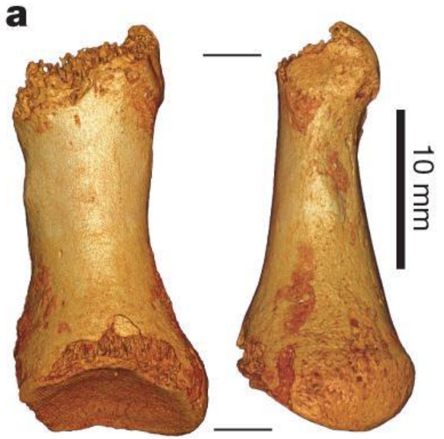
An Introduction to the Analysis of Ancient Sequencing Data

Pre-Processing the Unmapped Reads of the Altai Neanderthal Sequencing

Maximilian Arlt
Master Student
Chair of RNA Bioinformatics and High-Throughput Sequencing,
Prof. Dr. Manja Marz
FSU Jena

What is ancient DNA (aDNA) ?

- ancient = before the fall of Roman Empire
- extracted from ‘old’ remains (bones, tooth, soft tissue)



adapted from K Prüfer et al. Nature 000, 1-7 (2013)
doi:10.1038/nature12886

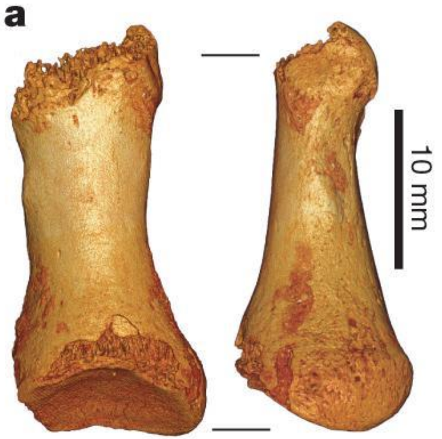
What is ancient DNA (aDNA) ?

- ancient = before the fall of Roman Empire
- extracted from 'old' remains (bones, tooth, soft tissue)

It is degraded.

It is modified.

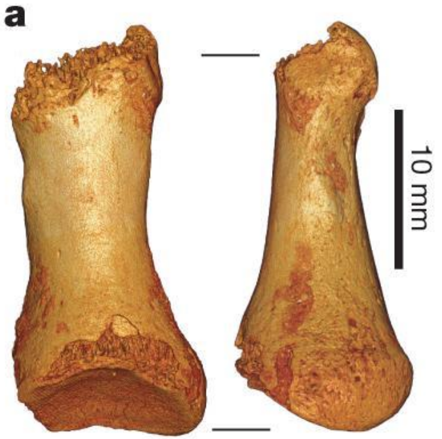
It is contaminated.



adapted from K Prüfer et al. Nature 000, 1-7 (2013)
doi:10.1038/nature12886

What is ancient DNA (aDNA) ?

- pedal toe phalanx, Neanderthal Individual
- ~ 50 000 years old
- Extraction, Amplification, PE - Sequencing (Illumina)
- ~ 70 % endogenous DNA



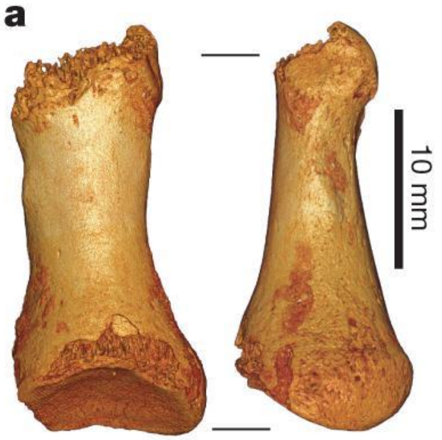
adapted from K Prüfer et al. Nature 000, 1-7 (2013)
doi:10.1038/nature12886

What is ancient DNA (aDNA) ?

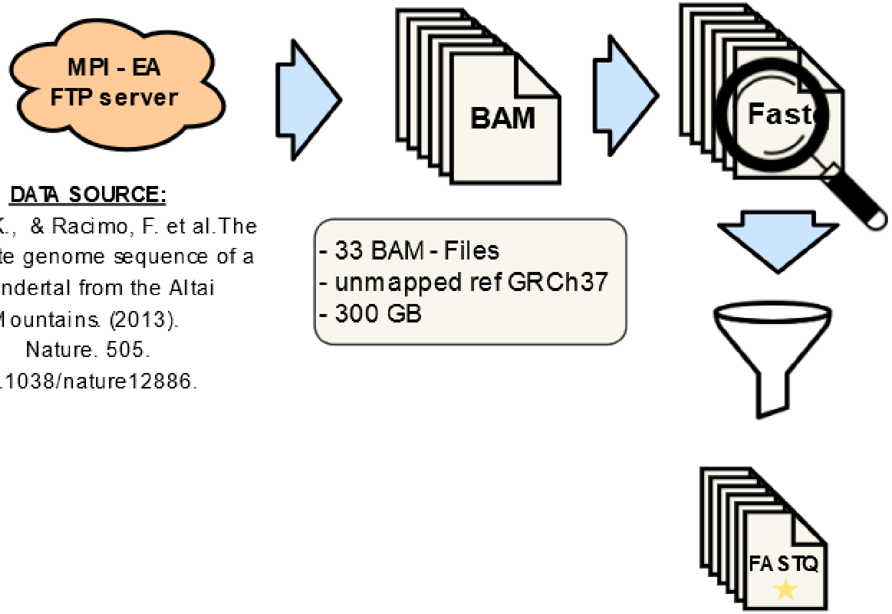
- pedal toe phalanx, Neanderthal Individual
- ~ 50 000 years old
- Extraction, Amplification, PE - Sequencing (Illumina)
- ~ 70 % endogenous DNA

↓

What about the other 30 % ?



adapted from K Prüfer et al. Nature 000, 1-7 (2013)
doi:10.1038/nature12886



DATA SOURCE:

Prüfer, K., & Racimo, F. et al. The complete genome sequence of a Neandertal from the Altai Mountains. (2013). Nature. 505. 10.1038/nature12886.

