Project 1: SARS-CoV-2 genome assembly from Illumina reads

Course: SARS-2 Bioinformatics & Data Science

Abhinav Mishra

29th September 2023

Background

(perspective is important !)

What's it good for ?

Solutions	Importance
• Lineage assignment • Genomic profiling • Surveillance • Outbreak & Cluster • MSA & Phylogeny • Incidence estimation	• Tracking the spread & evolution of the virus • Transmission patterns & outbreak detection • Identifying new variants & mutations • Monitoring the effectiveness of vaccines • Development of targeted interventions

Source: Pg. 5, Martin Hölzer (Establishing SARS-CoV-2 Genomic Surveillance at the National Level)

Data

& Notations

Sample **A**

Using 3 methods

6 Paired-end sequencing (**PE**) files

e.g. date_internalid_sample_*R{1,2}*.fastq.gz

reads

200408_20-04246_A_S1_L000_R1_001.fastq.gz

readme: file description (2 files per row)

Date	Sample	MF2ID	Method	Sequencer
200408	A	20-04246	CleanPlex SARS-CoV-2	IQ
200422	A	20-04444	Nextera Flex	IQ
200423	A	20-04411	Nextera_XT	NX

Files : BED ✓ fastQ ✓

IQ = Illumina Iseq

NX = Illumina NextSeq

Goal

Quality control

Mapping

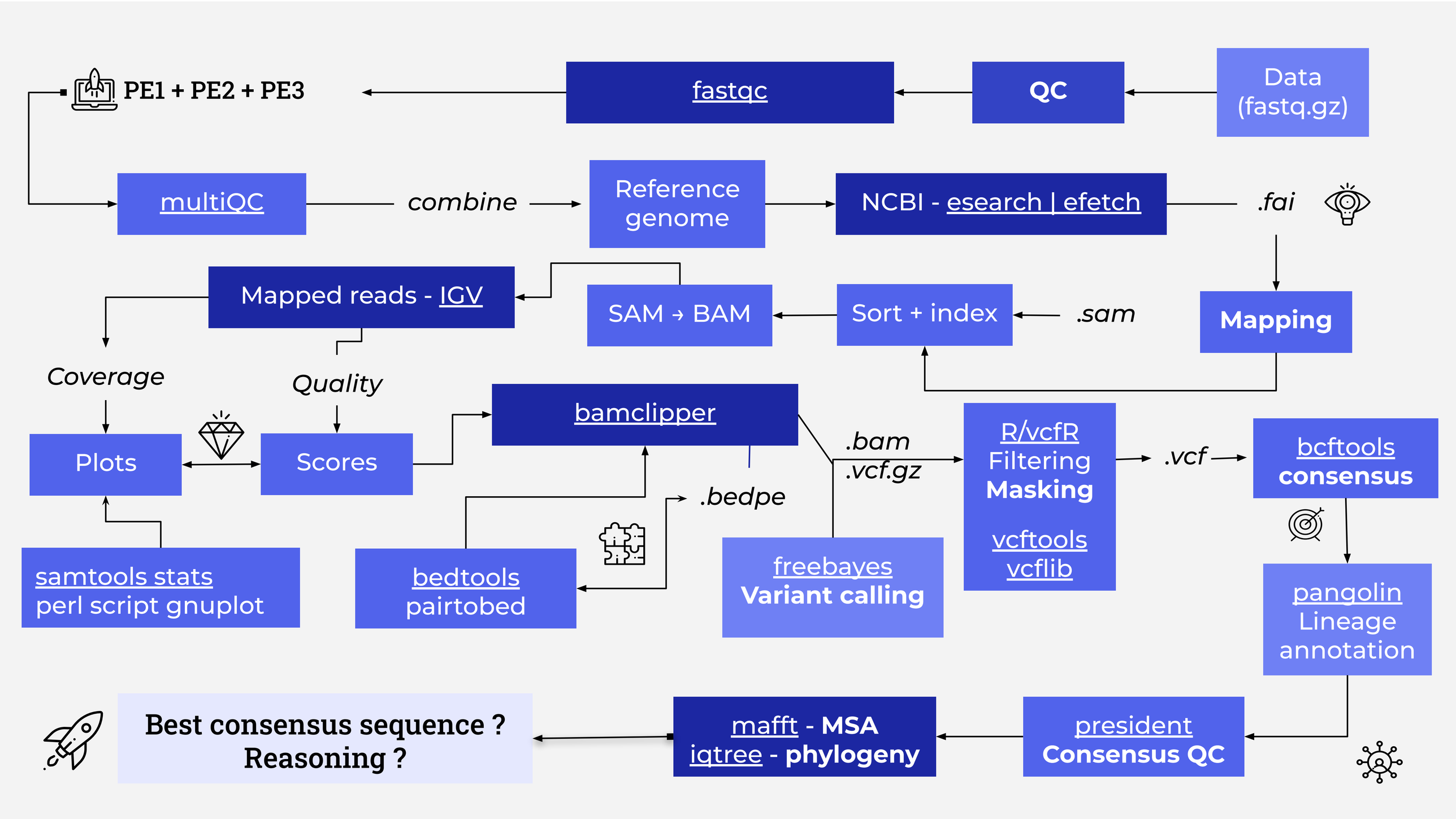
Consensus sequence

Lineage annotation

Consensus QC

Methods

(pipelines go clank !)


PE1 + PE2 + PE3
fastqc
QC
Data (fastq.gz)
multiQC
combine
Reference genome
NCBI - esearch | efetch
.fai
Mapped reads - IGV
SAM → BAM
Sort + index
.sam
Mapping
Coverage
Quality
bamclipper
Plots
Scores
.bam
.vcf.gz
R/vcfR Filtering Masking
vcftools vcflib
.vcf
bcftools consensus
.bedpe
samtools stats perl script gnuplot
bedtools pairtobed
freebayes Variant calling
pangolin Lineage annotation
Best consensus sequence ? Reasoning ?
mafft - MSA
iqtree - phylogeny
president Consensus QC

Tools & Steps

(everything is optimisation !)

Preliminary

Step 1

- Installing environment manager: mamba=1.4.2 and conda=23.3.1
- Adding channels: conda config - -add
- Creating environment with tools : multiqc, fastqc, fastp, minimap2, samtools, bcftools, igv, pangolin, president, bamclipper, freebayes, vcftools, vcflib, mafft, bedtools, iqtree, jalview : mamba create -y -p
- Activate & download data, unzip into PE: wget, tar -xvf
- Download reference genome: esearch -db | efetch -format fasta > *.fasta

fastQC - quality control

Step 2

For each PE-group

Basic Statistics: length, %GC ..etc.

Per Sequence Quality Scores

Per Base + Sequence GC content

Sequence length distribution

Overrepresented Sequences

Per Base Sequence Quality: box plots

Per Base Sequence Content

Per Base N content

Duplication level

Adapter content

fastqc -t 8 *…R1*.fastq.gz *…R2*.fastq.gz

fastp - preprocessing

Step 3

fastp /

--detect_adapter_for_pe /

--overrepresentation_analysis /

--correction --qualified_quality_phred 20 / → overlap correction - base quality, >Q20

--cut_right --thread 8 / → 5'->3' : meanQ <20 (drop bases-stop)

--html pair{1,2,3}.fastp.html /

--json pair1.fastp.json -i *…R1*.fastq.gz *…R2*.fastq.gz /

-o pair{1,2,3}.R1.clean.fastq.gz /

-O pair{1,2,3}.R2.clean.fastq.gz /

Run fastqc again ⟷ compare reports (*multiqc*) + improvements

minimap2 - Mapping & igv - Visualisation

Step 4

- aligning Illumina PE-reads with reference
- sequence alignment program that aligns DNA
- create *.sam* files

minimap2 -x sr -t 8 -a

alignment score ≥ 80 + 20,000 ≥ **bandwidth** ≥ 500

-o pair{1,2,3}/minimap2-illumina.sam /

reference.fasta /

pair{1,2,3}.R1.clean.fastq.gz /

pair{1,2,3}.R2.clean.fastq.gz

bamclipper - Primer clipping

Step 5

bamclipper - Primer clipping

Step 5

1st Try ✕

bamclipper - Primer clipping

Step 5

1st Try ×

Mapping from original files - converting .*sam* to .*bam* ≠ PE-reads .*bed* file

bamclipper - Primer clipping

Step 5

1st Try ×

Mapping from original files - converting .*sam* to .*bam* ≠ PE-reads .*bed* file

2nd Try ×

bamclipper - Primer clipping

Step 5

1st Try ×

Mapping from original files - converting *.sam* to *.bam* ≠ PE-reads *.bed* file

2nd Try ×

Converting *.bed* file to *.bam* (bedtools bedtobam -g reference.fasta -i *.bed > *.bam)

bamclipper - Primer clipping

Step 5

1st Try ×

Mapping from original files - converting *.sam* to *.bam* ≠ PE-reads *.bed* file

2nd Try ×

Converting *.bed* file to *.bam* (bedtools bedtobam -g reference.fasta -i *.bed > *.bam)

Converting *.bam* to *.bedpe* (bedtools bamtobed -bedpe -mate1 -i *bam > *.bedpe)

bamclipper - Primer clipping

Step 5

1st Try ×

Mapping from original files - converting *.sam* to *.bam* ≠ PE-reads *.bed* file

2nd Try ×

Converting *.bed* file to *.bam* (bedtools bedtobam -g reference.fasta -i *.bed > *.bam)

Converting *.bam* to *.bedpe* (bedtools bamtobed -bedpe -mate1 -i *bam > *.bedpe)

≠ PE-reads *.bam* file from *minimap2*

bamclipper - Primer clipping

Step 5

1st Try ×

Mapping from original files - converting *.sam* to *.bam* ≠ PE-reads *.bed* file

2nd Try ×

Converting *.bed* file to *.bam* (bedtools bedtobam -g reference.fasta -i *.bed > *.bam)

Converting *.bam* to *.bedpe* (bedtools bamtobed -bedpe -mate1 -i *bam > *.bedpe)

≠ PE-reads *.bam* file from *minimap2*

bamclipper - Primer clipping

Step 5

bamclipper - Primer clipping

Step 5

3rd Try ✓

Using *cleanplex.amplicons.bedpe* (Day 2 - hands-on), and checking **difference of locations**

bedtools pairtobed -a cleanplex.amplicons.bedpe -b *.bed -type neither

FASTA header and ID match

Perform primer clipping

bamclipper.sh -b pair{1,2,3}/minimap2-illumina.sorted.bam -p SARSCoV2.amplicons.bedpe -n 8

freebayes - Variant calling

Step 6

- bayesian genetic variant detector

    - SNPs, indels, MNPs, complex events

Input primerclipped.bam, reference.fasta

Output *.vcf* file

Parameters

min-alternate-count (**N** reads in a sample support an allele)

min-alternate-fraction (**N** fraction of observations in supporting an allele)

min-coverage, *pooled-continuous* (number of samples in the pool)

haplotype-length (short reads issue - avoid base quality recalibration)

freebayes - Variant calling

Step 6

Bottlenecks

Primer clipped sorted PE3-reads didn't had any variants so a quorum effort was done that gave a lot of unknown detections ~ 245 after using *report-monomorphic* and *pooled-continuous* parameters only. **Interpret ⟷ Iterate**

Naive variant calling

i.e. simply annotate observation counts of SNPs and indels

freebayes - - help : more insightful troubleshooting

freebayes - Variant calling

Step 6

Bottlenecks

Primer clipped sorted PE3-reads didn't had any variants so a quorum effort was done that gave a lot of unknown detections ~ 245 after using *report-monomorphic* and *pooled-continuous* parameters only. **Interpret ⟷ Iterate**

Naive variant calling

i.e. simply annotate observation counts of SNPs and indels

freebayes - - help : more insightful troubleshooting

freebayes - Variant calling

Step 6

Bottlenecks

Primer clipped sorted PE3-reads didn't had any variants so a quorum effort was done that gave a lot of unknown detections ~ 245 after using *report-monomorphic* and *pooled-continuous* parameters only. **Interpret ⟷ Iterate**

Naive variant calling

i.e. simply annotate observation counts of SNPs and indels

loci which appear to be monomorphic, and alleles, even those not present in called genotypes.

freebayes - - help : more insightful troubleshooting

R/vcfR - masking & filtering

Step 7

- understand and explore data with visualisation, quality control, and writing out the masked *.vcf* files

*../pair{1,2,3}/freebayes-illumina.vcf  *../pair{1,2,3}/masked-strict.vcf

- annotation for each PE data for analysis and composite plots (variants, reads, bases)
- masking low-coverage region based on *quality* and *depth* (low confidence)

bcftools - Consensus generation

Step 8

~~Assumption~~: we have already made the calls, normalised indels and filtered ✓

1. bcftools view *.vcf -Oz -o *.vcf.gz .....*vcf file zipping* (block compression)
2. bcftools index *.vcf.gz .....*creating index file*
3. bcftools consensus -f reference.fasta *.vcf.gz -o *-qc-strict.fasta
4. Replacing the reference genome header with

Consensus-Illumina-PE{1,2,3} | date, sample, internalID, method, sequencer

pangolin - Lineage Annotation

Step 9

- dynamic nomenclature of SARS-CoV-2 lineages a.k.a. pangloin nomenclature, assigns most likely lineage (Pango lineage) to query sequences

Source: https://cov-lineages.org/lineage_list.html

pangolin - Lineage Annotation

Step 9

- dynamic nomenclature of SARS-CoV-2 lineages a.k.a. pangloin nomenclature, assigns most likely lineage (Pango lineage) to query sequences

- updated routinely

Source: https://cov-lineages.org/lineage_list.html

pangolin - Lineage Annotation

Step 9

- dynamic nomenclature of SARS-CoV-2 lineages a.k.a. pangloin nomenclature, assigns most likely lineage (Pango lineage) to query sequences

- updated routinely

- always run update before using the tool

Source: https://cov-lineages.org/lineage_list.html

president - Consensus QC

Step 10

- calculate pairwise nucleotide identity and reports ambiguous 'N'

1. Put all three consensus sequences in *.fasta* file.
2. Run with default setting on the multiple query *FASTA* sequences.
3. Output the valid + invalid *.fasta*, and report with *PSL* alignment columns.

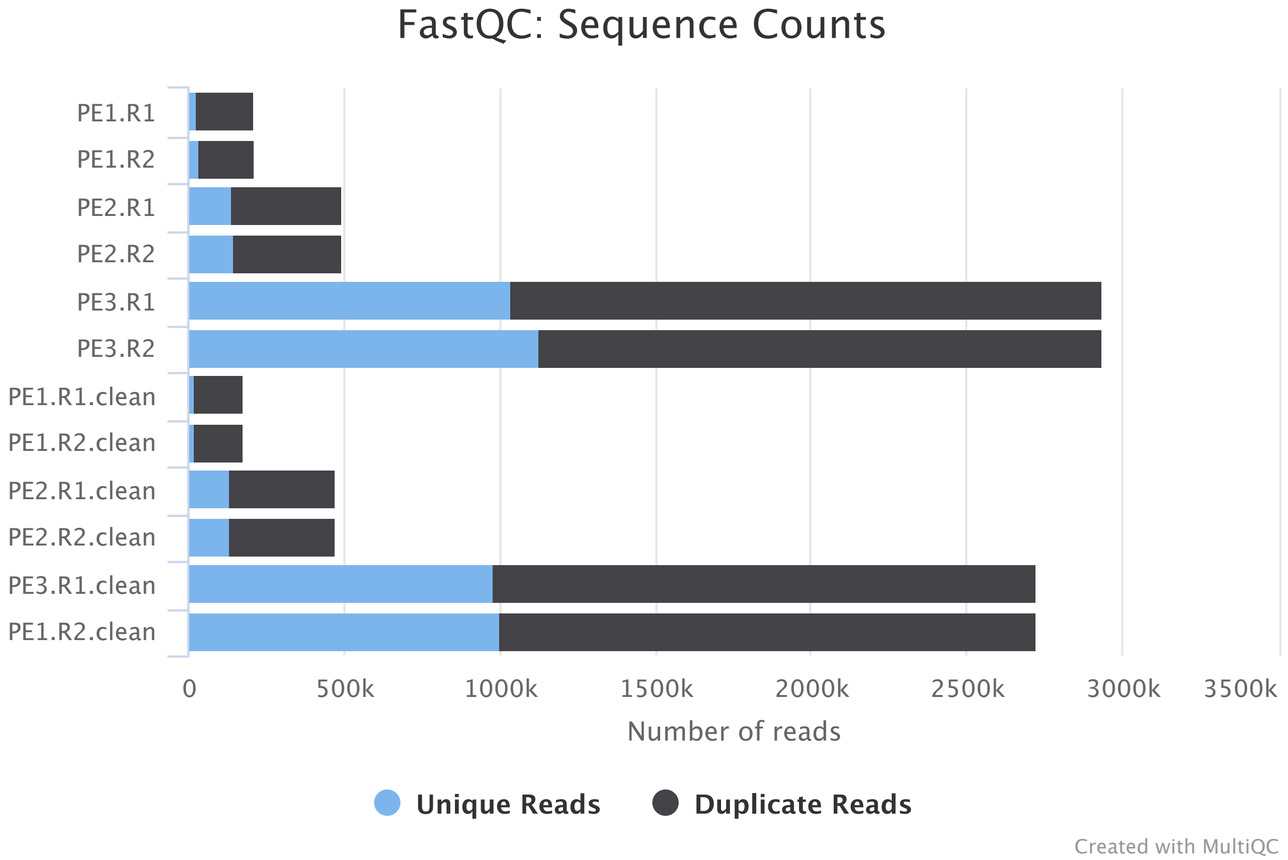
Aftermath

4. Create and visualise the progressive alignment using *mafft* and *jalview.*
5. Create and visualise phylogenetic tree using *iqtree* and IROKI.

