb 03 17:07:42 1.RawFiles.20150202]$ prinseq-lite.pl -verbose -fastq sar11_R1.fq -graph_data -graph_stats ld,gc,qd,ns,pt,de,da,sc -out_good null -out_bad null
        [pranjan6@bacphile2 Tue Feb 03 17:09:27 1.RawFiles.20150202]$ prinseq-lite.pl -verbose -fastq sar11_R2.fq -graph_data -graph_stats ld,gc,qd,ns,pt,de,da,sc -out_good null -out_bad null
        [pranjan6@bacphile2 Tue Feb 03 17:51:15 1.RawFiles.20150202]$ ls -1 sar11_*.gd|xargs -I FILE prinseq-graphs-noPCA.pl -i FILE -o FILE -html_all
        [pranjan6@bacphile2 Tue Feb 03 17:06:28 1.RawFiles.20150202]$ prinseq-lite.pl -verbose -fastq sar406_R1.fq -graph_data -graph_stats ld,gc,qd,ns,pt,de,da,sc -out_good null -out_bad null
        [pranjan6@bacphile2 Tue Feb 03 17:11:47 1.RawFiles.20150202]$ prinseq-lite.pl -verbose -fastq sar406_R2.fq -graph_data -graph_stats ld,gc,qd,ns,pt,de,da,sc -out_good null -out_bad null
        [pranjan6@bacphile2 Tue Feb 03 17:12:11 1.RawFiles.20150202]$ ls -1 sar406_*.gd|xargs -I FILE prinseq-graphs-noPCA.pl -i FILE -o FILE -html_all
```

SAR406 raw read statistics

Input Information

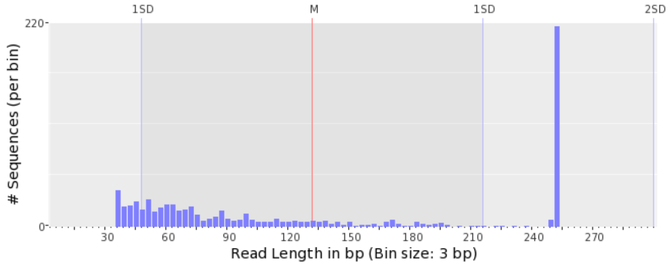
Input file(s):	sar406_R1.fq
Input format(s):	FASTQ
# Sequences:	778
Total bases:	101,667

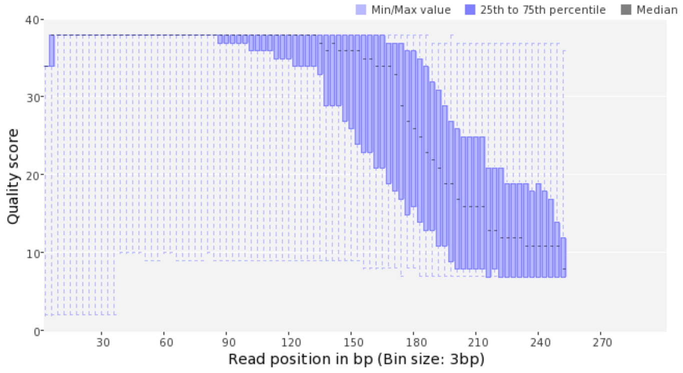




Input Information

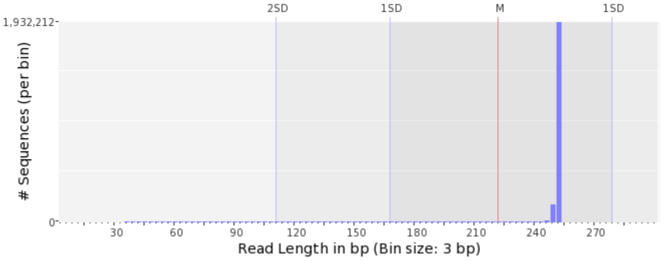
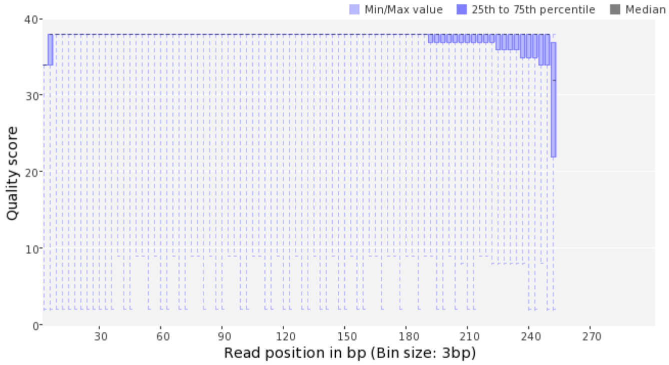
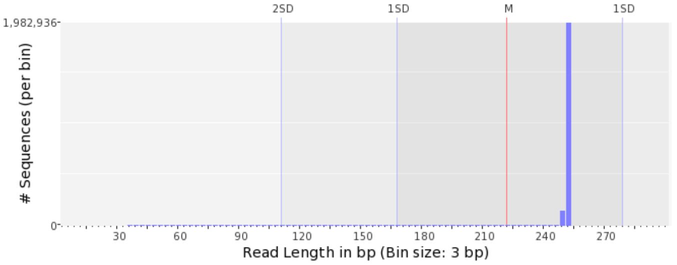
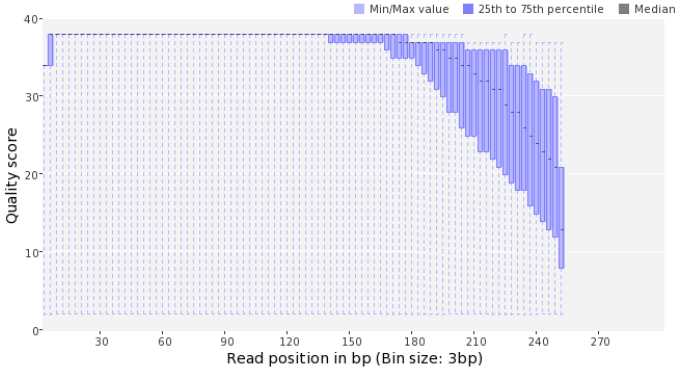
Input file(s):	sar406_R2.fq
Input format(s):	FASTQ
# Sequences:	778
Total bases:	102,190





sar406_R1.fq.gd	sar406_R2.fq.gd

SAR11 raw read statistics

Input Information Input file(s): sar11_R1.fq Input format(s): FASTQ # Sequences: 2,924,904 Total bases: 648,756,865   sar11_R1.fq.gd	Input Information Input file(s): sar11_R2.fq Input format(s): FASTQ # Sequences: 2,924,904 Total bases: 649,157,781   sar11_R2.fq.gd

Quality Processing:

Proceeding with only SAR11 reads for quality processing

Primer sequences used during sequencing for SAR11:

