RNA virus full genome sequencing and haplotype reconstruction

Sebastian Krautwurst

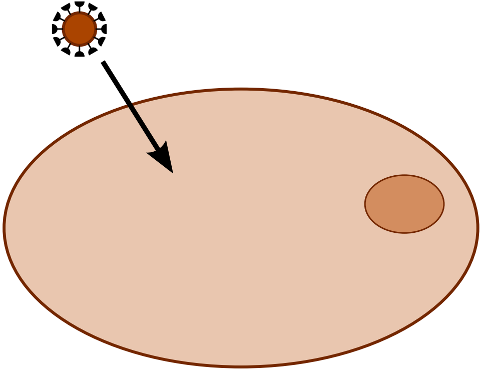
February 13, 2018
33rd TBI Winterseminar in Bled

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

Background

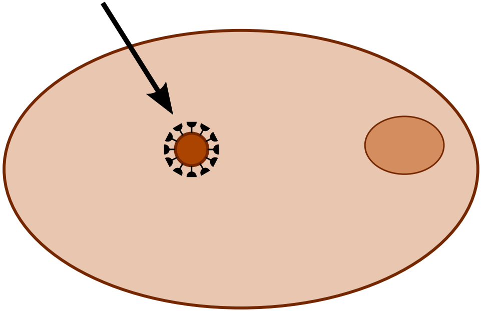
VIRAL HAPLOTYPES

- One species = one genome?



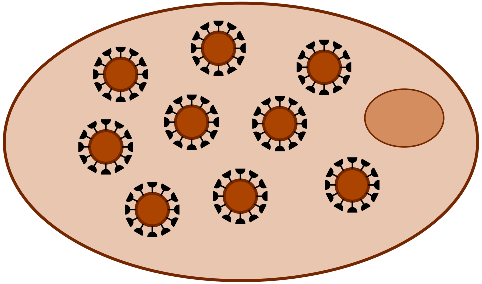
Viral haplotypes

- One species = one genome?



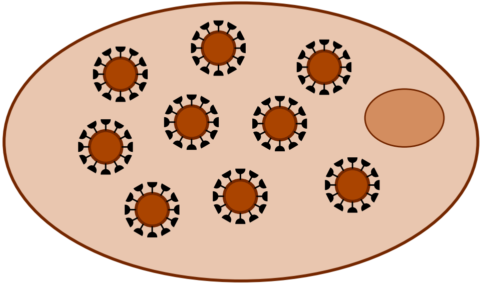
VIRAL HAPLOTYPES

- One species = one genome?



VIRAL HAPLOTYPES

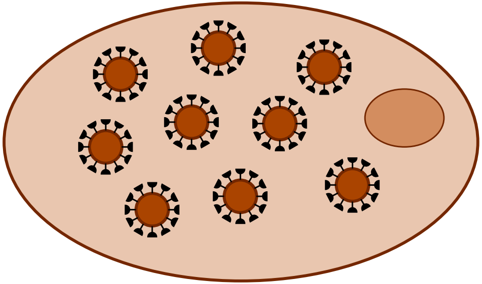
- One species = one genome?
- RNA viruses: error-prone replication



FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

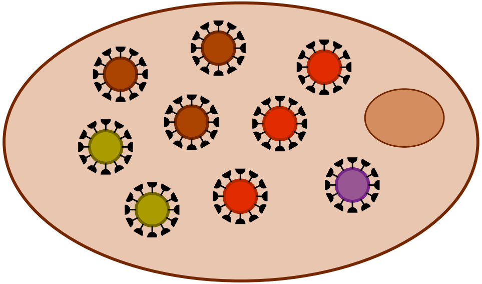
VIRAL HAPLOTYPES

- One species = one genome?
- RNA viruses: error-prone replication
- Mutation, recombination, segment reassortment



VIRAL HAPLOTYPES

- One species = one genome?
- RNA viruses: error-prone replication
- Mutation, recombination, segment reassortment
- Diverse spectrum of genomes
  ⇒ Quasispecies reconstruction



Nanopore sequencing

- ONT MinION

NANOPORE SEQUENCING

- ONT MinION

nanoporetech.com/sites/default/files/s3/minion-cutout.png

Nanopore sequencing

- ONT MinION
- 10-20 Gb per flow cell

`nanoporetech.com/sites/default/files/s3/minion-cutout.png`

Nanopore sequencing

- ONT MinION
- 10-20 Gb per flow cell
- Very long reads possible – up to 1 Mb

nanoporetech.com/sites/default/files/s3/minion-cutout.png

NANOPORE SEQUENCING

- ONT MinION
- 10-20 Gb per flow cell
- Very long reads possible – up to 1 Mb
- Noisy – 15% indels

nanoporetech.com/sites/default/files/s3/minion-cutout.png

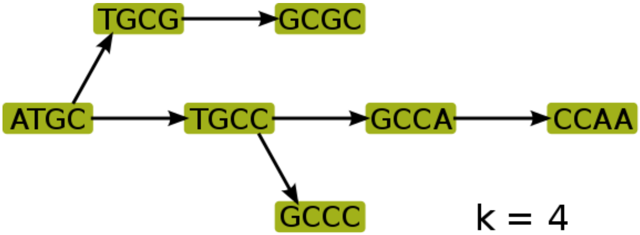
Nanopore sequencing

- ONT MinION
- 10-20 Gb per flow cell
- Very long reads possible – up to 1 Mb
- Noisy – 15% indels
- Direct RNA sequencing protocol kit

nanoporetech.com/sites/default/files/s3/minion-cutout.png

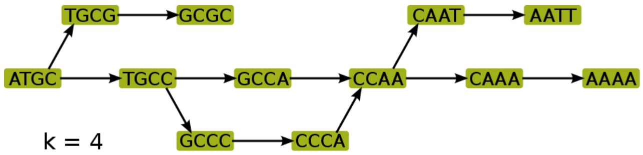
De Bruijn graph

- Constructed from overlapping k-mers



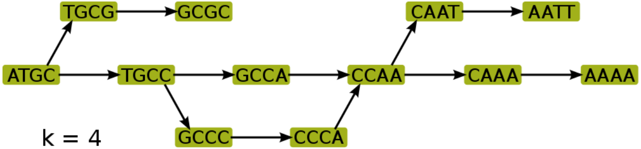
De Bruijn graph

- Constructed from overlapping k-mers
- Captures variants



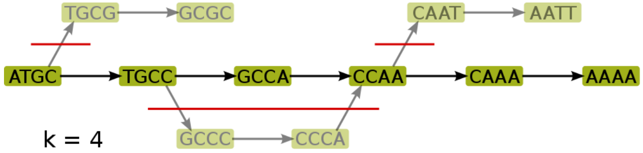
De Bruijn graph

- Constructed from overlapping k-mers
- Captures variants
- Assembly: consensus



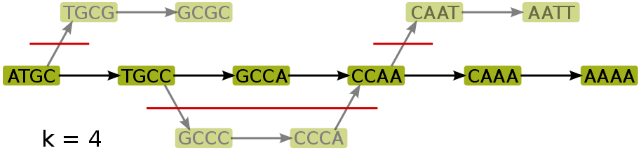
DE BRUIJN GRAPH

- Constructed from overlapping k-mers
- Captures variants
- Assembly: consensus
- Tip- and bulge removal