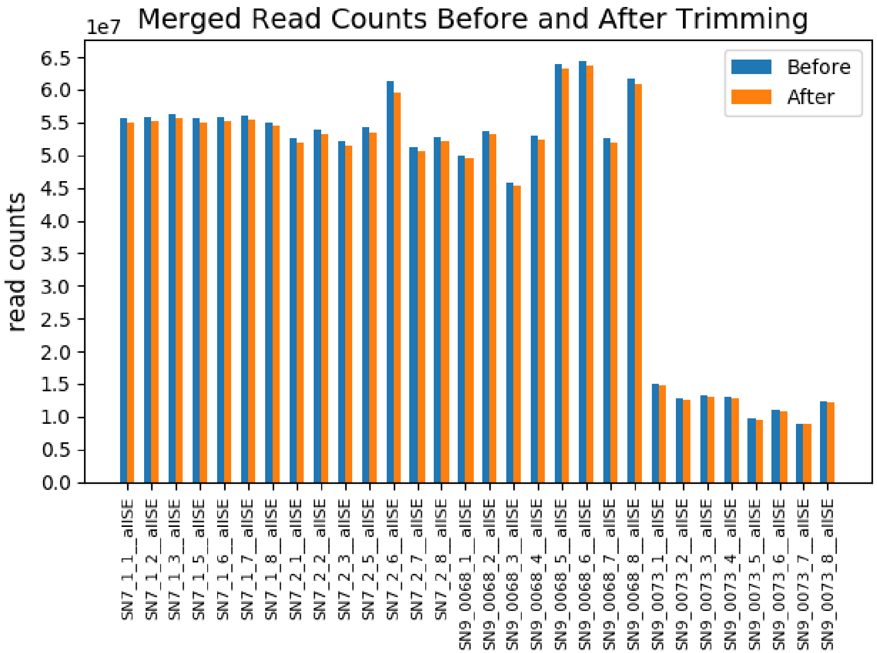
- 33 BAM - Files
- unmapped ref GRCh37
- 300 GB

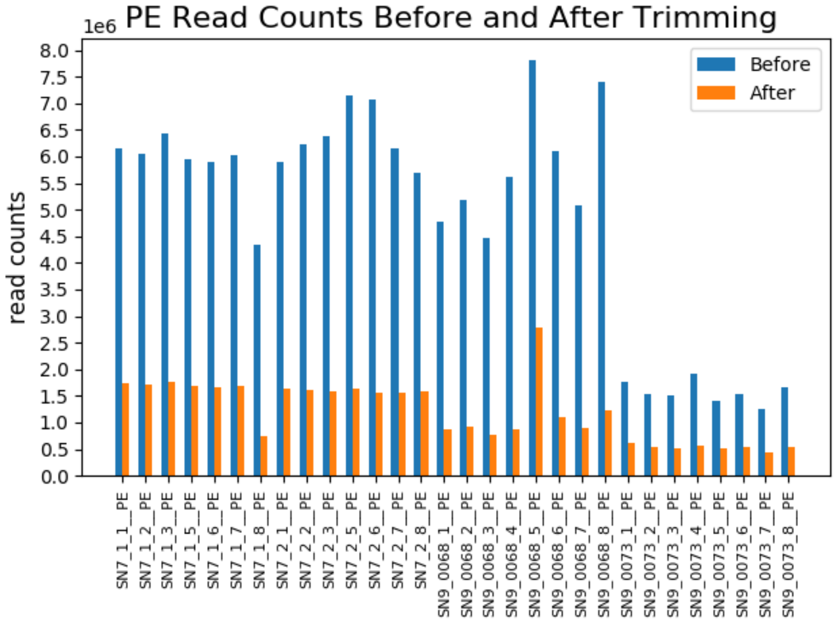
1. Quality assessment
- fastqc reports
- multiqc report

2. Quality enrichment
- fastp
- fastqc

file.bam

Pairs **Merged Pairs** **Singletons**





Before Filtering:

