FACULTY OF VETERINARY MEDICINE
accredited by EAEVE

DEPARTMENT VIROLOGY, PARASITOLOGY, AND IMMUNOLOGY
RESEARCH GROUP OF VIROLOGY

Linking Nanopore sequencing & High-Performance Computing

HPC UGent User Meeting

Nick Vereecke

28 June 2021

GHENT UNIVERSITY

PathoSense

Presentation Outline

- Nanopore sequencing & Applications
- SARS-CoV-2 Sequencing
- Bacterial Whole Genome Sequencing

Nanopore sequencing & Applications

Nanopore Sequencing

	Oxford NANOPORE Technologies	illumina®
Approach	Single Molecule	Sequencing by Synthesis
PCR-dependent	No, *but possible*	Yes
Read length	Up to Mbps	150-300 bp (x2)
Read Quality	Q20[2021]	Q30
Throughput	Real-Time	Days > Months
Instrument cost	$	$$$

Nanopore Sequencing

	Oxford NANOPORE Technologies	illumina®
Approach	Single Molecule	Sequencing by Synthesis
PCR-dependent	No, *but possible*	Yes
Read length	Up to Mbps	150-300 bp (x2)
Read Quality	Q20[2021]	Q30
Throughput	Real-Time	Days > Months
Instrument cost	$	$$$

Nanopore Sequencing

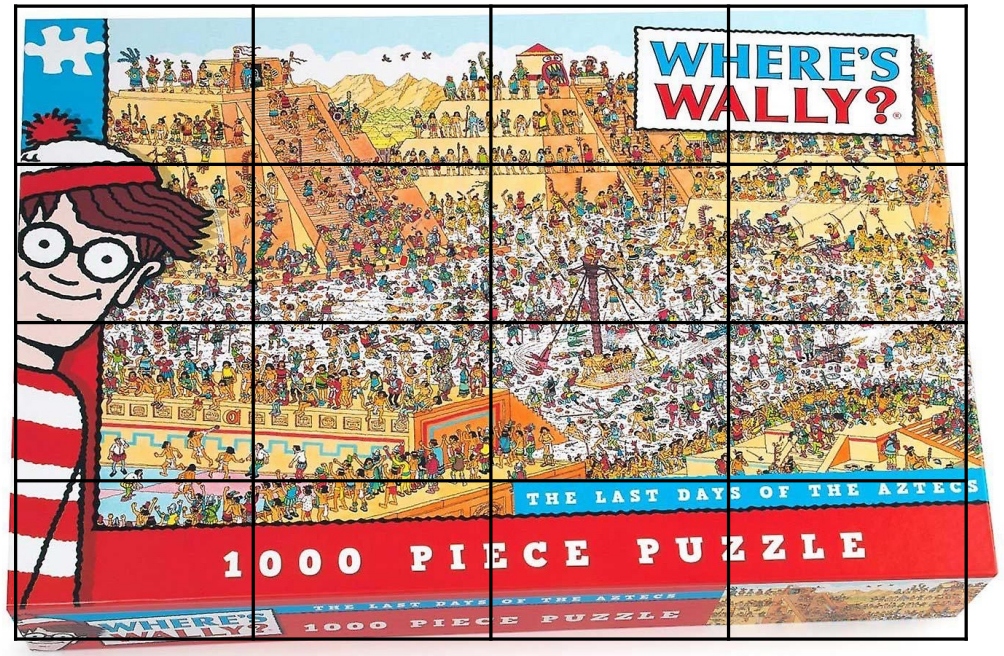
- Third-Generation Sequencing
- Single Molecule label-free sequencing
- Versatile

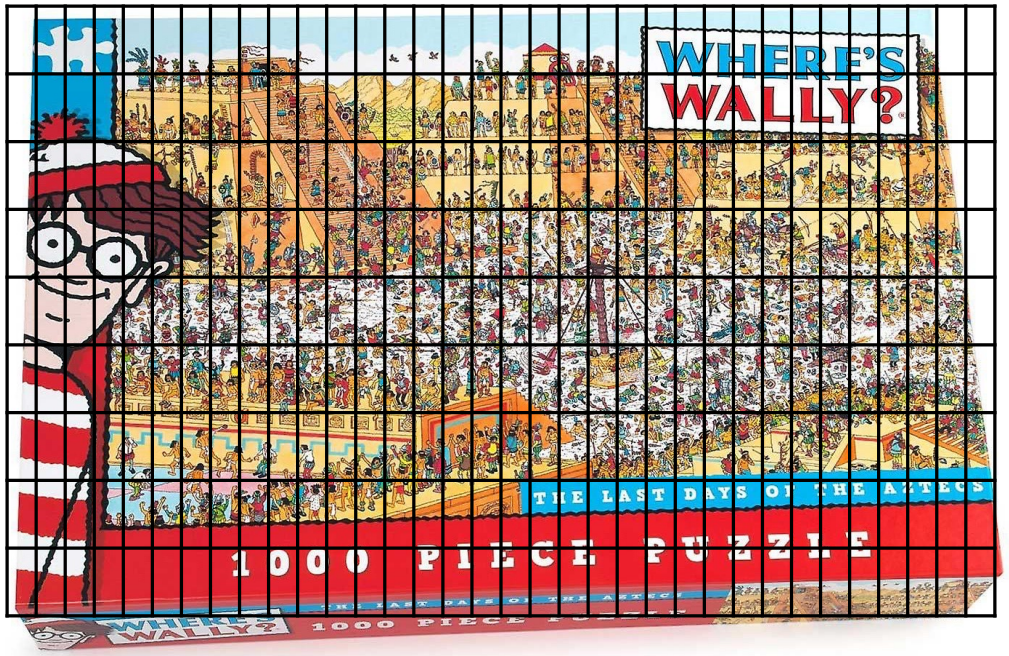
Oxford NANOPORE Technologies

Nanopore Sequencing

	Oxford NANOPORE Technologies	illumina®
Approach	Single Molecule	Sequencing by Synthesis
PCR-dependent	No, but possible	Yes
Read length	Up to Mbps	150-300 bp (x2)
Read Quality	Q20[2021]	Q30
Throughput	Real-Time	Days > Months
Instrument cost	$	$$$
Versatility	High	Medium
In-field	Yes	No

Nanopore Sequencing





Oxford NANOPORE Technologies

PACIFIC BIOSCIENCES

illumina

iontorrent by Thermo Fisher Scientific

MGI 华大智造

Dr. S. Theuns; 2019

Nanopore Sequencing

	Oxford NANOPORE Technologies	illumina®
Approach	Single Molecule	Sequencing by Synthesis
PCR-dependent	No, but possible	Yes
Read length	Up to Mbps	150-300 bp (x2)
Read Quality	Q20[2021]	Q30
Throughput	Real-Time	Days > Months
Instrument cost	$	$$$
Versatility	High	Medium
In-field	Yes	No

Nanopore Sequencing

Current raw read QC = 98.3%

= 7 mistakes in 400 bp amplicon
= 85 mistakes in 5,000 bp reads

Raw-read accuracy	99.3 %, > Q20
Duplex	99.8%, ~ Q29
SNP detection (F1 scores human)	SNV: 99.9 % [1] Indel: 98.5 % [2]
Assembly+ (Human)	80 Mbase N50 [3] Q 47
Assembly (Bacterial)	Circular > Q 50
SV detection (F1 scores human)	96%
Methylation included	6mA, 5mC, 5hmC

Quality score	Error probability	Accuracy
10	0.1 (1 in 10)	90%
20	0.01 (1 in 100)	99%
30	0.001 (1 in 1000)	99.9%
40	0.0001 (1 in 10.000)	99.99%

Nanopore Sequencing

	Oxford NANOPORE Technologies	illumina®
Approach	Single Molecule	Sequencing by Synthesis
PCR-dependent	No, but possible	Yes
Read length	Up to Mbps	150-300 bp (x2)
Read Quality	Q20[2021]	Q30
Throughput	Real-Time	Days > Months
Instrument cost	$	$$$
Versatility	High	Medium
In-field	Yes	No

- 5 M Flow Cells
- 512 channels/Flow cell
- Up to 250 GB data (72h)
- $$$

GridION

- 1 F Flow Cell
- 126 channels/Flow cell
- Up to 2.8 GB data (16h)
- $

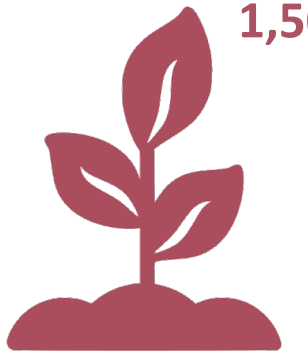
Flongle

PromethION

- 24 or 48 P Flow cells
- 3000 channels/Flow cell
- Up to 14 TB data (72h)
- $$$$

MinION

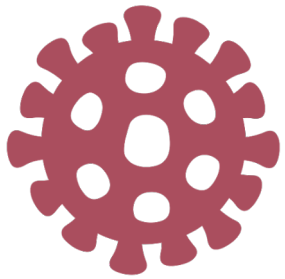
- 1 M Flow Cell
- 512 channels/Flow cell
- Up to 50 GB data (72h)
- $$