```
>nextera forward primer (index: S503)
AATGATACGGCGACCACCGAGATCTACACTATCCTCTTCGTCGGCAGCGTC
>nextera forward primer reverse complement
GACGCTGCCGACGAAGAGGATAGTGTAGATCTCGGTGGTCGCCGTATCATT
>nextera reverse primer (index: N704)
CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGG
>nextera reverse primer reverse complement
CCGAGCCCACGAGACTCCTGAGCATCTCGTATGCCGTCTTCTGCTTG
```

Running cutadapt to get rid of 3' adapters, with min 10 bp overlap anywhere in the sequence

```
[pranjan6@bacphile2 Tue Feb 03 18:13:23 2.QualityProcessedReads.20150203]$ cutadapt -f fastq -b
AATGATACGGCGACCACCGAGATCTACACTATCCTCTTCGTCGGCAGCGTC -b
GACGCTGCCGACGAAGAGGATAGTGTAGATCTCGGTGGTCGCCGTATCATT -b
CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGG -b
CCGAGCCCACGAGACTCCTGAGCATCTCGTATGCCGTCTTCTGCTTG -O 10 -o sar11_R1
_adapterTrimmed.fq ../1.RawFiles.20150202/sar11_R1.fq >sar11_R1_cutadaptTrimmingReport.txt
[pranjan6@bacphile2 Tue Feb 03 18:13:59 2.QualityProcessedReads.20150203]$ cutadapt -f fastq -b
AATGATACGGCGACCACCGAGATCTACACTATCCTCTTCGTCGGCAGCGTC -b
GACGCTGCCGACGAAGAGGATAGTGTAGATCTCGGTGGTCGCCGTATCATT -b
```

```
CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGG -b
CCGAGCCCACGAGACTCCTGAGCATCTCGTATGCCGTCTTCTGCTTG -O 10 -o sar11_R2
_adapterTrimmed.fq ../1.RawFiles.20150202/sar11_R2.fq >sar11_R2_cutadaptTrimmingReport.txt
```

Quality control using TrimGalore, with phred cutoff of 25

