Lab Journal

@October 20, 2022 12:00

Beginning

- Creted new virtual environment in conda named *antib_proj:*

```
conda create --name antib_proj
conda activate antib_proj
```

- Created directory for project called *antibiotics* in directory *projects* that is in *home directory*

```
mkdir antibotics_1
```

- Downloaded reference genome, annotation file and reads .fastq files directly from links and added them in project directory

Checked if files has a correct format - everything is okay

```
head -n 4 amp_res_1.fastq

#@SRR1363257.37 GWZHISEQ01:153:C1W31ACXX:5:1101:14027:2198 length=101
#GGTTGCAGATTCGCAGTGTCGCTGTTCCAGCGCATCACATCTTTGATGTTCACGCCGTGGCGTTTAGCAATGCTTGAAAGCGAATCGCCTTTGCCCACACG
#+
#@?:=:;DBFADH;CAECEE@@E:FFHGAE4?C?DE<BFGEC>?>FHE4BFFIIFHIBABEECA83;>>@>@CCCDC9@@CC08<@?@BB@9:CC#######

head -n 4 amp_res_2.fastq

#@SRR1363257.37 GWZHISEQ01:153:C1W31ACXX:5:1101:14027:2198 length=101
#GATCTAAGCTGAAGCCAGGCCAAAGTTTGACGATTGGTGCAGGCAGTAGCGCACAGCGACTGGCAAACAACAGCGATAGCATTACGTATCGTGTGCGCAAA
#+
#???BDB:DFHBFD@9;;+A;AFGH;ABHFHHGE@9:B:??@D>@;F?D8<<F8AA9EHHD8'..;5?A?A992(',(59CC3@C>22::A238+2>B<>B<
```

```
less GCF_000005845.2_ASM584v2_genomic.fna
less GCF_000005845.2_ASM584v2_genomic.gff
```

- Checked number of lines - they are equal for both of .fastq files

```
wc -l amp_res_1.fastq    #1823504
wc -l amp_res_2.fastq    #1823504
```

That's mean that there are $\frac{1823504}{4} = 455876$ reads

Quality control

- Installed **FastQC** and checked if it works properly - everything is okay

```
conda install fastqc
```

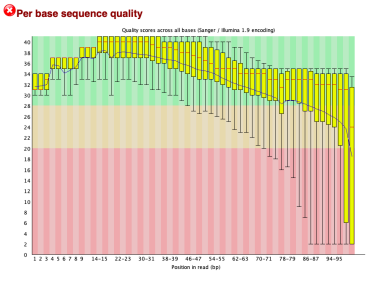
- Launched FastQC with our .fastq files

```
fastqc -o . amp_res_1.fastq amp_res_2.fastq
```

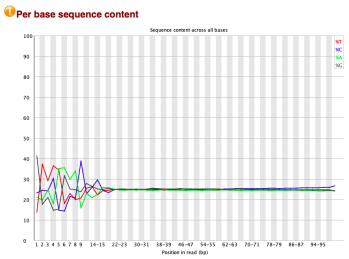
HTML files with reports were created

amp_res_1_fastqc.html

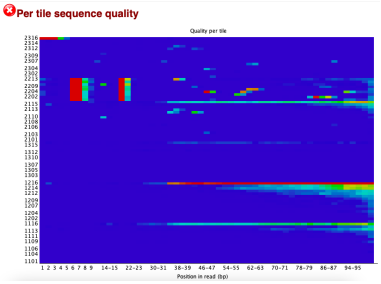
The most problem statistics:



Adapters and the quality of calls on most platforms will degrade as the run progresses



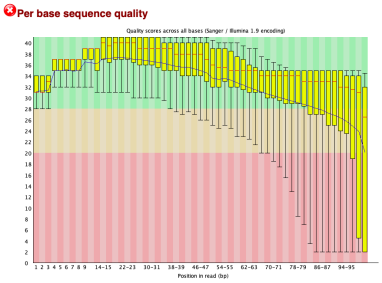
problem connected with Illumina pipeline (adapters)



