**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

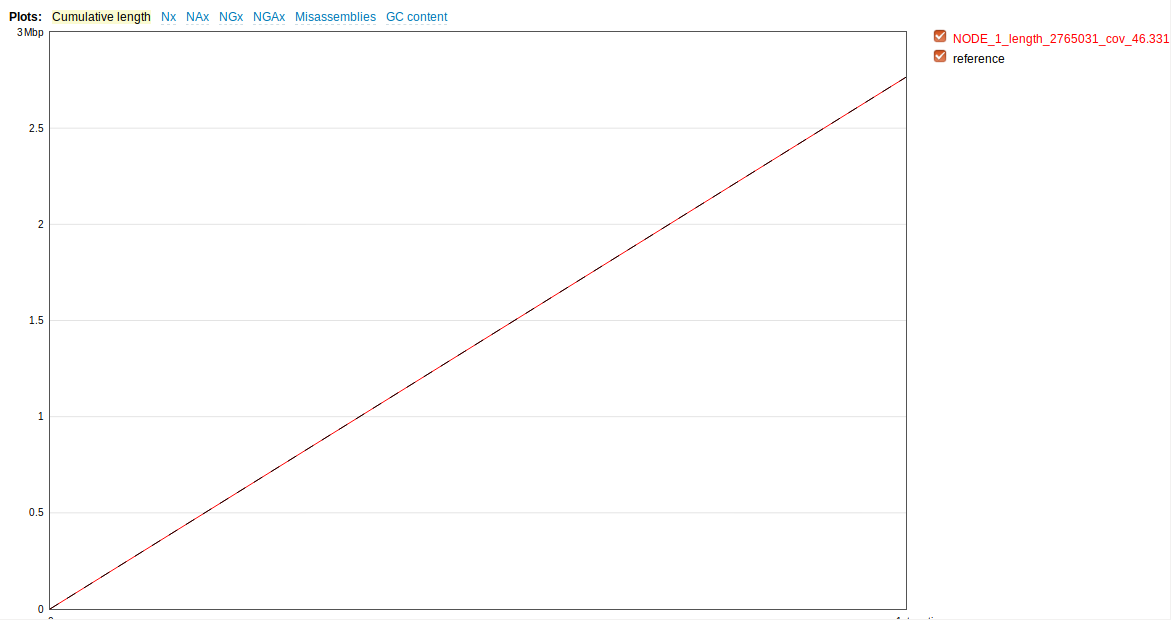
Create the full reference:

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

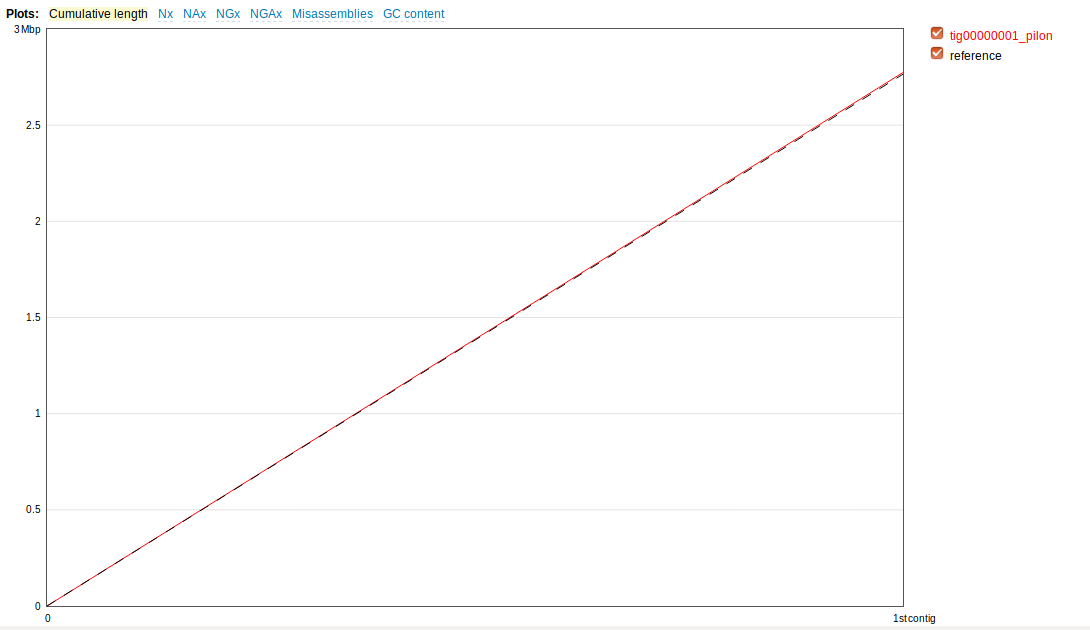
Compare with quast (canu+pilon vs reference chromosome, spades vs reference chromosome):

1.