1,500X *A. thaliana*

- 5 M Flow Cells
- 512 channels/Flow cell
- Up to 250 GB data (72h)
- $$$

GridION



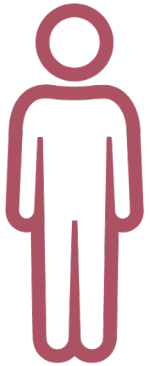
90,000x SARS-CoV-2

- 1 F Flow Cell
- 126 channels/Flow cell
- Up to 2.8 GB data (16h)
- $

Flongle

PromethION

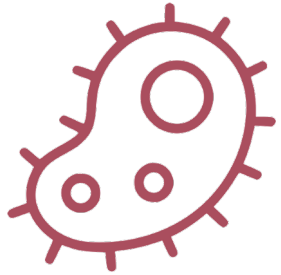
- 24 or 48 P Flow cells
- 3000 channels/Flow cell
- Up to 14 TB data (72h)
- $$$$



2,000X *Homo sapiens*

MinION

- 1 M Flow Cell
- 512 channels/Flow cell
- Up to 50 GB data (72h)
- $$



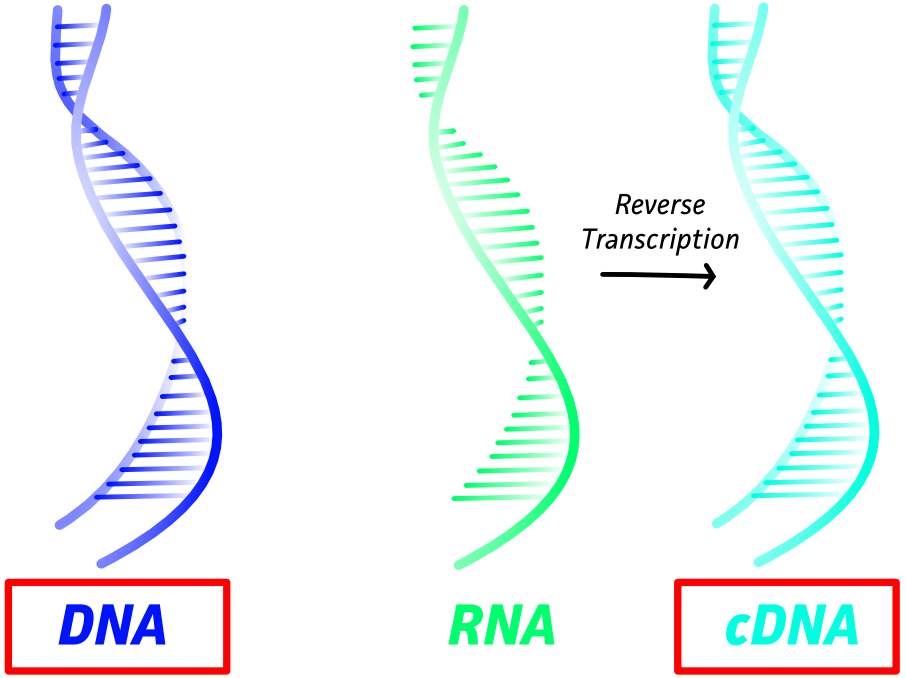
1,000X *E. coli*

Nanopore Sequencing

	Oxford NANOPORE Technologies	illumina®
Approach	Single Molecule	Sequencing by Synthesis
PCR-dependent	No, but possible	Yes
Read length	Up to Mbps	150-300 bp (x2)
Read Quality	Q20[2021]	Q30
Throughput	Real-Time	Days > Months
Instrument cost	$	$$$
Versatility	High	Medium
In-field	Yes	No

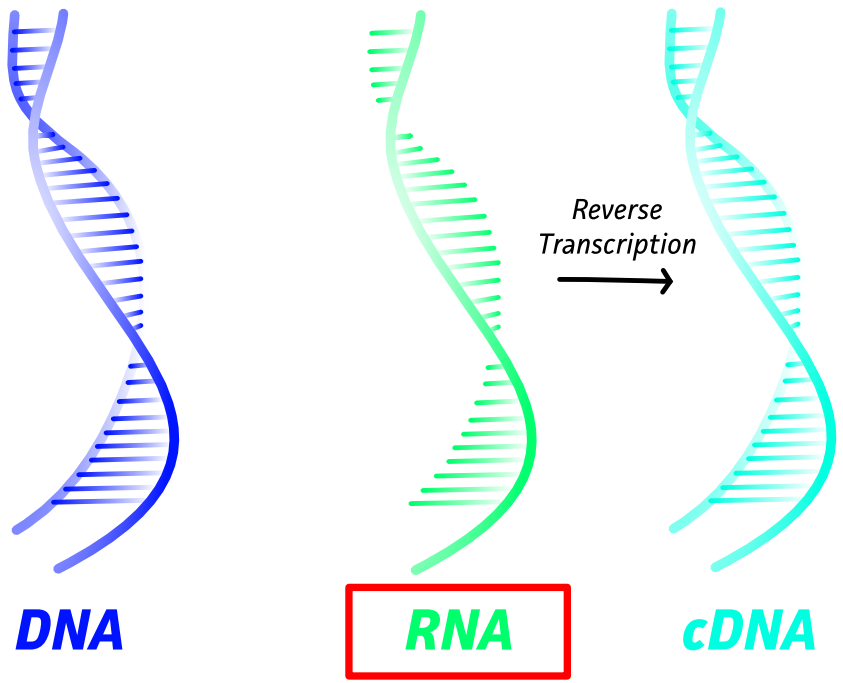
Nanopore Sequencing

- Third-Generation Sequencing
- Single Molecule label-free sequencing
- Versatile
  - *Native DNAseq*
  - *PCR-amplified DNAseq*
  - *PCR-amplified cDNAseq (from RNA)*



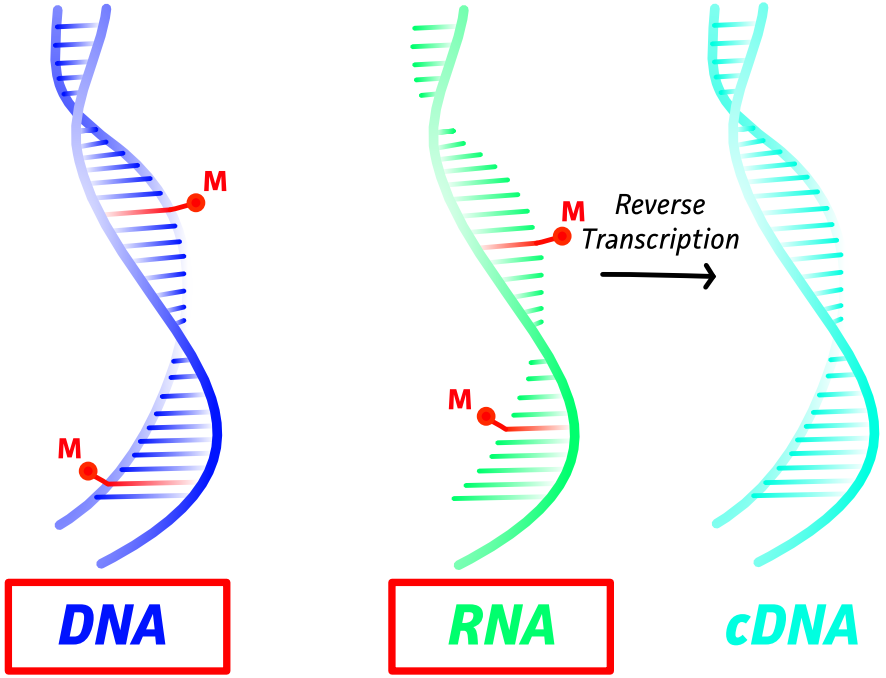
Nanopore Sequencing

- Third-Generation Sequencing
- Single Molecule label-free sequencing
- Versatile
  - *Native DNAseq*
  - *PCR-amplified DNAseq*
  - *PCR-amplified cDNAseq (from RNA)*
  - ***Native RNAseq***