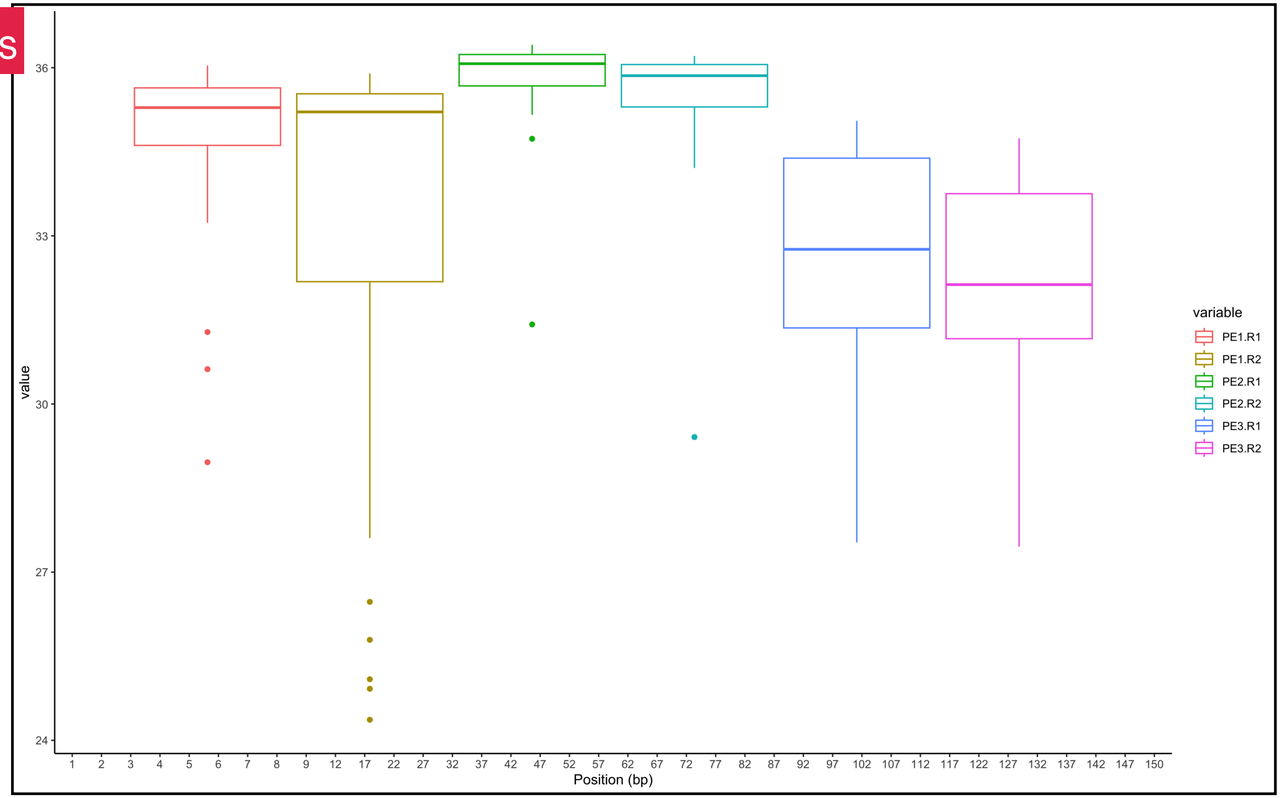
Results

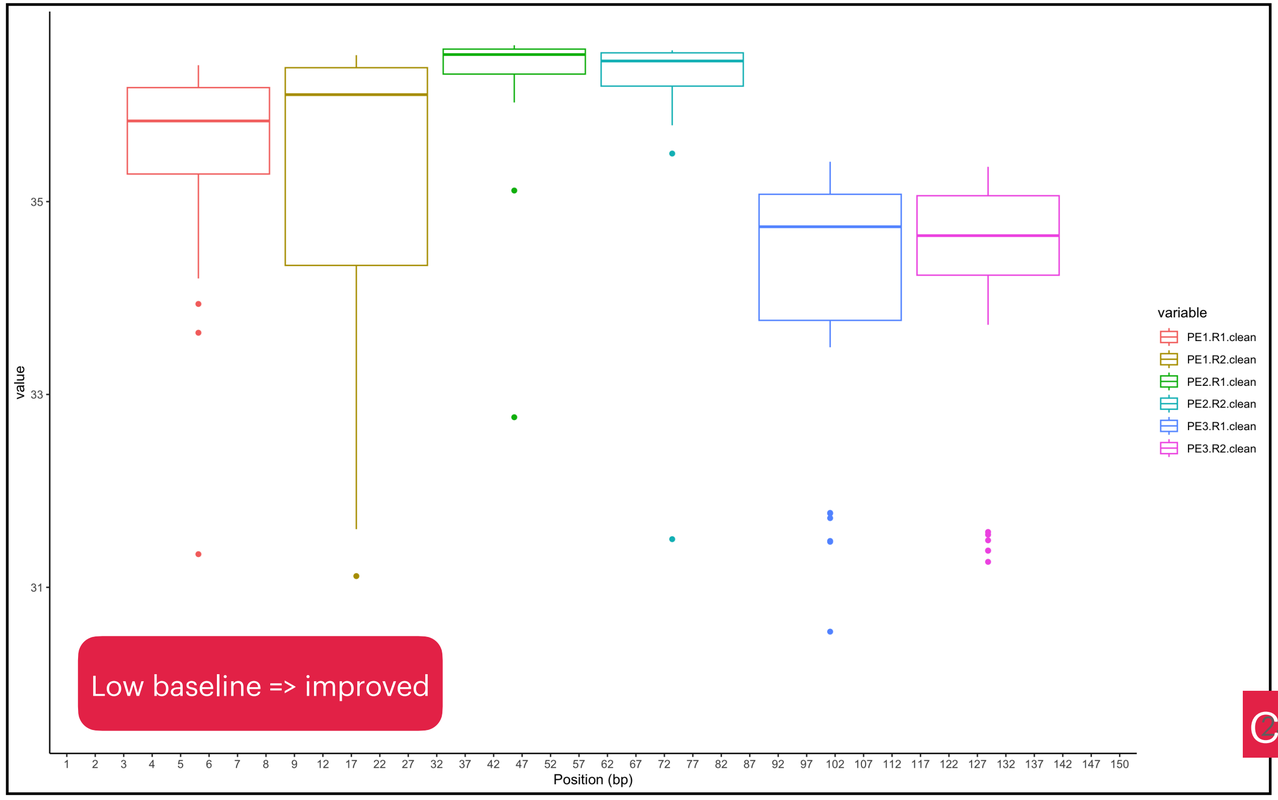
(to the point !)

Aligned Reads - raw (%)	Aligned Reads - clean (%)	Sample
14.6	11.9	PE1.R1
15.6	11.8	PE1.R2
28.2	28	PE2.R1
29.2	28.3	PE2.R2
35.4	36	PE3.R1
38.5	36.9	PE3.R2