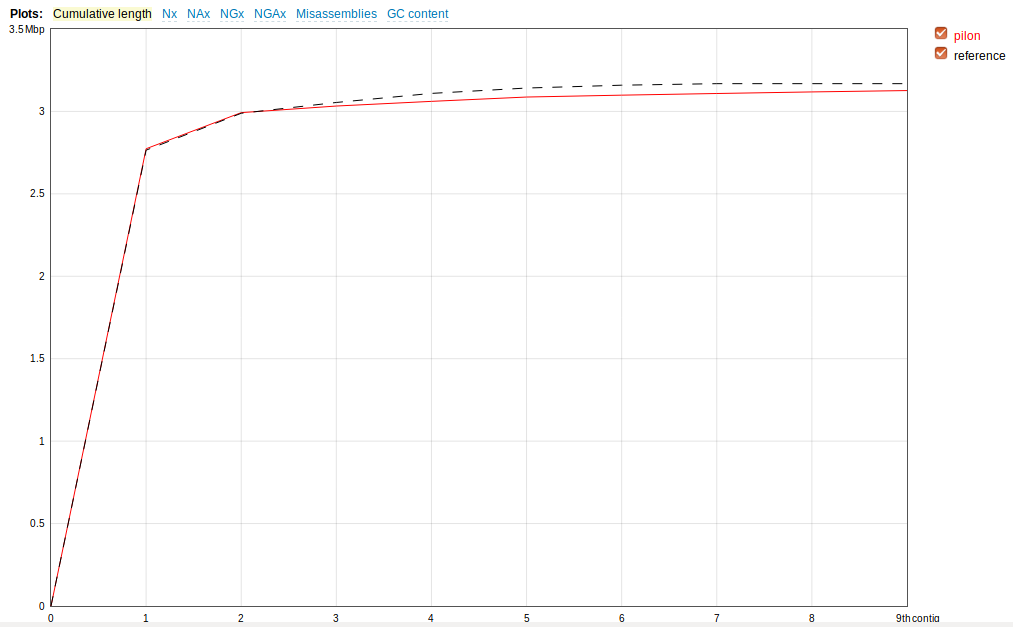
quast /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/assemblies/canu\_assembly\_corrected/pilon.fasta -r /home/erik/sweden/courses/2nd\_semester/Genome\_Analysis/reference/the\_full\_reference.fasta -o quast\_comparison\_canu\_vs\_reference\_chromosome



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.

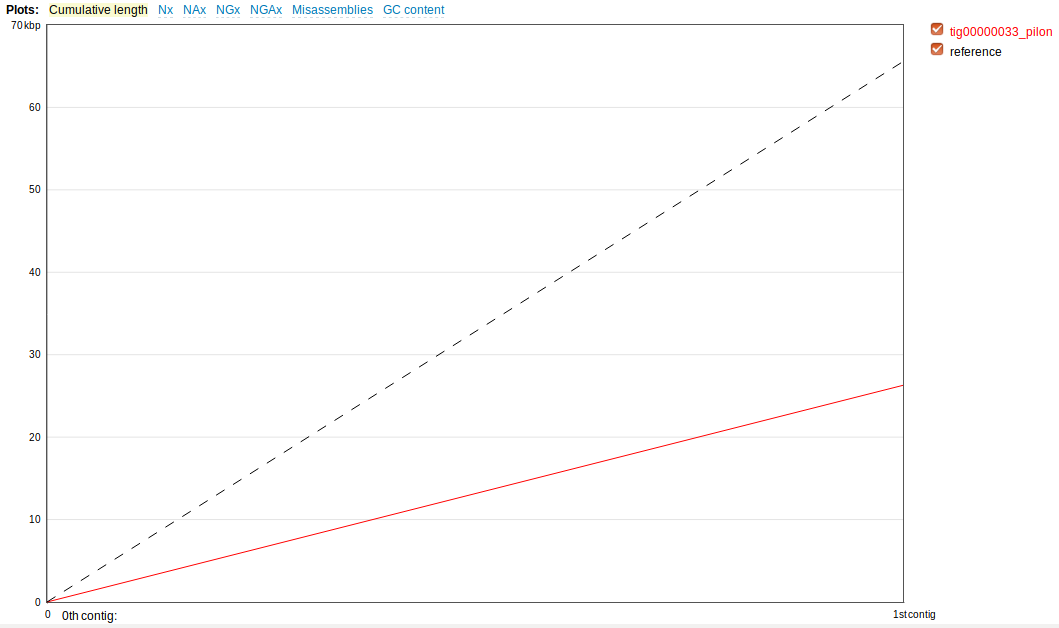
* tig00000025\_pilon (40017bp) uniquely aligns to plasmid 2 (33kb)
* tig00000030\_pilon (11400bp) uniquely aligns to plasmid 3 (9.5kb)
* tig00000031\_pilon (9546bp) uniquely aligns to plasmid 4 (17.6kb)
* tig00000026\_pilon (28307bp) and tig00000470\_pilon (218986bp) aligns to plasmid 5 (56kb)
* **tig00000033\_pilon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