De Bruijn graph

- Constructed from overlapping k-mers
- Captures variants
- Assembly: consensus
- Tip- and bulge removal
- Collapse unambiguous chains

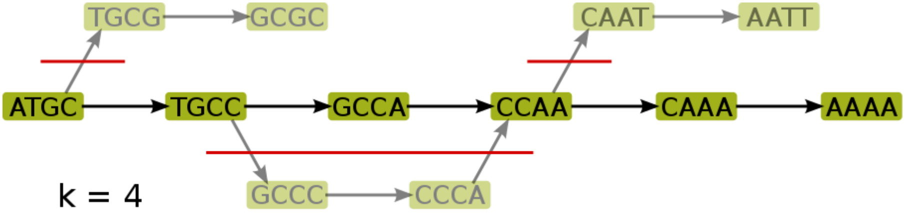


ASSEMBLY BY DE BRUIJN GRAPH

- Established de novo assembly method (Velvet, SPAdes)

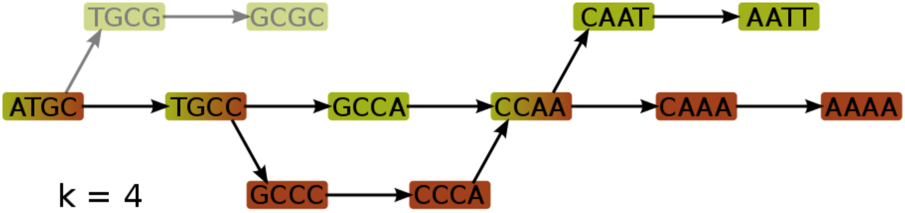
ASSEMBLY BY DE BRUIJN GRAPH

- Established de novo assembly method (Velvet, SPAdes)
- Goal: Enhance with focus on quasispecies reconstruction



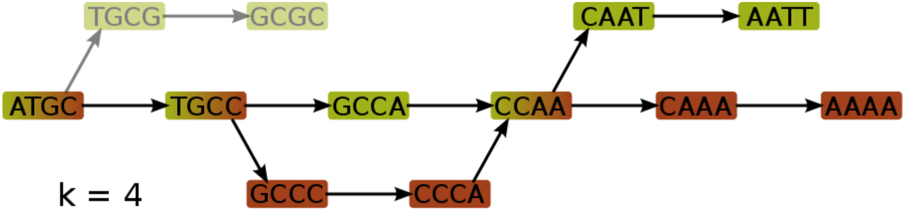
Assembly by de Bruijn graph

- Established de novo assembly method (Velvet, SPAdes)
- Goal: Enhance with focus on quasispecies reconstruction
- Separate haplotypes by graph manipulation, long read information

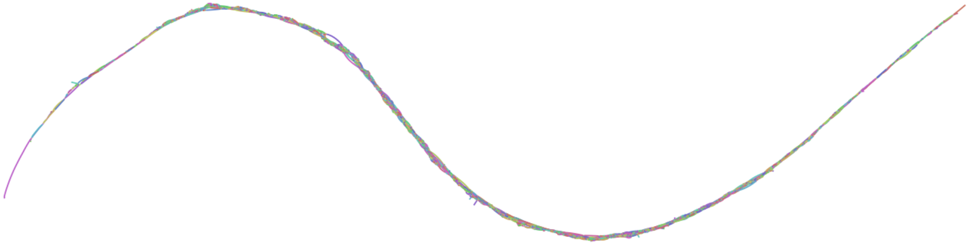


ASSEMBLY BY DE BRUIJN GRAPH

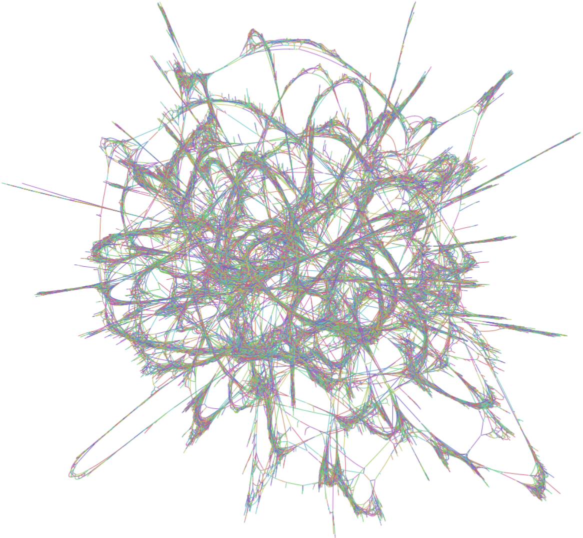
- Established de novo assembly method (Velvet, SPAdes)
- Goal: Enhance with focus on quasispecies reconstruction
- Separate haplotypes by graph manipulation, long read information
- Assemble haplotype consensus sequences



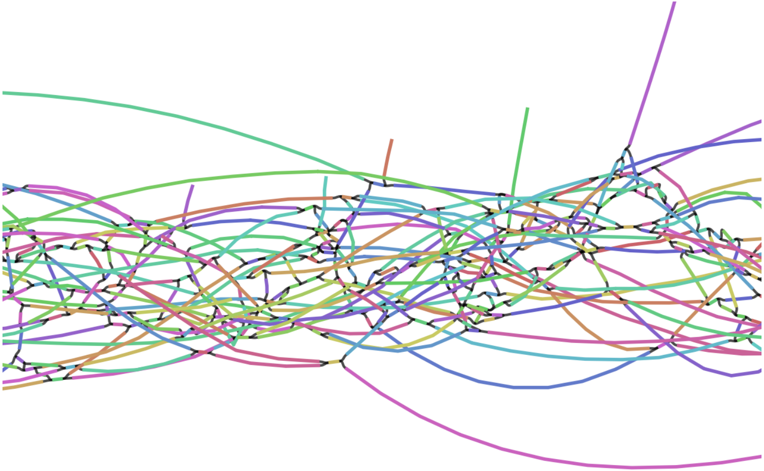
Results so far



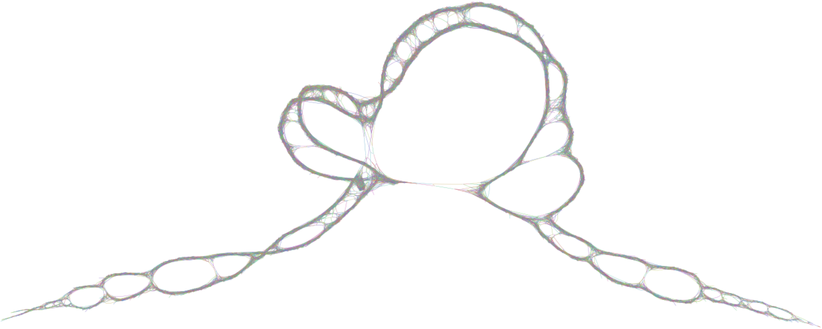
corona simul - 1 genome - 20 reads - 10000 nt - 10 % indels - k=25



