**29/03/2019 Project Plan**

**29/03/19: Download data now**

**04/04/19: Quality control with FASTQC tool (genomic data – canu (Illumina, PacBio, Nanopore)+ RNA-Seq). Find tools for assembly with Nanopore + PacBio, Nanopore + Illumina. Find tools to compate assemblies**

canu -d canu\_030419\_non\_corrected -p 030419\_non\_corrected genomeSize=3.3M -pacbio-raw PacBio/\*.fasta.gz

scp -r erzh4483@rackham.uppmax.uu.se:/home/erzh4483/genomics\_data/Illumina/ .

scp -r . erzh4483@rackham.uppmax.uu.se:/home/erzh4483/genomics\_data/assembly/

spades -o illumina\_and\_nanopore --nanopore Nanopore/E745\_all.fasta.gz --pe1-1 Illumina/E745-1.L500\_SZAXPI015146-56\_1\_clean.fq.gz --pe1-2 Illumina/E745-1.L500\_SZAXPI015146-56\_2\_clean.fq.gz

java -jar /sw/apps/bioinfo/trimmomatic/0.36/rackham/trimmomatic-0.36.jar PE -phred33 /home/erzh4483/genomics\_data/Illumina/E745-1.L500\_SZAXPI015146-56\_1\_clean.fq.gz /home/erzh4483/genomics\_data/Illumina/E745-1.L500\_SZAXPI015146-56\_2\_clean.fq.gz /home/erzh4483/genomics\_data/Illumina/E745-1.L500\_SZAXPI015146-56\_1\_clean\_trimmed.fq.gz /home/erzh4483/genomics\_data/Illumina/E745-1.L500\_SZAXPI015146-56\_1\_forward\_unpaired.fq.gz /home/erzh4483/genomics\_data/Illumina/E745-1.L500\_SZAXPI015146-56\_2\_clean\_trimmed.fq.gz /home/erzh4483/genomics\_data/Illumina/E745-1.L500\_SZAXPI015146-56\_2\_reverse\_unpaired.fq.gz ILLUMINACLIP:/sw/apps/bioinfo/trimmomatic/0.36/rackham/adapters/TruSeq3-PE-2.fa:2:30:10 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:36

ILLUMINACLIP: Using 1 prefix pairs, 4 forward/reverse sequences, 0 forward only sequences, 0 reverse only sequences

Input Read Pairs: 1666667 Both Surviving: 1666214 (99.97%) Forward Only Surviving: 453 (0.03%) Reverse Only Surviving: 0 (0.00%) Dropped: 0 (0.00%)

TrimmomaticPE: Completed successfully

cd /sw/apps/bioinfo/trimmomatic/0.36/rackham/adapters/TruSeq3-PE-2.fa

**05/04/19: Do correction with Illumina reads based on canu assembly**

**GENOME ASSEMBLY**

**Canu assembly:**

#tigID tigLen coordType covStat coverage tigClass sugReptsugCirc numChildren

1 2773605 ungapped 3036.76 38.72 contig no yes 13045

25 40017 ungapped 39.22 31.11 contig no yes 158

26 28297 ungapped 66.80 13.97 contig no no 51

29 9976 ungapped 16.50 2.99 contig no no 5

30 11396 ungapped 11.41 5.71 contig no no 9

31 9536 ungapped 2.90 3.57 contig no no 4

33 26278 ungapped 63.60 6.24 contig no no 20

72 6731 ungapped 0.00 1.00 unassm no no 1

73 6969 ungapped 0.00 1.00 unassm no no 1

1. Bwa index Illumina reads

2. Bwa MEM align illumina reads → sam file

3. Convert sam file to bam file

4. Pilon aligned reads in BAM file to contigs in fasta file. Pilon uses read alignment analysis to identify inconsistencies between the input genome and the evidence in the reads. It then attempts to make improvements to the input genome, including:

* Single base differences
* Small indels
* Larger indel or block substitution events
* Gap filling
* Identification of local misassemblies, including optional opening of new gaps

1. bwa index 030419\_non\_corrected.contigs.fasta 30419\_non\_corrected\_9\_contigs

2. bwa mem /home/erzh4483/genomics\_data/canu\_030419\_non\_corrected/bam\_reference/030419\_non\_corrected.contigs.fasta /home/erzh4483/genomics\_data/Illumina/E745-1.L500\_SZAXPI015146-56\_1\_clean\_trimmed.fq.gz /home/erzh4483/genomics\_data/Illumina/E745-1.L500\_SZAXPI015146-56\_2\_clean\_trimmed.fq.gz > /home/erzh4483/genomics\_data/canu\_040419\_corrected/bwa\_mem\_040419.sam