Nanopore Sequencing

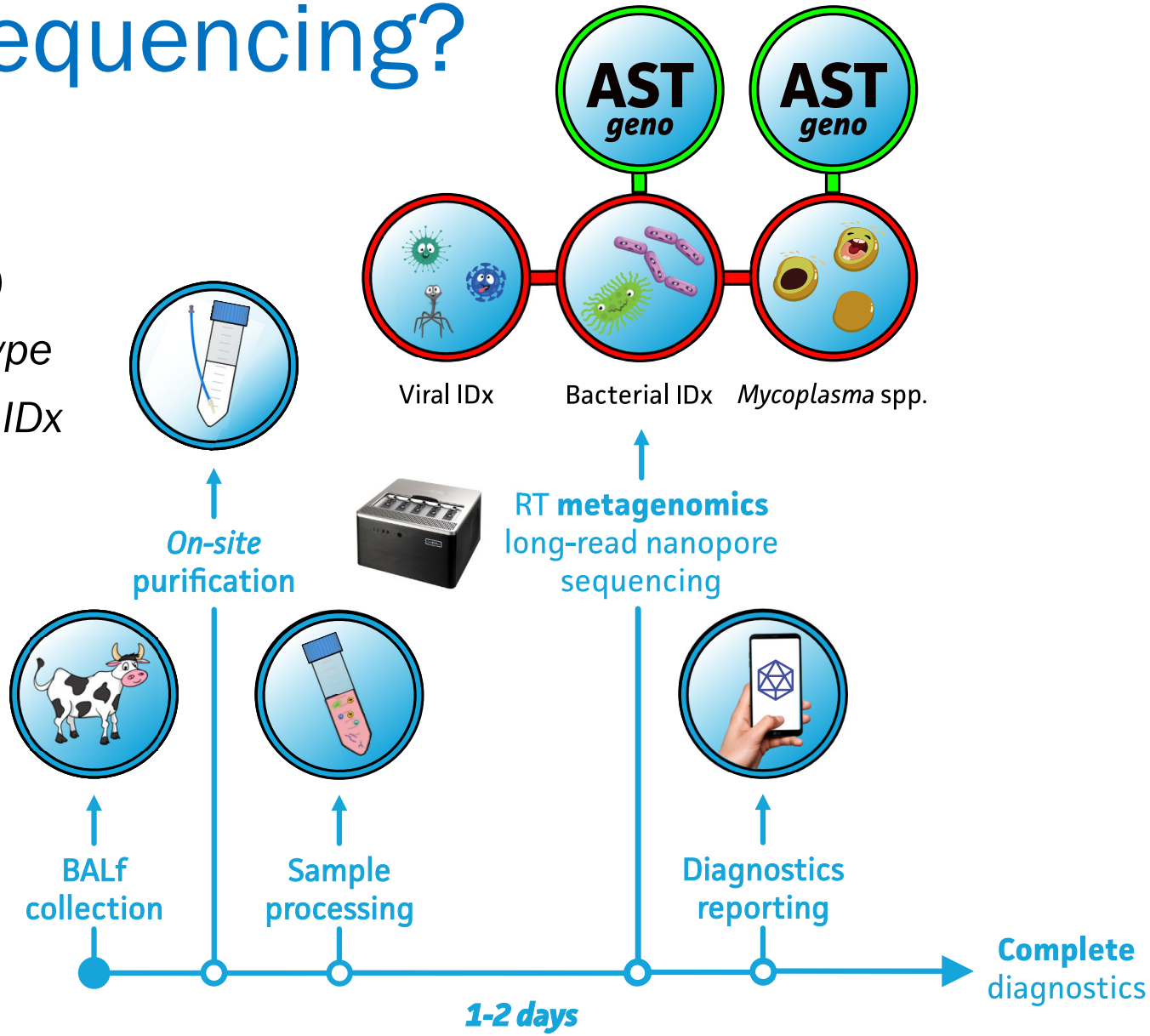
- Third-Generation Sequencing
- Single Molecule label-free sequencing
- Versatile
  - *Native DNAseq*
  - *PCR-amplified DNAseq*
  - *PCR-amplified cDNAseq (from RNA)*
  - *Native RNAseq*[New]
  - ***DNA & RNA Methylation***



Why Nanopore Sequencing?

- PathoSense BV
  - *Complete platform (1-2 days)*
  - *Any host species & sample type*
  - *CURRENT* > *Metagenomics IDx*
  - *FUTURE* > *Genetic AST*

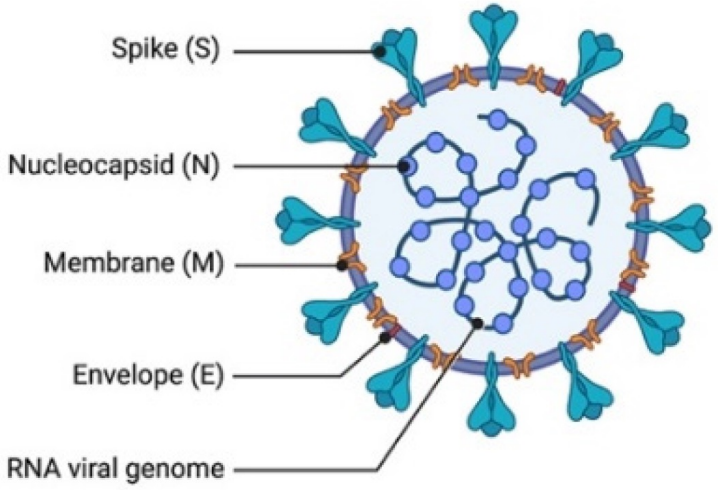
PathoSense



SARS-CoV-2 Sequencing

SARS-CoV-2

- Genome = ± **30 kbases +ssRNA**
- First complete sequence available **December 2019 (Wuhan-Hu-1)**
- Standardized protocols readily available through **ARTIC Network**
- Josh Quick, James Ferguson & Nick Loman

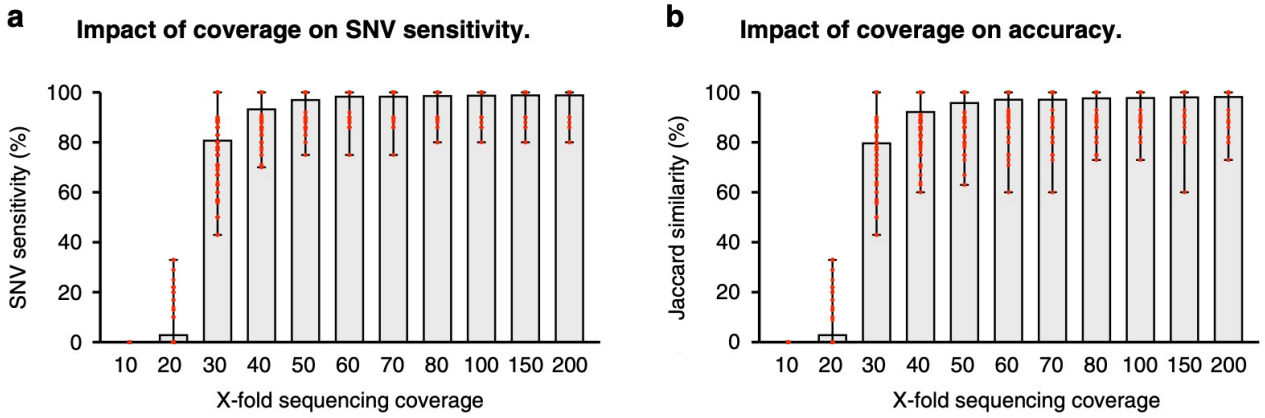


SARS-CoV 2 virion structure

Alanagreh et al., 2020 MDPI Pathogens

SARS-CoV-2

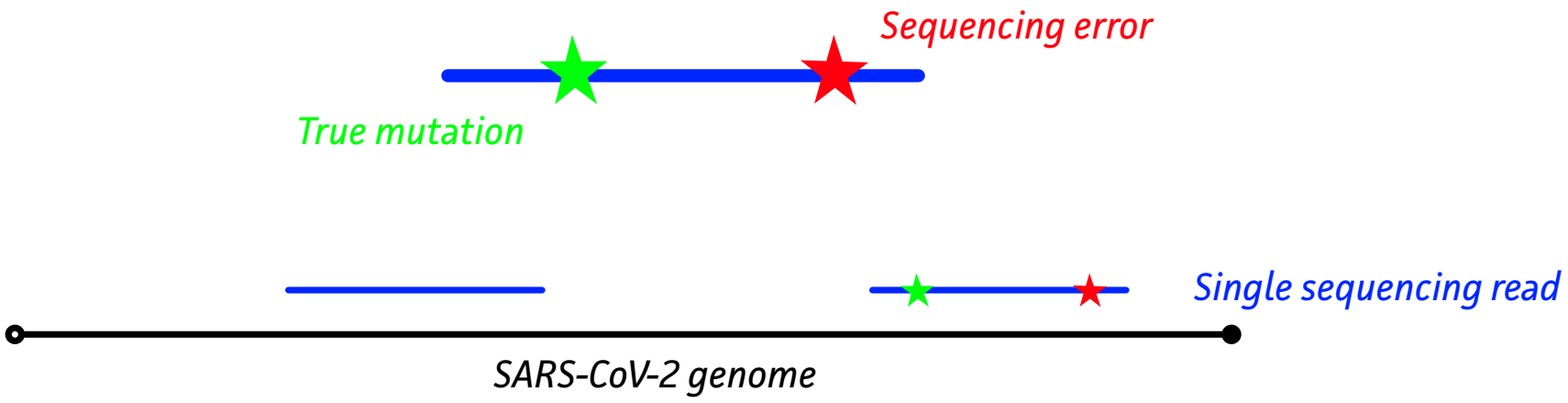
- Genome = ± **30 kbases +ssRNA**
- First complete sequence available **December 2019 (Wuhan-Hu-1)**
- Standardized protocols readily available through **ARTIC Network**
- Josh Quick, James Ferguson & Nick Loman
- **Nanopore Accuracy?**



Bull et al., 2021 Nature Com.