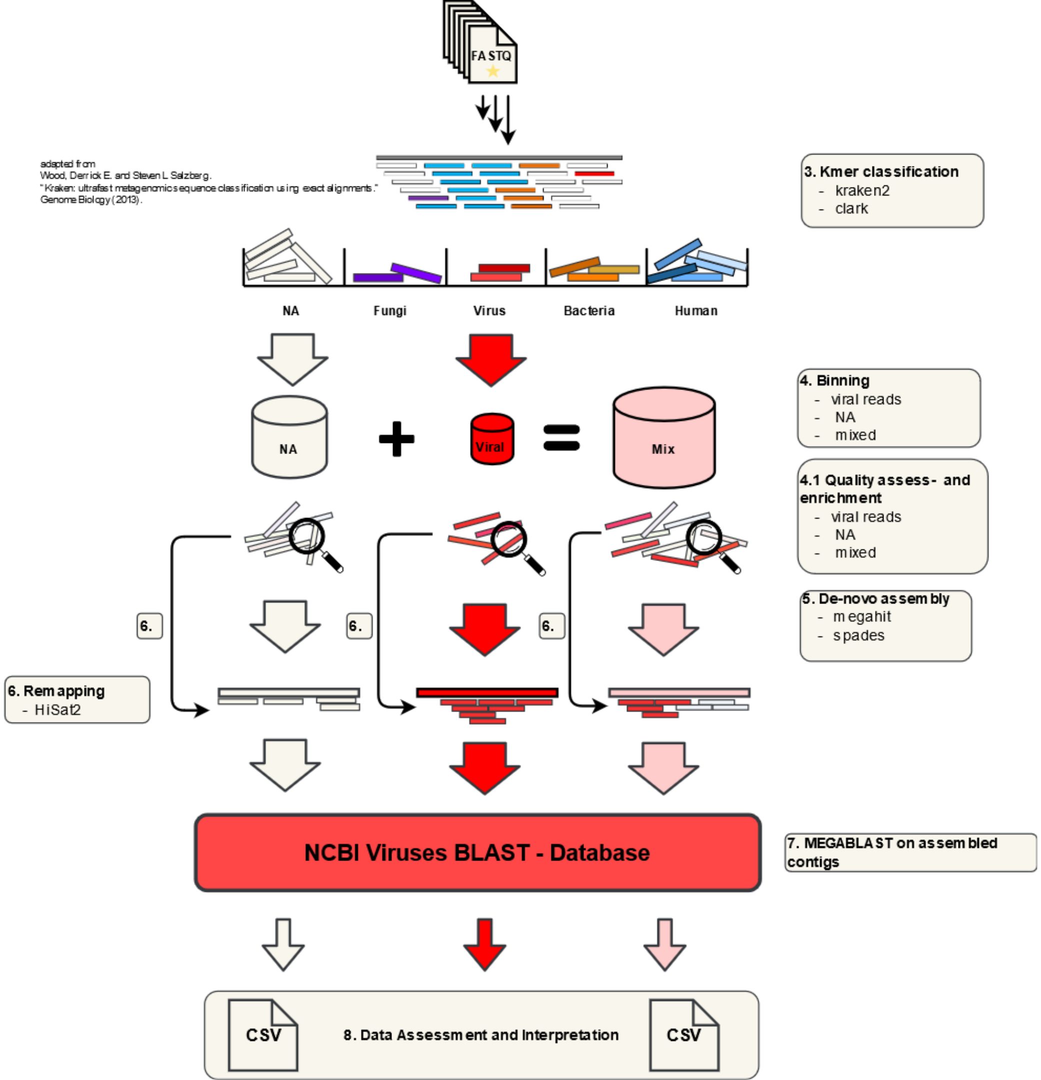
1,309,313,311	M
144,357,489	PE (2x)
1,598,028,289	**TOTAL**

Fastp parameters:

Phred >= 24
Length >= 20 nt
Adapter Trimming
Low complexity Filtering
PolyX Trimming >= 10 nt

After Filtering:

1,293,551,377	M
35,876,296	PE (2x)
1,365,303,969	**TOTAL**



adapted from
Wood, Derrick E. and Steven L Salzberg.
"Kraken: ultrafast metagenomics sequence classification using exact alignments."
Genome Biology (2013).

3. Kmer classification
- kraken2
- clark

4. Binning
- viral reads
- NA
- mixed

4.1 Quality assess- and enrichment
- viral reads
- NA
- mixed

5. De-novo assembly
- megahit
- spades

6. Remapping
- HiSat2

NCBI Viruses BLAST - Database

7. MEGABLAST on assembled contigs

8. Data Assessment and Interpretation

Thank you for your attention!

References

1. Prüfer, K., Racimo, F., Patterson, N. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505, 43–49 (2014).

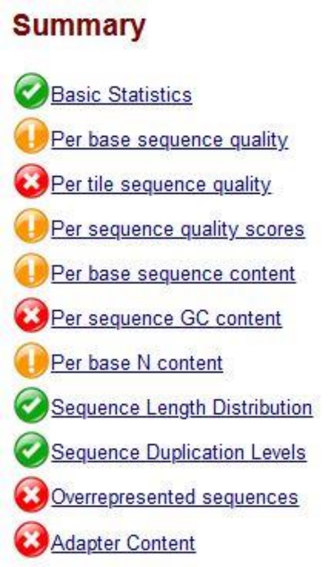
2. Prüfer, K., Racimo, F., Patterson, N. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505, 43–49 (2014). Supplementary information.

3. Gansauge MT., Meyer M. (2019) A Method for Single-Stranded Ancient DNA Library Preparation. In: Shapiro B., Barlow A., Heintzman P., Hofreiter M., Paijmans J., Soares A. (eds) Ancient DNA. Methods in Molecular Biology, vol 1963. Humana Press, New York, NY.

