Project 2 - Why did I get the flu? - lab notebook

Downloading roommate results file and inspecting it:

```
wget http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/SRR1705851.fastq.gz
```

```
zless SRR1705851.fastq.gz | wc -l
#Result
1433060 / 4 = 358265
```

```
fastqc '/home/anya/Рабочий стол/IB/workshop/project2/SRR1705851.fastq'
```

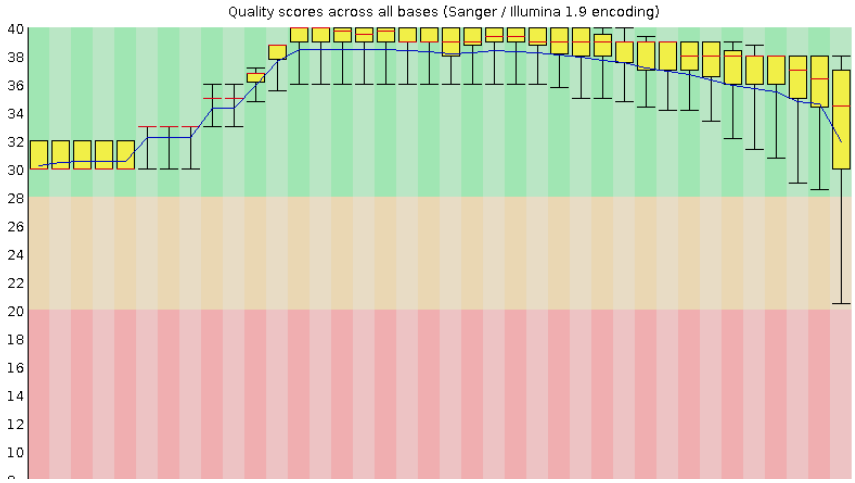
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	SRR1705851.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	358265
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	42

Per base sequence quality



Downloading and indexing reference sequence from NCBI Database

Influenza A virus (A/USA/RVD1_H3/2011(H3N2)) segment 4 hemagglutinin (HA) gene, partial cds

Creating fast pipeline with snakefile for aligning with reference sequence, creating mpileupfile and looking for common variants with VarScan

And creating config file, where we can change URL and percent of variant frequency

Snakefile and config.yaml file can be found in our repository on github BIOINF_Influenza

Depth for mpileup was chosen as -d 0 to detect all possible variants

Inspecting VarScan results for roommate data

```
percent: 0.95

Min coverage: 8
Min reads2: 2
Min var freq: 0.95
Min avg qual: 15
P-value thresh: 0.01
Reading input from my_d0.mpileup
1665 bases in pileup file
5 variant positions (5 SNP, 0 indel)
0 were failed by the strand-filter
5 variant positions reported (5 SNP, 0 indel)
```

```
percent: 0.001

Min coverage: 8
Min reads2: 2
Min var freq: 0.001
Min avg qual: 15
P-value thresh: 0.01
Reading input from my_d0.mpileup
1665 bases in pileup file
23 variant positions (21 SNP, 2 indel)
0 were failed by the strand-filter
21 variant positions reported (21 SNP, 0 indel)
```

Inspecting VarScan results for control samples

Just changing URL in yaml.config file

SRR1705858: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz

SRR1705859: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz

SRR1705860: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz

Copying results from vcf file into .xlsx table

Head of the table looks like this:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
KF848938.1	38	.	T	C	.	PASS	ADP=2740;W
KF848938.1	54	.	T	C	.	PASS	ADP=6393;W
KF848938.1	72	.	A	G	.	PASS	ADP=8919;W

Reading .xlsx table in R and selecting columns of interest

Main code points:

1. Separating column 14 with symbol “:” using library stringr. This step creates 15 new columns, from with only one column with percentage can be extracted for further analysis

   ```
   separated_column <- str_split_fixed(selected$Sample1, ":", 14)
   ```

2. Making column with percentage numeric

```
result$pr <- as.factor(result$pr)
result$pr <- as.character(result$pr)
result$pr <- gsub("%", "", result$pr)
result$pr <- gsub(",", ".", result$pr)
result$pr <- as.numeric(result$pr)
```

3. Counting standard deviations and frequencies for each control vcf file

```
vcf58 <- subset(result, vcf == 58)

mean(vcf58$pr)
sd(vcf58$pr)
```

Standart deviations and means for control:

vcf	58	59	60
mean	0,2565%	0,2369%	0,2503%
sd	0,0717%	0,0766%	0,0520%

Comparing roommate Varscan results with control results mean

Searching for positions in roommate results with frequencies that are more that 3 standard deviations away.

	pos	ref	alt	percent	
1	72	A	G	99,96%	T -> T
11	774	T	C	99,96%	F -> F
19	1260	A	C	99,94%	L -> L
15	999	C	T	99,86%	G -> G
2	117	C	T	99,82%	A -> A
5	307	C	T	0,94%	P -> S
21	1458	T	C	0,84%	Y -> Y
12	802	A	G	0,23%	
7	389	T	C	0,22%	
18	1213	A	G	0,22%	
17	1086	A	G	0,21%	
9	722	A	G	0,2%	
14	915	T	C	0,19%	
13	859	A	G	0,18%	
16	1043	A	G	0,18%	
20	1280	T	C	0,18%	
3	254	A	G	0,17%	
4	276	A	G	0,17%	
6	340	T	C	0,17%	
8	691	A	G	0,17%	
10	744	A	G	0,17%	

Two interesting mutations found, one of them changes amino-acid on the epitope of the virus.