3. samtools view -S -b bwa\_mem\_040419.sam > bwa\_mem\_040419.bam

4. samtools sort bwa\_mem\_040419.bam > bwa\_mem\_040419\_sorted.bam | samtools index

5. samtools index bwa\_mem\_040419\_sorted.bam

6. pilon --genome /home/erzh4483/genomics\_data/canu\_030419\_non\_corrected/bam\_reference/030419\_non\_corrected.contigs.fasta --frags /home/erzh4483/genomics\_data/canu\_040419\_corrected/bwa\_mem\_040419\_sorted.bam --outdir /home/erzh4483/genomics\_data/canu\_040419\_corrected/pilon\_040419 –changes

7. Compare with mummer:

mummerplot mummer\_assembly\_evalution.txt -t png

Compare with quast (canu+pilon vs reference chromosome - 1st contig):

quast /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/spades\_assembly/NODE\_1\_length\_2765031\_cov\_46.3316.fasta -r /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/reference/chromosme\_sequence.fasta -o quast\_comparison\_spades\_1st\_contig\_reference\_chromosome

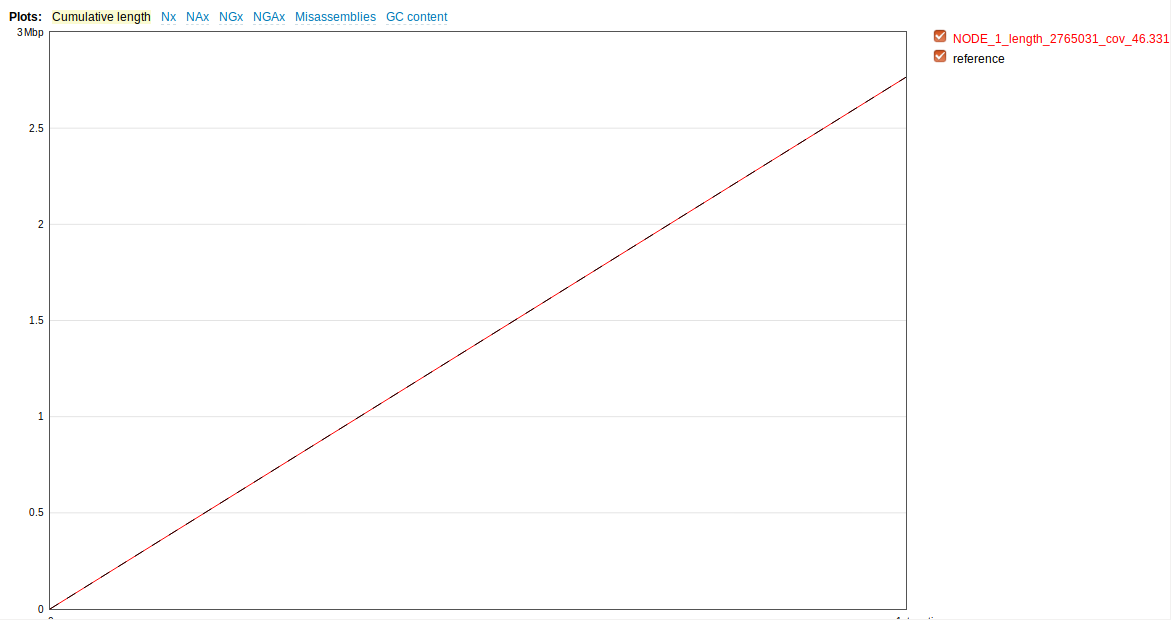
Create the full reference:

cat \*.fasta > the\_full\_referense.fasta

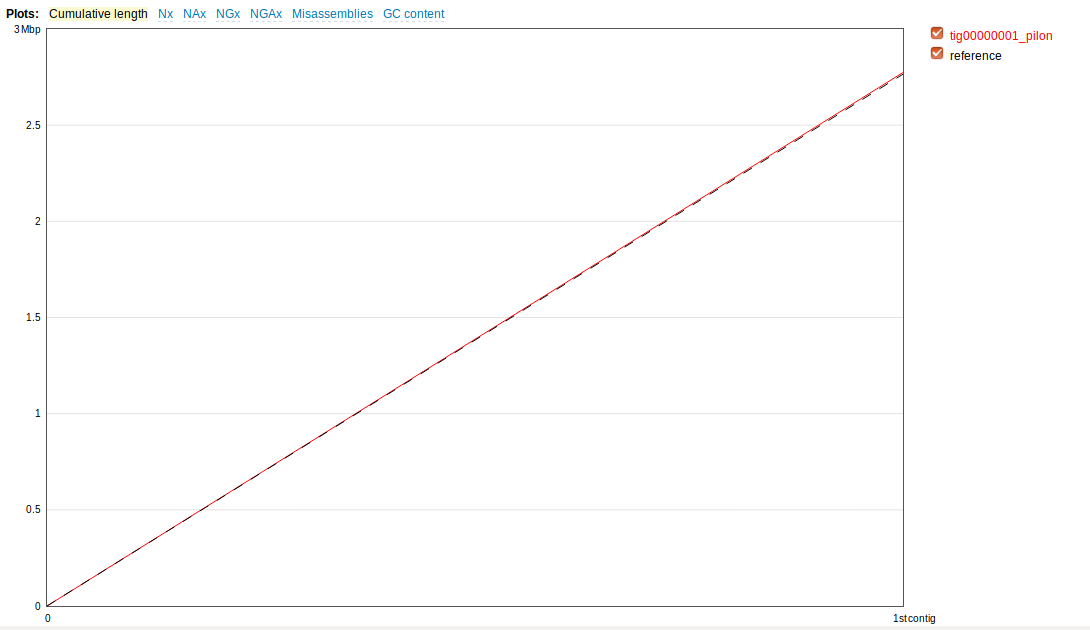
Compare with quast (canu+pilon vs reference chromosome, spades vs reference chromosome):

1.

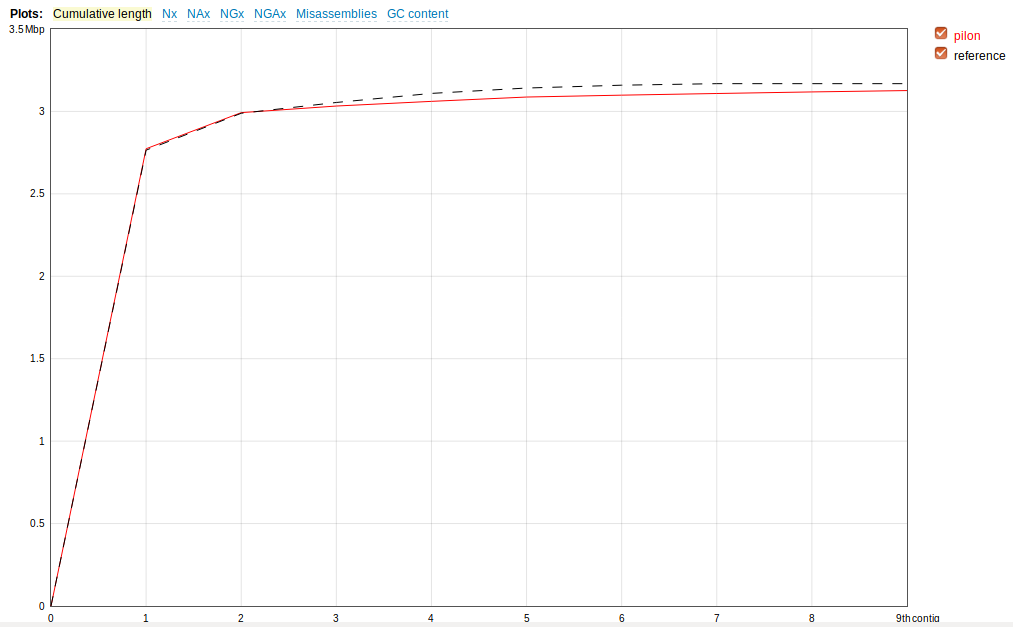
quast /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/assemblies/canu\_assembly\_corrected/pilon.fasta -r /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/reference/the\_full\_reference.fasta -o quast\_comparison\_canu\_vs\_reference\_chromosome



Compare with quast (spades vs reference chromosome):



Compare with quast (canu vs reference chromosome – all genome):



| Genome statistics | scaffolds |
| --- | --- |
| Genome fraction (%) | 98.892 |
| Duplication ratio | 1.002 |
| Largest alignment | 1899903 |
| Total aligned length | 3128539 |
| NG50 | 2765031 |

|  |  |
| --- | --- |
| Misassemblies |  |
| # misassemblies | 7 |
| Misassembled contigs length | 3052405 |
| Mismatches |  |
| # mismatches per 100 kbp | 14.24 |
| # indels per 100 kbp | 6.49 |
| # N's per 100 kbp | 0 |
| Statistics without reference |  |
| # contigs | 9 |
| Largest contig | 2773702 |
| Total length | 3126531 |
| Total length (>= 1000 bp) | 3126531 |
| Total length (>= 10000 bp) | 3098699 |
| Total length (>= 50000 bp) | 2992688 |

**Conclusion**: As we see the canu assembly out of PacBio reads polished with pilon using Illumina reads (filtered) gives the best result.