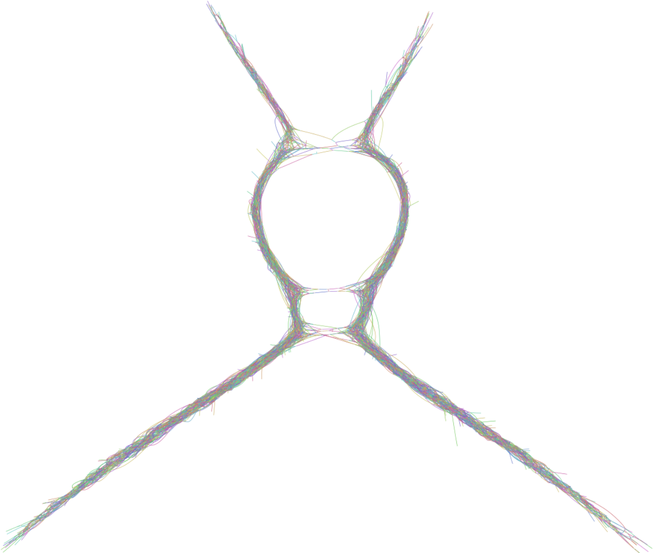
corona simul - 1 genome - 100 reads - 10000 nt - 10 % indels - **k=16**



corona - 1 genome - 40 reads - 10000 nt - **15 % indels** - k=30



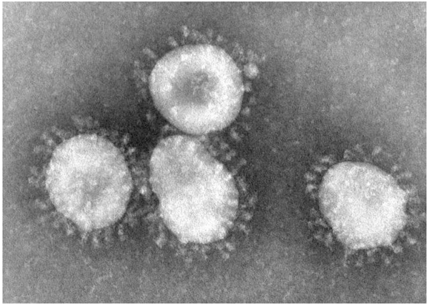
bvdv simul - **2 genomes** - 100/100 reads - 6000 nt - 10 % indels - k=20



bvdv simul - **2 genomes** - 100/100 reads - 6000 nt - 10 % indels - **k=30**

Real coronavirus read data

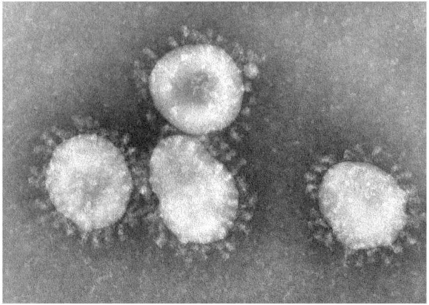
- HCoV 229E in human cell culture



en.wikipedia.org/wiki/Coronavirus#/media/File:Coronaviruses_004_lores.jpg

REAL CORONAVIRUS READ DATA

- HCoV 229E in human cell culture
- Direct RNA protocol kit on MinION

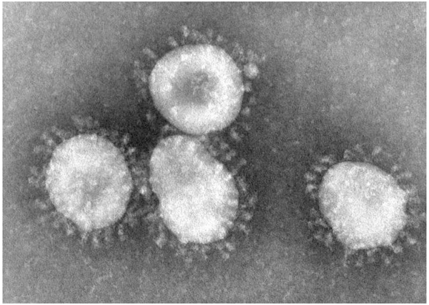


en.wikipedia.org/wiki/Coronavirus#/media/

File:Coronaviruses_004_lores.jpg

REAL CORONAVIRUS READ DATA

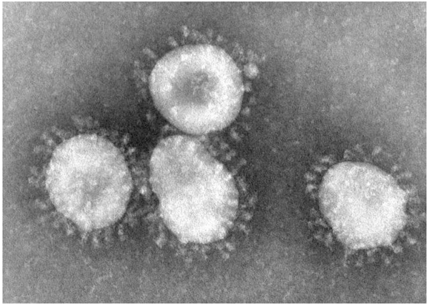
- HCoV 229E in human cell culture
- Direct RNA protocol kit on MinION
- 293406 reads, 27 % virus, rest human



en.wikipedia.org/wiki/Coronavirus#/media/File:Coronaviruses_004_lores.jpg

Real coronavirus read data

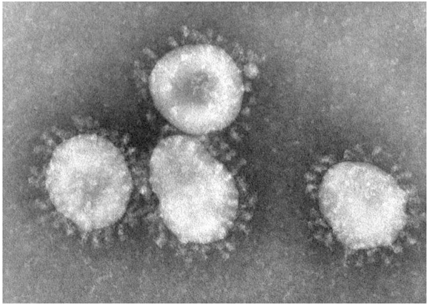
- HCoV 229E in human cell culture
- Direct RNA protocol kit on MinION
- 293406 reads, 27 % virus, rest human
- Median read length 2.5 kb



en.wikipedia.org/wiki/Coronavirus#/media/File:Coronaviruses_004_lores.jpg

Real coronavirus read data

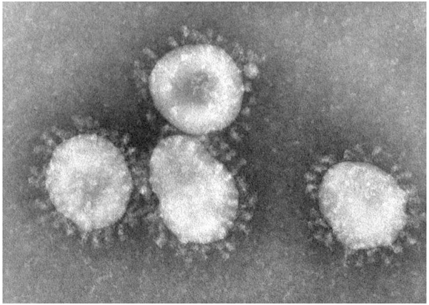
- HCoV 229E in human cell culture
- Direct RNA protocol kit on MinION
- 293406 reads, 27 % virus, rest human
- Median read length 2.5 kb
- Longest read: 26 kb (genome 27.3 kb)