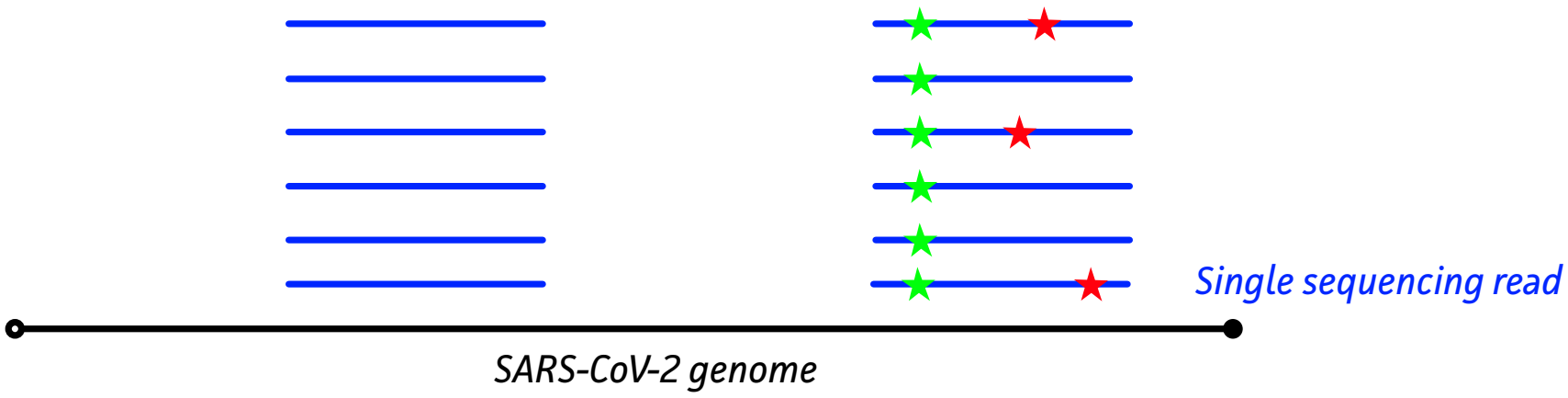
SARS-CoV-2

- Genome = ± **30 kbases +ssRNA**
- First complete sequence available **December 2019 (Wuhan-Hu-1)**
- Standardized protocols readily available through **ARTIC Network**
- Josh Quick, James Ferguson & Nick Loman
- **Nanopore Accuracy?**



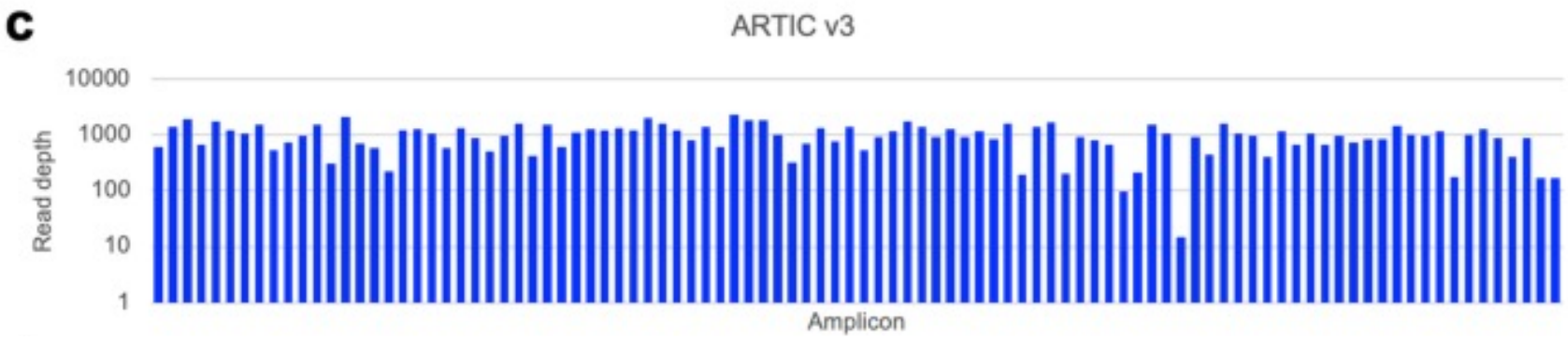
SARS-CoV-2

- Genome = ± **30 kbases +ssRNA**
- First complete sequence available **December 2019 (Wuhan-Hu-1)**
- Standardized protocols readily available through **ARTIC Network**
- Josh Quick, James Ferguson & Nick Loman
- **Nanopore Accuracy?**



SARS-CoV-2

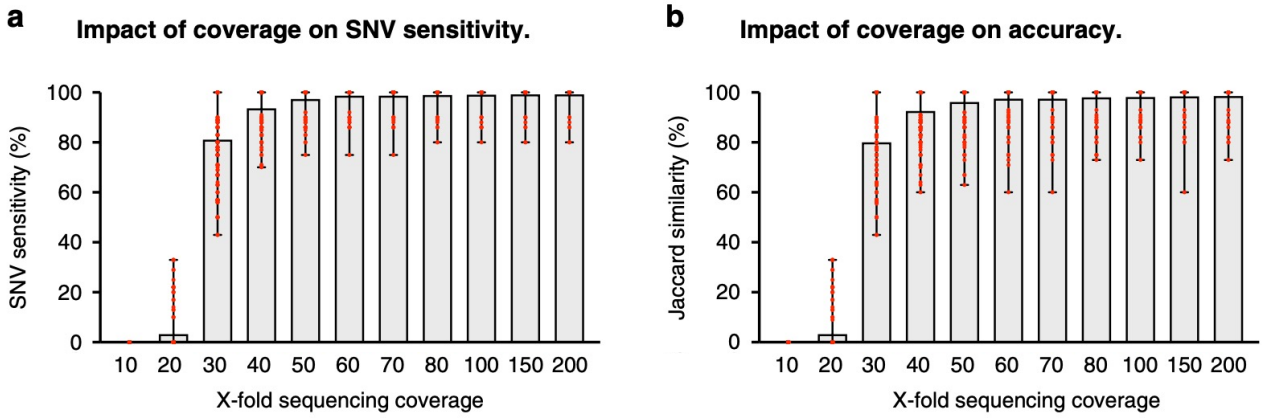
- Genome = ± **30 kbases +ssRNA**
- First complete sequence available **December 2019 (Wuhan-Hu-1)**
- Standardized protocols readily available through **ARTIC Network**
- Josh Quick, James Ferguson & Nick Loman
- **Nanopore Accuracy?**



Gohl et al., 2020 BMC Bioinformatics

SARS-CoV-2

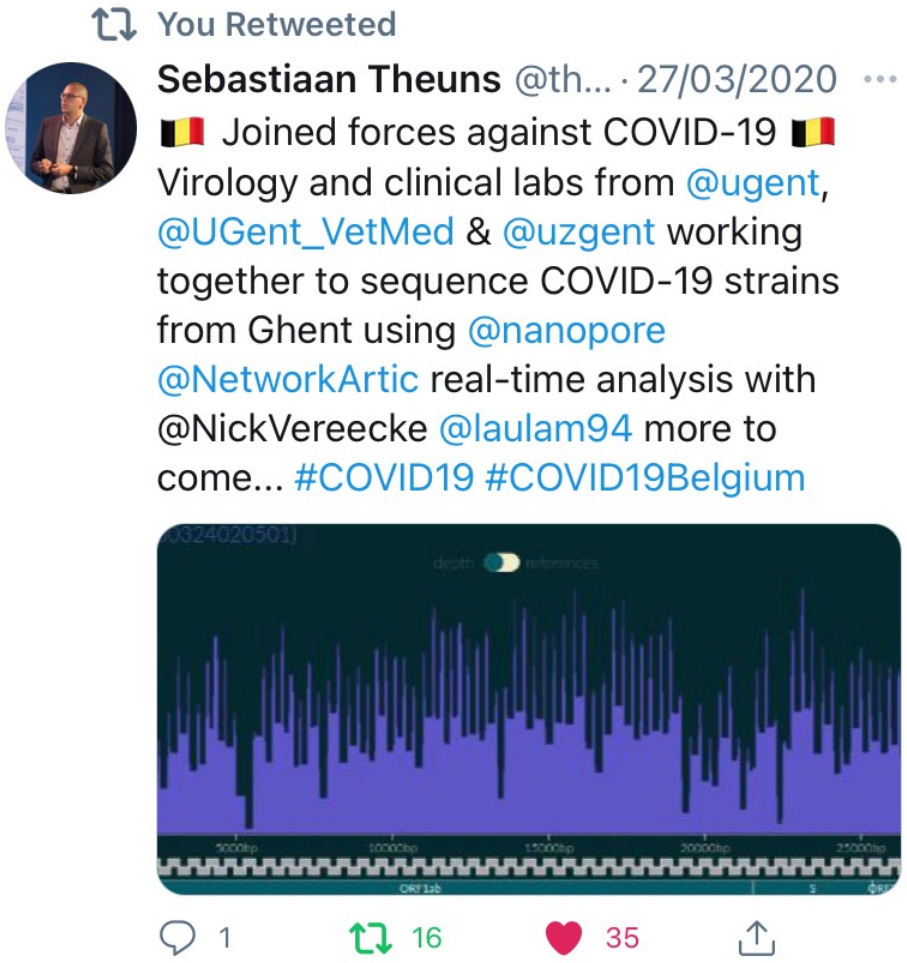
- Genome = ± **30 kbases +ssRNA**
- First complete sequence available **December 2019 (Wuhan-Hu-1)**
- Standardized protocols readily available through **ARTIC Network**
- Josh Quick, James Ferguson & Nick Loman
- **Nanopore Accuracy?**



Bull et al., 2021 Nature Com.