```
[pranjan6@bacphile2 Tue Feb 03 18:32:15 2.QualityProcessedReads.20150203]$ trim_galore -q 25 --length 90 --paired -a
AATGATACGGCGACCACCGAGATCTACACTATCCTCTTCGTCGGCAGCGTC -a2
CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGG --retain_unpaired -r1 100 -r2 100 sar11_R1_adapterTrimmed.fq
sar11_R2_adapterTrimmed.fq
```

Renamed files

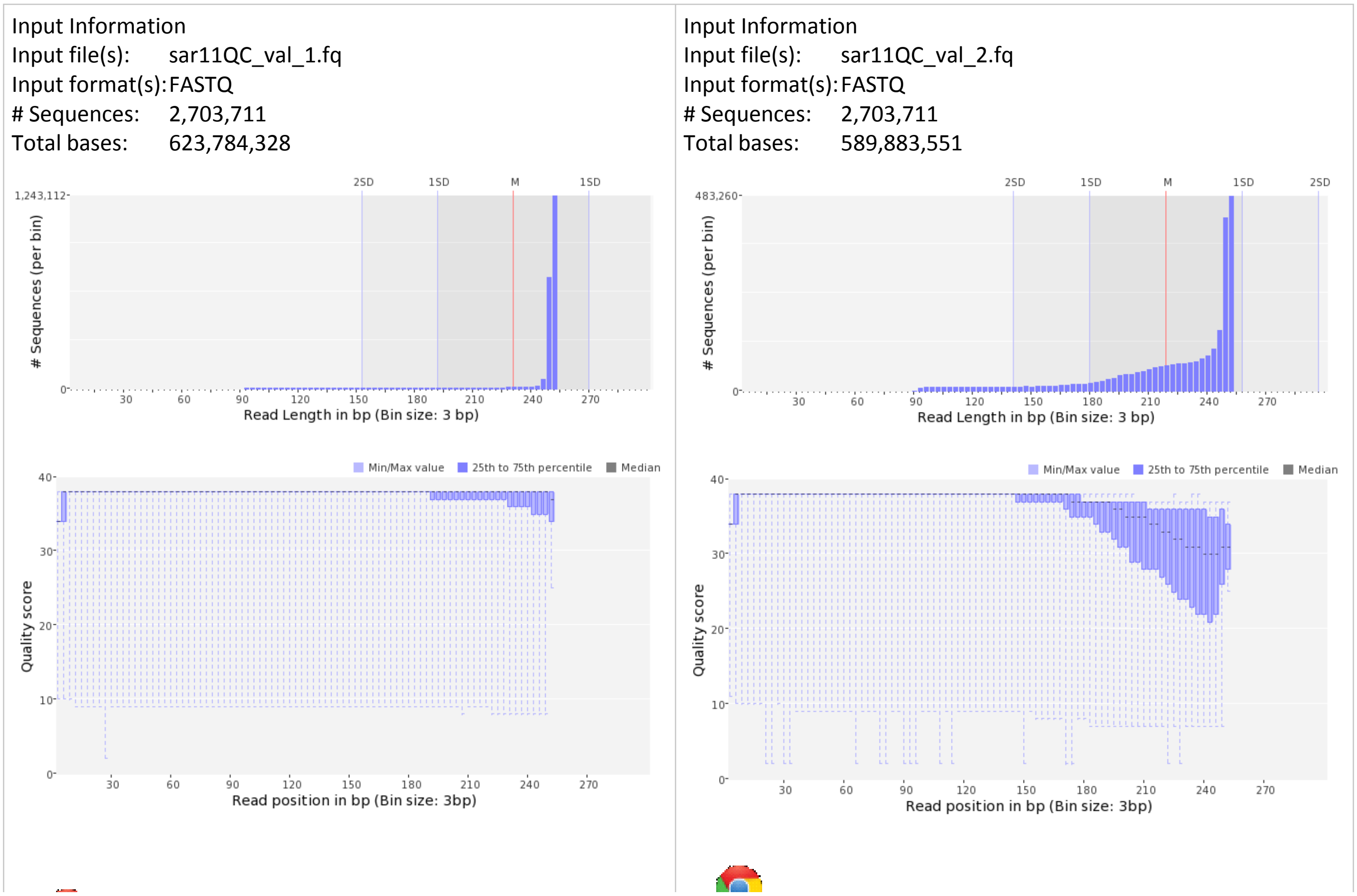
```
[pranjan6@bacphile2 Tue Feb 03 18:45:01 2.QualityProcessedReads.20150203]$ mv sar11_R1_adapterTrimmed_val_1.fq
sar11QC_val_1.fq
[pranjan6@bacphile2 Tue Feb 03 18:47:27 2.QualityProcessedReads.20150203]$ mv sar11_R2_adapterTrimmed_val_2.fq
sar11QC_val_2.fq
[pranjan6@bacphile2 Tue Feb 03 18:48:01 2.QualityProcessedReads.20150203]$ mv sar11_R1_adapterTrimmed_unpaired_1.fq
sar11QC_unpaired_1.fq
[pranjan6@bacphile2 Tue Feb 03 18:49:24 2.QualityProcessedReads.20150203]$ mv sar11_R2_adapterTrimmed_unpaired_2.fq
sar11QC_unpaired_2.fq
```

Generating read statistics after QC