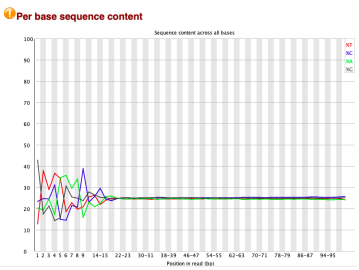
BUBBLES

amp_res_2_fastqc.html

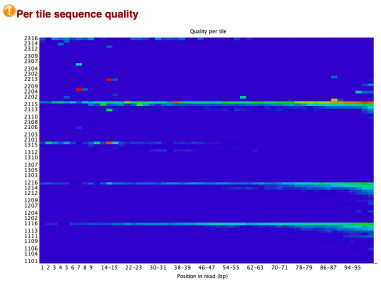
The most problem statistics:



Adapters and the quality of calls on most platforms will degrade as the run progresses



problem connected with Illumina pipeline (adapters)



BUBBLES

Filtering

- Installed **Trimmomatic**

```
conda install trimmomatic
```

- Filtered paired-sequences using Trimmomatic with options:
  - Cut bases off the start of a read if quality below 20
  - Cut bases off the end of a read if quality below 20
  - Trim reads using a sliding window approach, with window size 10 and average quality within the window 20.
  - Drop the read if it is below length 20.
  - Phred33 scale

```
trimmomatic PE -phred33 amp_res_1.fastq amp_res_2.fastq output_amp1_paired.fq.gz  output_amp1_unpaired.fq.gz output_amp2_paire
d.fq.gz output_amp2_unpaired.fq.gz LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:20

#Multiple cores found: Using 4 threads
#Input Read Pairs: 455876 Both Surviving: 446259 (97,89%) Forward Only Surviving: 9216 (2,02%) Reverse Only Surviving: 273 (0,0
6%) Dropped: 128 (0,03%)
#TrimmomaticPE: Completed successfully
```

- Checked output's number of sequences:

```
zcat < output_amp1_paired.fq.gz | wc -l    #1785036
zcat < output_amp1_unpaired.fq.gz | wc -l    #36864
zcat < output_amp2_paired.fq.gz | wc -l    #1785036
zcat < output_amp2_unpaired.fq.gz | wc -l    #1092
```

Number of sequences passed the filter:

1. amp1_paired = 446,259
2. amp1_unpaired = 9,216
3. amp2_paired = 446,259
4. amp2_unpaired = 273

- FastQC analysis on filtered paired data

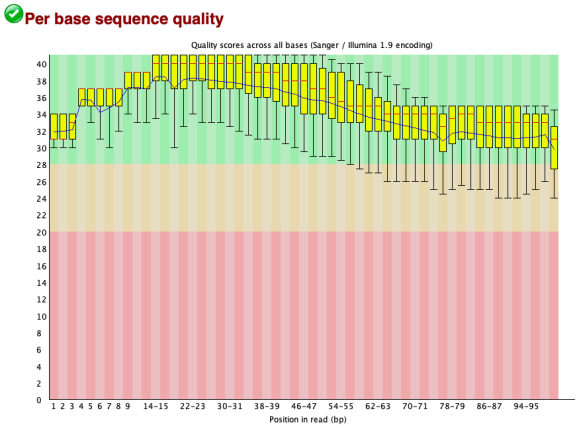
```
gunzip *_paired.fq.gz
fastqc -o . output_amp1_paired.fq output_amp2_paired.fq
```

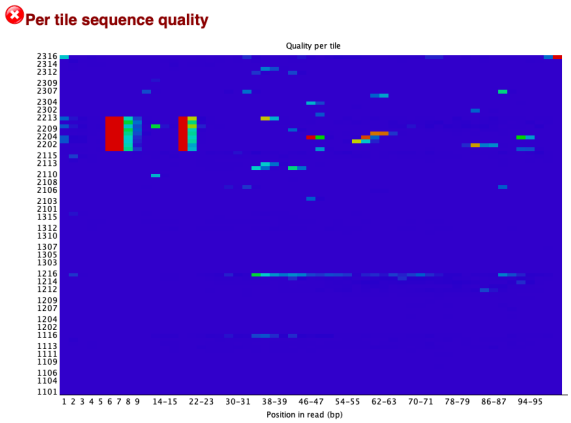
Cretaed .html files with report

And it is better now!