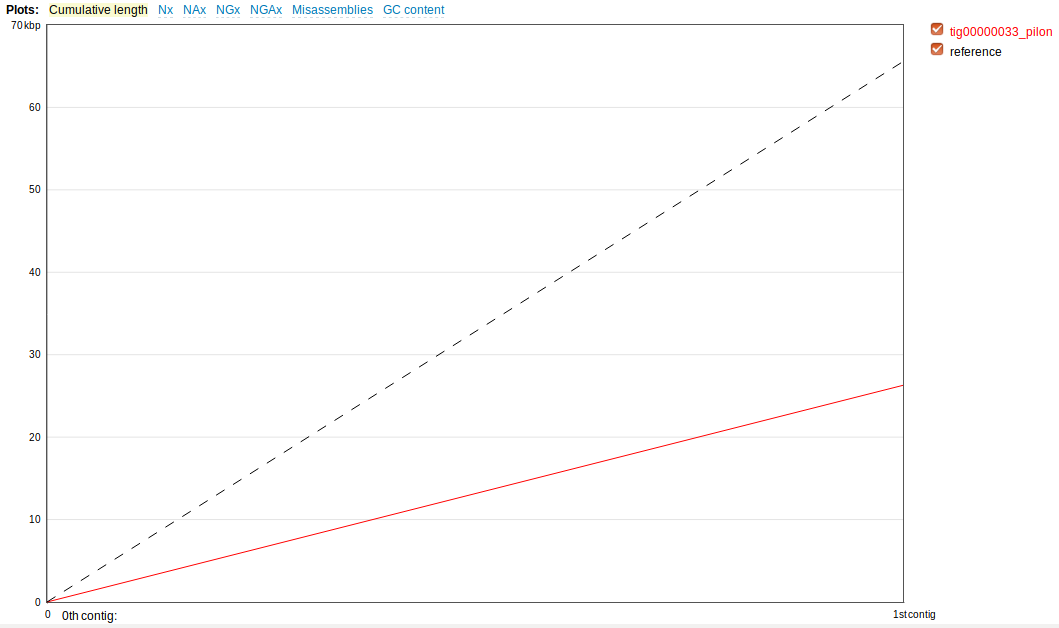
* tig00000025\_pilon (40017bp) uniquely aligns to plasmid 2 (33kb)
* tig00000030\_pilon (11400bp) uniquely aligns to plasmid 3 (9.5kb)
* tig00000031\_pilon (9546bp) uniquely aligns to plasmid 4 (17.6kb)
* tig00000026\_pilon (28307bp) and tig00000470\_pilon (218986bp) aligns to plasmid 5 (56kb)
* **tig00000033\_pilon (26287bp) uniquely aligns to plasmid 6 (66.6kb)**
* tig00000029\_pilon (9986bp) uniquely aligns and tig00000470\_pilon (218986bp) aligns to plasmid 1 (227kb)

However, the large part of plasmid 6 is missing (39271), as contig tig00000033\_pilon (26287bp) uniquely aligns to plasmid 6 (65558bp). Therefore, the assembly with Nanopore reads is still needed.

quast /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/assemblies/canu\_assembly\_corrected/tig00000033\_pilon.fasta -r /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/reference/plasmidpl6\_sequence.fasta -o quast\_comparison\_canu\_vs\_plasmid6