```
[pranjan6@bacphile2 Tue Feb 03 18:48:23 2.QualityProcessedReads.20150203]$ prinseq-lite.pl -verbose -fastq sar11QC_val_1.fq -
graph_data -graph_stats ld,gc,qd,ns,pt,de,da,sc -out_good null -out_bad null
[pranjan6@bacphile2 Tue Feb 03 18:53:38 2.QualityProcessedReads.20150203]$ prinseq-lite.pl -verbose -fastq sar11QC_val_2.fq -
graph_data -graph_stats ld,gc,qd,ns,pt,de,da,sc -out_good null -out_bad null
[pranjan6@bacphile2 Tue Feb 03 18:54:46 2.QualityProcessedReads.20150203]$ prinseq-lite.pl -verbose -fastq sar11QC_
unpaired_1.fq -graph_data -graph_stats ld,gc,qd,ns,pt,de,da,sc -out_good null -out_bad null
[pranjan6@bacphile2 Tue Feb 03 18:55:39 2.QualityProcessedReads.20150203]$ prinseq-lite.pl -verbose -fastq sar11QC_
unpaired_2.fq -graph_data -graph_stats ld,gc,qd,ns,pt,de,da,sc -out_good null -out_bad null
[pranjan6@bacphile2 Tue Feb 03 19:32:14 2.QualityProcessedReads.20150203]$ ls -1 sar11QC_val_*.gd|xargs -I FILE prinseq-
graphs-noPCA.pl -i FILE -o FILE -html_all
[pranjan6@bacphile2 Tue Feb 03 18:55:51 2.QualityProcessedReads.20150203]$ ls -1 sar11QC_unpaired_*.gd|xargs -I FILE
prinseq-graphs-noPCA.pl -i FILE -o FILE -html_all
```

SAR11 read statistics after QC

Paired-End reads:



Input Information
Input file(s): sar11QC_val_1.fq
Input format(s): FASTQ
# Sequences: 2,703,711
Total bases: 623,784,328

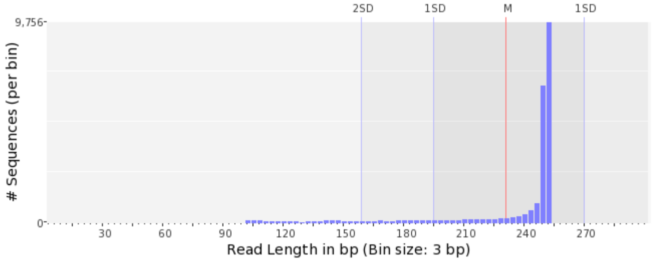
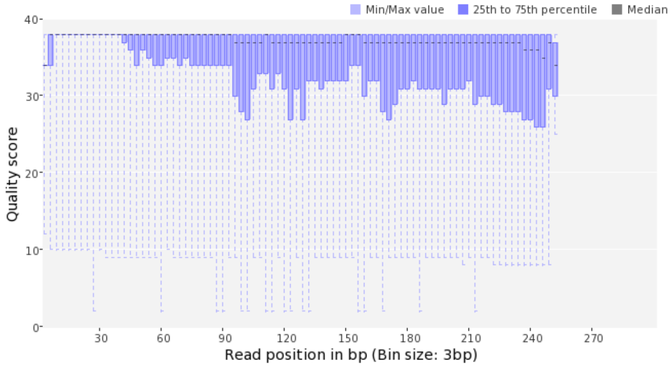
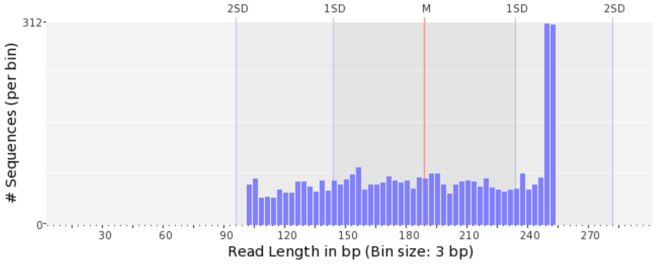
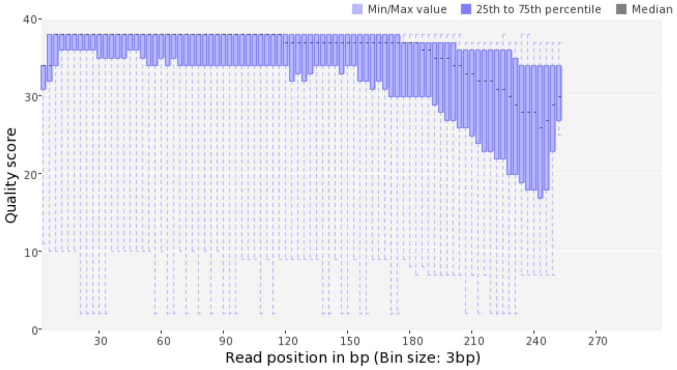
Input Information
Input file(s): sar11QC_val_2.fq
Input format(s): FASTQ
# Sequences: 2,703,711
Total bases: 589,883,551

sar11QC_va

sar11QC_va
l_2.fq.gd

sar11QC_val_1.fq.gd	sar11QC_val_2.fq.gd

Unpaired reads:

Input Information Input file(s): sar11QC_unpaired_1.fq Input format(s): FASTQ # Sequences: 24,969 Total bases: 5,766,063   sar11QC_unpaired_...	Input Information Input file(s): sar11QC_unpaired_2.fq Input format(s): FASTQ # Sequences: 3,706 Total bases: 693,272   sar11QC_unpaired_...

Merged paired end reads using FLASH on 3/6/2015