Raw reads

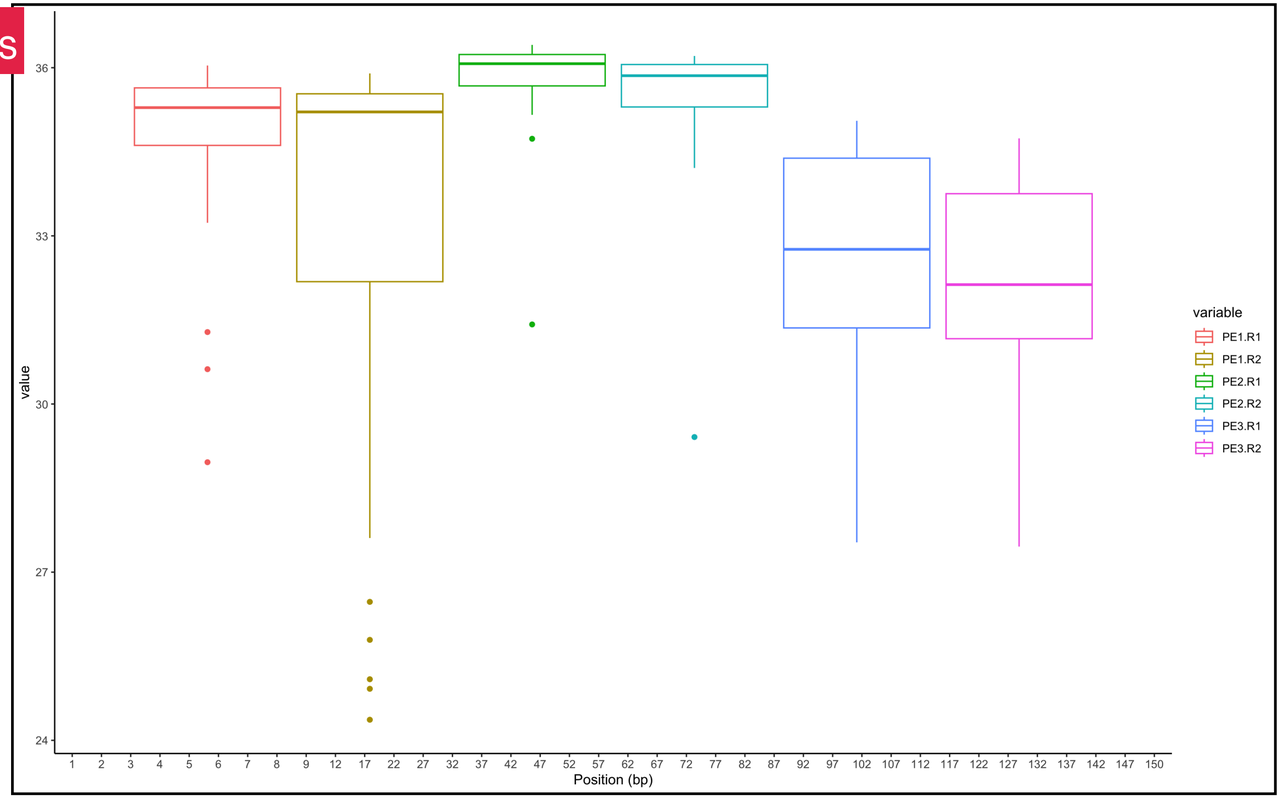


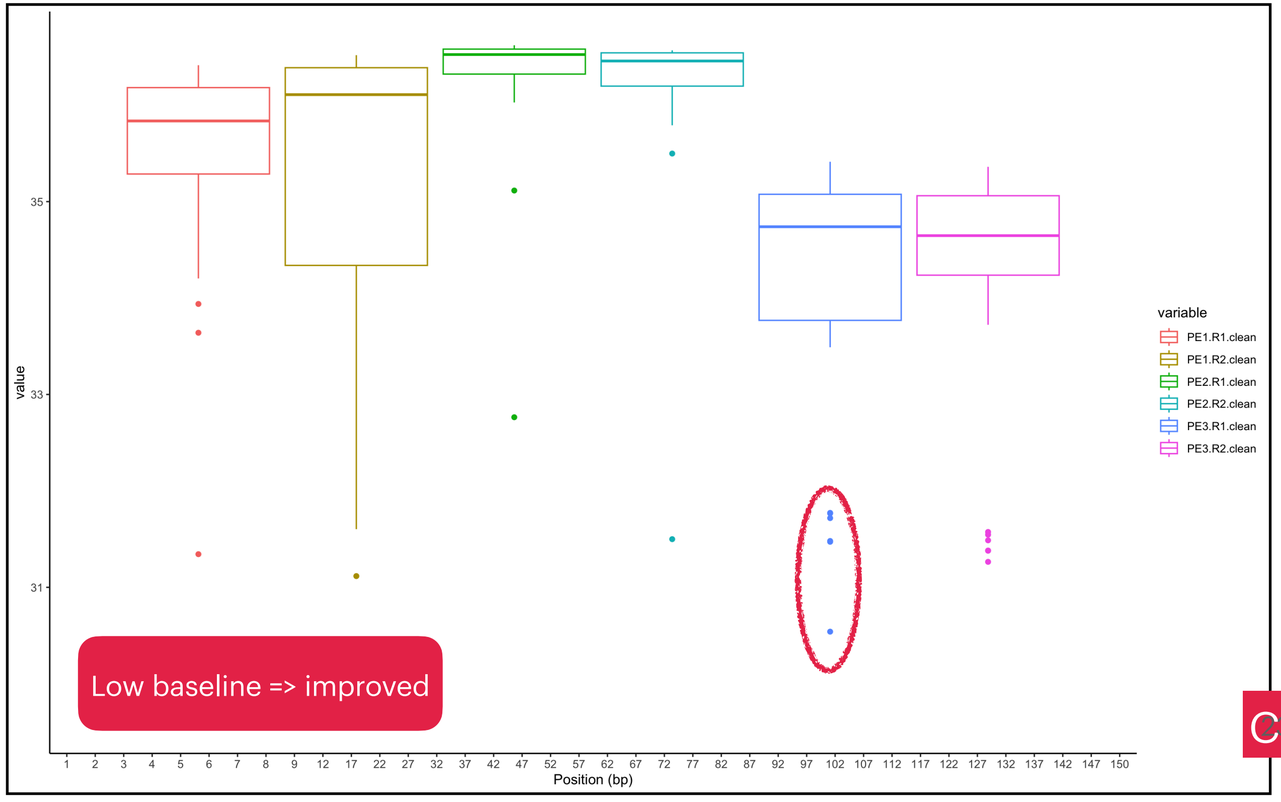


Low baseline => improved

Clean reads

Raw reads

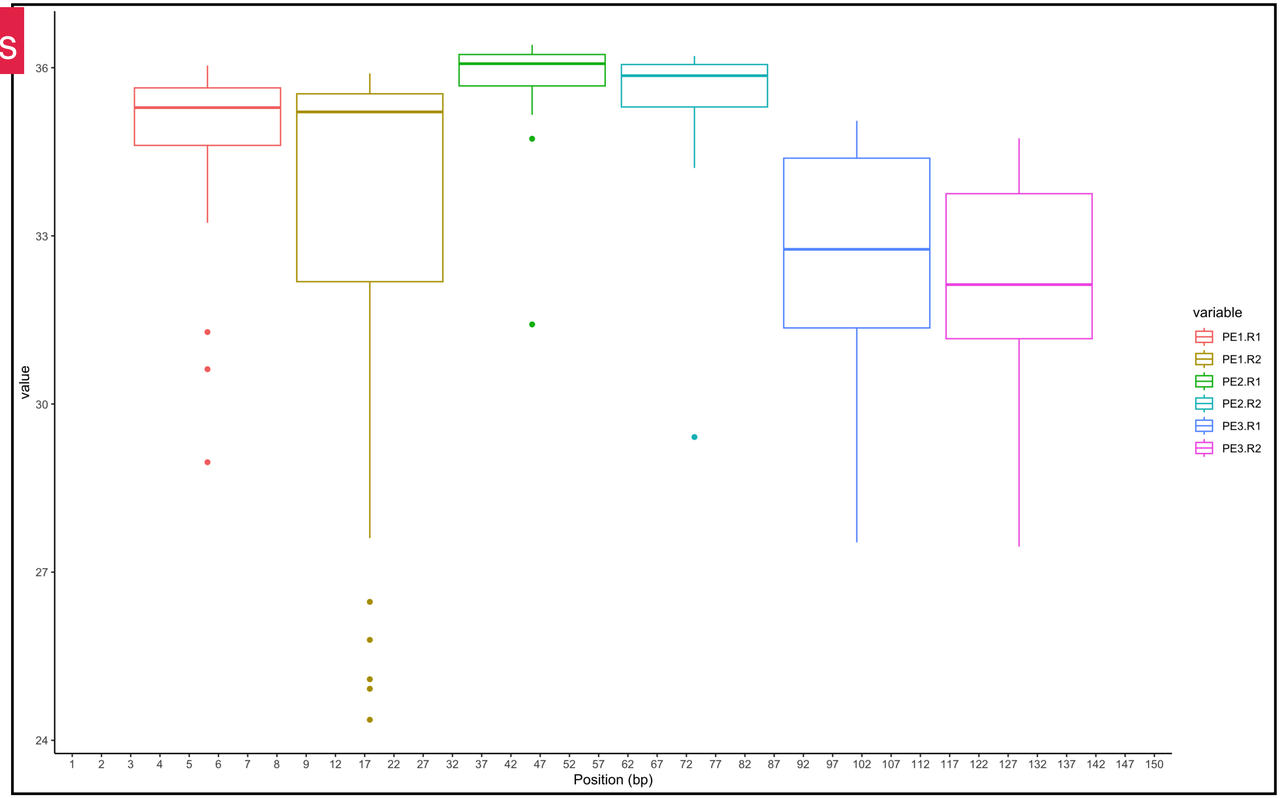


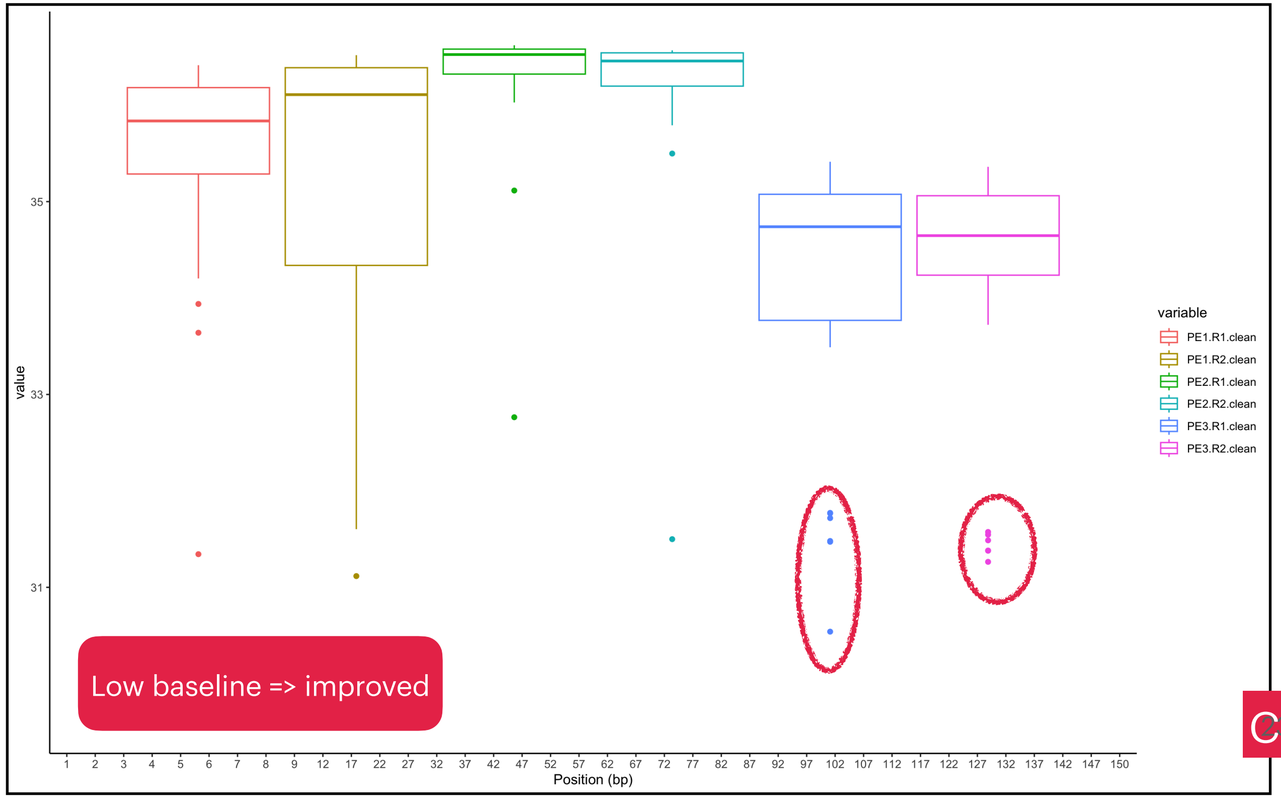


Low baseline => improved

Clean reads

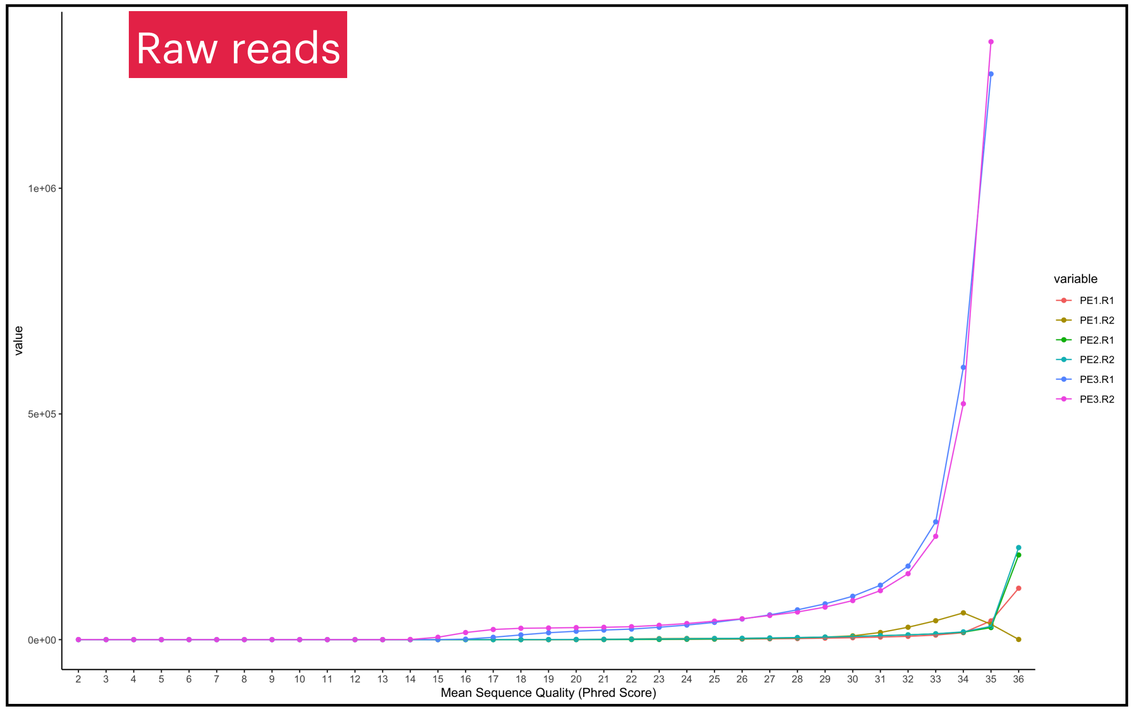
Raw reads

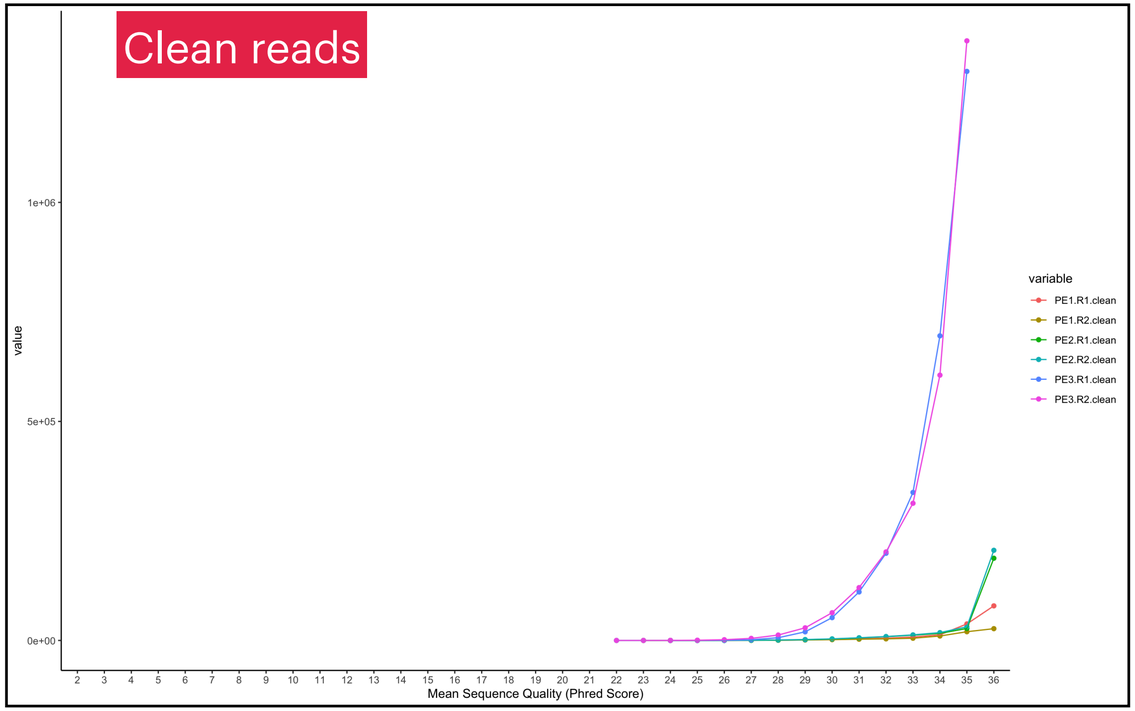


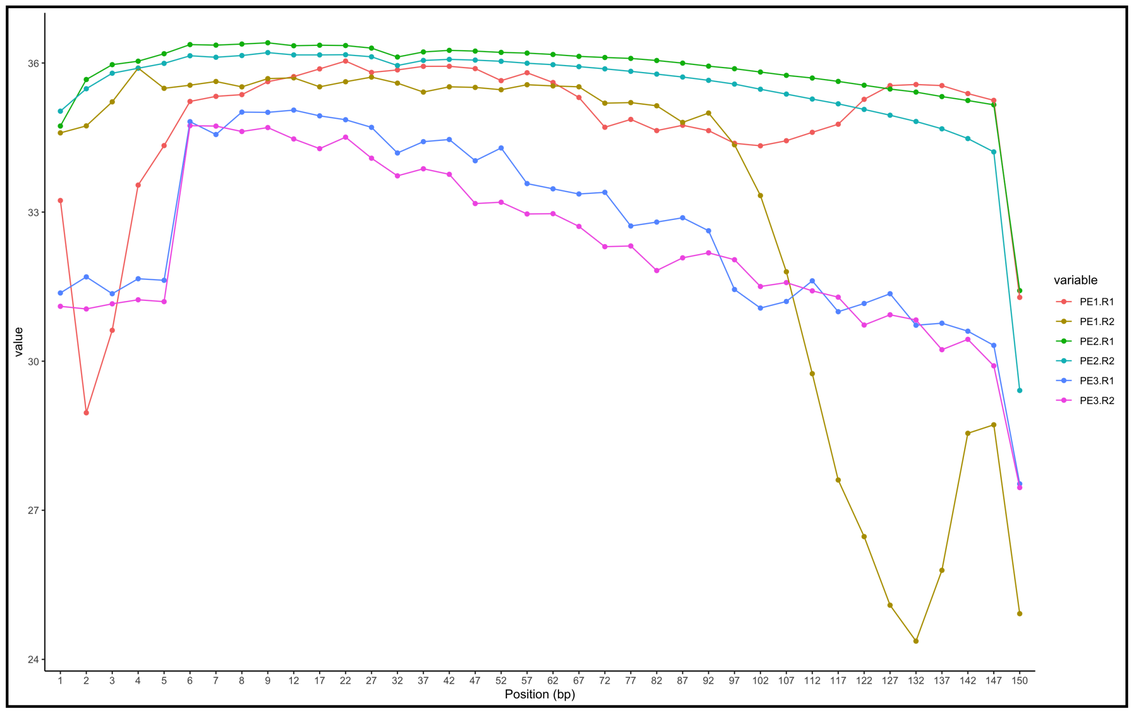


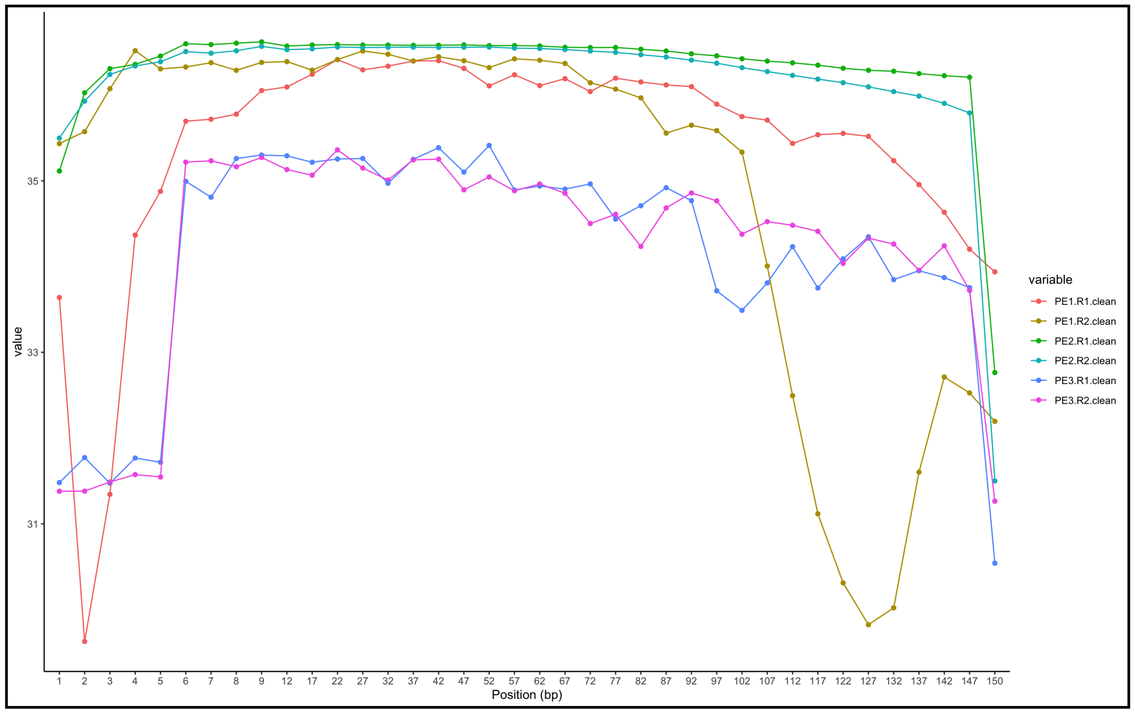
Low baseline => improved

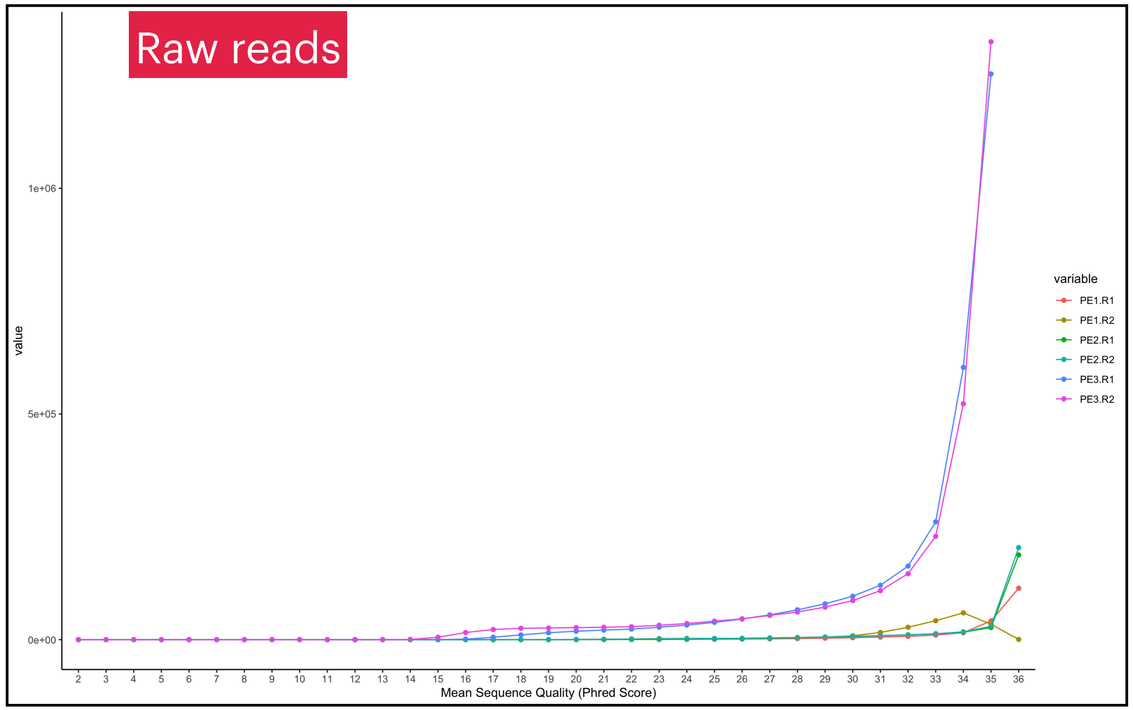
Clean reads

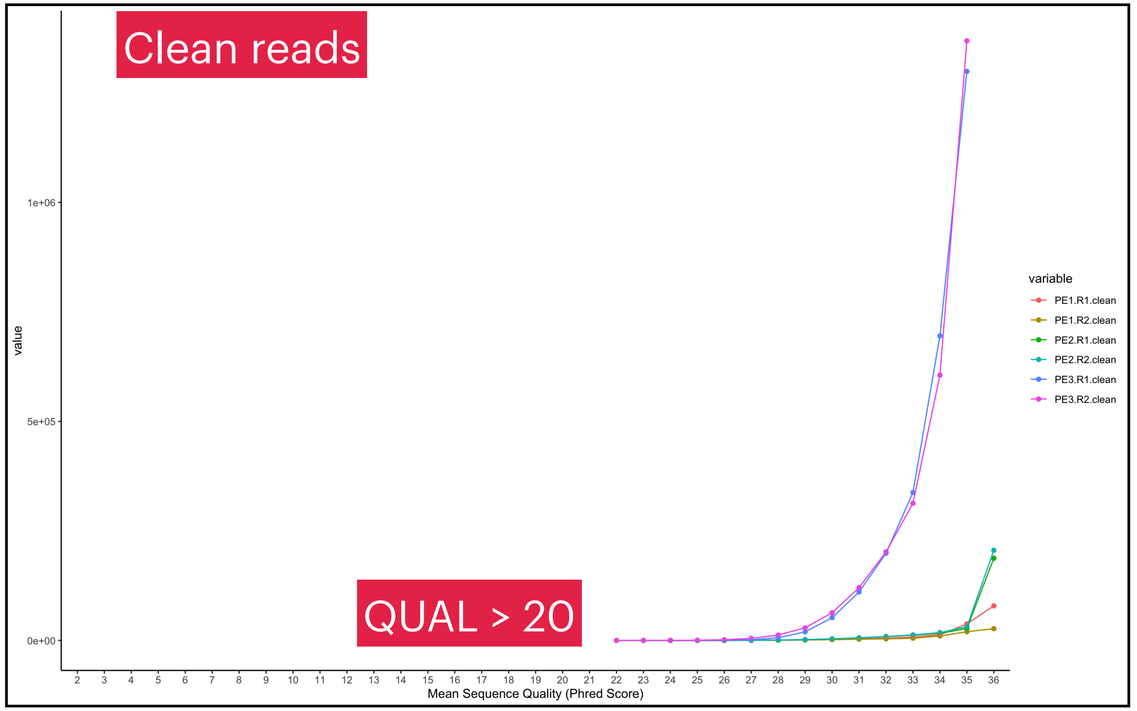

Raw reads
value
1e+06
5e+05
0e+00
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
Mean Sequence Quality (Phred Score)
variable
PE1.R1
PE1.R2
PE2.R1
PE2.R2
PE3.R1
PE3.R2

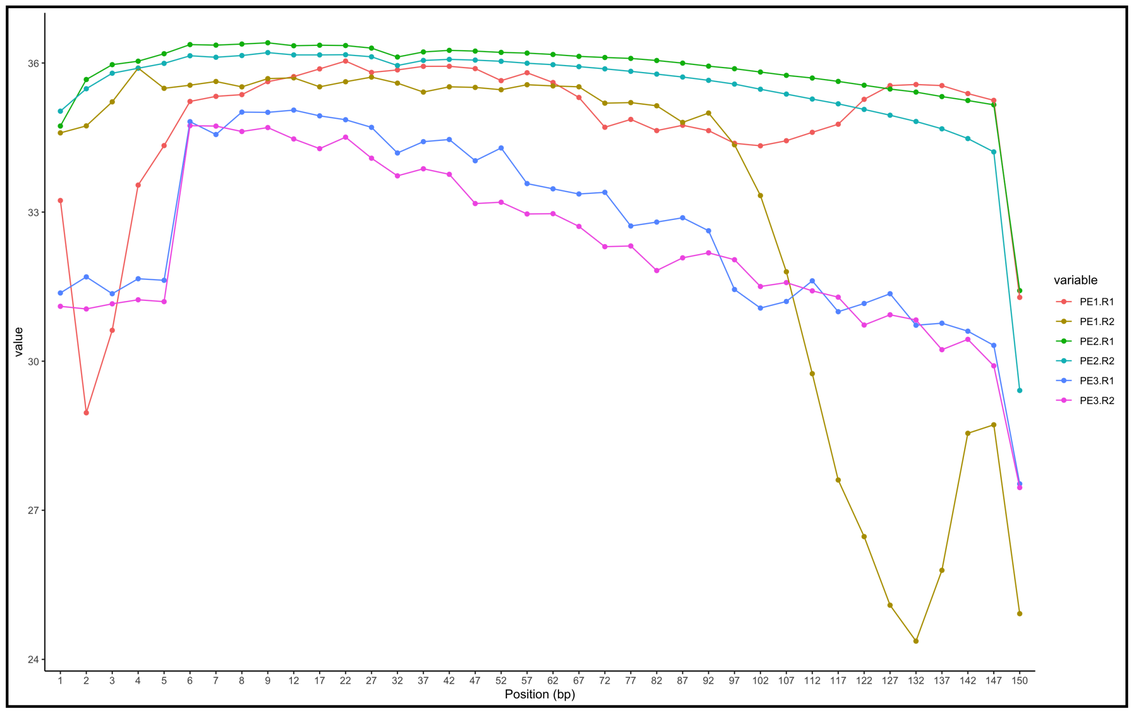

Clean reads
value
1e+06
5e+05
0e+00
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
Mean Sequence Quality (Phred Score)
variable
PE1.R1.clean
PE1.R2.clean
PE2.R1.clean
PE2.R2.clean
PE3.R1.clean
PE3.R2.clean

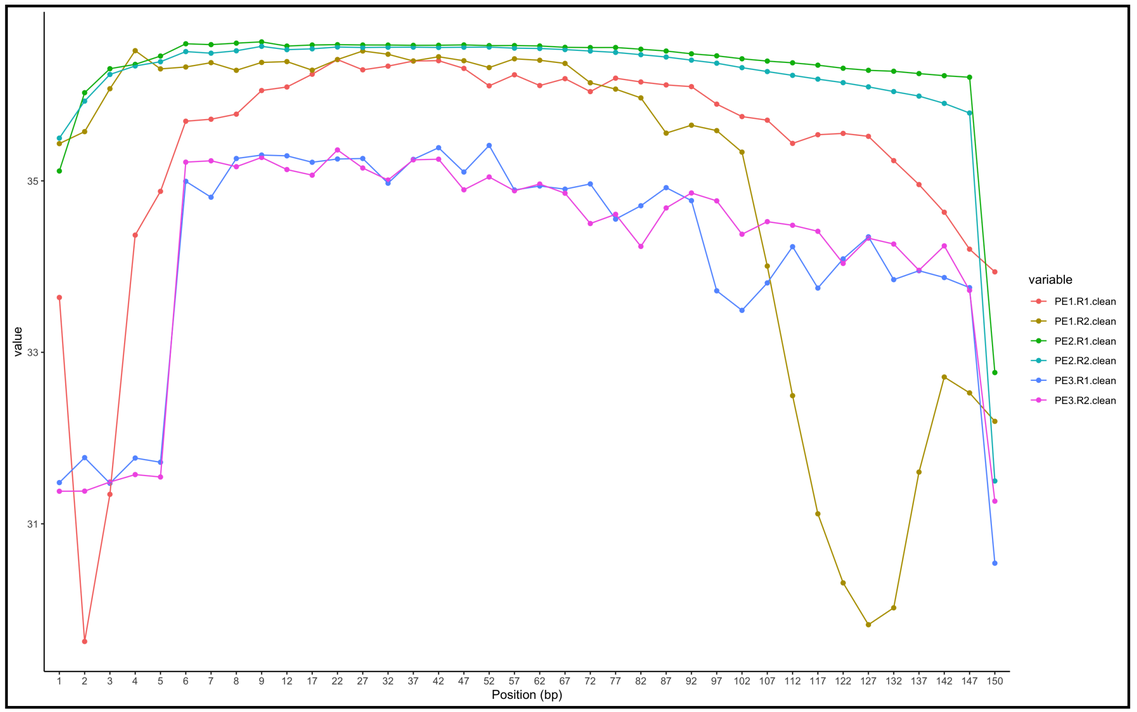

value
36
33
30
27
24
1 2 3 4 5 6 7 8 9 12 17 22 27 32 37 42 47 52 57 62 67 72 77 82 87 92 97 102 107 112 117 122 127 132 137 142 147 150
Position (bp)
variable
PE1.R1
PE1.R2
PE2.R1
PE2.R2
PE3.R1
PE3.R2