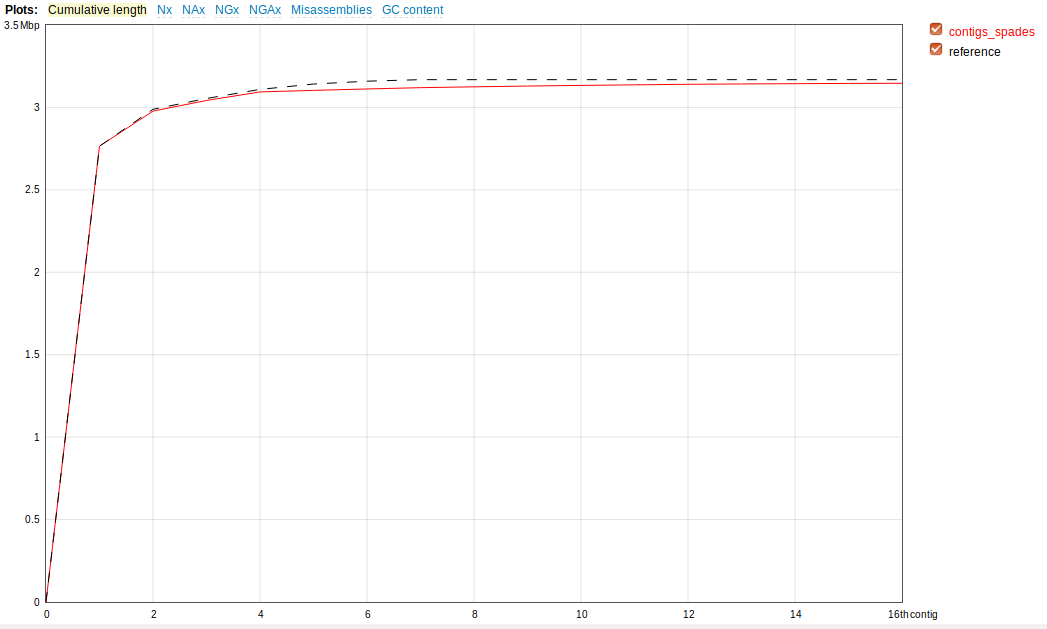
Aligned to "plasmidpl6\_sequence" | 65558 bp | 1 fragment | 32.05 % G+C

| 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 |

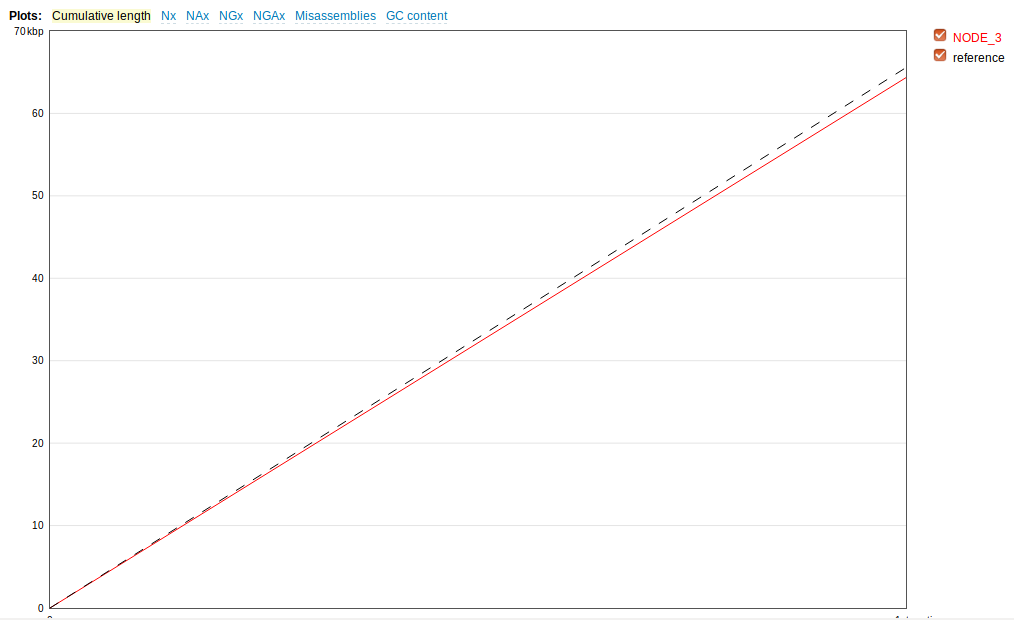
|  |  |
| --- | --- |
| Misassemblies |  |
| # misassemblies | 3 |
| Misassembled contigs length | 2816937 |
| Mismatches |  |
| # mismatches per 100 kbp | 29.97 |
| # indels per 100 kbp | 3.73 |
| # N's per 100 kbp | 0 |
| Statistics without reference |  |
| # contigs | 16 |
| Largest contig | 2765031 |
| 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.