Fastqc Reports

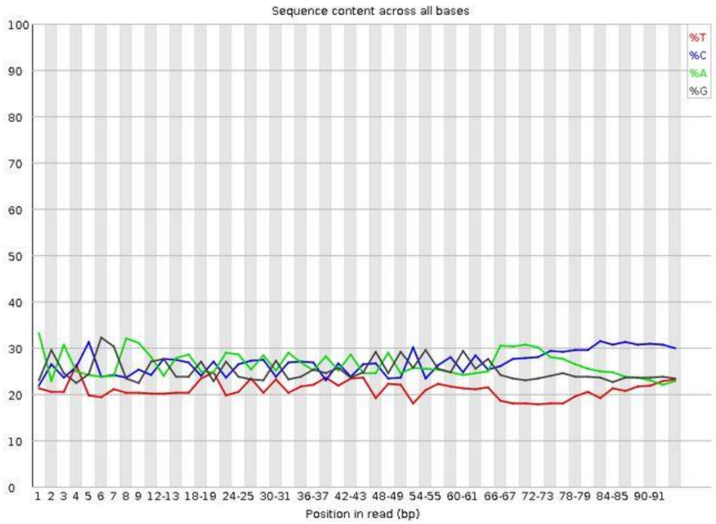
Quality Reports indicate artificial Sequence Content

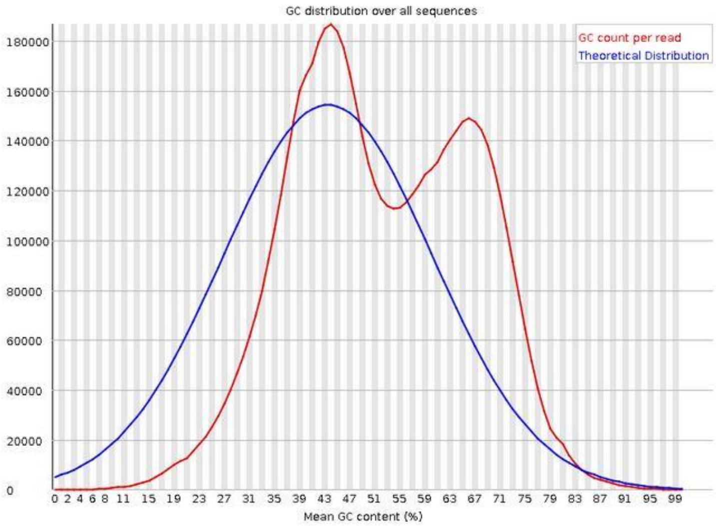
FastQC - Before



Summary

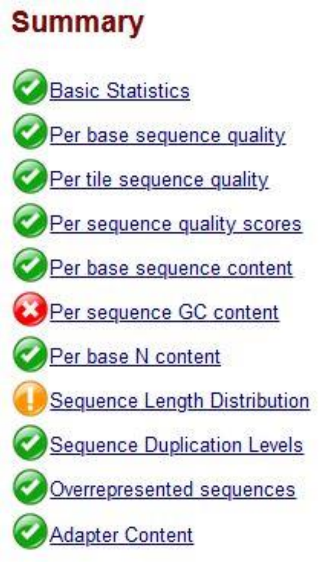
- ✓ Basic Statistics
- ! Per base sequence quality
- ✗ Per tile sequence quality
- ! Per sequence quality scores
- ! Per base sequence content
- ✗ Per sequence GC content
- ! Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✗ Adapter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





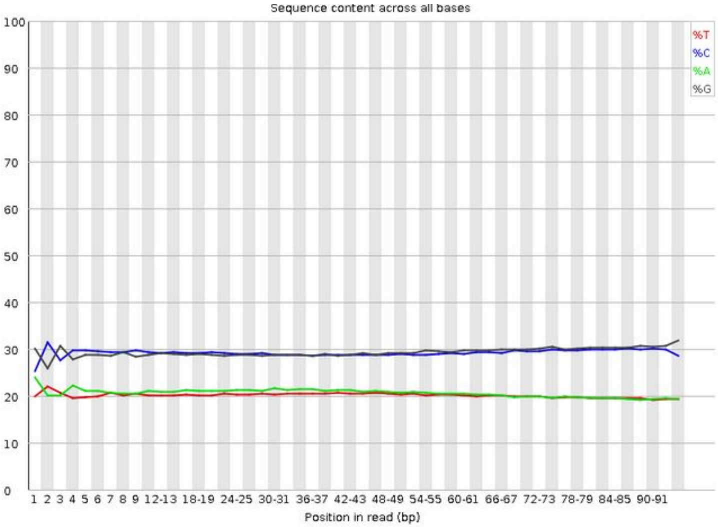
Quality Reports indicate artificial Sequence Content

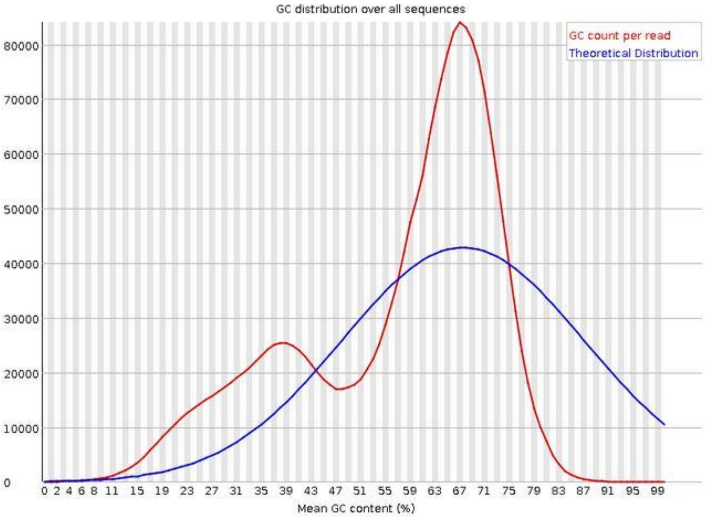
FastQC - After



Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile 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