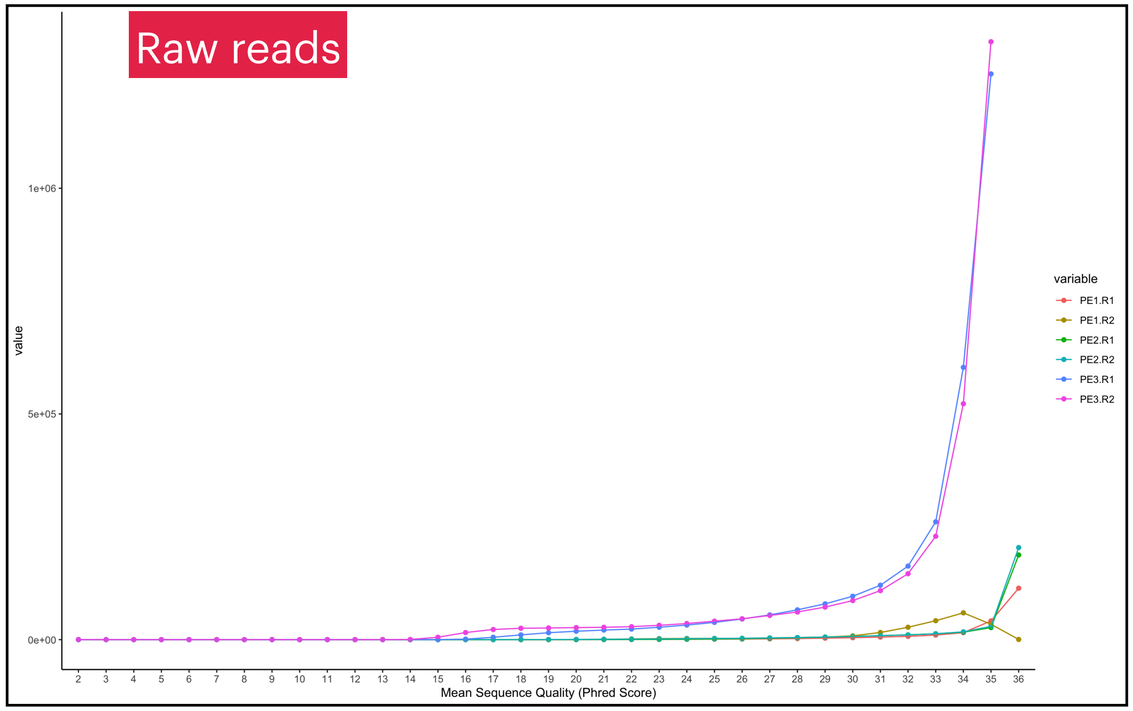

value
35
33
31
1 2 3 4 5 6 7 8 9 12 17 22 27 32 37 42 47 52 57 62 67 72 77 82 87 92 97 102 107 112 117 122 127 132 137 142 147 150
Position (bp)
variable
PE1.R1.clean
PE1.R2.clean
PE2.R1.clean
PE2.R2.clean
PE3.R1.clean
PE3.R2.clean

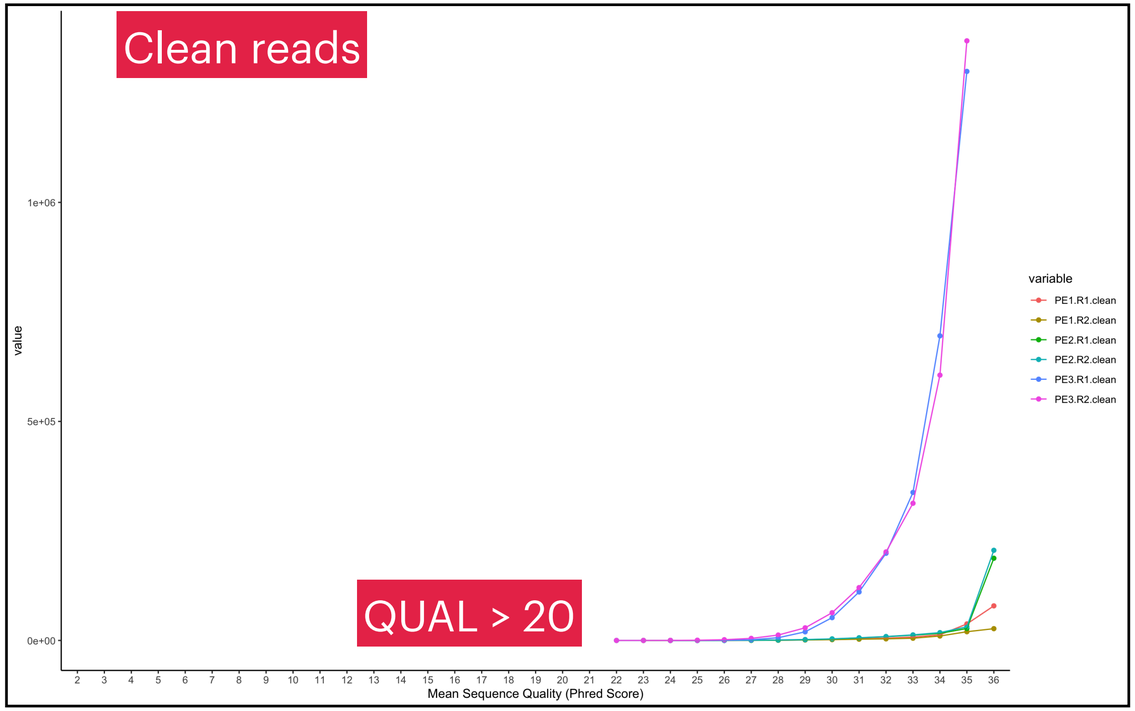

Raw reads
value
1e+06
5e+05
0e+00
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
Mean Sequence Quality (Phred Score)
variable
PE1.R1
PE1.R2
PE2.R1
PE2.R2
PE3.R1
PE3.R2

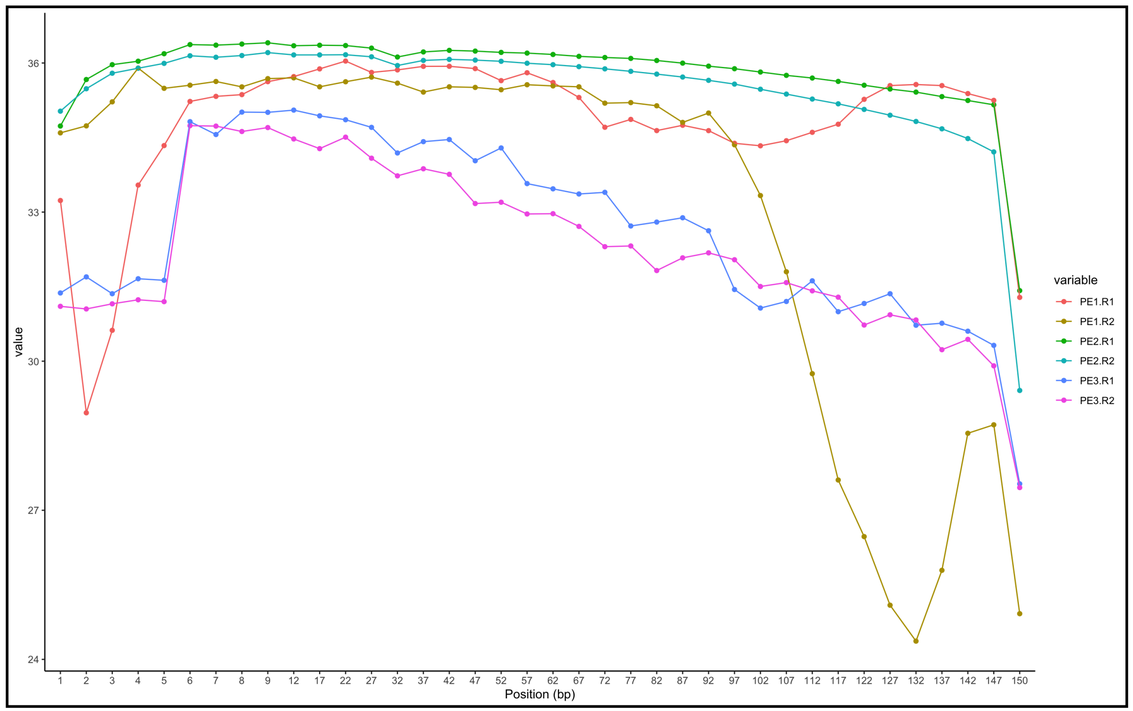

Clean reads
QUAL > 20
value
1e+06
5e+05
0e+00
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
Mean Sequence Quality (Phred Score)
variable
PE1.R1.clean
PE1.R2.clean
PE2.R1.clean
PE2.R2.clean
PE3.R1.clean
PE3.R2.clean

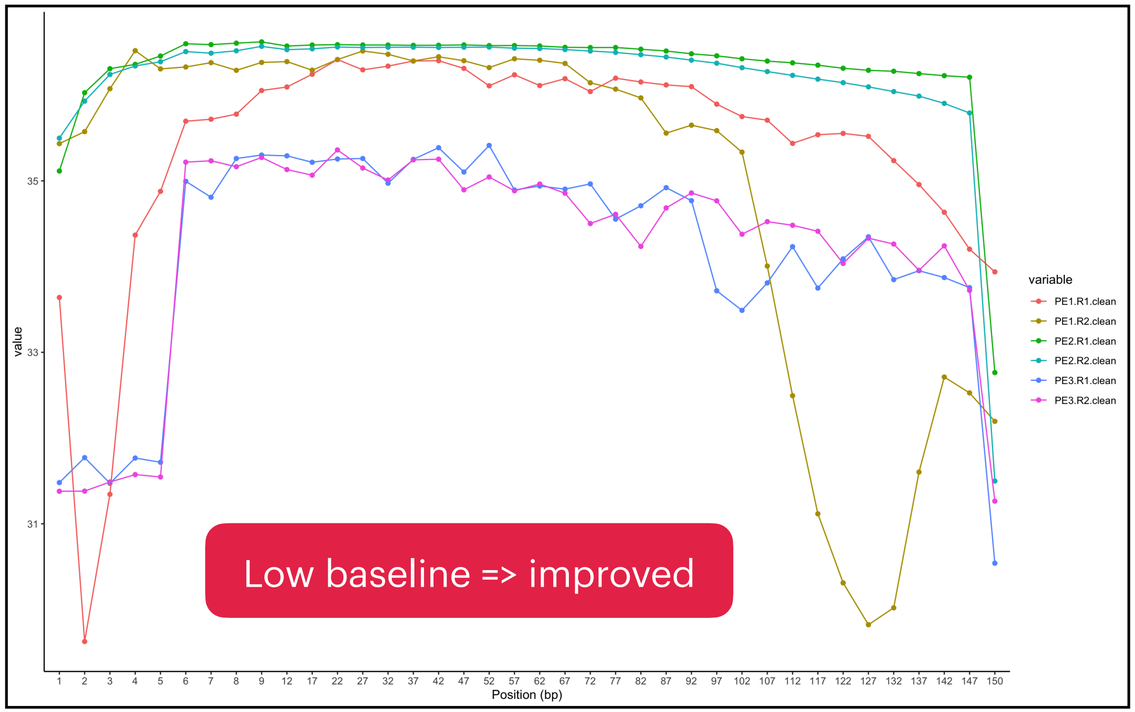

value
36
33
30
27
24
1 2 3 4 5 6 7 8 9 12 17 22 27 32 37 42 47 52 57 62 67 72 77 82 87 92 97 102 107 112 117 122 127 132 137 142 147 150
Position (bp)
variable
PE1.R1
PE1.R2
PE2.R1
PE2.R2
PE3.R1
PE3.R2


value
35
33
31
1 2 3 4 5 6 7 8 9 12 17 22 27 32 37 42 47 52 57 62 67 72 77 82 87 92 97 102 107 112 117 122 127 132 137 142 147 150
Position (bp)
variable
PE1.R1.clean
PE1.R2.clean
PE2.R1.clean
PE2.R2.clean
PE3.R1.clean
PE3.R2.clean

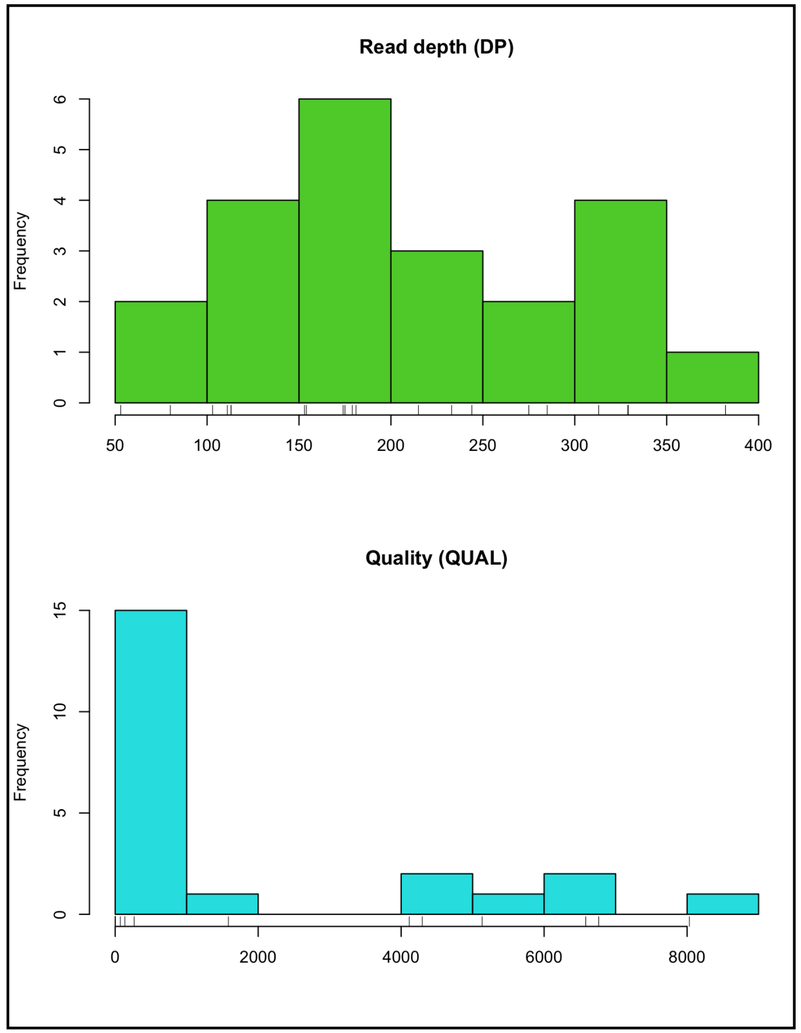

Raw reads
value
1e+06
5e+05
0e+00
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
Mean Sequence Quality (Phred Score)
variable
PE1.R1
PE1.R2
PE2.R1
PE2.R2
PE3.R1
PE3.R2

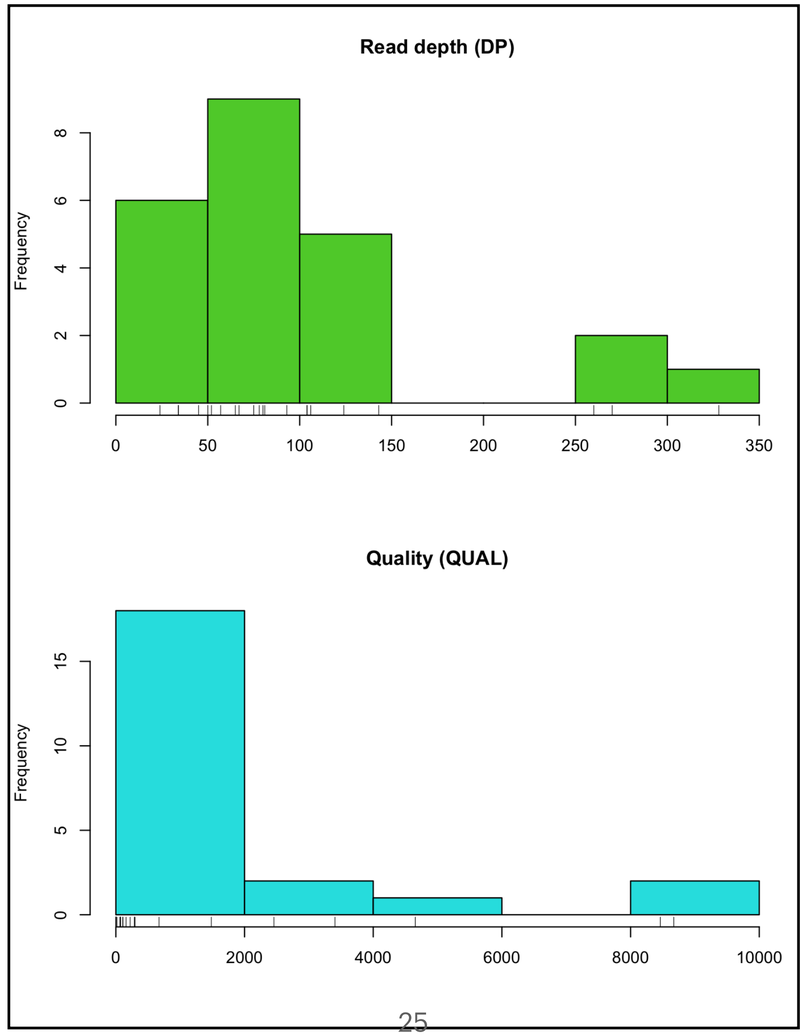

Clean reads
QUAL > 20
value
1e+06
5e+05
0e+00
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
Mean Sequence Quality (Phred Score)
variable
PE1.R1.clean
PE1.R2.clean
PE2.R1.clean
PE2.R2.clean
PE3.R1.clean
PE3.R2.clean