output_amp1_paired_fastqc.html

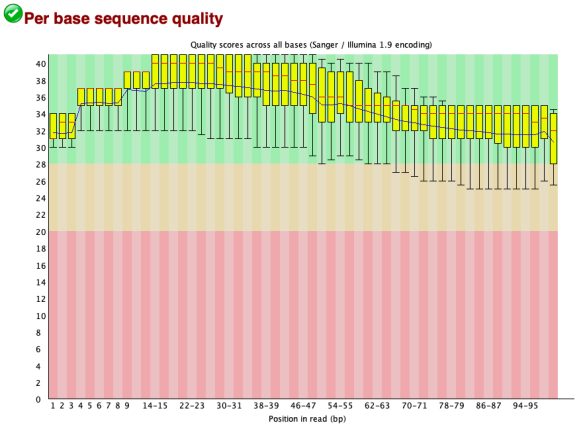
New statistics:

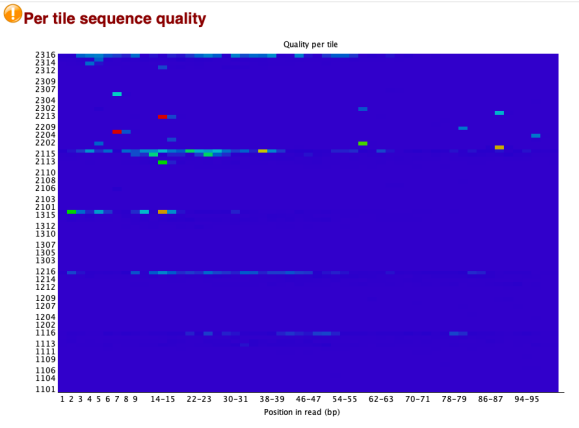




output_amp2_paired_fastqc.html

New statistics:





Alignment

- Installed a **BWA** package, made shorter link reference genome file and made an index of it

```
conda install bwa
ln -s GCF_000005845.2_ASM584v2_genomic.fna GCF.fna
bwa index GCF.fna
```

Created new files

GCF.fna.amb GCF.fna.ann GCF.fna.bwt GCF.fna.pac GCF.fna.sa

- Aligned reads

```
bwa mem GCF.fna output_amp1_paired.fq output_amp2_paired.fq > alignment.sam
```

Created file **alignemnts.sam**

- Made a **.bam** file with **samtools**

```
conda install samtools
samtools view -b alignment.sam > alignment.bam
```

- Looked up statistics

```
samtools flagstat alignment.bam

#892776 + 0 in total (QC-passed reads + QC-failed reads)
#892518 + 0 primary
#0 + 0 secondary
#258 + 0 supplementary
#0 + 0 duplicates
#0 + 0 primary duplicates
#891649 + 0 mapped (99.87% : N/A)
#891391 + 0 primary mapped (99.87% : N/A)
#892518 + 0 paired in sequencing
#446259 + 0 read1
#446259 + 0 read2
#888554 + 0 properly paired (99.56% : N/A)
#890412 + 0 with itself and mate mapped
#979 + 0 singletons (0.11% : N/A)
#0 + 0 with mate mapped to a different chr
#0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Mapped 99.56% of paired-end reads

- Sorted a .bam file and indexed it

```
samtools sort alignment.bam -o alignment_sorted.bam
samtools index alignment_sorted.bam
```

Created file alignment_sorted.bam.bai

- downloaded IGV browser of version 2.14.1 and visualized alignment

@October 20, 2022 18:00

SNP calling

- Generated text pileup output for alignment_sorted.bam file. (index needed!)