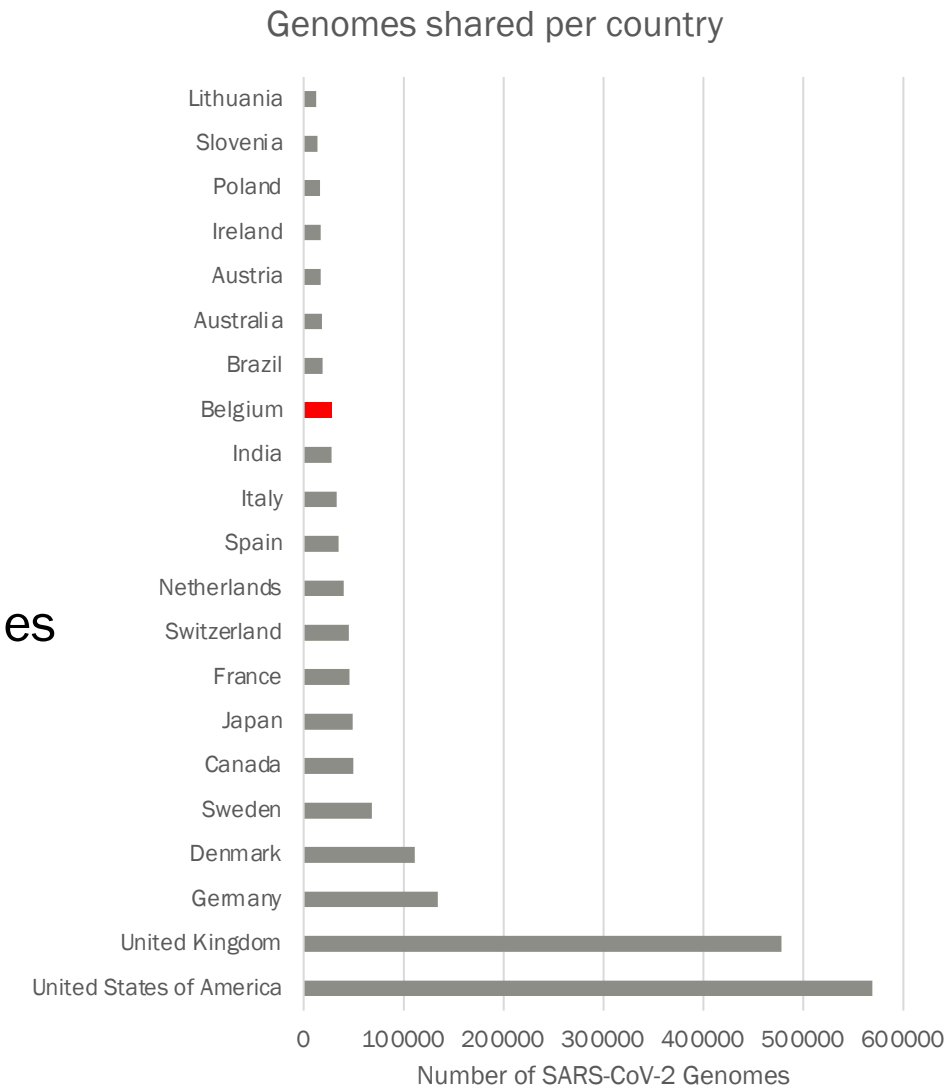
SARS-CoV-2

- **Belgium?**
  - *March 2020 - small scale*
  - *January 2021 - upscaling*
  - *June 2021 = **27,926 genomes (2.59%)***



SARS-CoV-2

- **Belgium?**
  - *March 2020 - small scale*
  - *January 2021 - upscaling*
  - *June 2021 = **27,926 genomes (2.59%)***
- **UK & Denmark** biggest sequencing efforts
- Today = 2.019.497 complete SARS-CoV-2 genomes



GISAID on 24.06.2021

Importance of HPC

- More genomes = More data = More **computational power** required
- **2.019.497** genomes (30 kbases$_{/genome}$) = **61 gigabases** = **± 65 gigabytes data**
- HPC implementation
  - *SARS-CoV-2 genome construction*
  - *Phylogenetic analyses (bootstrapping = repeating for significance)*
  - *Time-guided phylogenetic analyses (e.g. beast = **GPU version** available)*

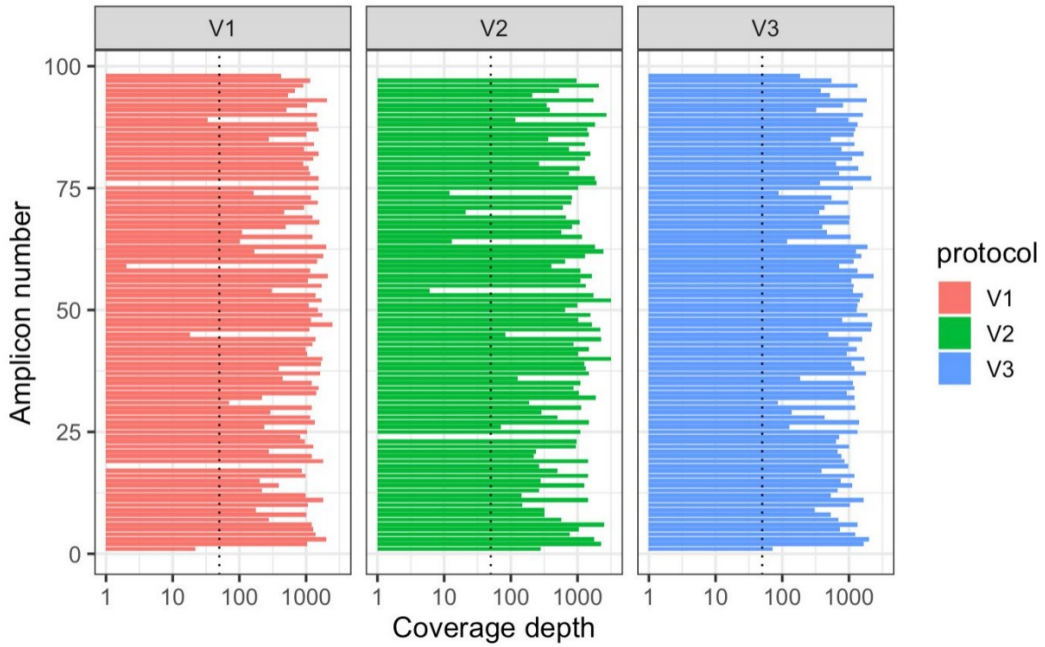
Importance of HPC – Genome Construction

- Each SARS-CoV-2 genome generated from ± 90 amplicons (*400 bases*)
- V3 protocol requires (only) 100,000 sequenced amplicons to get **50X overall coverage**

  = Each amplicon (400 bp) x 50

  = min. 1.8 million bases/genome

  = min. 2 gigabytes data/genome
- Multiplexing to save costs per genome
- 24 - 96 genomes per sequencing run (24h)
- **48 - 200 gigabytes data**

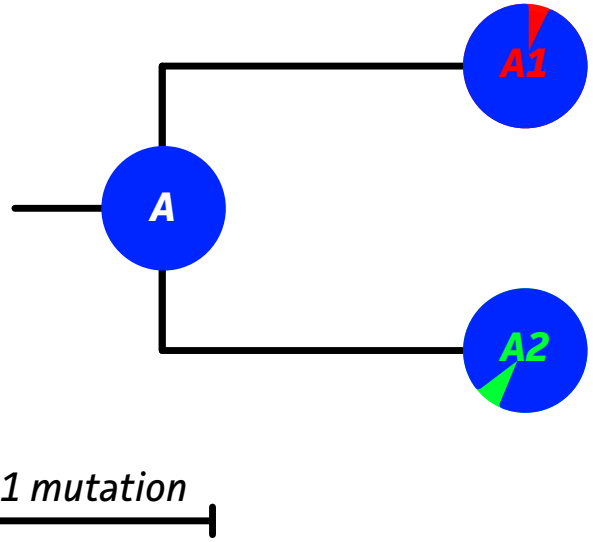


Loman & Ferguson; 2021

CPU & GPU

Importance of HPC – Phylogenetic Analyses

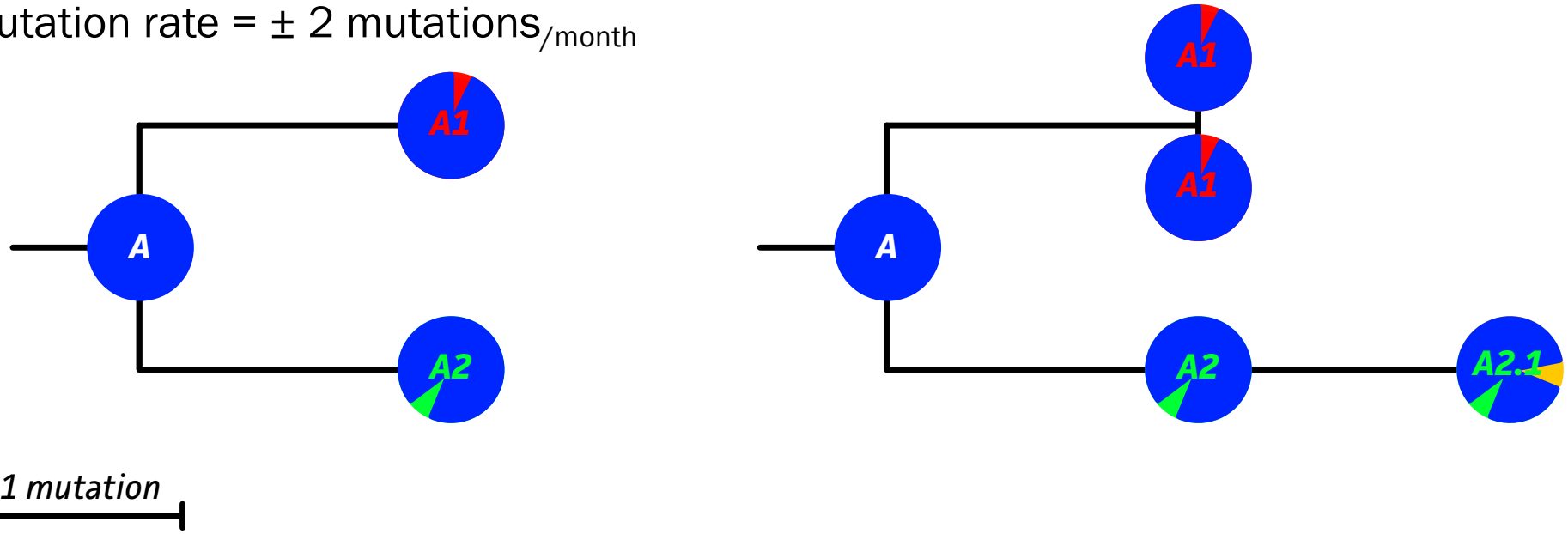
- Phylogenetic analyses
  1. *Compare sequences **one by one (Multiple-Sequence Alignment)***
  2. *Cluster closest/similar genomes together (**Phylogeny**)*
  3. *Add significant power of lineages (**Bootstrapping**)*
  4. *Add time of sampling (time-guided phylogenetic analysis)*
- Mutation rate = ± 2 mutations$_{/month}$