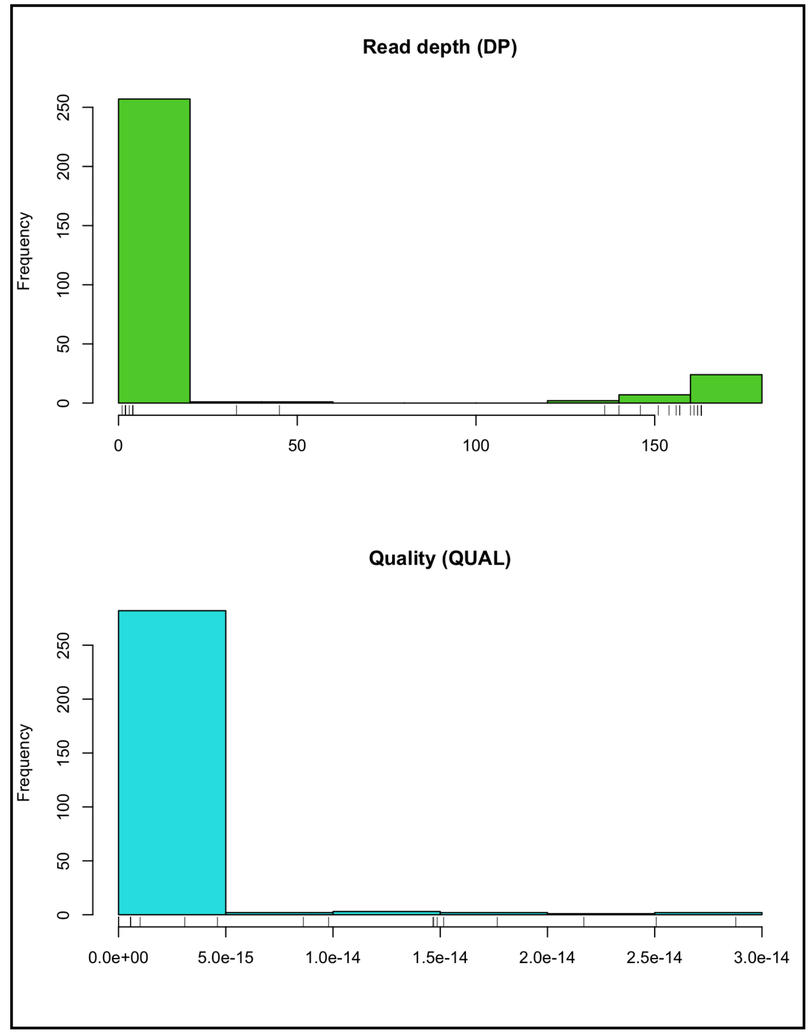

value
36
33
30
27
24
1 2 3 4 5 6 7 8 9 12 17 22 27 32 37 42 47 52 57 62 67 72 77 82 87 92 97 102 107 112 117 122 127 132 137 142 147 150
Position (bp)
variable
PE1.R1
PE1.R2
PE2.R1
PE2.R2
PE3.R1
PE3.R2


Low baseline => improved
value
35
33
31
1 2 3 4 5 6 7 8 9 12 17 22 27 32 37 42 47 52 57 62 67 72 77 82 87 92 97 102 107 112 117 122 127 132 137 142 147 150
Position (bp)
variable
PE1.R1.clean
PE1.R2.clean
PE2.R1.clean
PE2.R2.clean
PE3.R1.clean
PE3.R2.clean