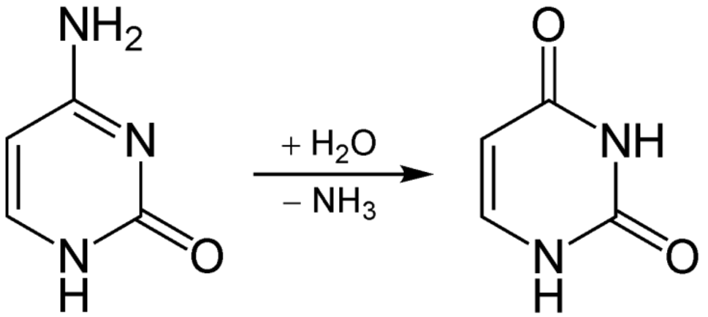


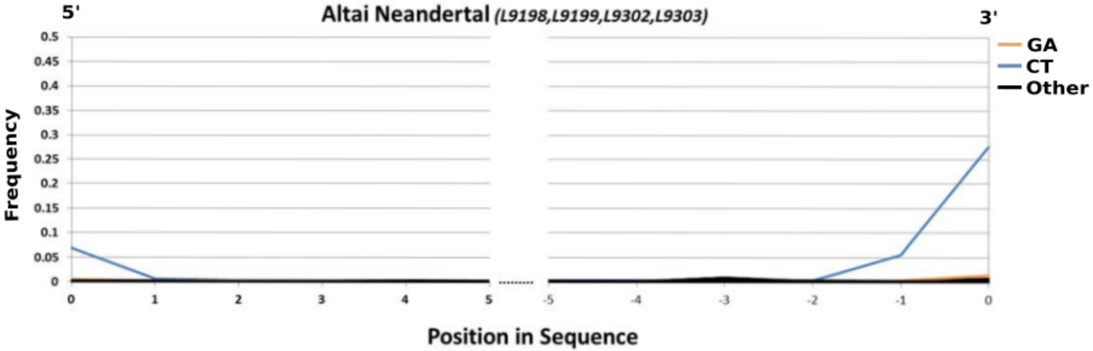
Ancient DNA is commonly:

degraded

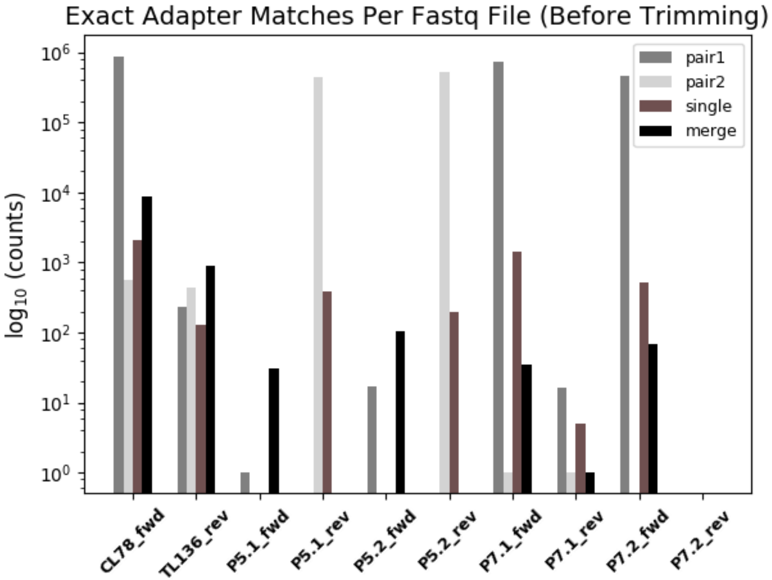
modified

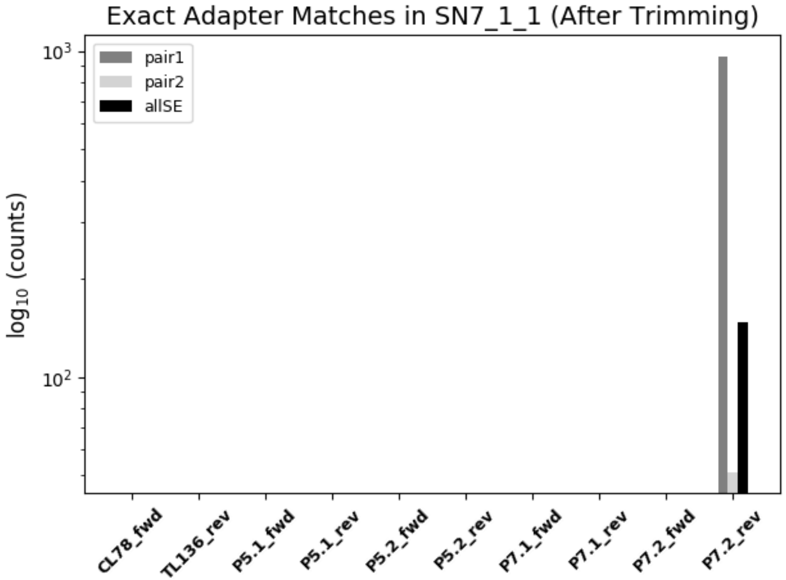
contaminated

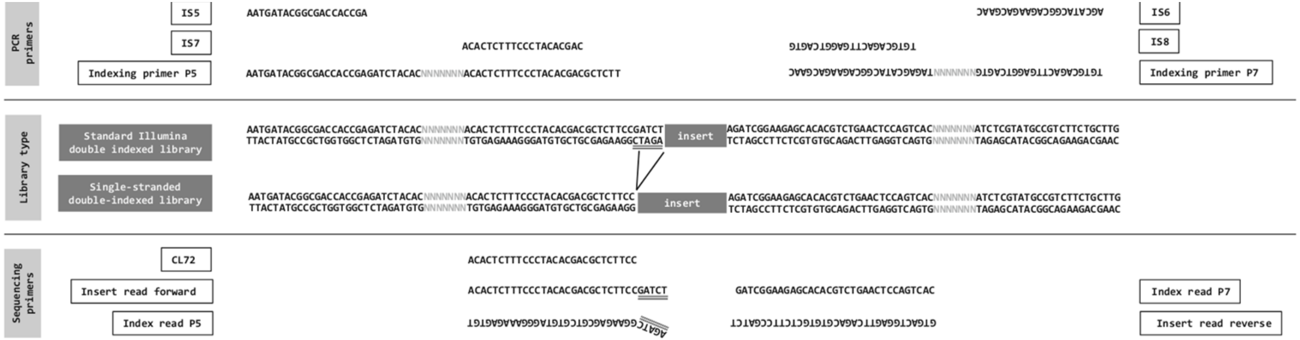




adapted from: K Prüfer et al. Nature 000, 1-7 (2013); doi:10.1038/nature12886; supplements Figure S5a.3







PCR primers

IS5 AATGATACGGCGACCACCGA

IS6 AGCATACGGCAGAAGACGAAC

IS7 ACACTCTTTCCCTACACGAC

IS8 TGTGCAGACTTGAGGTCAGTG

Indexing primer P5 AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTT

Indexing primer P7 TGTGCAGACTTGAGGTCAGTGNNNNNNNTAGAGCATACGGCAGAAGACGAAC

Library type

Standard Illumina double indexed library

AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTTCCGATCT insert AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTAGATGTGNNNNNNNTGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA insert TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNNTAGAGCATACGGCAGAAGACGAAC

Single-stranded double-indexed library

AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTTCC insert AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTAGATGTGNNNNNNNTGTGAGAAAGGGATGTGCTGCGAGAAGG insert TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNNTAGAGCATACGGCAGAAGACGAAC

Sequencing primers

CL72 ACACTCTTTCCCTACACGACGCTCTTCC

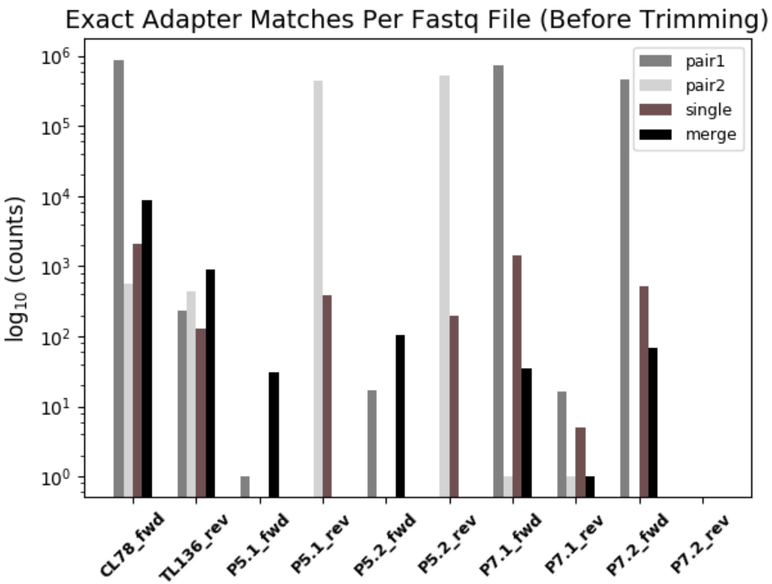
Insert read forward ACACTCTTTCCCTACACGACGCTCTTCCGATCT

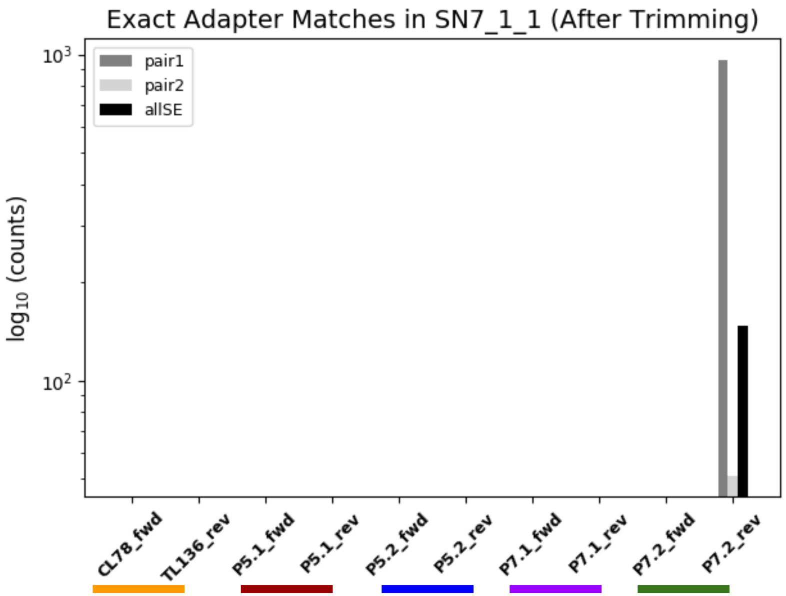
Index read P5 AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