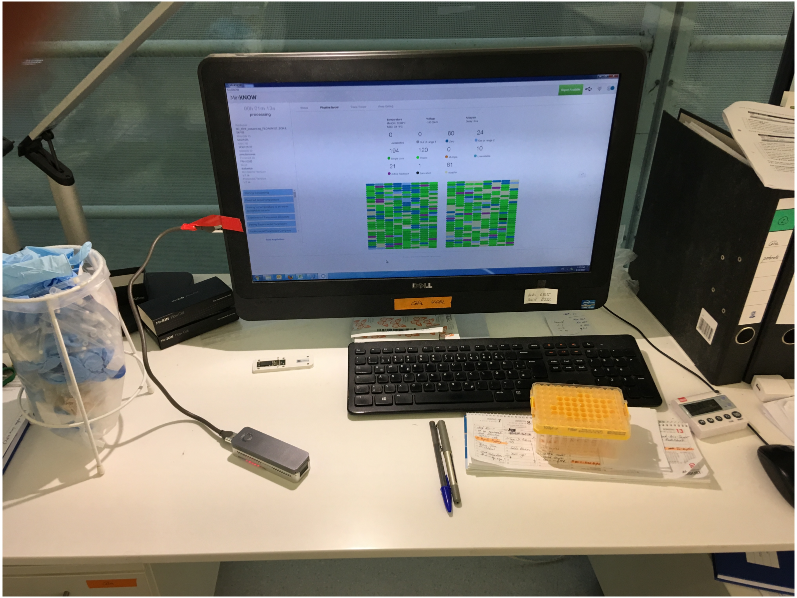
en.wikipedia.org/wiki/Coronavirus#/media/File:Coronaviruses_004_lores.jpg

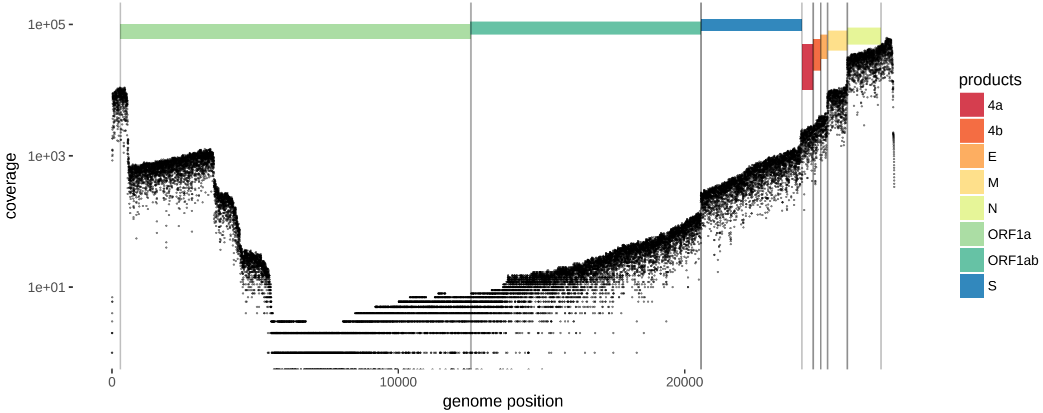
REAL CORONAVIRUS READ DATA

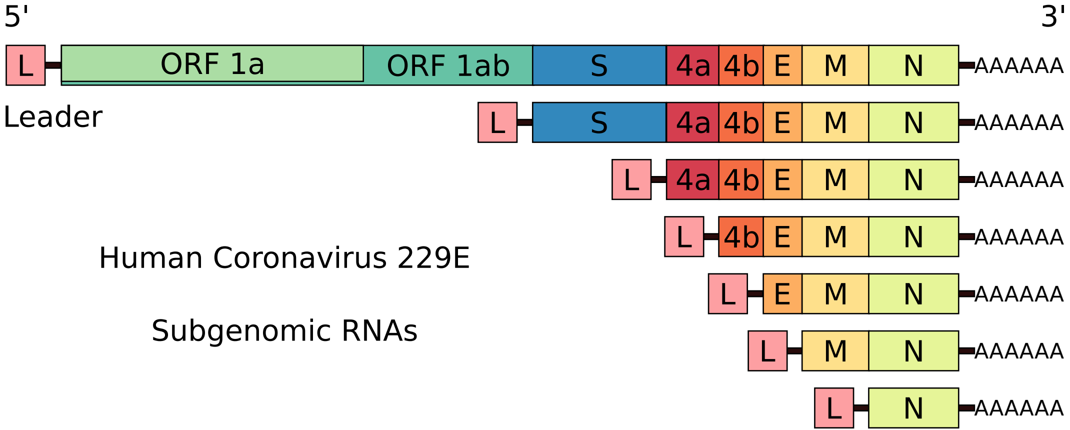
- HCoV 229E in human cell culture
- Direct RNA protocol kit on MinION
- 293406 reads, 27 % virus, rest human
- Median read length 2.5 kb
- Longest read: 26 kb (genome 27.3 kb)
- Error rate 15 % - mainly indels



en.wikipedia.org/wiki/Coronavirus#/media/ File:Coronaviruses_004_lores.jpg

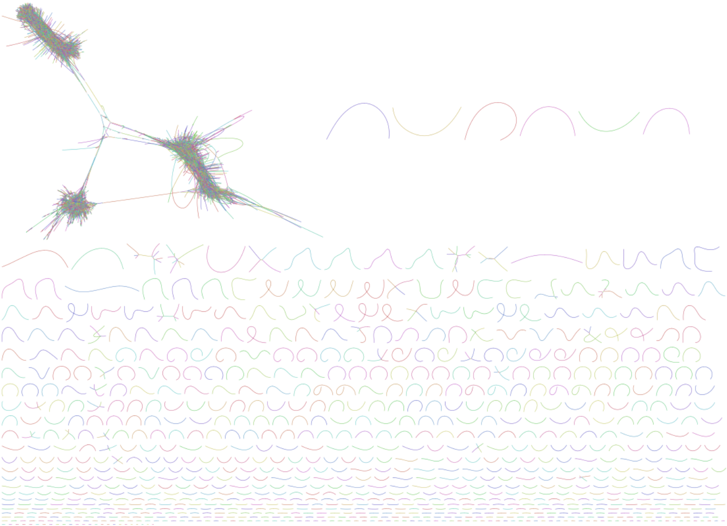

MinKNOW
processing
MinION Flow Cell
MinION Flow Cell
DELL


1e+05
1e+03
1e+01
coverage
0
10000
20000
genome position
products
4a
4b
E
M
N
ORF1a
ORF1ab
S