.vcf PE1 PE2 PE3

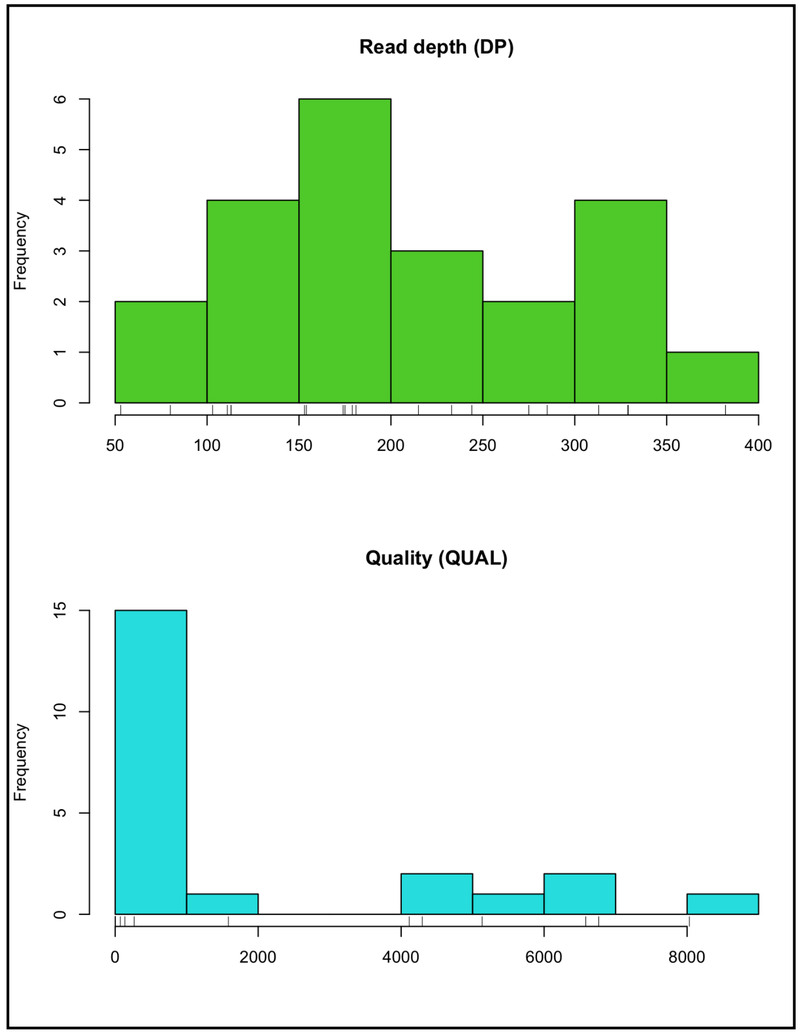




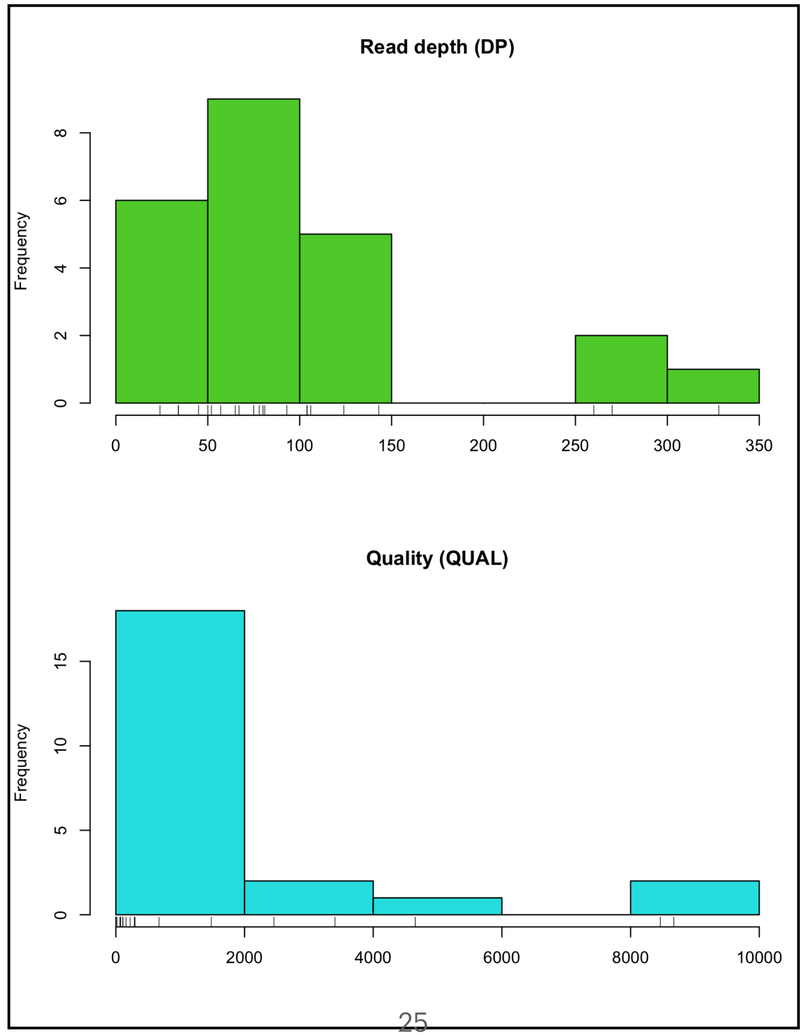


.vcf

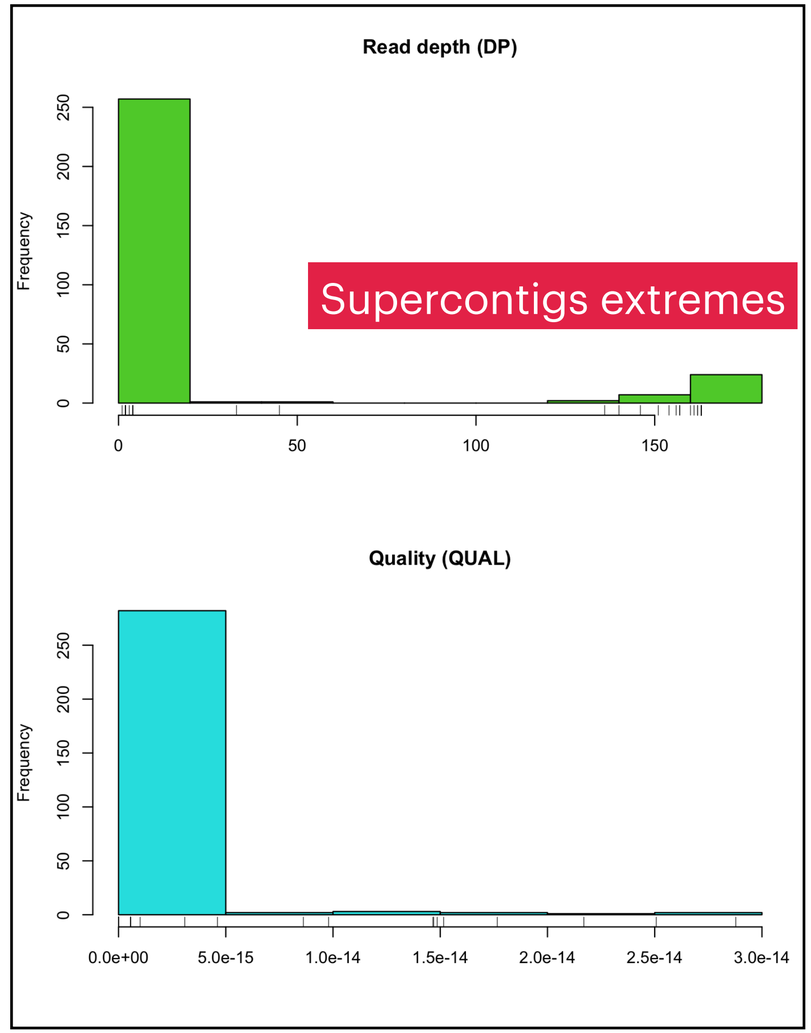
PE1



PE2

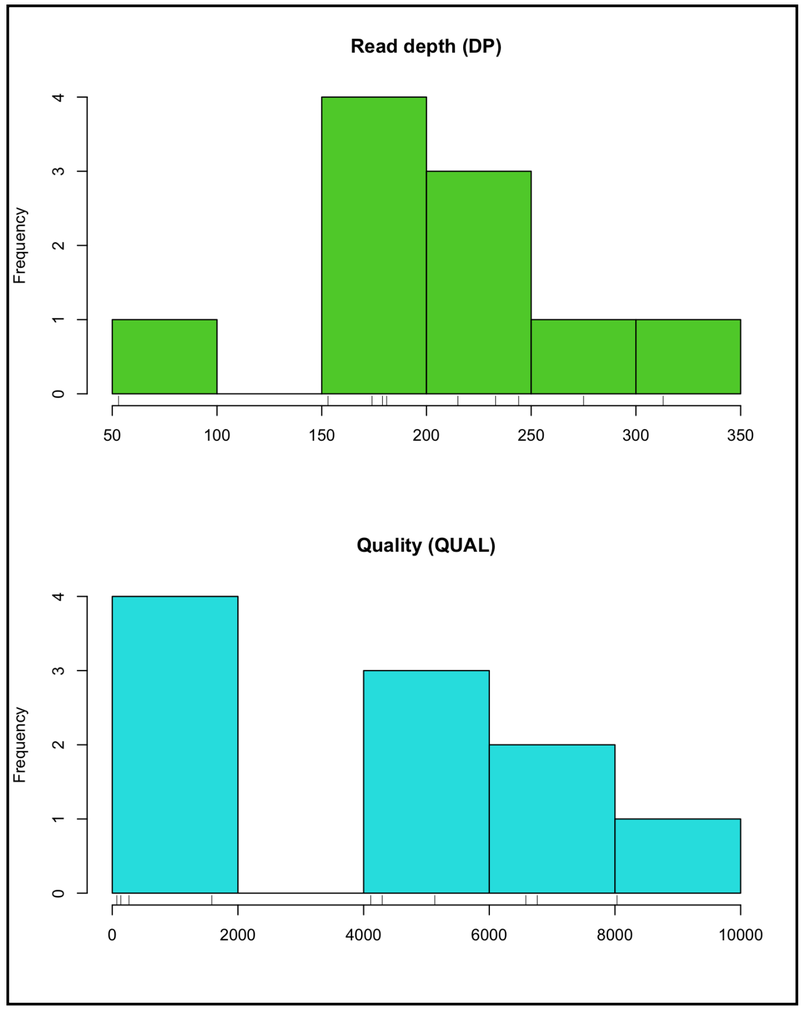


PE3

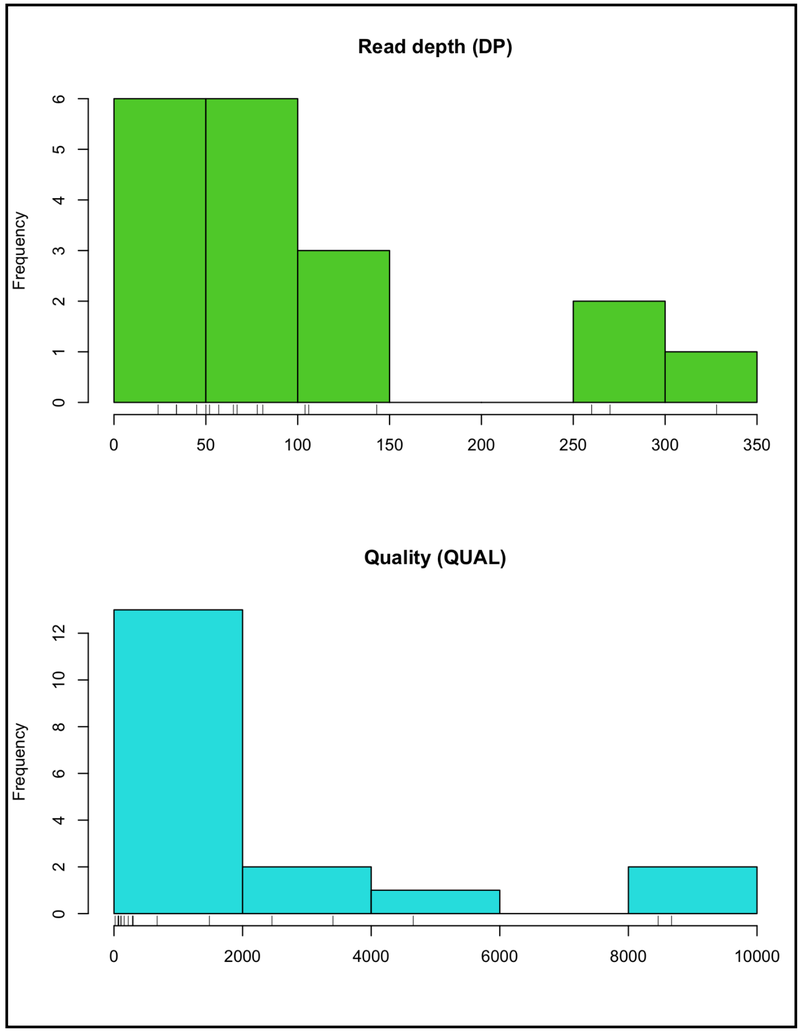


mask.vcf

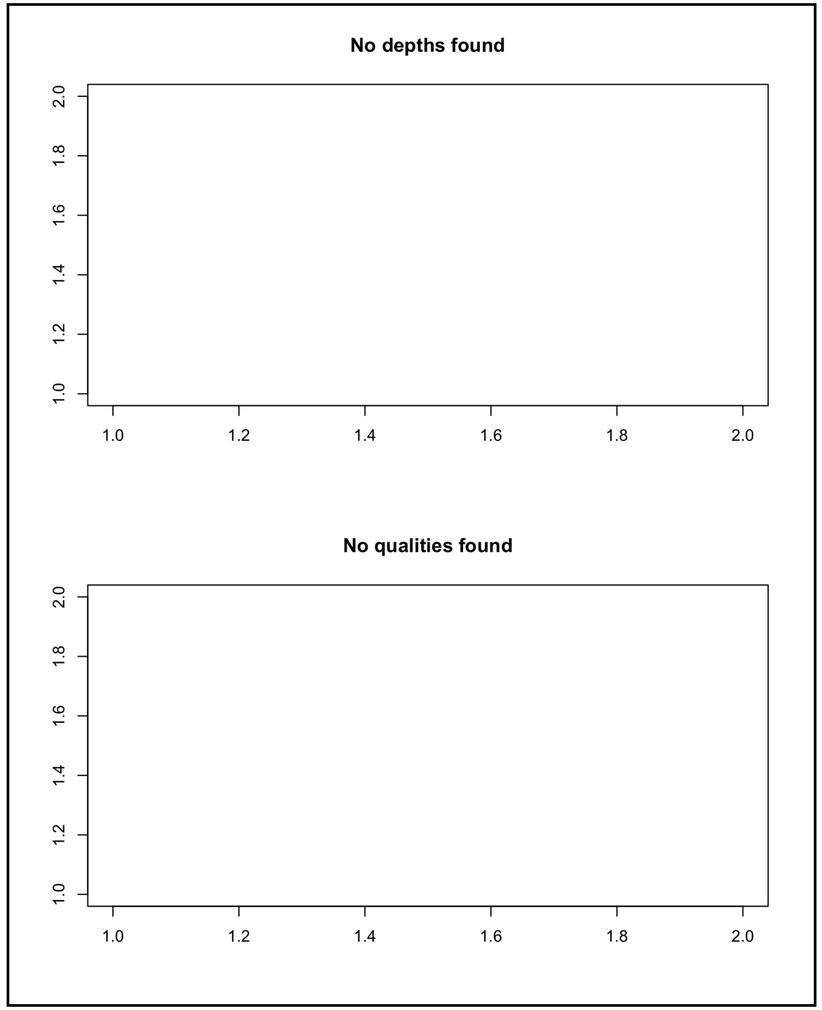
PE1



PE2

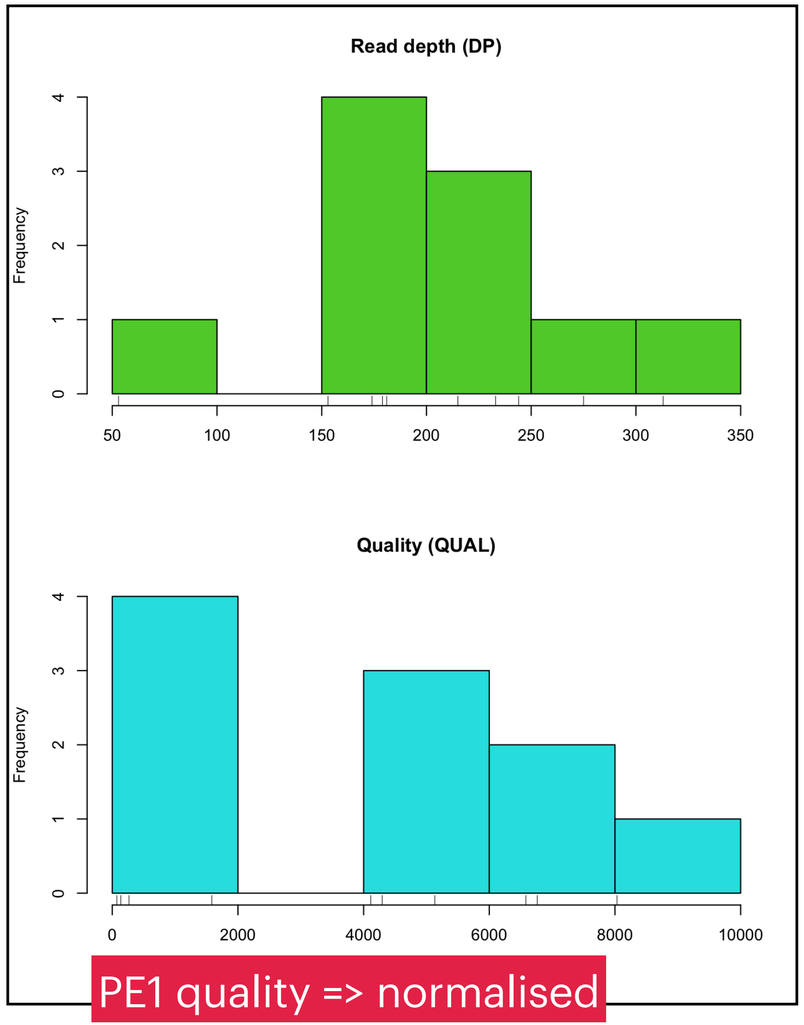


PE3



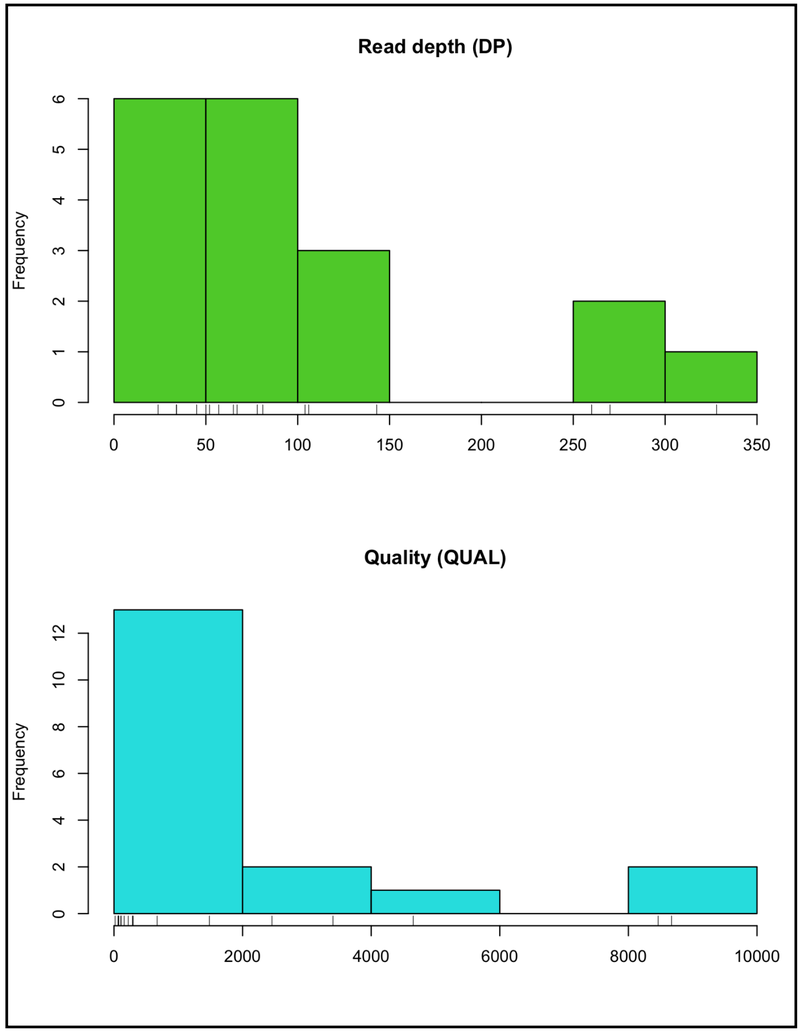
mask.vcf

PE1

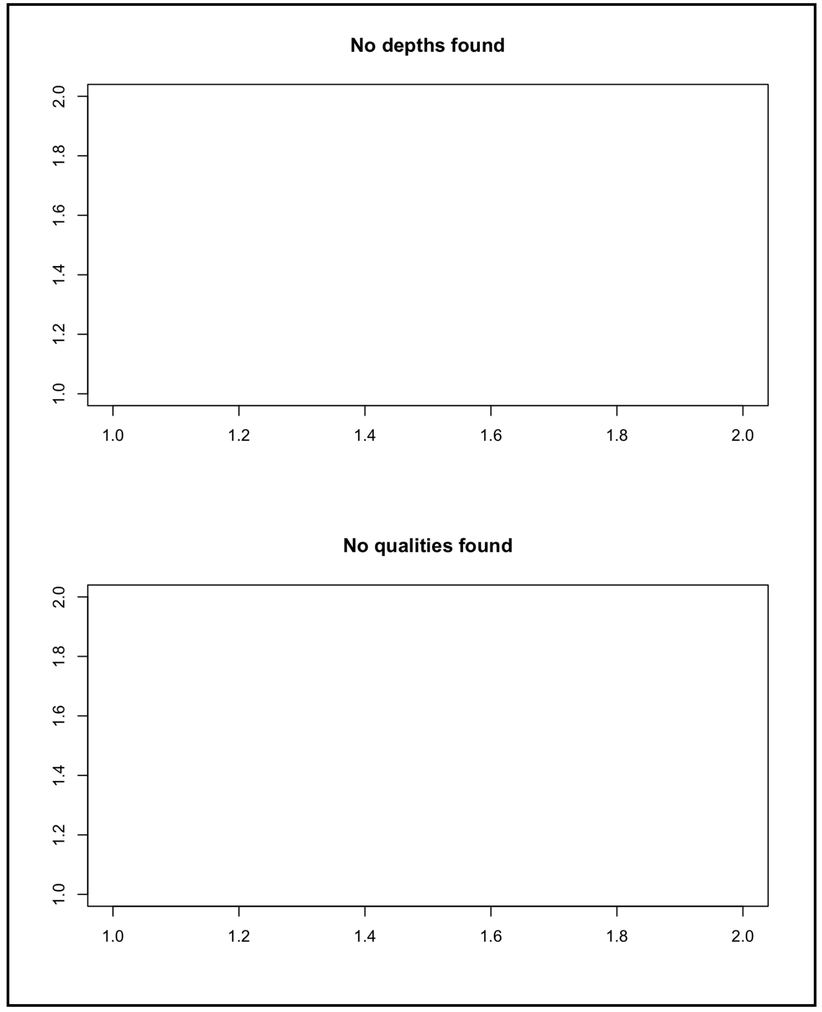


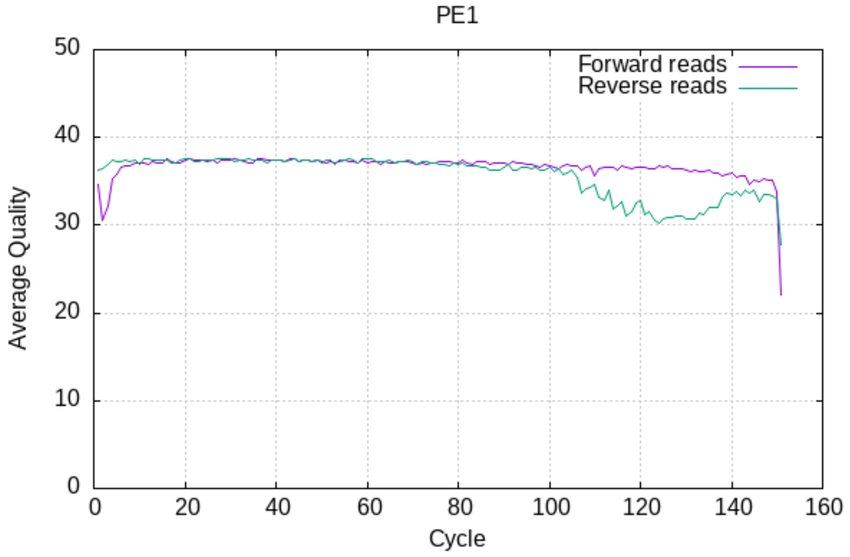
PE1 quality => normalised

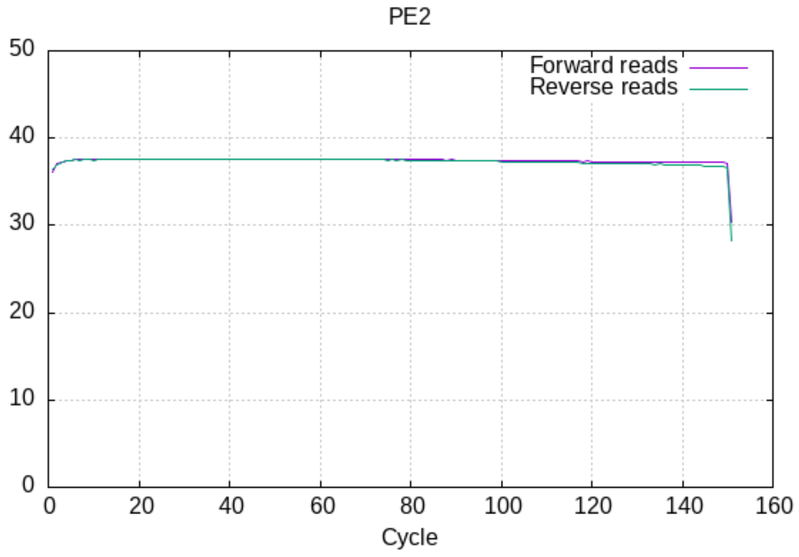
PE2

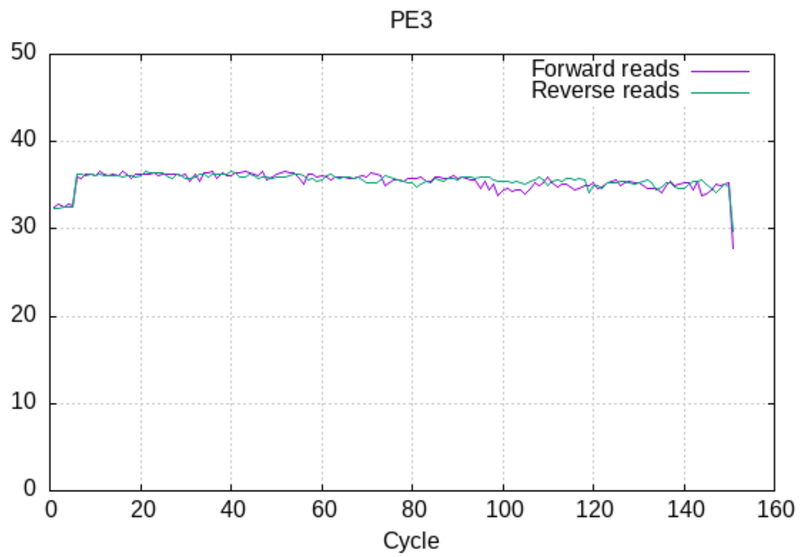


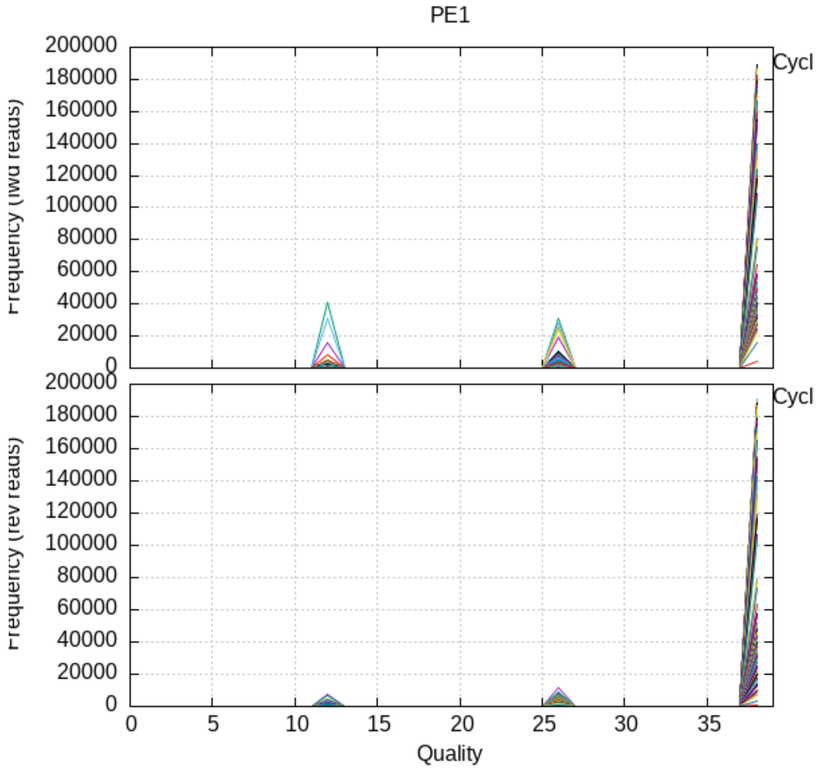
PE3

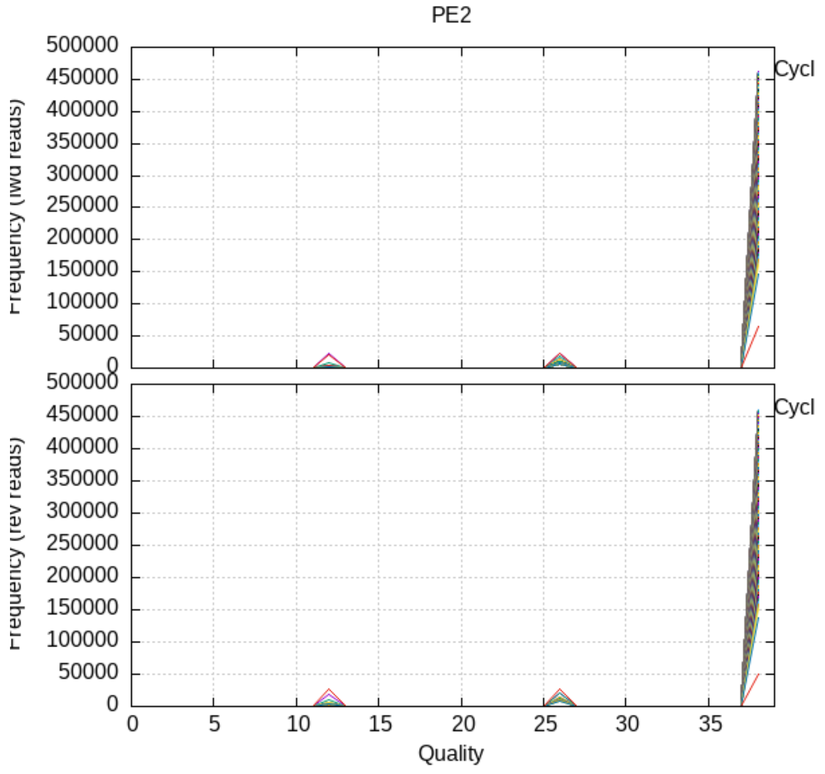


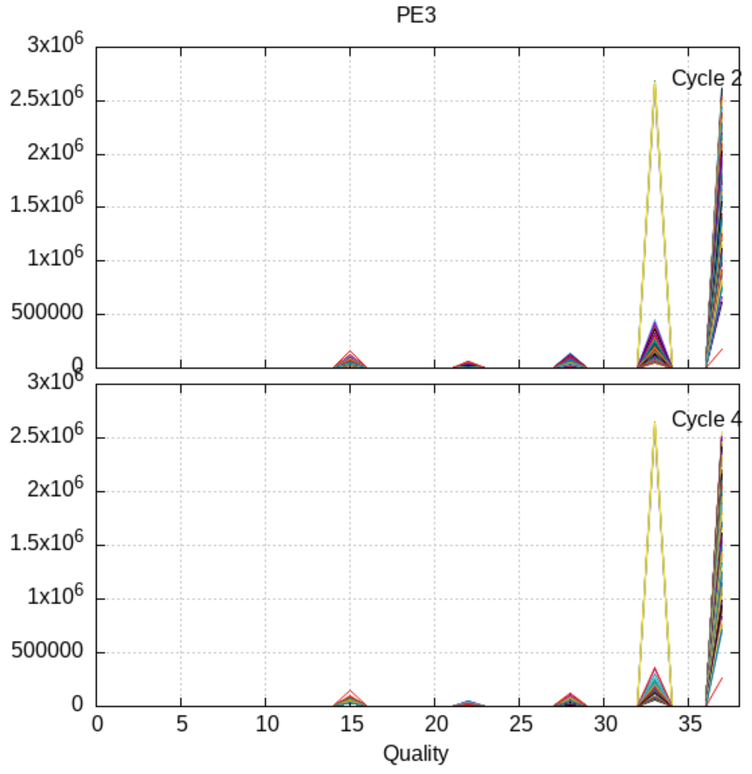

PE1
Forward reads
Reverse reads
Average Quality
Cycle

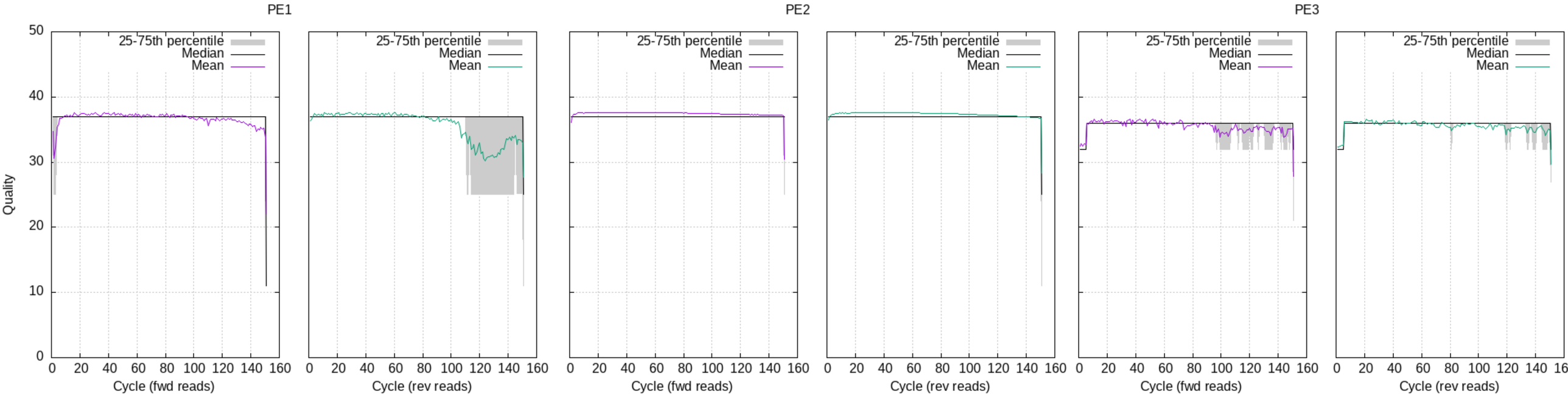

PE2
Forward reads
Reverse reads
Cycle


PE3
Forward reads
Reverse reads
Cycle

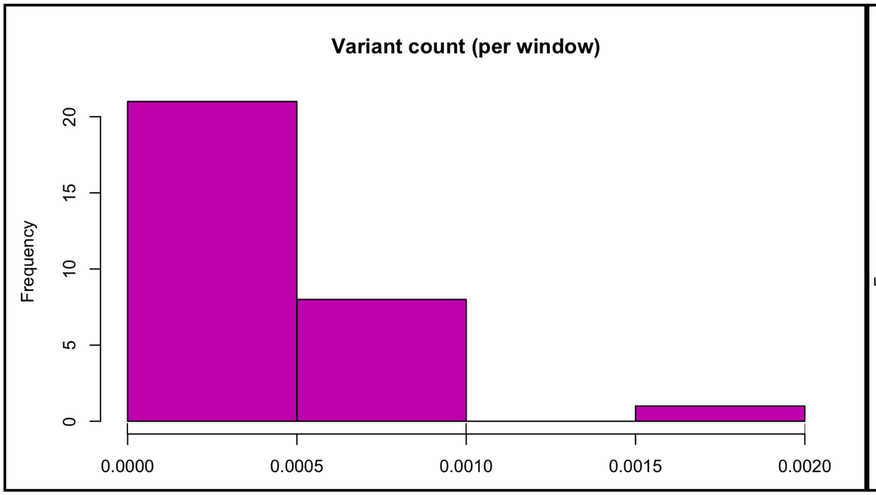

PE1
Frequency (fwd reads)
Frequency (rev reads)
Cycl
Quality


PE2
Frequency (fwd reads)
Frequency (rev reads)
Cycl
Quality

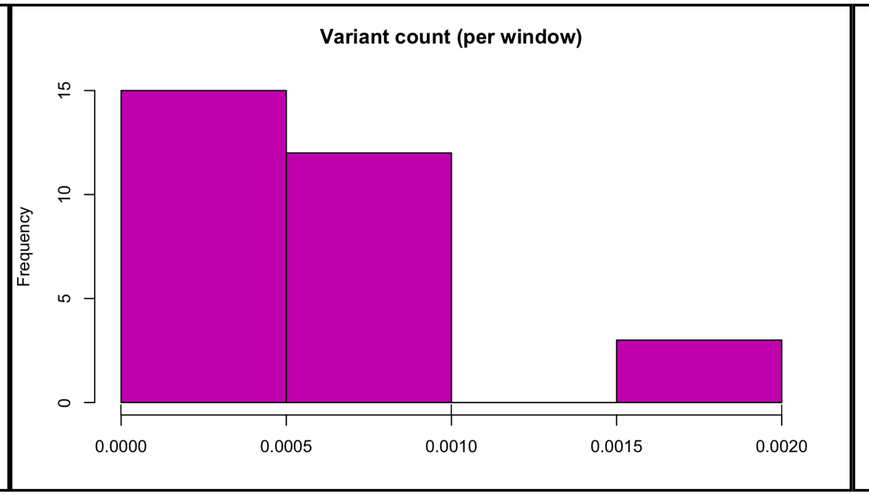

PE3
Cycle 2
Cycle 4
Quality


PE1
PE2
PE3
25-75th percentile
Median
Mean
Quality
Cycle (fwd reads)
Cycle (rev reads)
0
10
20
30
40
50
0 20 40 60 80 100 120 140 160