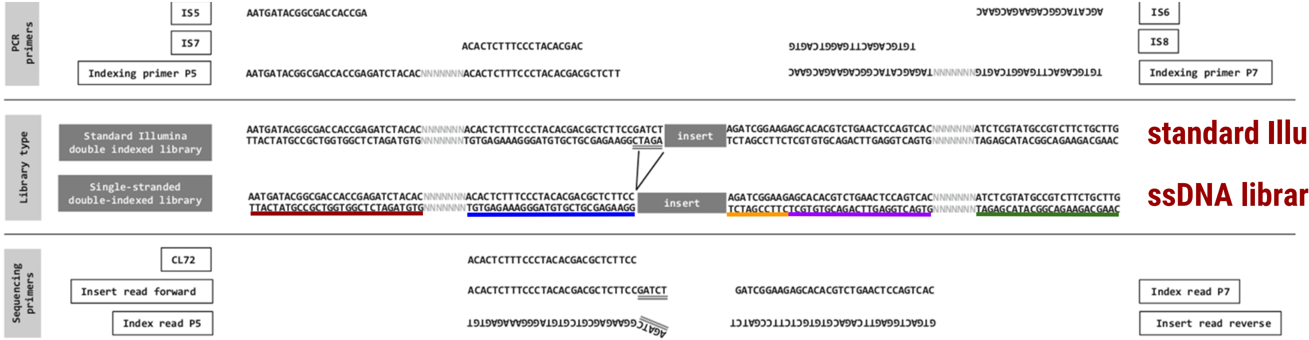
Index read P7 GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

Insert read reverse TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG

adapted from MT. Gansauge , M. Meyer, Methods in Molecular Biology(2019), https://doi.org/10.1007/978-1-4939-9176-1_9







IS5 AATGATACGGCGACCACCGA

IS7 ACACTCTTTCCCTACACGAC

Indexing primer P5 AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTT

IS6 CAAGCAGAAGACGGCATACGA

IS8 GTGACTGGAGTTCAGACGTGT

Indexing primer P7 CAAGCAGAAGACGGCATACGAGATNNNNNNNGTGACTGGAGTTCAGACGTGT

Standard Illumina double indexed library
AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTTCCGATCT insert AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTAGATGTGNNNNNNNTGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNNTAGAGCATACGGCAGAAGACGAAC

Single-stranded double-indexed library
AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTTCC insert AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTAGATGTGNNNNNNNTGTGAGAAAGGGATGTGCTGCGAGAAGG TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNNTAGAGCATACGGCAGAAGACGAAC

CL72 ACACTCTTTCCCTACACGACGCTCTTCC

Insert read forward ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Index read P5 AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Index read P7 GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