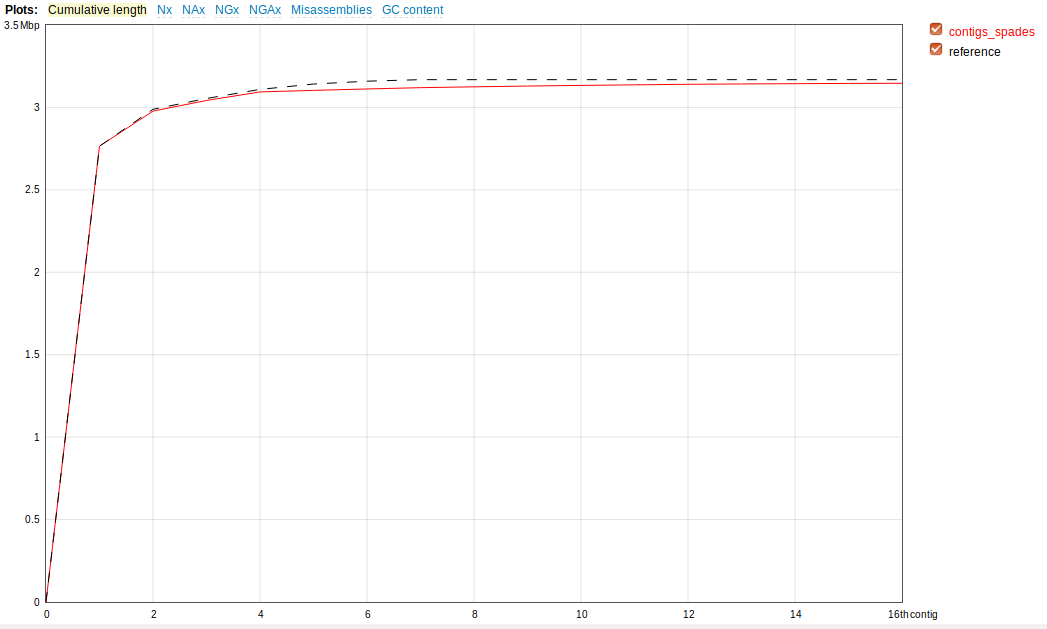
Aligned to "plasmidpl6\_sequence" | 65558 bp | 1 fragment | 32.05 % G+C

| Genome statistics | tig00000033\_pilon |
| --- | --- |
| Genome fraction (%) | 40.099 |



Compare with quast (spades vs reference chromosome – all genome):

2. quast /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/assemblies/spades\_assembly/contigs\_spades.fasta -r /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/reference/the\_full\_reference.fasta -o quast\_comparison\_spades\_vs\_reference\_chromosome



|  |  |
| --- | --- |
| Genome fraction (%) | 98.892 |
| Duplication ratio | 1.002 |
| Largest alignment | 1899903 |
| Total aligned length | 3128539 |
| NGA50 | 1899903 |
| LGA50 | 1 |

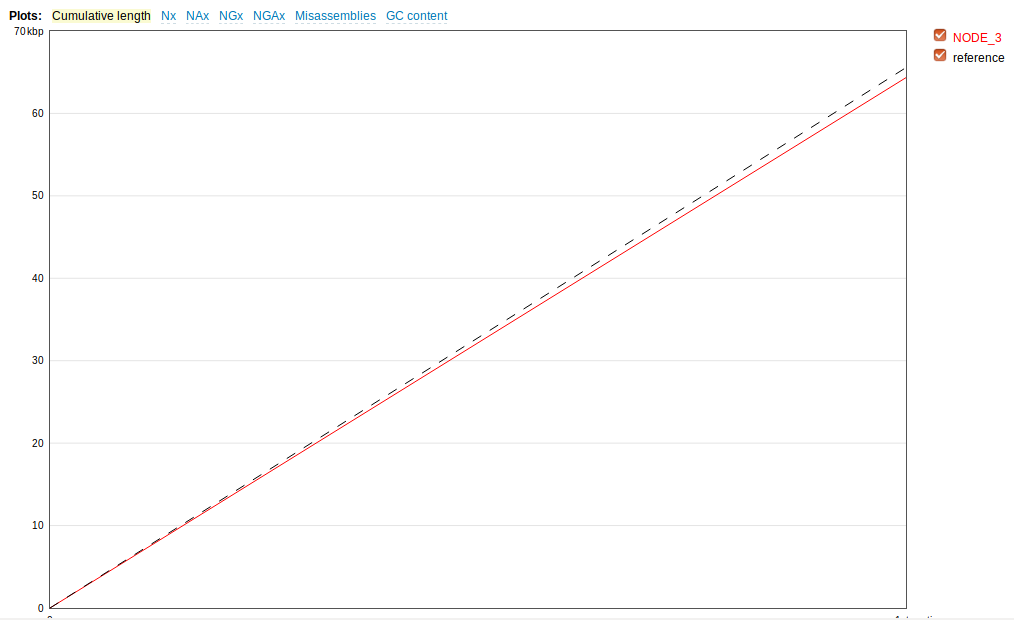
|  |  |
| --- | --- |
| Misassemblies |  |
| # misassemblies | 3 |
| Misassembled contigs length | 2816937 |
| Mismatches |  |
| # mismatches per 100 kbp | 29.97 |
| # indels per 100 kbp | 3.73 |
| # N's per 100 kbp | 0 |
| Statistics without reference |  |
| # contigs | 16 |
| Largest contig | 2765031 |
| Total length | 3146525 |
| Total length (>= 1000 bp) | 3145695 |
| Total length (>= 10000 bp) | 3094039 |
| Total length (>= 50000 bp) | 3094039 |