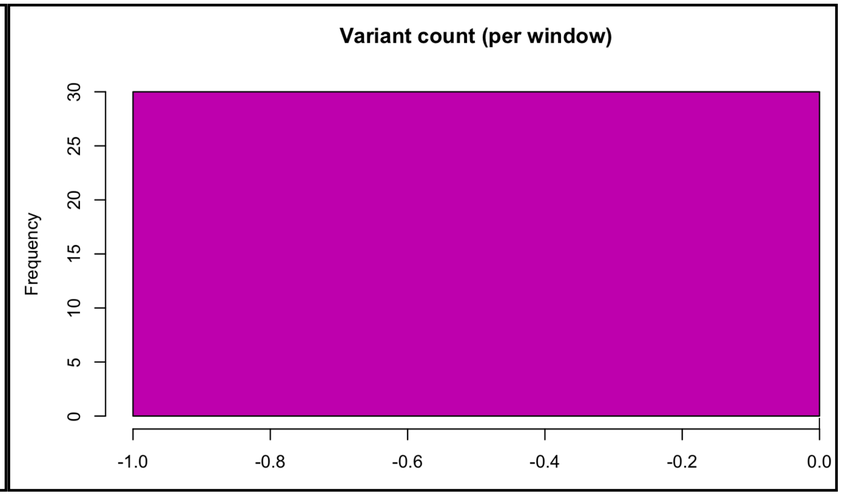
PE1


Variant count (per window)
Frequency
0
5
10
15
20
0.0000
0.0005
0.0010
0.0015
0.0020

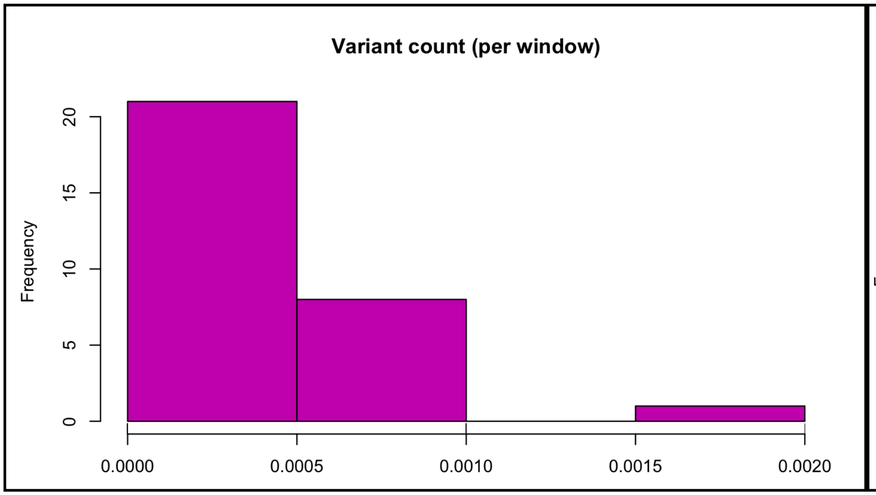
PE2


Variant count (per window)
Frequency
0
5
10
15
0.0000
0.0005
0.0010
0.0015
0.0020

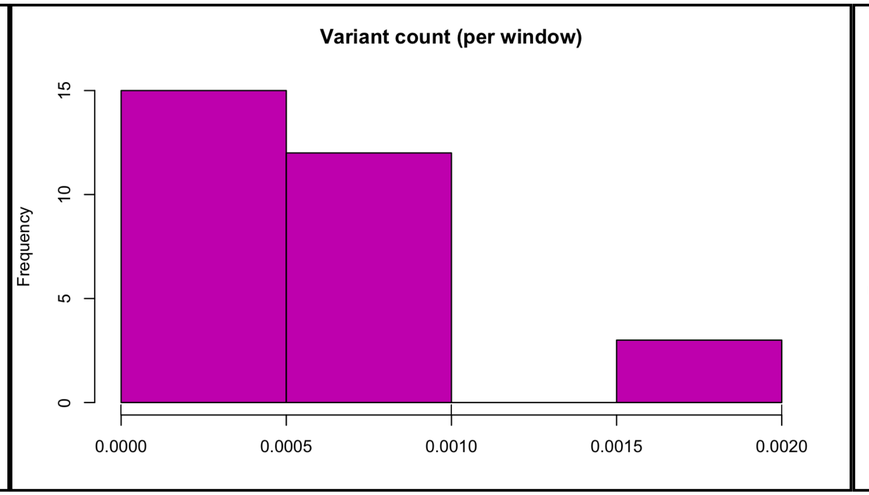
PE3


Variant count (per window)
Frequency
0
5
10
15
20
25
30
-1.0
-0.8
-0.6
-0.4
-0.2
0.0

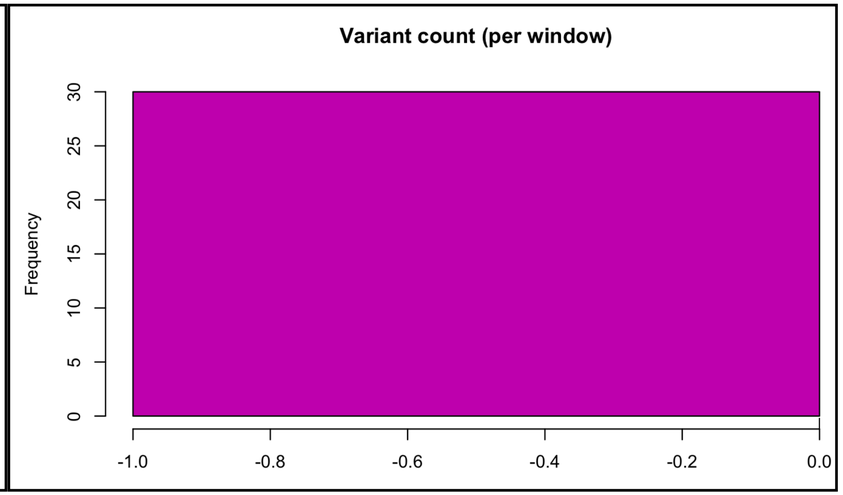
PE1



PE2

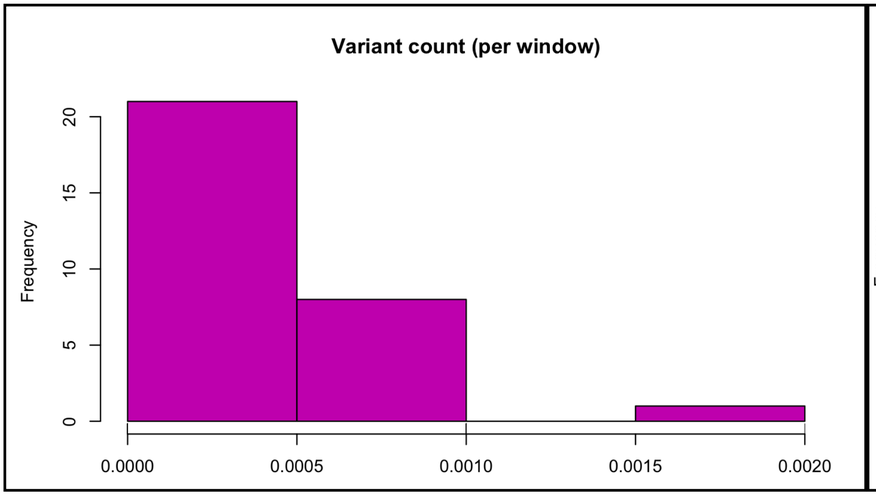


PE3

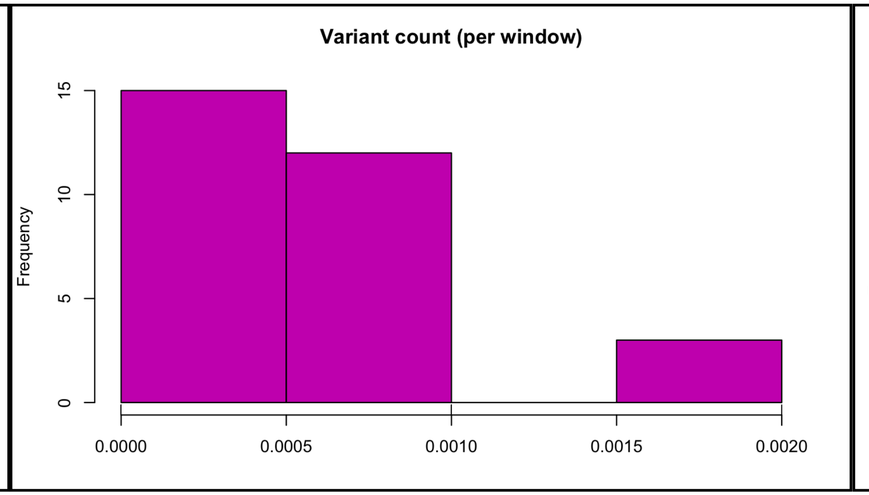


Remember the variant call ?

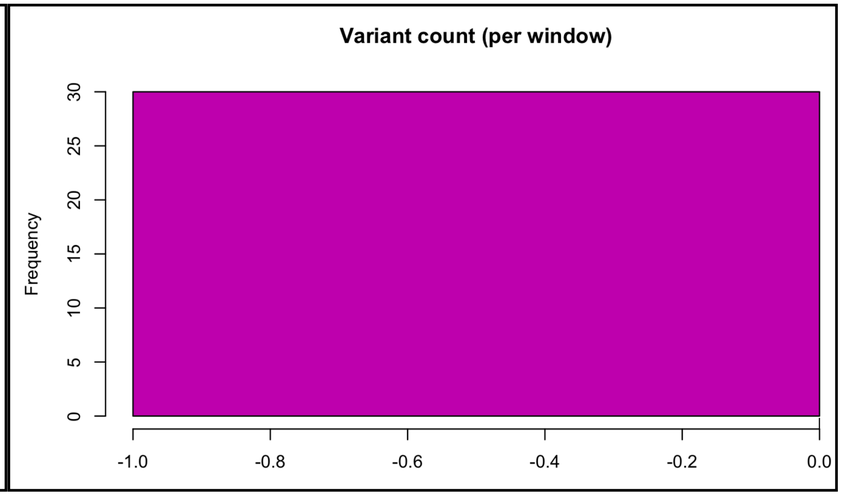
PE1



PE2



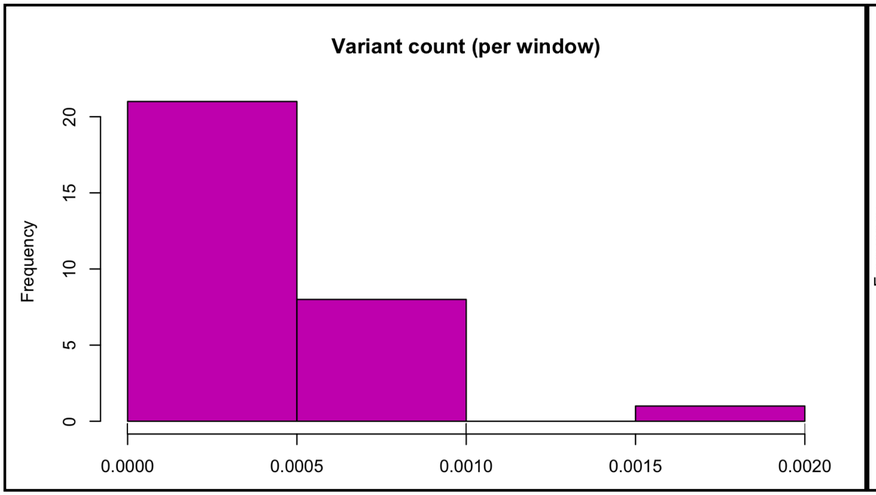
PE3



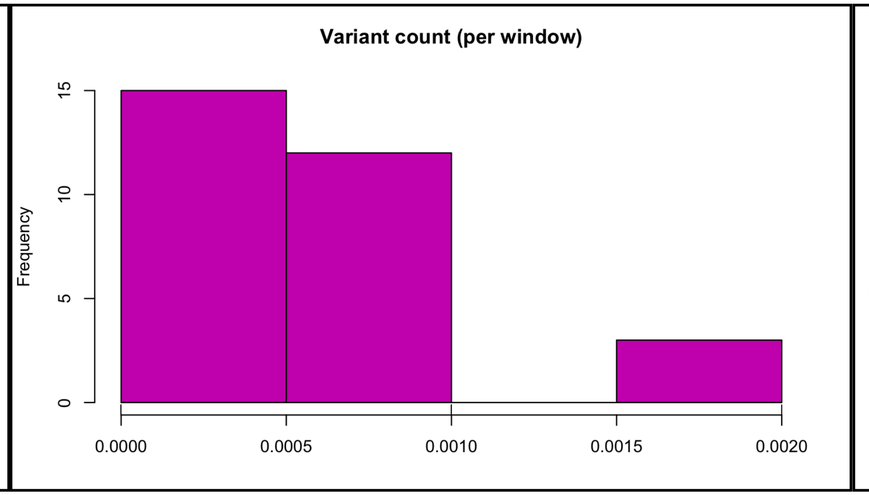
Remember the variant call ?

Unknown ? Monomorphic ? (hail mary attempt !!!)