CPU

Importance of HPC – Phylogenetic Analyses

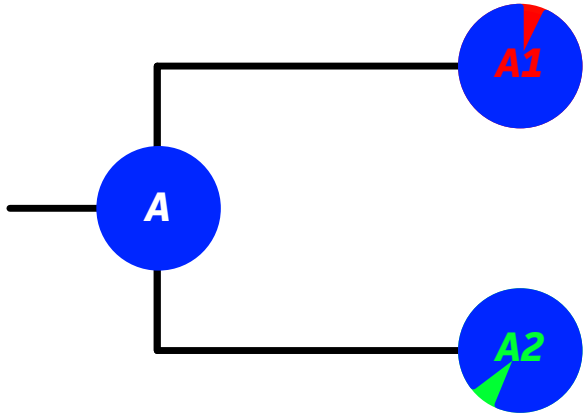
- Phylogenetic analyses
  1. *Compare sequences **one by one (Multiple-Sequence Alignment)***
  2. *Cluster closest/similar genomes together (**Phylogeny**)*
  3. *Add significant power of lineages (**Bootstrapping**)*
  4. *Add time of sampling (time-guided phylogenetic analysis)*
- Mutation rate = ± 2 mutations/month

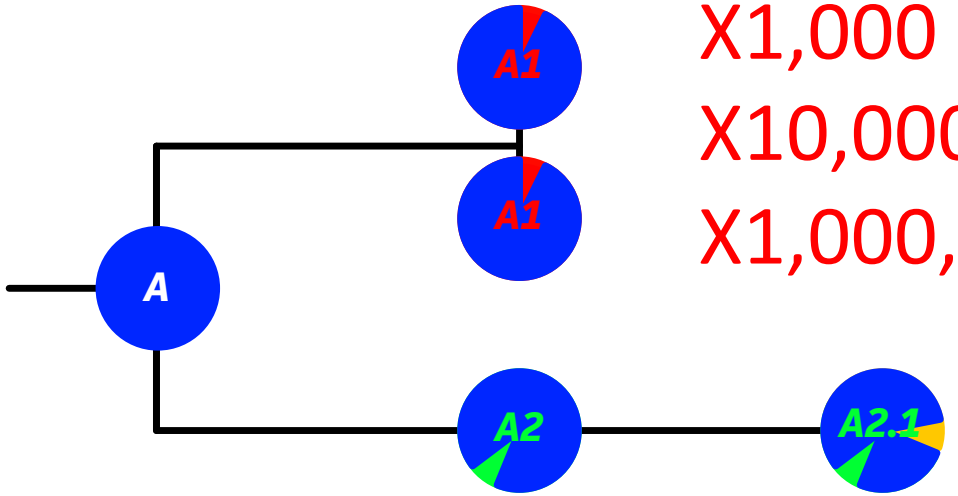


CPU

Importance of HPC – Phylogenetic Analyses

- Phylogenetic analyses
  1. *Compare sequences **one by one (Multiple-Sequence Alignment)***
  2. *Cluster closest/similar genomes together (**Phylogeny**)*
  3. *Add significant power of lineages (**Bootstrapping**)*
  4. *Add time of sampling (time-guided phylogenetic analysis)*
- Mutation rate = ± 2 mutations/month





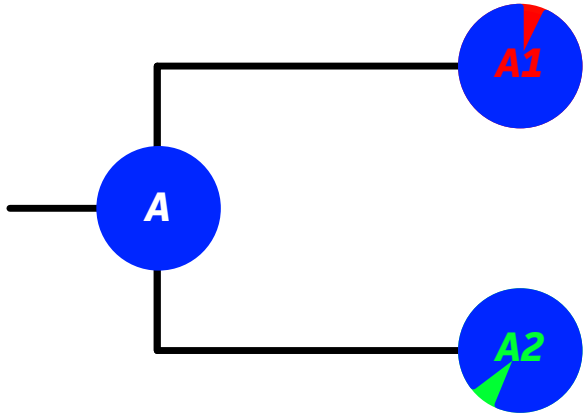
X1,000
X10,000
X1,000,000

1 mutation

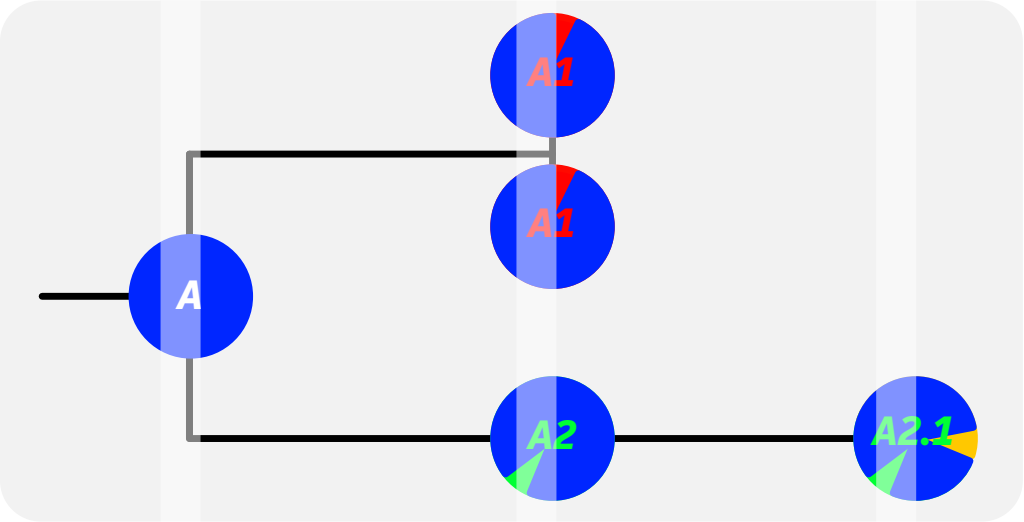
CPU

Importance of HPC – Phylogenetic Analyses

- Phylogenetic analyses
  1. *Compare sequences **one by one (Multiple-Sequence Alignment)***
  2. *Cluster closest/similar genomes together (**Phylogeny**)*
  3. *Add significant power of lineages (**Bootstrapping**)*
  4. *Add time of sampling (**time-guided** phylogenetic analysis)*
- Mutation rate = ± 2 mutations/month