```
[pranjan6@bacphile2 Fri Mar 06 16:15:54 2.QualityProcessedReads.20150203]$ flash -m 50 -o
sar11QCMerged -d ./ sar11QC_val_1.fq sar11QC_val_2.fq
[FLASH] Read combination statistics:
[FLASH]     Total reads:      2703711
[FLASH]     Combined reads:   1808996
[FLASH]     Uncombined reads: 894715
[FLASH]     Percent combined: 66.91%
[pranjan6@bacphile2 Fri Mar 06 16:18:43 2.QualityProcessedReads.20150203]$ flash -m 30 -o
sar11QCMerged30 -d ./ sar11QC_val_1.fq sar11QC_val_2.fq
[FLASH] Read combination statistics:
[FLASH]     Total reads:      2703711
[FLASH]     Combined reads:   1927873
[FLASH]     Uncombined reads: 775838
[FLASH]     Percent combined: 71.30%
```

2/5/2015

IDBA assembly for SAR11

Converting paired-end fastq to fasta interleaved

```
[pranjan6@bacphile2 Thu Feb 05 13:40:56 1.Sar11IDBAAssembly.20150205]$ fq2fa --merge --filter ../../2.QualityProcessedReads.20150203/sar11QC_val_1.fq ../../2.QualityProcessedReads.20150203/sar11QC_val_2.fq sar11QC.fa
```

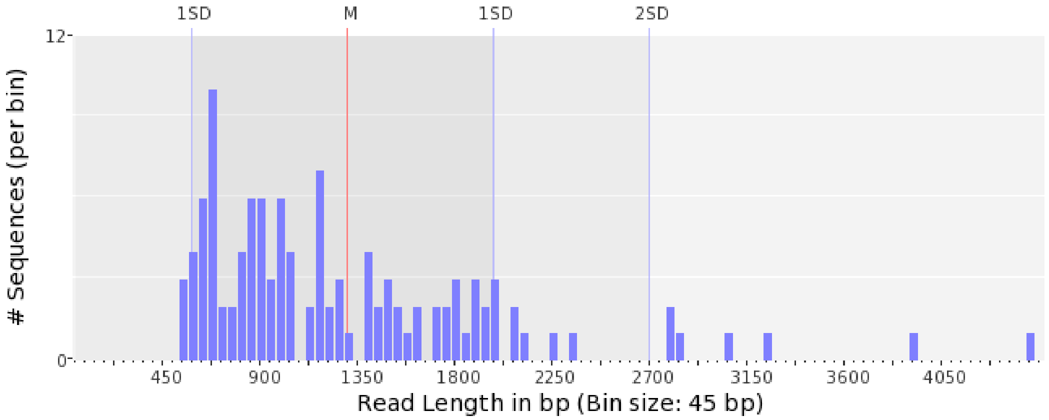
Running IDBA_UD on Bacphile2 for preliminary assembly