Insert read reverse GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

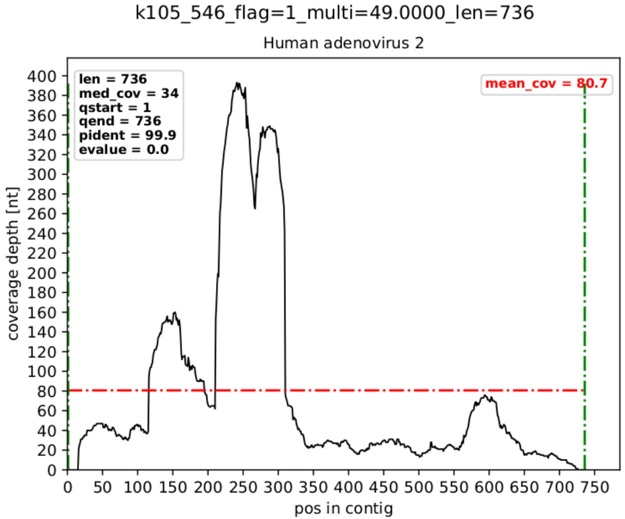
standard Illumina

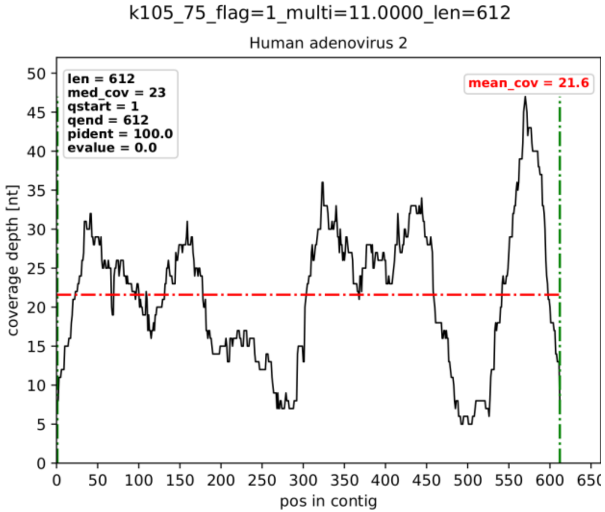
ssDNA library

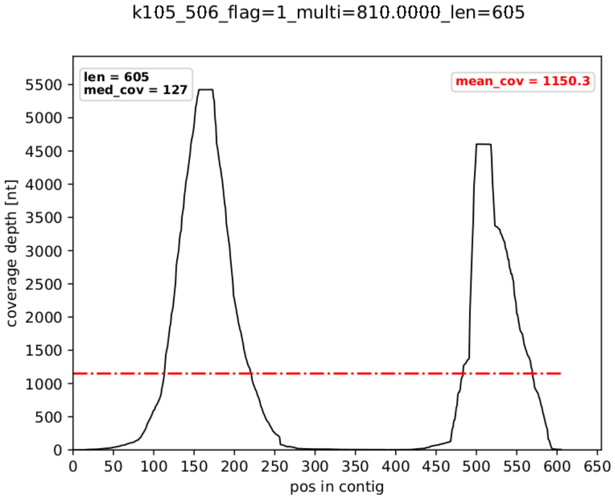
adapted from MT. Gansauge , M. Meyer, Methods in Molecular Biology(2019), https://doi.org/10.1007/978-1-4939-9176-1_9

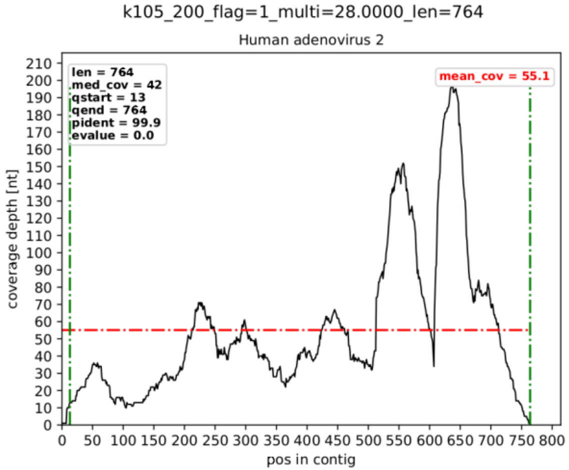
Maximilian Arlt

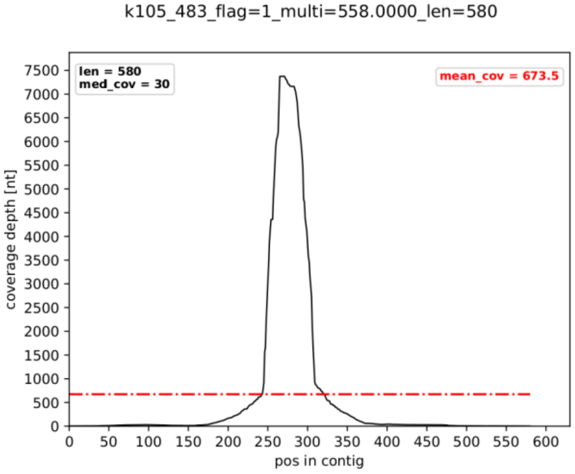
Increasing resolution - coverage depth plots on MEGAHIT viral candidates

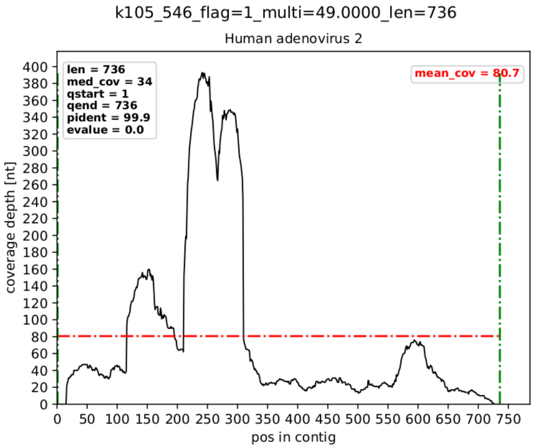

k105_546_flag=1_multi=49.0000_len=736
Human adenovirus 2
len = 736
med_cov = 34
qstart = 1
qend = 736
pident = 99.9
evalue = 0.0
mean_cov = 80.7
coverage depth [nt]
pos in contig


k105_75_flag=1_multi=11.0000_len=612
Human adenovirus 2
len = 612
med_cov = 23
qstart = 1
qend = 612
pident = 100.0
evalue = 0.0
mean_cov = 21.6
coverage depth [nt]
pos in contig


k105_506_flag=1_multi=810.0000_len=605
len = 605
med_cov = 127
mean_cov = 1150.3
coverage depth [nt]
pos in contig


k105_200_flag=1_multi=28.0000_len=764
Human adenovirus 2
len = 764
med_cov = 42
qstart = 13
qend = 764
pident = 99.9
evalue = 0.0
mean_cov = 55.1
coverage depth [nt]
pos in contig


k105_483_flag=1_multi=558.0000_len=580
len = 580
med_cov = 30
mean_cov = 673.5
coverage depth [nt]
pos in contig


k105_546_flag=1_multi=49.0000_len=736
Human adenovirus 2
len = 736
med_cov = 34
qstart = 1
qend = 736
pident = 99.9
evalue = 0.0
mean_cov = 80.7
coverage depth [nt]
pos in contig