Final assembly assessment:

quast ~/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419.fasta -r / ~/ga\_course/reference/the\_full\_reference.fasta -o ~/ga\_course/final\_assembly\_comparisson

**GENOME ANNOTATION:**

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

organism: Genus species strain

contigs: 9

bases: 3126531

tmRNA: 1

tRNA: 70

**CDS: 3036**

rRNA: 18

prokka ~/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419.fasta --outdir prokka\_for\_spades\_gff

organism: Genus species strain

contigs: 43

bases: 3151876

tmRNA: 1

tRNA: 70

rRNA: 18

**CDS: 3060**

prokka ~/ga\_course/final\_assemblies/spades\_assembly\_110419.fasta prokka\_for\_canu\_and\_spades\_node3\_gff

organism: Genus species strain

contigs: 9

bases: 3164587

tRNA: 70

**CDS: 3083**

rRNA: 18

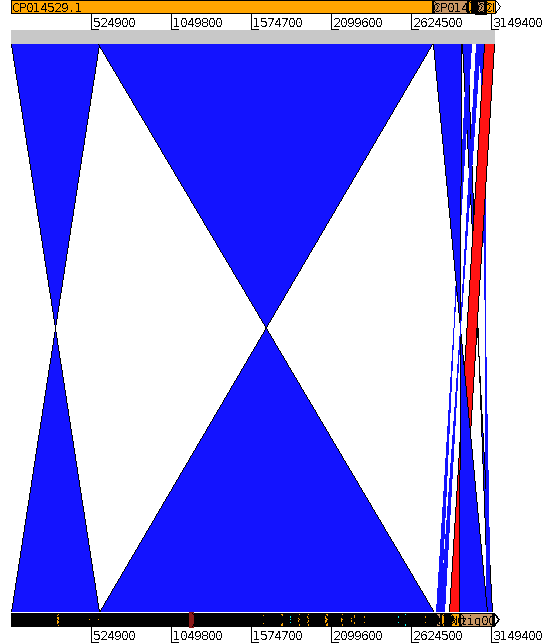
tmRNA: 1

**Syntheny**

./merge\_gbk.py \*.gb > combined.gbk

makeblastdb -in paper\_reference\_whole\_sequence\_1.fasta -dbtype nucl

blastn -query ~/sweden/courses/2nd\_semester/Genome\_Analysis/ga\_course/synteny/PROKKA\_04122019\_1.fna -db ~/sweden/courses/2nd\_semester/Genome\_Analysis/ga\_course/synteny/paper\_reference\_whole\_sequence\_1.fasta -evalue 1 -task megablast -outfmt 6 > syntheny\_comparison\_between\_assemblies\_ours\_vs\_paper.crunch

act

**Synteny conclusion: as we see all annotated genes are present in the same order in chromosome and plasmids. The reason why there are inversions displayed is that our sequence assembly starts in different position in comparison to reference. The same about plasmids, in our sequence they are just in different order.**

**Antibiotic resistance genes**

ResFinder result file

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Trimethoprim

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

RES. GENE IDENTITY QUERY/HSP CONTIG POSITION ACC NO.

=================================================================================

dfrG 100.00 498/498 1 2210..2707 AB205645

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Aminoglycoside

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

RES. GENE IDENTITY QUERY/HSP CONTIG POSITION ACC NO.

=================================================================================

aac(6')-Ii 99.64 549/549 2 2471910..2472458 L12710

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Glycopeptide

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

RES. GENE IDENTITY QUERY/HSP CONTIG POSITION ACC NO.

=================================================================================

VanH 91.85 969/969 3 37402..38370 1\_Y15705

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Glycopeptide

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

RES. GENE IDENTITY QUERY/HSP CONTIG POSITION ACC NO.

=================================================================================

VanA 93.51 1032/1032 4 38363..39394 1\_Y15704

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Glycopeptide

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

RES. GENE IDENTITY QUERY/HSP CONTIG POSITION ACC NO.

=================================================================================

VanX 96.79 249/249 5 39400..39648 1\_Y15708

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Glycopeptide

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

RES. GENE IDENTITY QUERY/HSP CONTIG POSITION ACC NO.

=================================================================================

VanHAX 100.00 2607/2607 6 4980..7586 M97297

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Macrolide

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

RES. GENE IDENTITY QUERY/HSP CONTIG POSITION ACC NO.