Human Coronavirus 229E

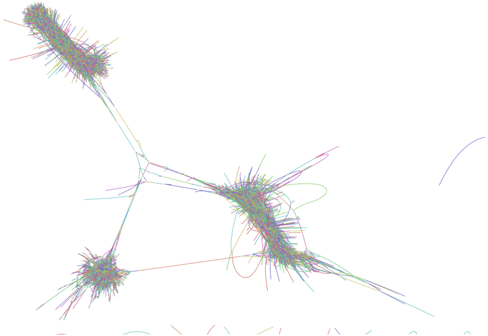
Subgenomic RNAs



corona - sequenced data - **2000 reads** - **k=40**

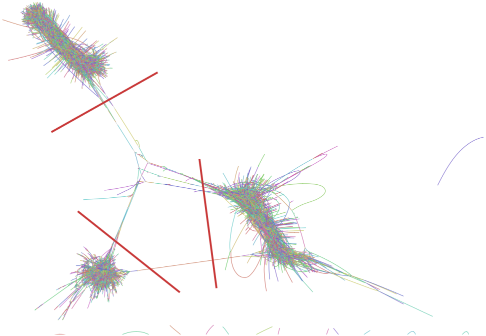
SUBGRAPH CONSENSUS

- Needed: subgraph separation



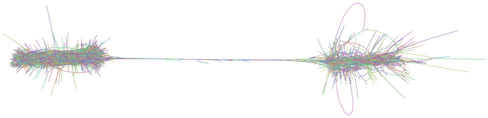
SUBGRAPH CONSENSUS

- Needed: subgraph separation



SUBGRAPH CONSENSUS

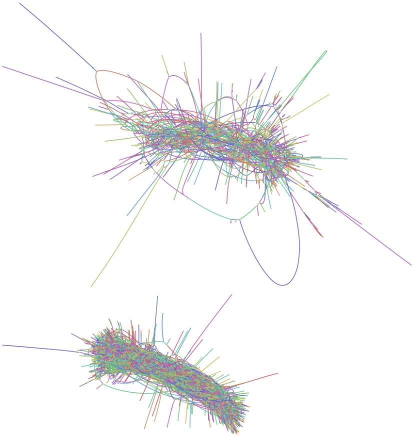
- Needed: subgraph separation
- Implemented with min-cut



FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

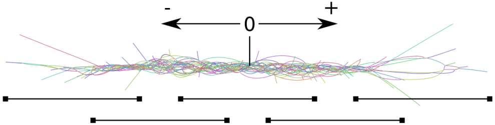
Subgraph consensus

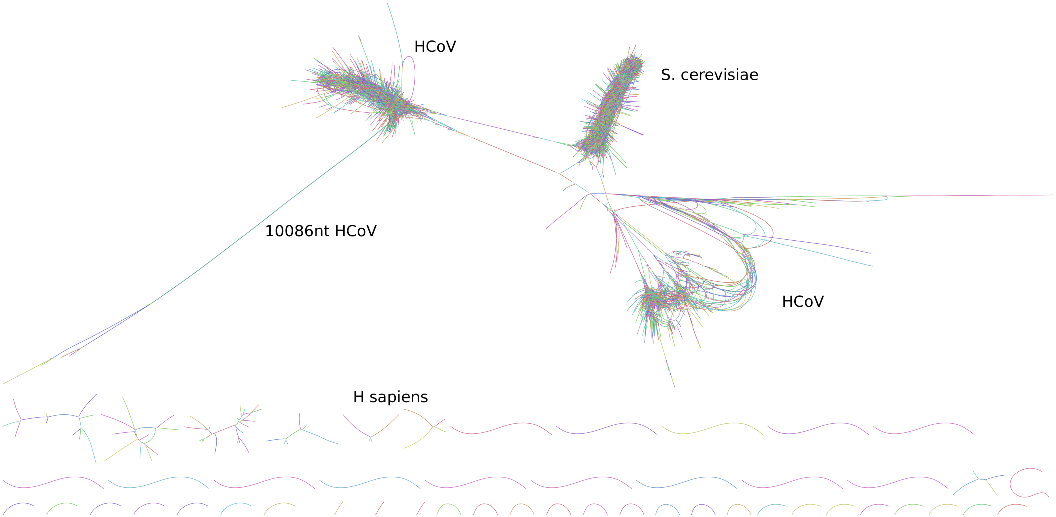
- Needed: subgraph separation
- Implemented with min-cut
- Separates clusters that are minimally connected



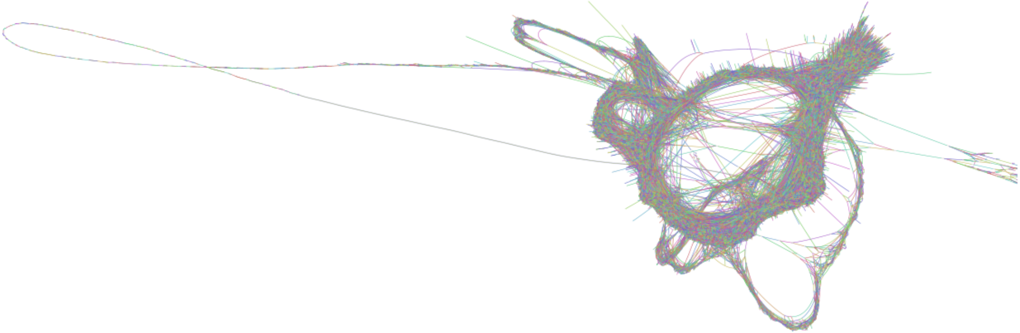
Subgraph consensus

- Needed: subgraph separation
- Implemented with min-cut
- Separates clusters that are minimally connected
- Subgraph consensus is implemented





Yeast enolase is included in the direct RNA kit as a positive control



corona - sequenced data - **73533 reads** - **k=40**

SEQUENCING ERRORS

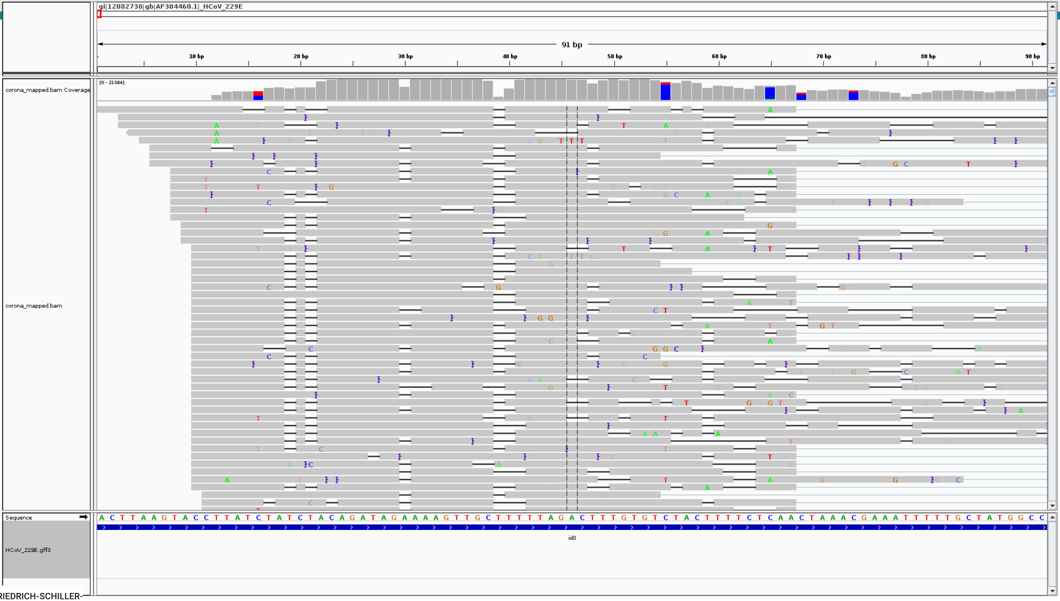
- Mostly insertions and deletions

SEQUENCING ERRORS

- Mostly insertions and deletions
- Of those: 70-80% deletions

Sequencing Errors

- Mostly insertions and deletions
- Of those: 70-80% deletions
- Deletions happen systematically at homopolymers