```
samtools mpileup -f GCF.fna alignment_sorted.bam > my.mpileup
```

- Calling SNP variants using **VarScan**

```
conda install varscan
varscan mpileup2snp my.mpileup --min-var-freq 0.3 --variants --output-vcf 1 > VarScan_results.vcf

#Only SNPs will be reported
#Warning: No p-value threshold provided, so p-values will not be calculated
#Min coverage:  8
#Min reads2:  2
#Min var freq:  0.3
#Min avg qual:  15
#P-value thresh:  0.01
#Reading input from my.mpileup

#4641343 bases in pileup file
#9 variant positions (6 SNP, 3 indel)
#0 were failed by the strand-filter
#6 variant positions reported (6 SNP, 0 indel)
```

The minimum % of non-reference bases was get in article (0.15) and than been increasing. Result of calling did not changed.

- Loaded in IGV browser .vcf ang .gff files (SNPs and anotations)
- Found genes with SNPs:
  - ftsI - ncbi
  - acrB - ncbi
  - rybA - ncbi
  - mntP -biocyc ncbi
  - envZ - biocyc article
  - rsgA - biocyc ncbi

@October 23, 2022 16:00

Annotation

- installed SnpEff

```
conda install snpeff
```

- Created directory data/k12
- Downloaded GCF_000005845.2_ASM584v2_genomic.gbff.gz file and moved it to this directory (renamed it to genes.gbk after unarchiving)
- created database and annotated SNPs:

```
snpeff build -genbank -v k12
snpeff ann k12 VarScan_results.vcf > snp_ann.vcf
```

- Looking up the results in IGV browser

Gene	Position	Reference nucleotide	Alternative Nucleotide	Reference aminoacid	Altternative aminoacid	Variants Effect
ftsI	93043	C	G	Ala	Gly	Missence
acrB	482698	A	T	Gln	Leu	Missence
rybA	852762	T	C	Phe	Ser	Missence
mntP	1905761	G	A	Gly	Asp	Missence
envZ	3535147	T	G	Val	Gly	Missence

Gene	Position	Reference nucleotide	Alternative Nucleotide	Reference aminoacid	Altternative aminoacid	Variants Effect
rsgA	4390754	C	A	Ala	Ala	Synonymous

@October 24, 2022 18:00

Analysis

Gene	Encoded protein	Function
ftsI	Encode Peptidoglycan D,D-transpeptidase	Essential cell division protein that catalyzes cross-linking of the peptidoglycan cell wall at the division septum
acrB	Multidrug efflux pump subunit AcrB	AcrA-AcrB-AcrZ-TolC is a drug efflux protein complex with broad substrate specificity
rybA (mntS)	Small protein MntS	Synthesis of MntS increases the total intracellular levels of Mn2+, likely by directly or indirectly inhibiting manganese export through MntP. Under peroxide stress conditions it is upregulated.
mntP	Probable manganese efflux pump MntP	Manganese ion transmembrane exporter
envZ	Sensor histidine kinase EnvZ	Member of the two-component regulatory system EnvZ/OmpR involved in osmoregulation. Involved in acid stress response
rsgA	Small ribosomal subunit biogenesis GTPase RsgA	One of proteins that assist in the late maturation steps of the functional core of the 30S ribosomal subunit.

ftsI: mutation in ftsI protein enables peptidoglycan synthesis in presence of ampicillin.

acrB: Increased efflux of ampicillin from periplasm space could prevent ampicillin from binding to PBPs.

envZ: OmpR-EnvZ complex is responsible for osmotic pressure regulation. Control of this process happens via the regulation of genes of porins (ompF and ompC). If porins form in smaller quantities, diffusion of antibiotics across the membrane gets worse. Therefore, a cell with such a mutation may have greater resistance to ampicillin.

mntS and mntP regulate magnesium homeostasis in the cell. It is unlikely that these mutations cause antibiotic resistance.

rsgA gene that assists in the late maturation steps of the functional core of the 30S ribosomal subunit had an synonymous mutation. Have no effect (?)