=================================================================================

msr(C) 98.99 1479/1479 7 743512..744990 AY004350

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**RNA SEQ ANALYSIS:**

**Read trimming:**

BH reads 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

Serum reads trimming:

trimmomatic PE -phred33 ~/transcriptomics\_data/RNA-Seq\_Serum/untrimmed/ERR1797969\_pass\_1.fastq.gz ~/transcriptomics\_data/RNA-Seq\_Serum/untrimmed/ERR1797969\_pass\_2.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797969\_pass\_1\_trimmed.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797969\_pass\_1\_unpaired.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797969\_pass\_2\_trimmed.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797969\_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

trimmomatic PE -phred33 ~/transcriptomics\_data/RNA-Seq\_Serum/untrimmed/ERR1797970\_pass\_1.fastq.gz ~/transcriptomics\_data/RNA-Seq\_Serum/untrimmed/ERR1797970\_pass\_2.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797970\_pass\_1\_trimmed.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797970\_pass\_1\_unpaired.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797970\_pass\_2\_trimmed.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797970\_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

trimmomatic PE -phred33 ~/transcriptomics\_data/RNA-Seq\_Serum/untrimed/ERR1797971\_pass\_1.fastq.gz ~/transcriptomics\_data/RNA-Seq\_Serum/untrimed/ERR1797971\_pass\_2.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797971\_pass\_1\_trimmed.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797971\_pass\_1\_unpaired.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797971\_pass\_2\_trimmed.fastq.gz ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797971\_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

Library type determination:

scp -r [erzh4483@rackham.uppmax.uu.se](mailto:erzh4483@rackham.uppmax.uu.se):/home/erzh4483/transcriptomics\_data/example/ .

Alignment:

bowtie2 -x final\_assembly\_canu\_and\_spades\_node3\_120419 -1 ~/transcriptomics\_data/trimed\_reads/BH\_trim\_paired\_ERR1797972\_pass\_1\_trimmed.fastq.gz -2 ~/transcriptomics\_data/trimed\_reads/BH\_trim\_paired\_ERR1797972\_pass\_2\_trimmed.fastq.gz -S ERR1797972.sam

bowtie2 -x final\_assembly\_canu\_and\_spades\_node3\_120419 -1 ~/transcriptomics\_data/trimed\_reads/BH\_trim\_paired\_ERR1797973\_pass\_1\_trimmed.fastq.gz -2 ~/transcriptomics\_data/trimed\_reads/BH\_trim\_paired\_ERR1797973\_pass\_2\_trimmed.fastq.gz -S ERR1797973.sam

bowtie2 -x final\_assembly\_canu\_and\_spades\_node3\_120419 -1 ~/transcriptomics\_data/trimed\_reads/BH\_trim\_paired\_ERR1797974\_pass\_1\_trimmed.fastq.gz -2 ~/transcriptomics\_data/trimed\_reads/BH\_trim\_paired\_ERR1797974\_pass\_2\_trimmed.fastq.gz -S ERR1797974.sam

bowtie2 -x final\_assembly\_canu\_and\_spades\_node3\_120419 -1 ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797969\_pass\_1\_trimmed.fastq.gz -2 ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797969\_pass\_2\_trimmed.fastq.gz -S ERR1797969.sam

bowtie2 -x final\_assembly\_canu\_and\_spades\_node3\_120419 -1 ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797970\_pass\_1\_trimmed.fastq.gz -2 ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797970\_pass\_2\_trimmed.fastq.gz -S ERR1797970.sam

bowtie2 -x final\_assembly\_canu\_and\_spades\_node3\_120419 -1 ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797971\_pass\_1\_trimmed.fastq.gz -2 ~/transcriptomics\_data/trimed\_reads/Serum\_trim\_paired\_ERR1797971\_pass\_2\_trimmed.fastq.gz -S ERR1797971.sam

rm \*.sam

Sort:

samtools sort ERR1797969.bam -o ERR1797969\_sorted.bam

rm ERR1797969.bam

samtools sort ERR1797970.bam -o ERR1797970\_sorted.bam

rm ERR1797970.bam

samtools sort ERR1797971.bam -o ERR1797971\_sorted.bam

rm ERR1797971.bam

samtools sort ERR1797972.bam -o ERR1797972\_sorted.bam

rm ERR1797972.bam

samtools sort ERR1797973.bam -o ERR1797973\_sorted.bam

rm ERR1797973.bam

samtools sort ERR1797974.bam -o ERR1797974\_sorted.bam

rm ERR1797974.bam

Alignment rate:

12628611 reads; of these:

12628611 (100.00%) were paired; of these:

1524397 (12.07%) aligned concordantly 0 times

10935663 (86.59%) aligned concordantly exactly 1 time

168551 (1.33%) aligned concordantly >1 times

----

1524397 pairs aligned concordantly 0 times; of these:

1380207 (90.54%) aligned discordantly 1 time

----

144190 pairs aligned 0 times concordantly or discordantly; of these:

288380 mates make up the pairs; of these:

221166 (76.69%) aligned 0 times

27013 (9.37%) aligned exactly 1 time

40201 (13.94%) aligned >1 times

99.12% overall alignment rate

12642446 reads; of these:

12642446 (100.00%) were paired; of these:

1315197 (10.40%) aligned concordantly 0 times

11160537 (88.28%) aligned concordantly exactly 1 time

166712 (1.32%) aligned concordantly >1 times

----

1315197 pairs aligned concordantly 0 times; of these:

1172520 (89.15%) aligned discordantly 1 time

----

142677 pairs aligned 0 times concordantly or discordantly; of these:

285354 mates make up the pairs; of these:

225904 (79.17%) aligned 0 times

25616 (8.98%) aligned exactly 1 time

33834 (11.86%) aligned >1 times

99.11% overall alignment rate

11394287 reads; of these:

11394287 (100.00%) were paired; of these:

1290999 (11.33%) aligned concordantly 0 times

9951140 (87.33%) aligned concordantly exactly 1 time

152148 (1.34%) aligned concordantly >1 times

----

1290999 pairs aligned concordantly 0 times; of these:

1166790 (90.38%) aligned discordantly 1 time

----

124209 pairs aligned 0 times concordantly or discordantly; of these:

248418 mates make up the pairs; of these:

190616 (76.73%) aligned 0 times

23514 (9.47%) aligned exactly 1 time

34288 (13.80%) aligned >1 times

99.16% overall alignment rate

139664 reads; of these:

139664 (100.00%) were paired; of these:

18663 (13.36%) aligned concordantly 0 times

118258 (84.67%) aligned concordantly exactly 1 time

2743 (1.96%) aligned concordantly >1 times

----

18663 pairs aligned concordantly 0 times; of these:

16763 (89.82%) aligned discordantly 1 time

----

1900 pairs aligned 0 times concordantly or discordantly; of these:

3800 mates make up the pairs; of these:

2813 (74.03%) aligned 0 times

357 (9.39%) aligned exactly 1 time

630 (16.58%) aligned >1 times

98.99% overall alignment rate

13601980 reads; of these:

13601980 (100.00%) were paired; of these:

1548674 (11.39%) aligned concordantly 0 times

11757103 (86.44%) aligned concordantly exactly 1 time

296203 (2.18%) aligned concordantly >1 times

----

1548674 pairs aligned concordantly 0 times; of these:

1373036 (88.66%) aligned discordantly 1 time

----

175638 pairs aligned 0 times concordantly or discordantly; of these:

351276 mates make up the pairs; of these:

257934 (73.43%) aligned 0 times

36753 (10.46%) aligned exactly 1 time

56589 (16.11%) aligned >1 times

99.05% overall alignment rate

12947386 reads; of these:

12947386 (100.00%) were paired; of these:

1581646 (12.22%) aligned concordantly 0 times

11096437 (85.70%) aligned concordantly exactly 1 time

269303 (2.08%) aligned concordantly >1 times

----

1581646 pairs aligned concordantly 0 times; of these:

1410695 (89.19%) aligned discordantly 1 time

----

170951 pairs aligned 0 times concordantly or discordantly; of these:

341902 mates make up the pairs; of these:

253158 (74.04%) aligned 0 times

34765 (10.17%) aligned exactly 1 time

53979 (15.79%) aligned >1 times

99.02% overall alignment rate

Index:

#!/bin/bash

for i in \*.bam

do

echo "Indexing: "$i

samtools index $i $i".bai"

done

IGV visualization on server:

[erzh4483@rackham2 aligned\_reads]$ module load java

[erzh4483@rackham2 aligned\_reads]$ module load IGV

[erzh4483@rackham2 aligned\_reads]$ igv-node

htseq counting:

htseq-count ~/transcriptomics\_data/aligned\_reads/ERR1797970\_sorted.bam PROKKA\_04122019.gtf -f bam -t CDS > ERR1797970\_transcripts.txt

**Tn-Seq Analysis**

module load bioinfo-tools

module load bowtie2

Trimming:

trimmomatic SE -phred33 ~/tnseq\_data/Tn-Seq\_BHI\_trim\_ERR1801012\_pass.fastq.gz ~/tnseq\_data/12\_bp/Tn-Seq\_BHI\_trim\_ERR1801012\_pass.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 CROP:16

Mapping:

bowtie2-build /home/erzh4483/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419.fasta final\_assembly\_canu\_and\_spades\_node3\_120419

bowtie2 -x ~/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419 -U ~/tnseq\_data/Tn-Seq\_BHI/trim\_ERR1801012\_pass\_passed.fastq.gz -S ~/tnseq\_data/aligned\_reads/Tn-Seq\_BHI\_ERR1801012.sam

6463683 reads; of these:

6463683 (100.00%) were unpaired; of these:

3427867 (53.03%) aligned 0 times

1529926 (23.67%) aligned exactly 1 time

1505890 (23.30%) aligned >1 times

46.97% overall alignment rate

run\_trimmo.sh

run\_bowtie2.sh

Do DE2-Seq analysis.