gi|12082738|gb|AF304460.1|_HCoV_229E

corona_mapped.bam Coverage

[0 - 21084]

corona_mapped.bam

Sequence

A C T T A A G T A C C T T A T C T A T C T A C A G A T A G A A A A G T T G C T T T T T A G A C T T T G T G T C T A C T T T T C T C A A C T A A A C G A A A T T T T T G C T A T G G C C

HCoV_229E.gff3

Long Read Error Correction

- Self-correction:
  Systematic errors are problematic

Long Read Error Correction

- Self-correction:
  Systematic errors are problematic
- Hybrid correction with i.e. Illumina data:
  Alignment to noisy long reads difficult

Long Read Error Correction

- Self-correction:
  Systematic errors are problematic
- Hybrid correction with i.e. Illumina data:
  Alignment to noisy long reads difficult
- `HG-CoLoR` by P. Morisse et al.

Long Read Error Correction

- Self-correction:
  Systematic errors are problematic
- Hybrid correction with i.e. Illumina data:
  Alignment to noisy long reads difficult
- `HG-CoLoR` by P. Morisse et al.
- Longest read (25932 nt)
  Identity to reference: 84% → 99%
  Gap of 407 nt, 90 min runtime

Reference-based Indel Correction

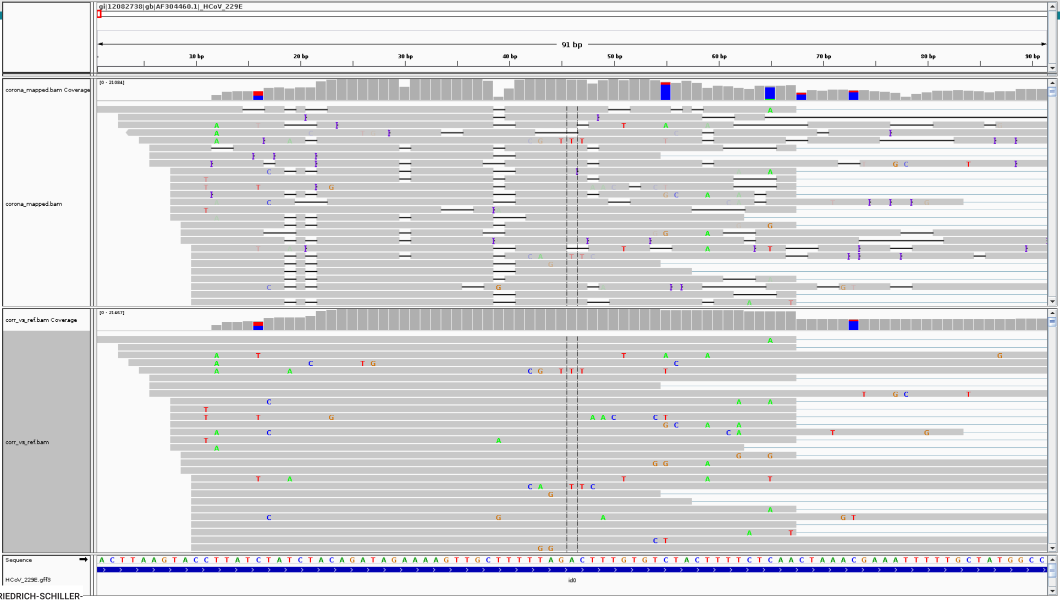
- Reference from `nanopolish` by J. Simpson – RNA NYI

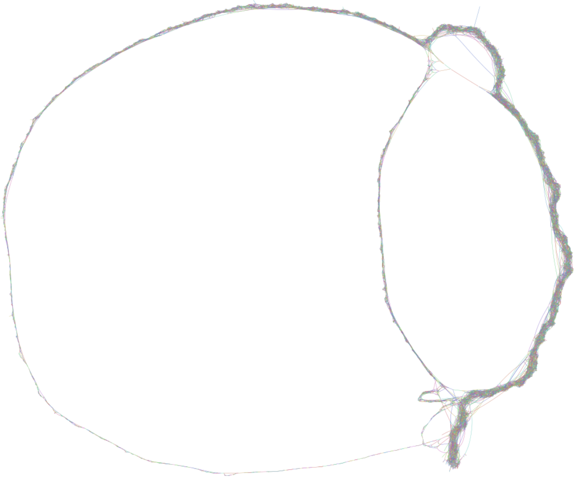
Reference-based Indel Correction

- Reference from `nanopolish` by J. Simpson – RNA NYI
- Align long reads to reference

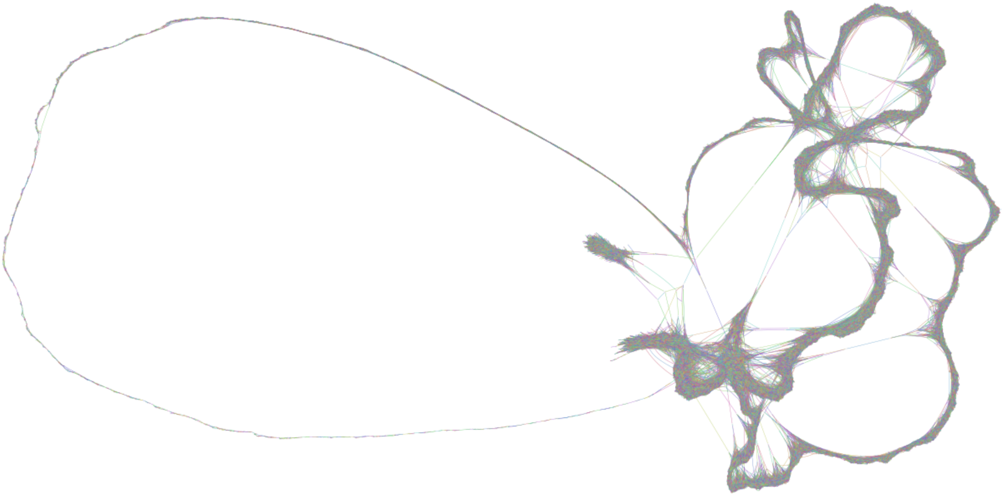
Reference-based Indel Correction

- Reference from `nanopolish` by J. Simpson – RNA NYI
- Align long reads to reference
- Parse CIGAR string to remove insertions, fill deletions