```
[pranjan6@bacphile2 Thu Feb 05 14:48:39 1.Sar11IDBAAssembly.20150205]$ idba_ud -o ./ -r sar11QC.fa --mink 40 --maxk 70 --step 10 --min_count 4 --min_support 2 --num_threads 4 --min_contig 500
```

Finished in ~1 hour, 4 processors being used nearly all the time but very less memory.

Stats for IDBA_UD preliminary assembly on bacphile2

```
[pranjan6@bacphile2 Thu Feb 05 17:06:16 1.Sar11IDBAAssembly.20150205]$ prinseq-lite.pl -fasta contig.fa -stats_assembly
stats_assembly  N50     1474
stats_assembly  N75     975
stats_assembly  N90     710
stats_assembly  N95     624
[pranjan6@bacphile2 Thu Feb 05 16:51:48 1.Sar11IDBAAssembly.20150205]$ prinseq-lite.pl -fasta contig.fa -graph_data -graph_stats ld,gc,ns,pt,aq,de,da,sc -out_good null -out_bad null
[pranjan6@bacphile2 Thu Feb 05 16:54:37 1.Sar11IDBAAssembly.20150205]$ prinseq-graphs-noPCA.pl -i contig.fa.gd -o contig.fa.gd -html_all
```



```
Input Information
Input file(s):      contig.fa
Input format(s):FASTA
# Sequences:        113
Total bases:        144,063
```

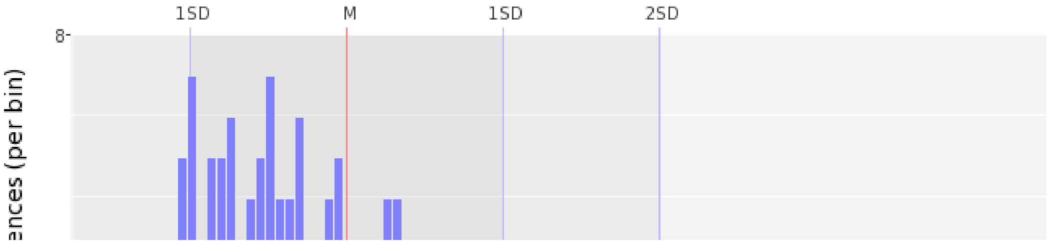
contig.fa.gd

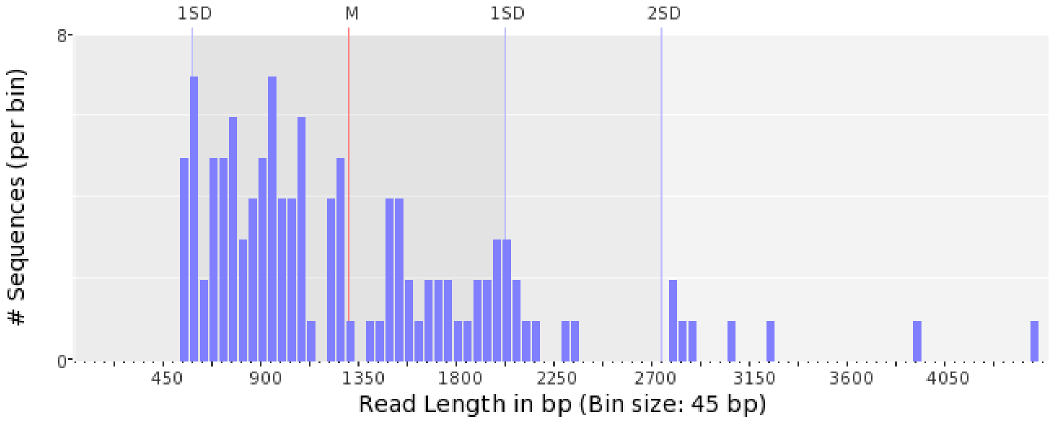
Running IDBA_UD on Bacphile2 with increased parameter space (Kmer)