**SNP Calling**

bowtie2 -x ~/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419 -1 ~/genomics\_data/Illumina/E745-1.L500\_SZAXPI015146-56\_1\_clean\_trimmed.fq.gz -2 ~/genomics\_data/Illumina/E745-1.L500\_SZAXPI015146-56\_2\_clean\_trimmed.fq.gz -S Illumina\_reads\_for\_SNP.sam

samtools view -Sb Illumina\_reads\_for\_SNP.sam > Illumina\_reads\_for\_SNP.bam

rm \*.sam

samtools sort Illumina\_reads\_for\_SNP.bam -o Illumina\_reads\_for\_SNP\_sorted.bam

samtools index Illumina\_reads\_for\_SNP\_sorted.bam Illumina\_reads\_for\_SNP\_sorted.bai

rm Illumina\_reads\_for\_SNP.bam

bcftools mpileup Illumina\_reads\_for\_SNP\_sorted.bam -o Illumina\_reads\_for\_SNP\_sorted.vcf -O v -f ~/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419.fasta

**bcftools mpileup Illumina\_reads\_for\_SNP\_sorted.bam -Ou -B -f ~/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419.fasta --min-MQ 30 -o Illumina\_reads\_for\_SNP\_sorted.vcf**

**bcftools call Illumina\_reads\_for\_SNP\_sorted.vcf -Ou -v -m -o Illumina\_reads\_for\_SNP\_sorted\_1.vcf**

**bcftools norm Illumina\_reads\_for\_SNP\_sorted\_1.vcf -Ou -f ~/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419.fasta -d all -o Illumina\_reads\_for\_SNP\_sorted\_2.vcf**

**bcftools filter Illumina\_reads\_for\_SNP\_sorted\_2.vcf -Ov -e 'QUAL<40 || DP<10 || GT!="1/1"' -o Illumina\_reads\_for\_SNP\_sorted\_3.vcf**

# SN, Summary numbers:

# SN [2]id [3]key [4]value

SN 0 number of samples: 1

SN 0 number of records: 3165922

SN 0 number of no-ALTs: 2693341

SN 0 number of SNPs: 470442

SN 0 number of MNPs: 0

SN 0 number of indels: 2139

SN 0 number of others: 0

SN 0 number of multiallelic sites: 470469

SN 0 number of multiallelic SNP sites: 470442

**Let’s do the same with McGann data!**

**1. QC (Phred score=30):**

java -jar /sw/apps/bioinfo/trimmomatic/0.36/rackham/trimmomatic-0.36.jar PE -phred33 /home/erzh4483/mcgann/SRR3306347\_1.fastq.gz /home/erzh4483/mcgann/SRR3306347\_2.fastq.gz /home/erzh4483/mcgann/SRR3306347\_1\_trimmed.fastq.gz /home/erzh4483/mcgann/SRR3306347\_1\_unpaired.fastq.gz /home/erzh4483/mcgann/SRR3306347\_2\_trimmed.fastq.gz /home/erzh4483/mcgann/SRR3306347\_2\_unpaired.fastq.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

java -jar /sw/apps/bioinfo/trimmomatic/0.36/rackham/trimmomatic-0.36.jar PE -phred33 /home/erzh4483/mcgann/SRR3306348\_1.fastq.gz /home/erzh4483/mcgann/SRR3306348\_2.fastq.gz /home/erzh4483/mcgann/SRR3306348\_1\_trimmed.fastq.gz /home/erzh4483/mcgann/SRR3306348\_1\_unpaired.fastq.gz /home/erzh4483/mcgann/SRR3306348\_2\_trimmed.fastq.gz /home/erzh4483/mcgann/SRR3306348\_2\_unpaired.fastq.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

java -jar /sw/apps/bioinfo/trimmomatic/0.36/rackham/trimmomatic-0.36.jar PE -phred33 /home/erzh4483/mcgann/SRR3306349\_1.fastq.gz /home/erzh4483/mcgann/SRR3306349\_2.fastq.gz /home/erzh4483/mcgann/SRR3306349\_1\_trimmed.fastq.gz /home/erzh4483/mcgann/SRR3306349\_1\_unpaired.fastq.gz /home/erzh4483/mcgann/SRR3306349\_2\_trimmed.fastq.gz /home/erzh4483/mcgann/SRR3306349\_2\_unpaired.fastq.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

**bcftools mpileup SRR3306347\_1\_trimmed.fastq.gz.sorted.bam -Ou -B -f ~/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419.fasta --min-MQ 60 -o SRR3306347\_1\_trimmed.fastq.gz.sorted.bam.vcf**