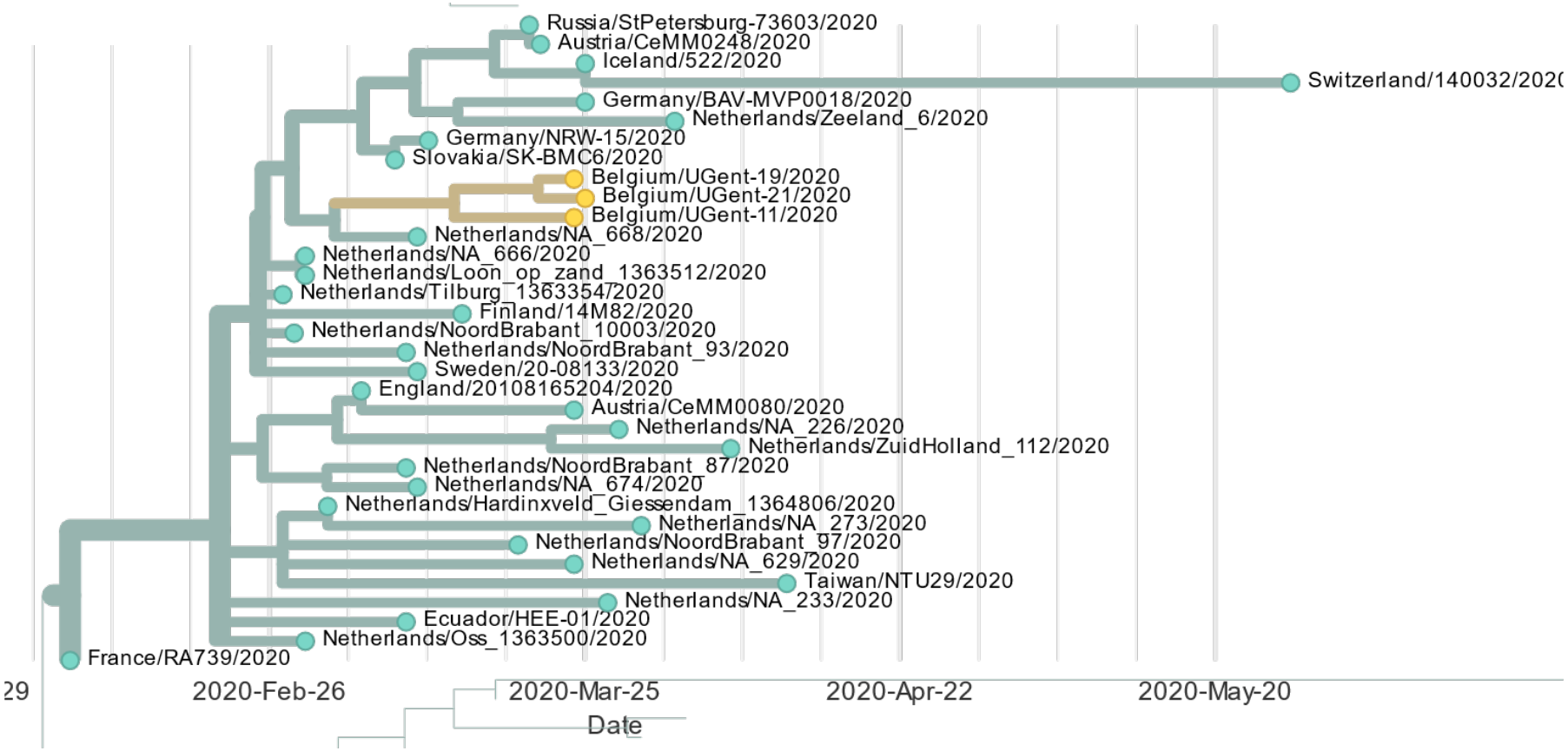
1 mutation



1 April '20 *15 April '20* *29 April '20*

GPU

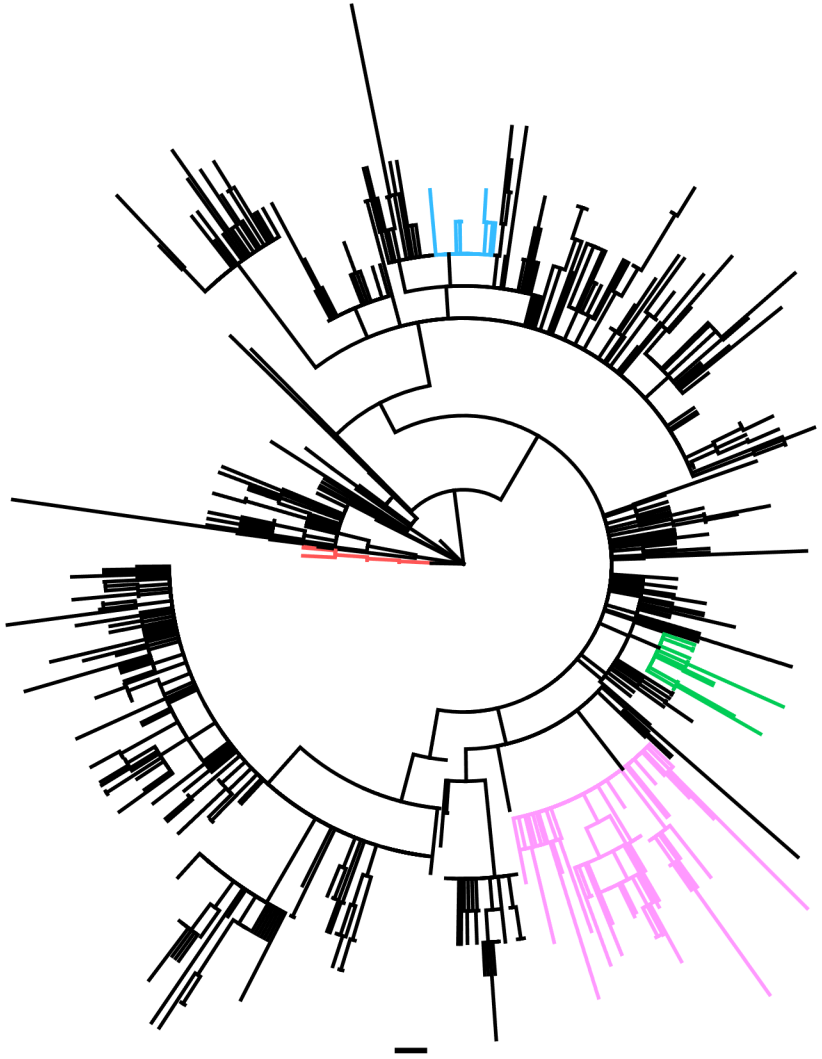
Importance of HPC – Phylogenetic Analyses



Laurens Lambrechts; 2020

CPU

Importance of HPC – Phylogenetic Analyses



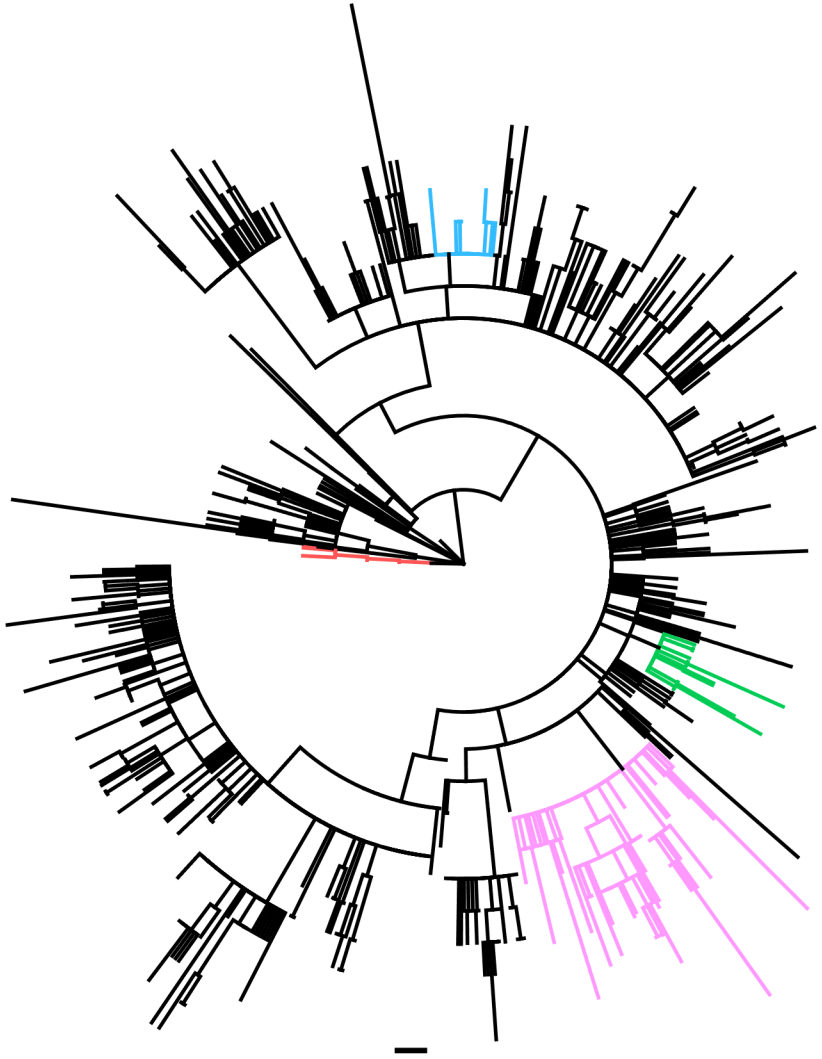
$3.3x10^{-5}$ subst. per site

- **Outbreak** analyses

Laurens Lambrechts; 2020

CPU

Importance of HPC – Phylogenetic Analyses



$3.3x10^{-5}$ subst. per site

- **Outbreak** analyses

vrt NWS

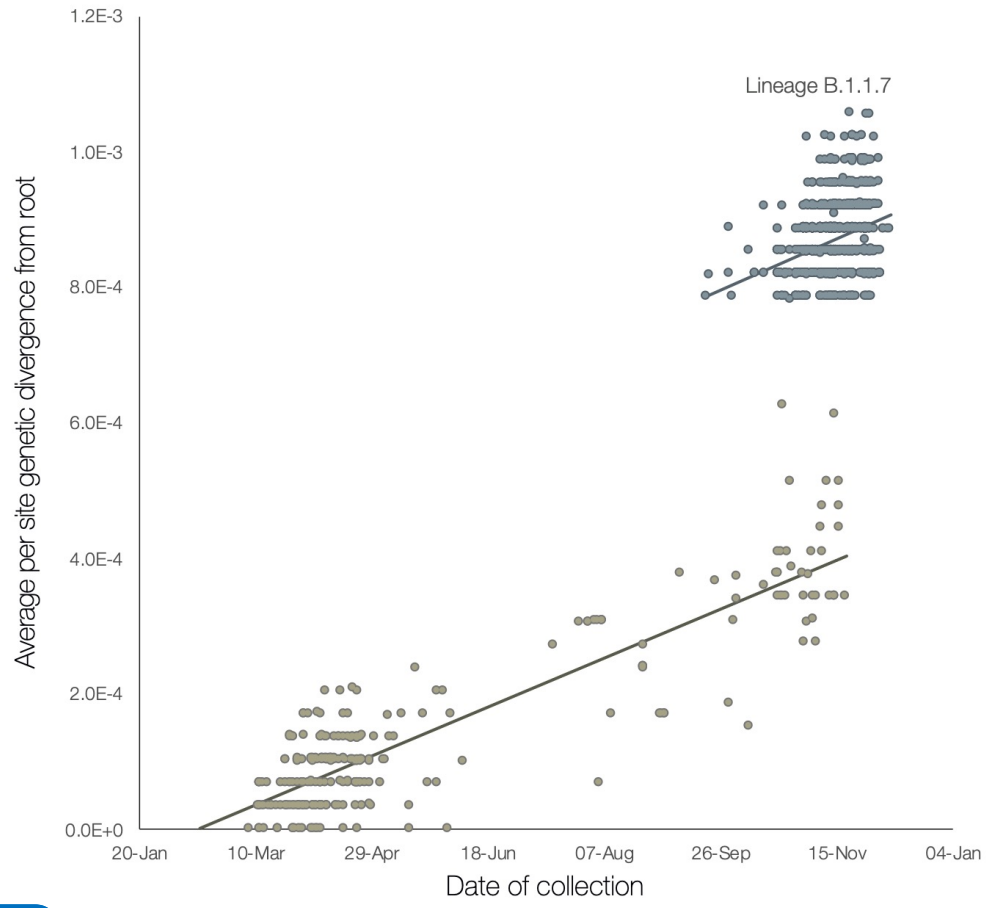
Al 18 bewoners dood in woonzorgcentrum in Mol, meesten besmet geraakt met zelfde virusstam