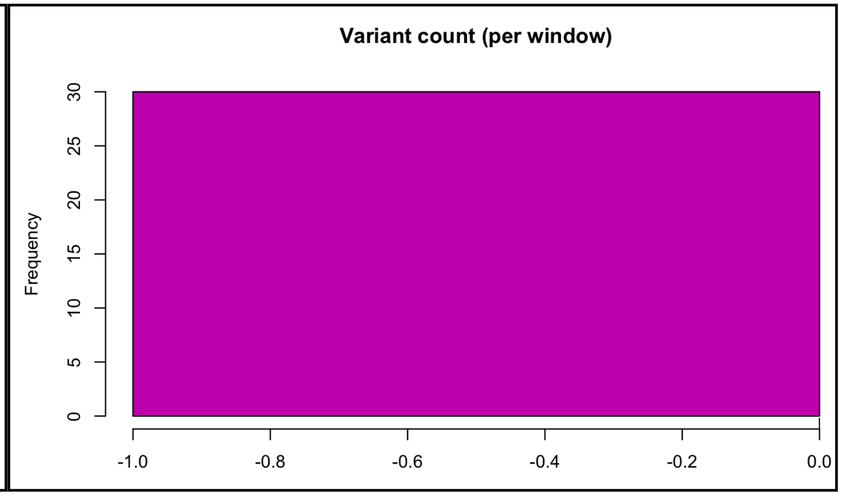
PE1



PE2



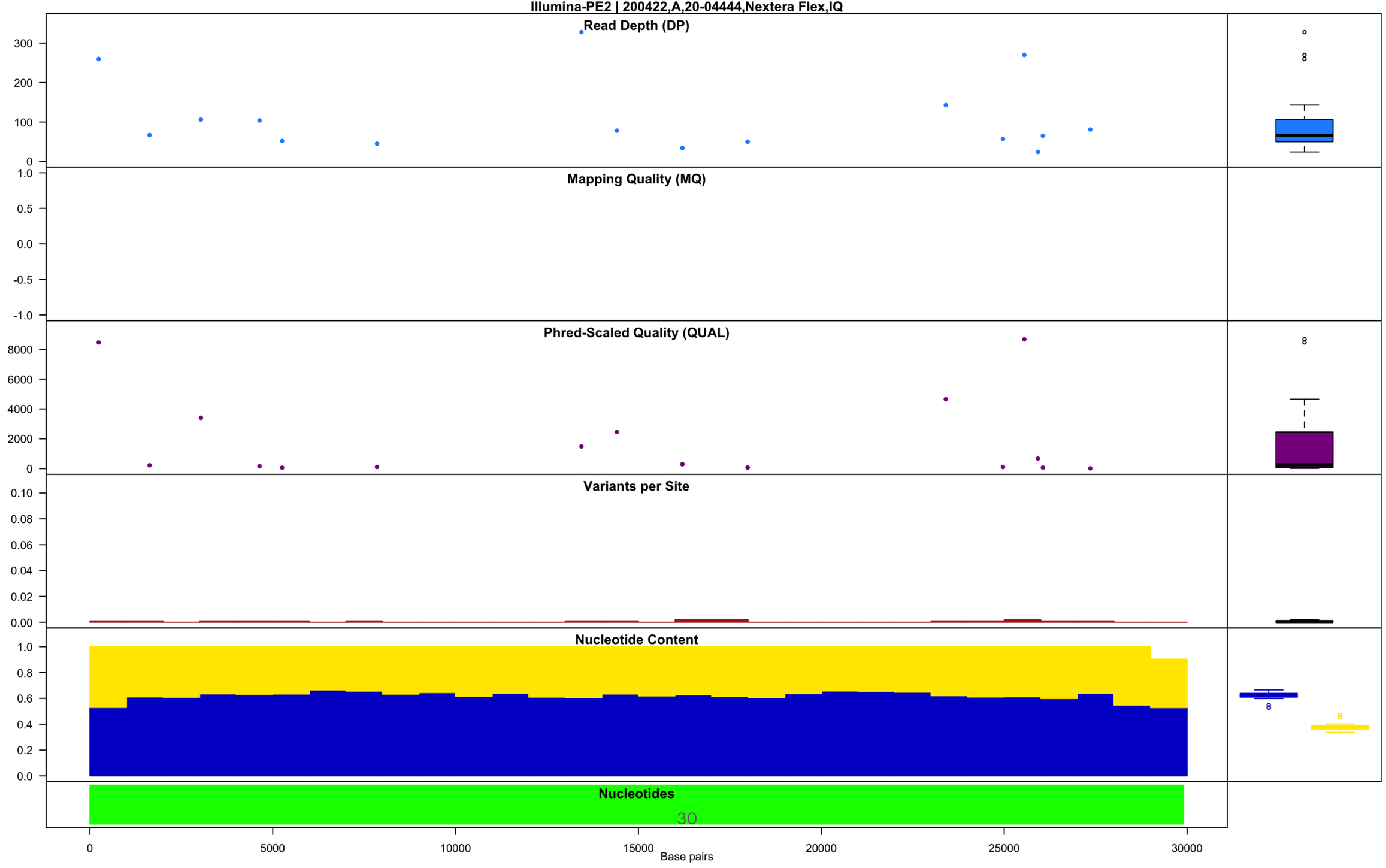
PE3

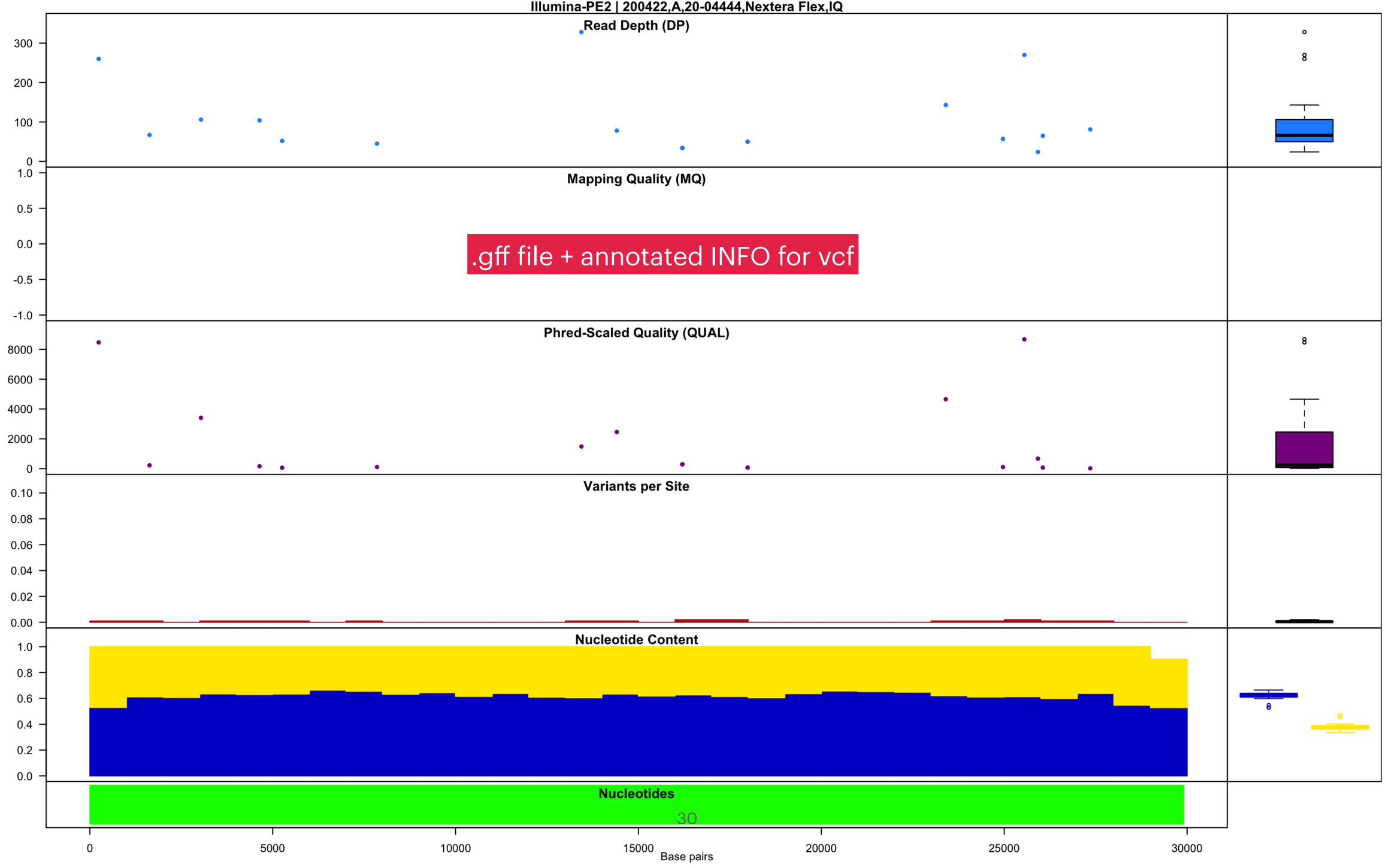


Remember the variant call ?

Unknown ? Monomorphic ? (hail mary attempt !!!)

vcfR ✓

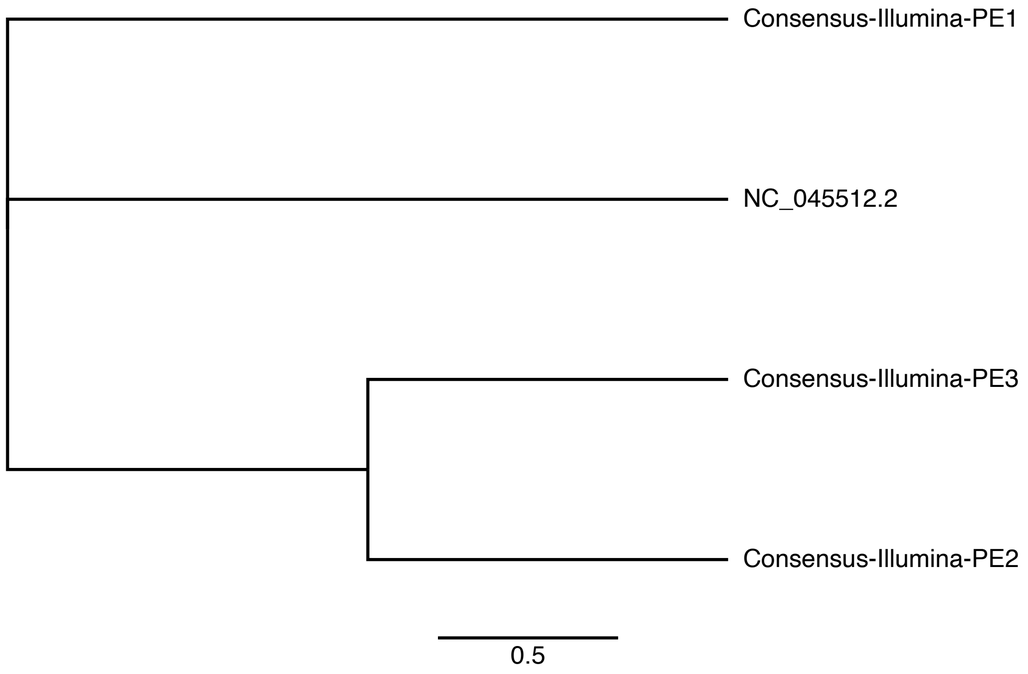

Illumina-PE2 | 200422,A,20-04444,Nextera Flex,IQ
Read Depth (DP)
300
200
100
0
Mapping Quality (MQ)
1.0
0.5
0.0
-0.5
-1.0
Phred-Scaled Quality (QUAL)
8000
6000
4000
2000
0
Variants per Site
0.10
0.08
0.06
0.04
0.02
0.00
Nucleotide Content
1.0
0.8
0.6
0.4
0.2
0.0
Nucleotides
30
0
5000
10000
15000
20000
25000
30000
Base pairs


Illumina-PE2 | 200422,A,20-04444,Nextera Flex,IQ
Read Depth (DP)
300
200
100
0
Mapping Quality (MQ)
1.0
0.5
0.0
-0.5
-1.0
.gff file + annotated INFO for vcf
Phred-Scaled Quality (QUAL)
8000
6000
4000
2000
0
Variants per Site
0.10
0.08
0.06
0.04
0.02
0.00
Nucleotide Content
1.0
0.8
0.6
0.4
0.2
0.0
Nucleotides
30
0
5000
10000
15000
20000
25000
30000
Base pairs

lineage B.1

0.02 % ambiguous content

(2 in 100 base pairs !)



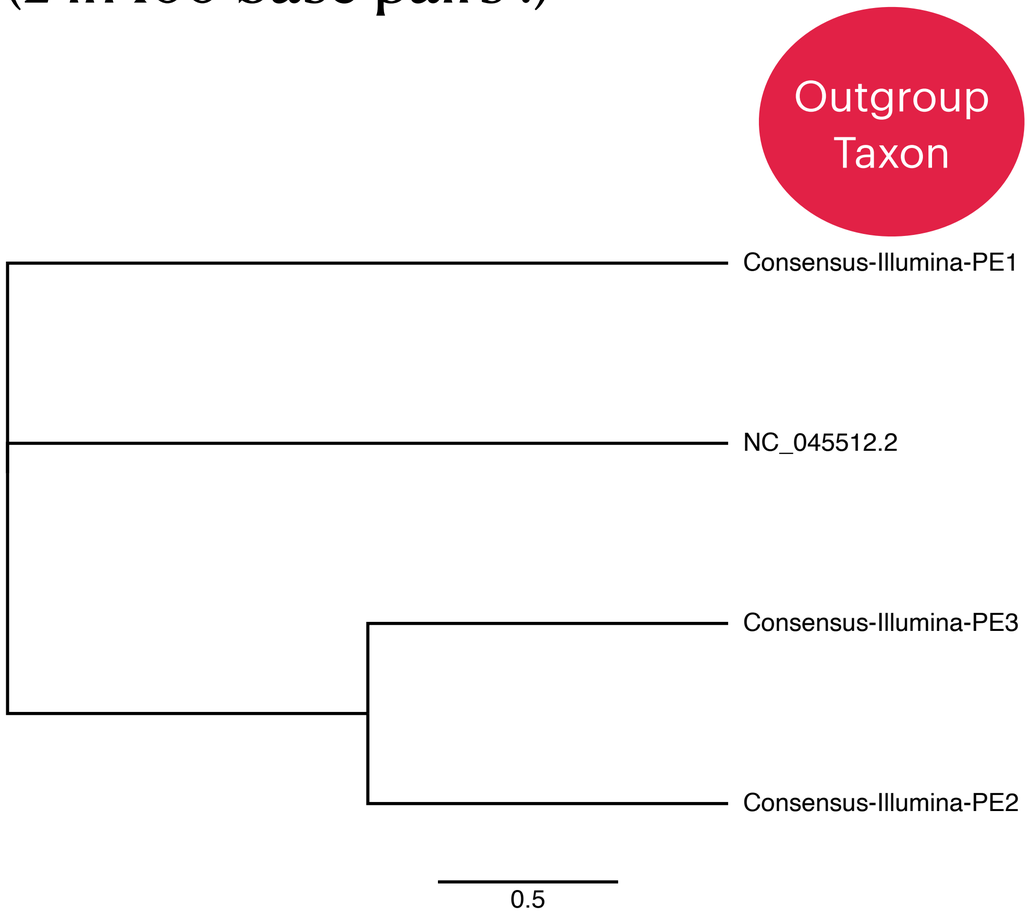
Cladogram (IROKI)

query_name	Consensus-Illumina-PE1	Consensus-Illumina-PE2	Consensus-Illumina-PE3
ACGT Nucleotide identity	0.9997	0.9994	0.9994
ACGT Nucleotide identity (ignoring Ns)	0.9997	0.9994	0.9994
ACGT Nucleotide identity (ignoring non-ACGTNs)	0.9998	0.9998	0.9998
qc_all_valid	TRUE	TRUE	TRUE
qc_post_align_pass_threshold	TRUE	TRUE	TRUE
qc_post_aligned	TRUE	TRUE	TRUE
qc_post_aligned_all_valid	TRUE	TRUE	TRUE
qc_valid_length	TRUE	TRUE	TRUE
qc_valid_nucleotides	TRUE	TRUE	TRUE
qc_valid_pass_nthreshold	TRUE	TRUE	TRUE
acgt_bases	29898	29891	29891
iupac_bases	5	12	12
non_iupac_bases	0	0	0
N_bases	0	0	0
length_query	29903	29903	29903
length_reference	29903	29903	29903
LongestNGap	0	0	0
Matches	29893	29885	29885
Mismatches	5	6	6
PSL_Mismatches	5	6	6
PSL_Ns	5	12	12

lineage B.1

0.02 % ambiguous content

(2 in 100 base pairs !)



Cladogram (IROKI)

query_name	Consensus-Illumina-PE1	Consensus-Illumina-PE2	Consensus-Illumina-PE3
ACGT Nucleotide identity	0.9997	0.9994	0.9994
ACGT Nucleotide identity (ignoring Ns)	0.9997	0.9994	0.9994
ACGT Nucleotide identity (ignoring non-ACGTNs)	0.9998	0.9998	0.9998
qc_all_valid	TRUE	TRUE	TRUE
qc_post_align_pass_threshold	TRUE	TRUE	TRUE
qc_post_aligned	TRUE	TRUE	TRUE
qc_post_aligned_all_valid	TRUE	TRUE	TRUE
qc_valid_length	TRUE	TRUE	TRUE
qc_valid_nucleotides	TRUE	TRUE	TRUE
qc_valid_pass_nthreshold	TRUE	TRUE	TRUE
acgt_bases	29898	29891	29891
iupac_bases	5	12	12
non_iupac_bases	0	0	0
N_bases	0	0	0
length_query	29903	29903	29903
length_reference	29903	29903	29903
LongestNGap	0	0	0
Matches	29893	29885	29885
Mismatches	5	6	6
PSL_Mismatches	5	6	6
PSL_Ns	5	12	12

Interpretation

Summary

1. B.1 is a large European lineage the origin of which roughly corresponds to the Northern Italian outbreak early in 2020. (Earliest date: 2020-01-01)
2. United States of America 46.0%, Turkey 11.0%, United Kingdom 6.0%, Canada 4.0%, France 3.0%

Source: https://cov-lineages.org/lineage_list.html

Conclusion

Which data set achieved the best consensus sequence ?

Based on the coverage, aligned reads, quality, depth plots along with quality control of consensus sequences

(**PE1**) Paired End read files for

200408,A,20-04246,CleanPlex SARS-CoV-2,IQ

Discussion

& Improvement

- Variant calling needs **different parameters or different tools** entirely (*bcftools mpileup* | *call*) ———> Fine tuning !
- Interpretation could have been done more carefully - **naive command running is foolish**
- For future, **understanding file types** in detail based on the experience with .bed and .bedpe files would be better
- Learning and trying more **plug and push** across different pipelines

Fun project!
Thanks !

(~ 64 hours *_*)

References

1. bamstats.pl script, 2012-2014 Genome Research Ltd (Author: Petr Danecek <pd3@sanger.ac.uk>)

2. Anon, 2020. *Anaconda Software Distribution*, Anaconda Inc. Available at: https://docs.anaconda.com/.

3. Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, *Bioinformatics*, Volume 32, Issue 19, October 2016, Pages 3047–3048, https://doi.org/10.1093/bioinformatics/btw354

4. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].

5. Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, *Bioinformatics*, Volume 34, Issue 17, September 2018, Pages i884–i890, https://doi.org/10.1093/bioinformatics/bty560

6. Heng Li, Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics*, Volume 34, Issue 18, September 2018, Pages 3094–3100, https://doi.org/10.1093/bioinformatics/bty191

7. Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li, Twelve years of SAMtools and BCFtools, *GigaScience*, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008

8. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

9. O'Toole Á, Hill V, Pybus OG et al. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 [version 1; peer review: 3 approved]. Wellcome Open Res 2021, 6:121 (https://doi.org/10.12688/wellcomeopenres.16661.1)

10. Áine O'Toole, Emily Scher, Anthony Underwood, Ben Jackson, Verity Hill, John T McCrone, Rachel Colquhoun, Chris Ruis, Khalil Abu-Dahab, Ben Taylor, Corin Yeats, Louis du Plessis, Daniel Maloney, Nathan Medd, Stephen W Attwood, David M Aanensen, Edward C Holmes, Oliver G Pybus, Andrew Rambaut, Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool, *Virus Evolution*, Volume 7, Issue 2, December 2021, veab064, https://doi.org/10.1093/ve/veab064

11. Rambaut, A., Holmes, E.C., O'Toole, Á. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403–1407 (2020). https://doi.org/10.1038/s41564-020-0770-5

12. Tool for QC with consensus sequences https://github.com/rki-mf1/president

13. Au, C., Ho, D., Kwong, A. *et al.* BAMClipper: removing primers from alignments to minimize false-negative mutations in amplicon next-generation sequencing. *Sci Rep* **7**, 1567 (2017). https://doi.org/10.1038/s41598-017-01703-6

14. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]* 2012

15. Vcflib and tools for processing the VCF variant call format. Erik Garrison, Zev N. Kronenberg, Eric T. Dawson, Brent S. Pedersen, Pjotr Prins. bioRxiv 2021.05.21.445151; doi: https://doi.org 10.1101/2021.05.21.445151

16. Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools, *Bioinformatics*, Volume 27, Issue 15, August 2011, Pages 2156–2158, https://doi.org/10.1093/bioinformatics/btr330

17. Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M., Barton, G.J (2009), "Jalview version 2: A Multiple Sequence Alignment and Analysis Workbench," *Bioinformatics* **25** (9) 1189-1191 doi: 10.1093/bioinformatics/btp033

18. Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

19. Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, Robert Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era, *Molecular Biology and Evolution*, Volume 37, Issue 5, May 2020, Pages 1530–1534, https://doi.org/10.1093/molbev/msaa015

20. Aaron R. Quinlan, Ira M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, Volume 26, Issue 6, March 2010, Pages 841–842, https://doi.org/10.1093/bioinformatics/btq033

21. Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

22. R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

23. Knaus, B.J. and Grünwald, N.J. (2017), vcfr: a package to manipulate and visualize variant call format data in R. Mol Ecol Resour, 17: 44-53. https://doi.org/10.1111/1755-0998.12549

24. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2023 Sep 29]. Available from: https://www.ncbi.nlm.nih.gov/

Credits

1. The used Pg. 8 template for workflow is from slidesgo
2. This presentation was made on Keynote version 13.2
3. The data files were provided by instructors
4. The project is available on google drive here. It will be there until formatted into a repo on GitHub.
5. Available: https://github.com/bibymaths/SARSCoV2

Device Info

Post Analysis

OS Fedora Linux 38

Kernel Linux 6.4.15-200.fc38.x86_64

Processor Intel i5-8250U (8 slots), with CUDA support

Graphics UHD 620 (KBL GT2)

Memory 8 GB

I uploaded the whole project folder to google drive, for transferring data to Mac and used Google Takeout to download the folder in .tgz format. While making the zipped file, google adds the information about files, and directories, so we need to delete them. Also, using the QualMap tool, it creates meta-data and replicated files so we need to delete them as well.