```
[pranjan6@bacphile2 Thu Feb 05 17:29:31 2.Sar11IDBAAssembly.20150205]$ time idba_ud -o ./ -r sar11QC.fa --mink 40 --maxk 100 --step 5 --min_count 4 --min_support 2 --num_threads 12 --min_contig 500
Time statistics:
        real    53m41.624s
        user    257m59.029s
        sys     0m50.929s
```

Stats for IDBA_UD enhanced assembly on bacphile2

```
[pranjan6@bacphile2 Fri Feb 06 14:50:27 2.Sar11IDBAAssembly.20150205]$ prinseq-lite.pl -fasta contig.fa --stats_assembly
stats_assembly  N50     1529
stats_assembly  N75     984
stats_assembly  N90     729
stats_assembly  N95     622
[pranjan6@bacphile2 Fri Feb 06 14:50:31 2.Sar11IDBAAssembly.20150205]$ prinseq-lite.pl -fasta contig.fa -graph_data -graph_stats ld,gc,ns,pt,aq,de,da,sc -out_good null -out_bad null
[pranjan6@bacphile2 Fri Feb 06 14:50:54 2.Sar11IDBAAssembly.20150205]$ prinseq-graphs-noPCA.pl -i contig.fa.gd -o contig.fa.gd -html_all
```





Input Information

Input file(s):	contig.fa
Input format(s):	FASTA
# Sequences:	119
Total bases:	154,238

contig.fa.gd

3/3/2015

Mapping processed reads to multiple whole genomes available in NCBI using Bowtie2

Genomes chosen: NC_007205.1.fasta NC_018643.1.fasta NC_018644.1.fasta

Making Bowtie-index

```
[pranjan6@bacphile2 Tue Mar 03 16:29:09 indexes]$ bowtie2-build ../references/NC_007205.1.fasta NC_007205.1
[pranjan6@bacphile2 Tue Mar 03 16:38:48 indexes]$ bowtie2-build ../references/NC_018643.1.fasta NC_018643.1
[pranjan6@bacphile2 Tue Mar 03 16:39:31 indexes]$ bowtie2-build ../references/NC_018644.1.fasta NC_018644.1
```

Tested bowtie-indexes

```
[pranjan6@bacphile2 Tue Mar 03 16:41:21 indexes]$ bowtie2-inspect -n NC_007205.1
gi|71082709|ref|NC_007205.1| Candidatus Pelagibacter ubique HTCC1062, complete genome
[pranjan6@bacphile2 Tue Mar 03 16:41:54 indexes]$ bowtie2-inspect -n NC_018643.1
gi|406705597|ref|NC_018643.1| Alpha proteobacterium HIMB5, complete genome
[pranjan6@bacphile2 Tue Mar 03 16:42:08 indexes]$ bowtie2-inspect -n NC_018644.1
gi|406707029|ref|NC_018644.1| Alpha proteobacterium HIMB59, complete genome
```

Aligning processed reads to references

```
[pranjan6@bacphile2 Tue Mar 03 18:02:30 indexes]$ bowtie2 -t -x NC_007205.1 -1 ../../2.QualityProcessedReads.20150203/sar11QC_val_1.fq -2 ../../2.QualityProcessedReads.20150203/sar11QC_val_2.fq -S ../alignments/sar11ReadsAlign_NC_007205.sam --un-conc ../alignments/sar11ReadsNoAlign_NC_007205.fq -p 8
```

No hits found in the alignment

Running with very sensitive local alignment

```
[pranjan6@bacphile2 Tue Mar 03 18:10:58 indexes]$ bowtie2 -t -x NC_007205.1 -1 ../../2.QualityProcessedReads.20150203/sar11QC_val_1.fq -2 ../../2.QualityProcessedReads.20150203/sar11QC_val_2.fq -S ../alignments/sar11ReadsAlign_NC_007205.sam --un-conc ../alignments/sar11ReadsNoAlign_NC_007205.fq -p 8 --very-sensitive-local
```

Only 18 hits found.

3/5/2015

Mapping Processed reads using Bowtie to SAR11 SPADes contigs to check percentage of reads included in assembly