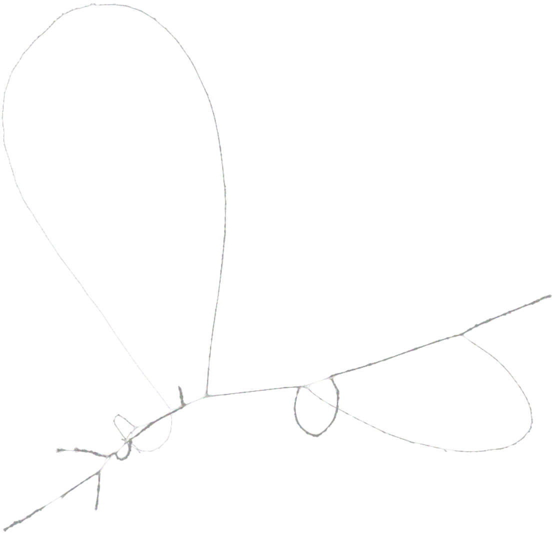

gi|12082738|gb|AF304460.1|_HCoV_229E
91 bp
10 bp
20 bp
30 bp
40 bp
50 bp
60 bp
70 bp
80 bp
90 bp
corona_mapped.bam Coverage
[0 - 21084]
corona_mapped.bam
corr_vs_ref.bam Coverage
[0 - 21467]
corr_vs_ref.bam
Sequence
A C T T A A G T A C C T T A T C T A T C T A C A G A T A G A A A A G T T G C T T T T T A G A C T T T G T G T C T A C T T T T C T C A A C T A A A C G A A A T T T T T G C T A T G G C C
HCoV_229E.gff3
id0



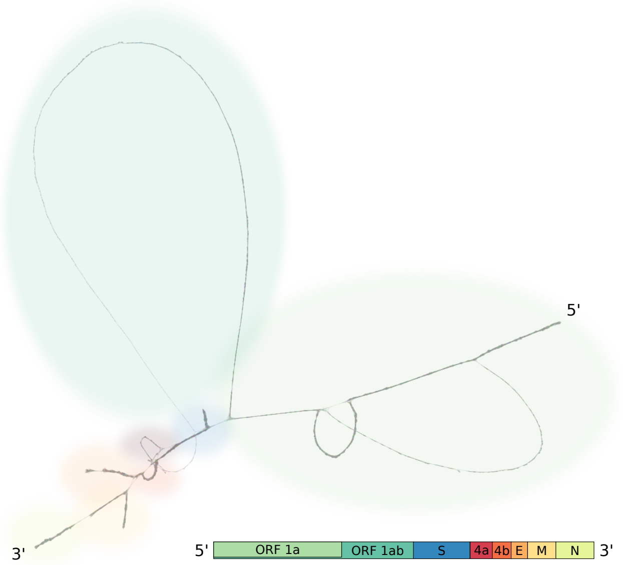
corona - indels corrected - **1% best nucleotides** - **k=40**



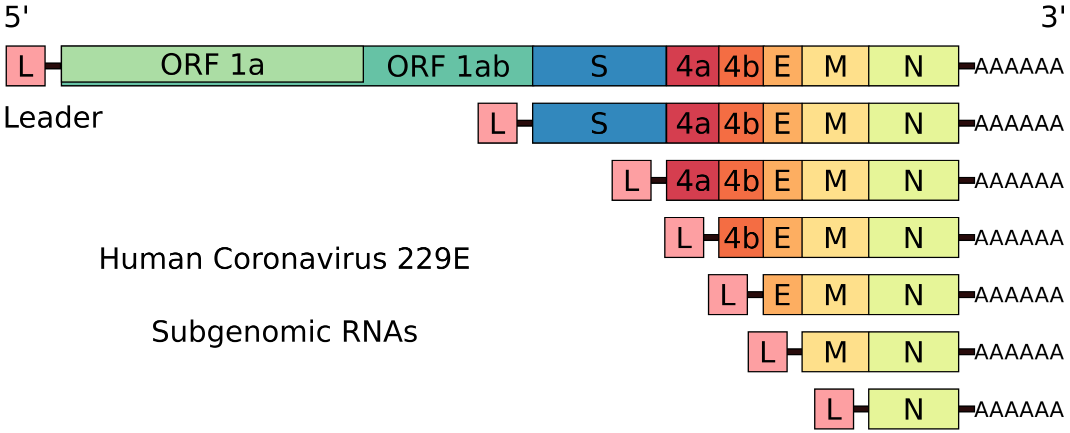
corona - indels corrected - **10% best nucleotides** - **k=30**



corona - indels corrected - **1% best nucleotides** - **k=20**


5'
3'
5'
ORF 1a
ORF 1ab
S
4a
4b
E
M
N
3'

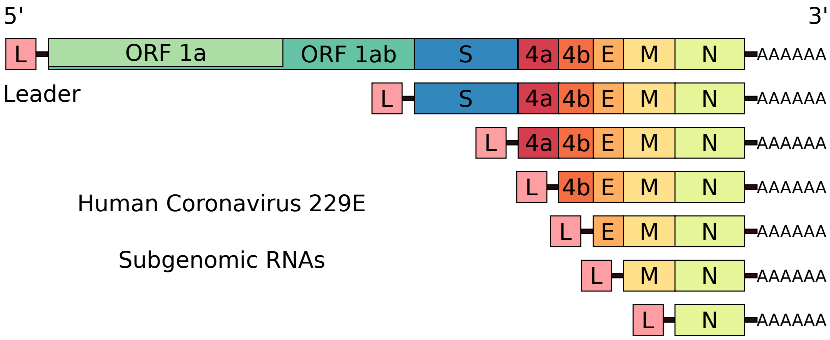
SUBGENOMIC TYPES



Human Coronavirus 229E

Subgenomic RNAs

ANNOTATION BASED CLASSIFICATION



assigned:

2

141

1077

85

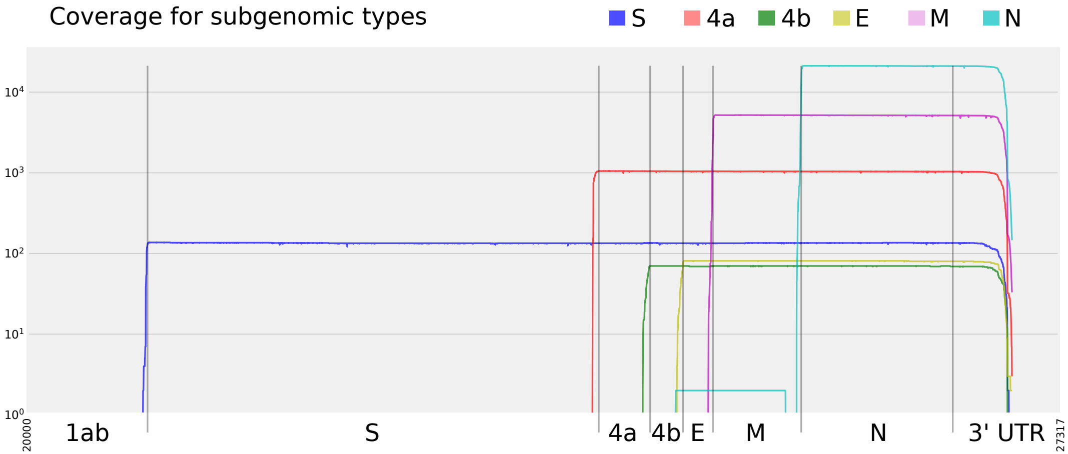
105

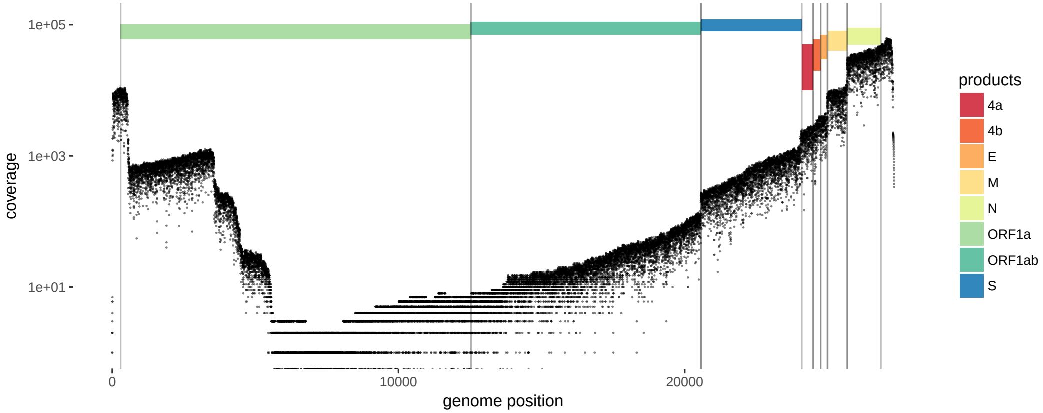
5407

21896

Total: 28711

Unassigned: 44822

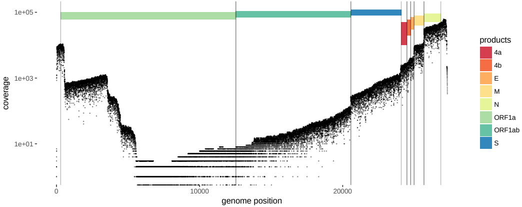

Coverage for subgenomic types
S
4a
4b
E
M
N
10^4
10^3
10^2
10^1
10^0
20000
1ab
S
4a
4b
E
M
N
3' UTR
27317



FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

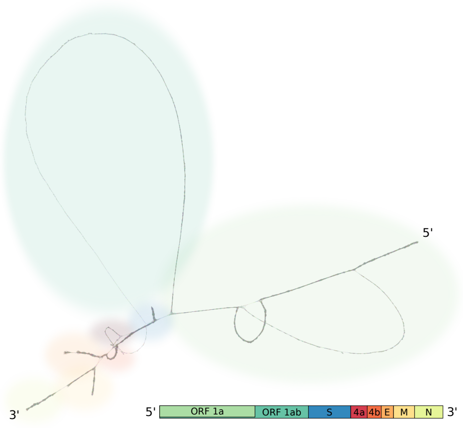
Conclusions

- Viral full genome sequencing



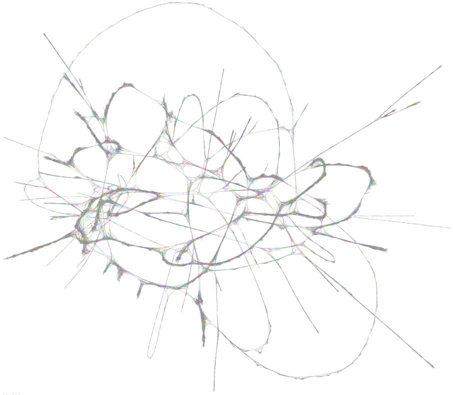
Conclusions

- Viral full genome sequencing
- Structure is visible in graph



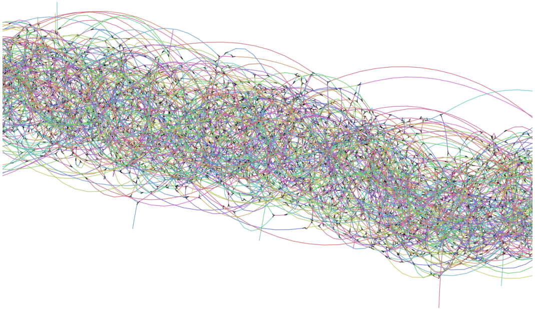
CONCLUSIONS

- Viral full genome sequencing
- Structure is visible in graph
- Importance of k



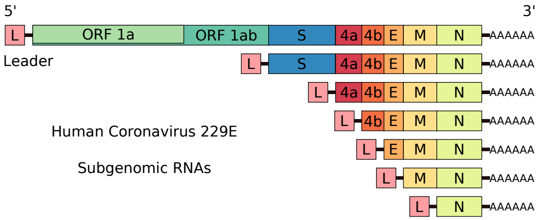
Conclusions

- Viral full genome sequencing
- Structure is visible in graph
- Importance of k
- Indel correction required



Conclusions

- Viral full genome sequencing
- Structure is visible in graph
- Importance of k
- Indel correction required
- Coronavirus is ...complicated



Outlook

NEXT STEPS

- Test on new data (more HCoV, plum pox virus)

NEXT STEPS

- Test on new data (more HCoV, plum pox virus)
- Improve error correction

NEXT STEPS

- Test on new data (more HCoV, plum pox virus)
- Improve error correction
- Find robust way to extract haplotypes

NEXT STEPS

- Test on new data (more HCoV, plum pox virus)
- Improve error correction
- Find robust way to extract haplotypes
- Utilize long read information

Andreas Goral

Adrian Viehweger
Celia Diezel

Manja Marz

Ramakanth Madhugiri
John Ziebuhr

All of my group!

Thank you!