**bcftools call SRR3306347\_1\_trimmed.fastq.gz.sorted.bam.vcf -Ou -v -m -o SRR3306347\_1\_trimmed.fastq.gz.sorted.bam.vcf\_1**

**bcftools norm SRR3306347\_1\_trimmed.fastq.gz.sorted.bam.vcf\_1 -Ou -f ~/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419.fasta -d all -o** **SRR3306347\_1\_trimmed.fastq.gz.sorted.bam.vcf\_2**

**Lines total/split/realigned/skipped: 8006/0/183/0**

**[erzh4483@rackham3 maping]$ bcftools norm SRR3306348\_1\_trimmed.fastq.gz.sorted.bam.vcf\_1 -Ou -f ~/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419.fasta -d all -o SRR3306348\_1\_trimmed.fastq.gz.sorted.bam.vcf\_2**

Lines total/split/realigned/skipped: 8484/0/184/0

[erzh4483@rackham3 maping]$ bcftools norm SRR3306349\_1\_trimmed.fastq.gz.sorted.bam.vcf\_1 -Ou -f ~/ga\_course/final\_assemblies/final\_assembly\_canu\_and\_spades\_node3\_120419.fasta -d all -o SRR3306349\_1\_trimmed.fastq.gz.sorted.bam.vcf\_2

Lines total/split/realigned/skipped: 8288/0/169/0

**bcftools filter SRR3306347\_1\_trimmed.fastq.gz.sorted.bam.vcf\_2 -Ov -e 'QUAL<40 || DP<10 || GT!="1/1"' -o SRR3306347\_1\_trimmed.fastq.gz.sorted.bam.vcf\_3**

**Annotation of vcf files:**

PicardCommandLine CreateSequenceDictionary R=final\_assembly\_canu\_and\_spades\_node3\_120419.fasta O=final\_assembly\_canu\_and\_spades\_node3\_120419.fasta.dict

java -jar picard.jar AddOrReplaceReadGroups -I SRR3306347\_1\_trimmed.fastq.gz.sorted.bam -O SRR3306347\_1\_trimmed.fastq.gz.sorted.groups.bam

java -jar picard.jar AddOrReplaceReadGroups -I SRR3306348\_1\_trimmed.fastq.gz.sorted.bam -O SRR3306348\_1\_trimmed.fastq.gz.sorted.groups.bam

java -jar picard.jar AddOrReplaceReadGroups -I SRR3306349\_1\_trimmed.fastq.gz.sorted.bam -O SRR3306349\_1\_trimmed.fastq.gz.sorted.groups.bam

java -jar GenomeAnalysisTK.jar \

-R final\_assembly\_canu\_and\_spades\_node3\_120419.fasta \

-T VariantAnnotator \

-I SRR3306347\_1\_trimmed.fastq.gz.sorted.groups.bam \

-V SRR3306347\_1\_trimmed.fastq.gz.sorted.bam.vcf \

-o SRR3306347\_1\_trimmed.fastq.gz.sorted.bam.annotated.vcf \

-A Coverage \

-L SRR3306347\_1\_trimmed.fastq.gz.sorted.bam.vcf \

java -jar GenomeAnalysisTK.jar \

-R final\_assembly\_canu\_and\_spades\_node3\_120419.fasta \

-T VariantAnnotator \

-I SRR3306348\_1\_trimmed.fastq.gz.sorted.groups.bam \

-V SRR3306348\_1\_trimmed.fastq.gz.sorted.bam.vcf \

-o SRR3306348\_1\_trimmed.fastq.gz.sorted.bam.annotated.vcf \

-A Coverage \

-L SRR3306348\_1\_trimmed.fastq.gz.sorted.bam.vcf \

java -jar GenomeAnalysisTK.jar \

-R final\_assembly\_canu\_and\_spades\_node3\_120419.fasta \

-T VariantAnnotator \

-I SRR3306349\_1\_trimmed.fastq.gz.sorted.groups.bam \

-V SRR3306349\_1\_trimmed.fastq.gz.sorted.bam.vcf \

-o SRR3306349\_1\_trimmed.fastq.gz.sorted.bam.annotated.vcf \

-A Coverage \

-L SRR3306349\_1\_trimmed.fastq.gz.sorted.bam.vcf \

**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)

Try LastZ for pairwise WholeGenomeAlignment

Busco: coding genome representation

5. An essential first step in comparative genomics analysis is to align whole genomes of

multiple species and identify orthologous regions. However, this essential step is not as

straightforward as one might think, particularly when dealing with eukaryotes. Please

explain what would be potentially confounding factors that makes this step difficult in

eukaryotes. List at least 4 factors and explain them. (4p)

* Polyploidy
* Introns(?)
* Sex chromosomes