```
[pranjan6@bacphile2 Thu Mar 05 18:04:45 alignmentBackToSpadesAssembly]$ bowtie2 -t -x
spadesContigs -1 ../../2.QualityProcessedReads.20150203/sar11QC_val_
1.fq -2 ../../2.QualityProcessedReads.20150203/sar11QC_val_2.fq -S sar11ReadsA
lignSpadesAssembly.sam -p 8 --very-sensitive-local --un-conc sar11ReadsNoAlignSpadesAssembly
Time loading reference: 00:00:00
Time loading forward index: 00:00:00
Time loading mirror index: 00:00:00
Multiseed full-index search: 01:04:36
2703711 reads; of these:
  2703711 (100.00%) were paired; of these:
    218501 (8.08%) aligned concordantly 0 times
    172838 (6.39%) aligned concordantly exactly 1 time
    2312372 (85.53%) aligned concordantly >1 times
    ----
    218501 pairs aligned concordantly 0 times; of these:
      1917 (0.88%) aligned discordantly 1 time
    ----
    216584 pairs aligned 0 times concordantly or discordantly; of these:
      433168 mates make up the pairs; of these:
        1172 (0.27%) aligned 0 times
        2463 (0.57%) aligned exactly 1 time
        429533 (99.16%) aligned >1 times
99.98% overall alignment rate
Time searching: 01:04:36
Overall time: 01:04:36
```

Realigning to contigs with length greater than 500

```
    [pranjan6@bacphile2 Fri Mar 06 12:49:17 alignmentBackToSpadesAssemblyLenGt500]$ bowtie2-
    build ../spades_contigs_k95_lenGt500.fasta spadesContigsGtLen500
    [pranjan6@bacphile2 Fri Mar 06 12:52:06 alignmentBackToSpadesAssemblyLenGt500]$ bowtie2 -t -
    x spadesContigsGtLen500 -1 ../../2.QualityProcessedReads.20150203/sar11QC_val_
    1.fq -2 ../../2.QualityProcessedReads.20150203/sar11QC_val_2.fq -S
    sar11ReadsAlignSpadesAssemblyLenGt500.sam -p 8 --very-sensitive-local --un-conc
    sar11ReadsNoAlignSpadesAssemblyLenGt500
    Time loading reference: 00:00:00
    Time loading forward index: 00:00:00
    Time loading mirror index: 00:00:00
    Multiseed full-index search: 00:13:20
    2703711 reads; of these:
      2703711 (100.00%) were paired; of these:
        1396928 (51.67%) aligned concordantly 0 times
        706890 (26.15%) aligned concordantly exactly 1 time
        599893 (22.19%) aligned concordantly >1 times
        ----
        1396928 pairs aligned concordantly 0 times; of these:
          110169 (7.89%) aligned discordantly 1 time
        ----
        1286759 pairs aligned 0 times concordantly or discordantly; of these:
          2573518 mates make up the pairs; of these:
            2181506 (84.77%) aligned 0 times
            211038 (8.20%) aligned exactly 1 time
            180974 (7.03%) aligned >1 times
    59.66% overall alignment rate
    Time searching: 00:13:20
    Overall time: 00:13:20
```

3/6/2015

Aligning Processed Merged reads using BLASTn to 16S microbial sequences to check presence of 16S DNA (reads)

Converted fastq to fasta

```
    [pranjan6@bacphile2 Fri Mar 06 17:08:41 6.BlastReadsWith16S.20150306]$
    cat ../2.QualityProcessedReads.20150203/sar11 QCMerged30.extendedFrags.fastq | awk '{if(NR%4
    ==1) {printf(">%s\n",substr($0,2));} else if(NR%4==2) print;}' > sar11QCMerged30.fa
```

BLASTn merged reads to 16S microbial sequences

```
    [pranjan6@bacphile2 Fri Mar 06 19:23:57 6.BlastReadsWith16S.20150306]$ blastn -query
    sar11QCMerged30.fa -db /data/BLASTdbs/nt16SMicrobial_2015-01-30/16SMicrobial -out
    sar11QCMerged30.blast -evalue .01 -outfmt '7 qseqid sseqid pident length mismatch gapopen
    qstart qend sstart send evalue bitscore stitle' -max_target_seqs 5 -num_threads 12
```

BLAST gave no hits

Assembling SAR11 SAG reads with EULER-SR and Velvet-SC

Reads are required as interleaved fasta format (already available beforehand, did not used Illumina2Fastq.pl utility in EULER-SR package)

Running error correction through Euler-SR

[pranjan6@bacphile2 Thu Mar 05 16:18:07 3.Sar11EVSCAssembly.20150305]$ EulerEC.pl sar11QC_interleavedReads.fa 55 -minMult 10

Finished with error corrected reads in sub-folder "fixed"