**Conclusion**: Even though the number of contigs is higher (16 vs 9) and NG50 is worse (1899903 vs 2765031) we were finally able to find the missed part of the plasmid 6.

* NODE\_3\_length\_64343\_cov\_75.9457 (64343bp) (the 3rd largest contig) uniquely aligns to plasmid 6 (66.6kb)

quast /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/assemblies/spades\_assembly/NODE\_3.fasta -r /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/reference/plasmidpl6\_sequence.fasta -o quast\_comparison\_spades\_vs\_plasmid6

| Genome statistics | NODE\_3 |
| --- | --- |
| Genome fraction (%) | 98.147 |
| Duplication ratio | 1 |
| Largest alignment | 64343 |
| Total aligned length | 64343 |
| NGA50 | 64343 |
| LGA50 | 1 |
| Misassemblies |  |
| # misassemblies | 0 |
| Misassembled contigs length | 0 |
| Mismatches |  |
| # mismatches per 100 kbp | 1.55 |
| # indels per 100 kbp | 0 |
| # N's per 100 kbp | 0 |
| Statistics without reference |  |
| # contigs | 1 |
| Largest contig | 64343 |
| Total length | 64343 |
| Total length (>= 1000 bp) | 64343 |
| Total length (>= 10000 bp) | 64343 |
| Total length (>= 50000 bp) | 64343 |



At this point with both these assemblies we could manually assemble contigs from both assemblies to complete chromosomes.

**GENOME ANNOTATION:**

prokka ~/ga\_course/canu\_pilon\_assembly\_110419.fasta --outdir prokka\_for\_canu\_gff --plasmid

**RNA SEQ ANALYSIS:**

**Read trimming:**

trimmomatic PE -phred33 ~/transcriptomics\_data/RNA-Seq\_BH/trim\_paired\_ERR1797974\_pass\_1.fastq.gz ~/transcriptomics\_data/RNA-Seq\_BH/trim\_paired\_ERR1797974\_pass\_2.fastq.gz ~/transcriptomics\_data/trimed\_reads/BH\_trim\_paired\_ERR1797974\_pass\_1\_trimmed.fastq.gz ~/transcriptomics\_data/trimed\_reads/BH\_trim\_paired\_ERR1797974\_pass\_1\_unpaired.fastq.gz ~/transcriptomics\_data/trimed\_reads/BH\_trim\_paired\_ERR1797974\_pass\_2\_trimmed.fastq.gz ~/transcriptomics\_data/trimed\_reads/BH\_trim\_paired\_ERR1797974\_pass\_2\_unpaired.fastq.gz ILLUMINACLIP:/sw/apps/bioinfo/trimmomatic/0.36/rackham/adapters/TruSeq3-PE-2.fa:2:25:10 LEADING:25 TRAILING:25 SLIDINGWINDOW:4:25 MINLEN:36

**11/04/19: Genome assembly assessment**

**24/04/19: Structural and functional annotation**

**25/04/19: RNA-Seq**

Basic analyses:

● Genome assembly with PacBio reads.

● Assembly evaluation.

21● Structural and functional annotation.

● Synteny comparison with a closely related genome.

● Reads preprocessing: trimming + quality check (before and after)

● RNA-Seq reads alignment against assembled genome.

● Differential expression analysis between rich medium and heat-inactivated serum

conditions.

Extra analyses:

● Genome assembly with Illumina and Nanopore reads.

● Assembly evaluation (extra methods).

● Plasmid identification.

● SNPs calling.

● Evaluate antibiotic resistance potential.

● Identify essential genes for growth in human serum based on the Tn-Seq data

analysis.

Day Hours

29/3 2 Genomics projects

4/4 4 Metagenomics project

5/4 4 Project planning

11/4 4 Project planning

12/4 4 Genome Assembly + Genome

annotation (Bact) Assembly

24/4 4 Transcriptome assembly (Euk) Binning

25/4 4 Comparative genomics (Bact)