- Not to identify a single "culprit"
- Use to evaluate/adjust local & global **safety measures**

Laurens Lambrechts; 2020

CPU

Importance of HPC – Phylogenetic Analyses

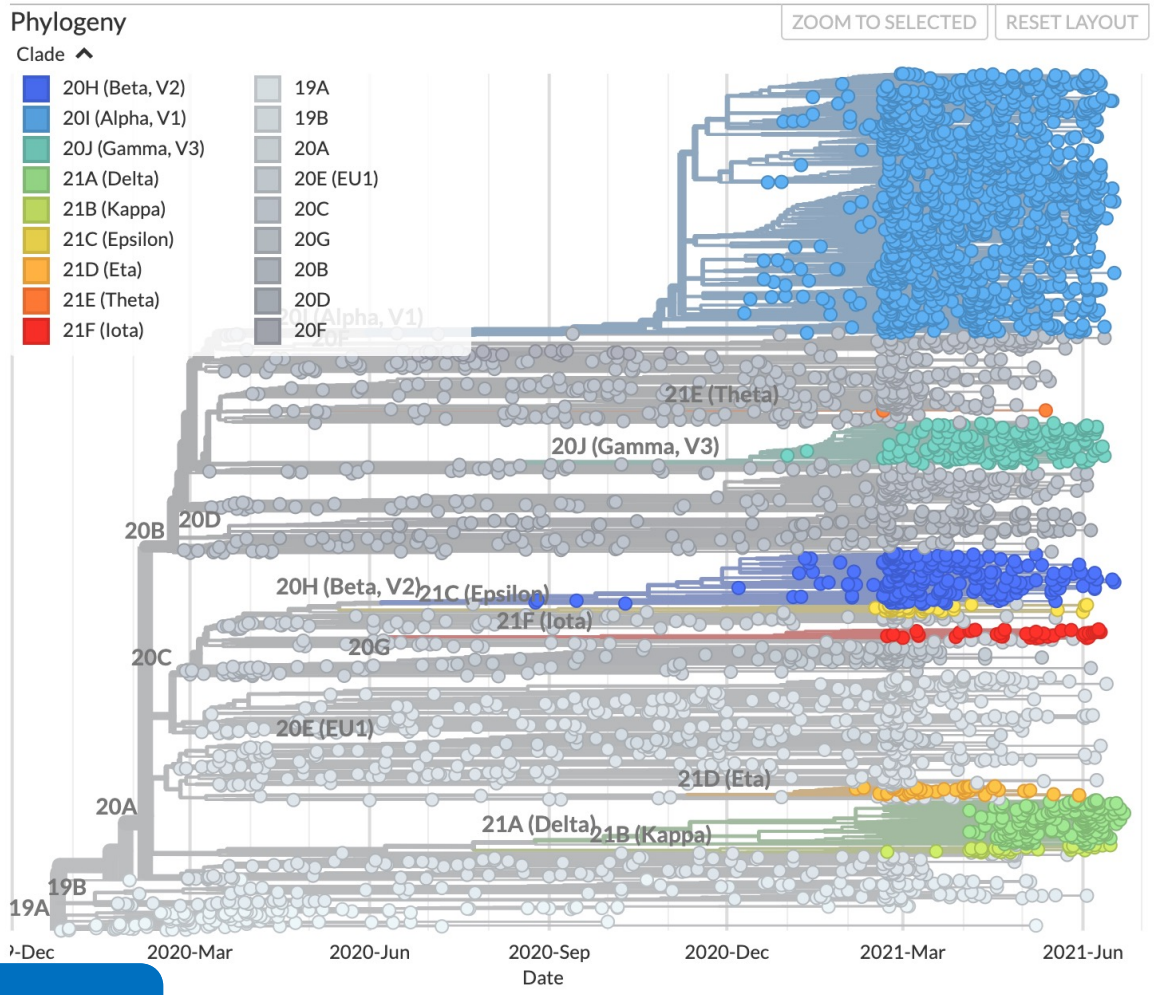


- **Outbreak** analyses
- **Variant of Concern** analyses

Rambaut et al.; 2020

CPU

Importance of HPC – Phylogenetic Analyses



- **Outbreak** & variant analyses
- **Variant of Concern** analyses
- Try it yourself?

  https://nextstrain.org/ncov/global
- Belgian builds > **Prof. G. Baele** (KULeuven)

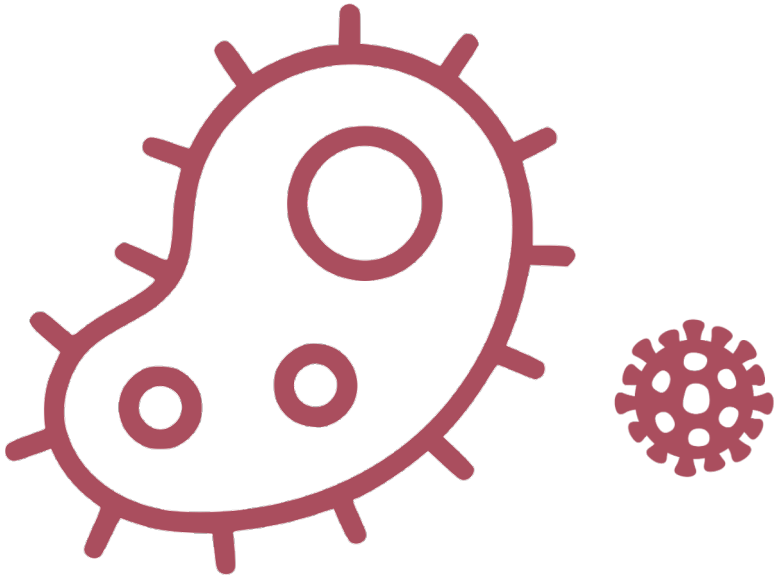
NextStrain on 26.06.2021

CPU

Bacterial Whole Genome Sequencing

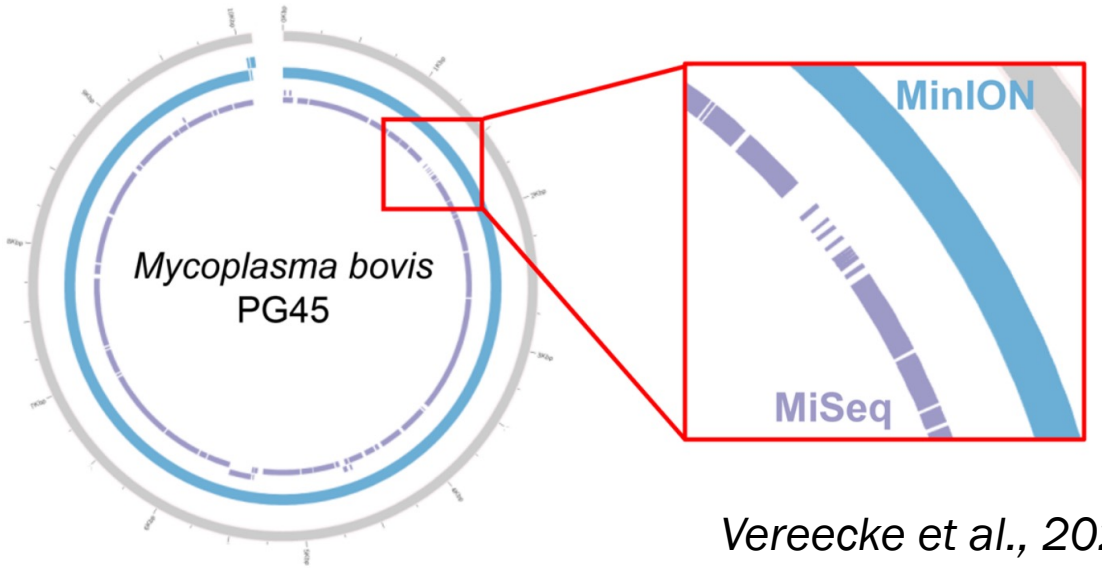
Bacterial Whole Genome Sequencing

- Bacterial genomes = **1-5 Mbp (+/- plasmids)**
- **33x up to 166x bigger genomes (*Vs. SARS-CoV-2*)**
- Standard Bacterial Genomics workflows:
  1. *Basecalling*
  2. *Demultiplexing & Quality Filtering*
  3. **De novo** *genome construction*
  4. *Downstream analyses*
     - Phylogenetic analysis
     - Bacterial Identification & Typing
     - Identification of Virulence & Antimicrobial Resistance Markers
     - Genome Annotation
     - …



Bacterial Whole Genome Sequencing

- Long-read sequencing data
  - *easily resolves **complete** bacterial genomes & plasmids*
  - *Applicable to **diverse** bacterial species*
  - ***No PCR bias** due to GC content*
  - *Resolving **repetitive** regions*
  - *Advantages in **Metagenomics***



Vereecke et al., 2020 BMC Bioinformatics

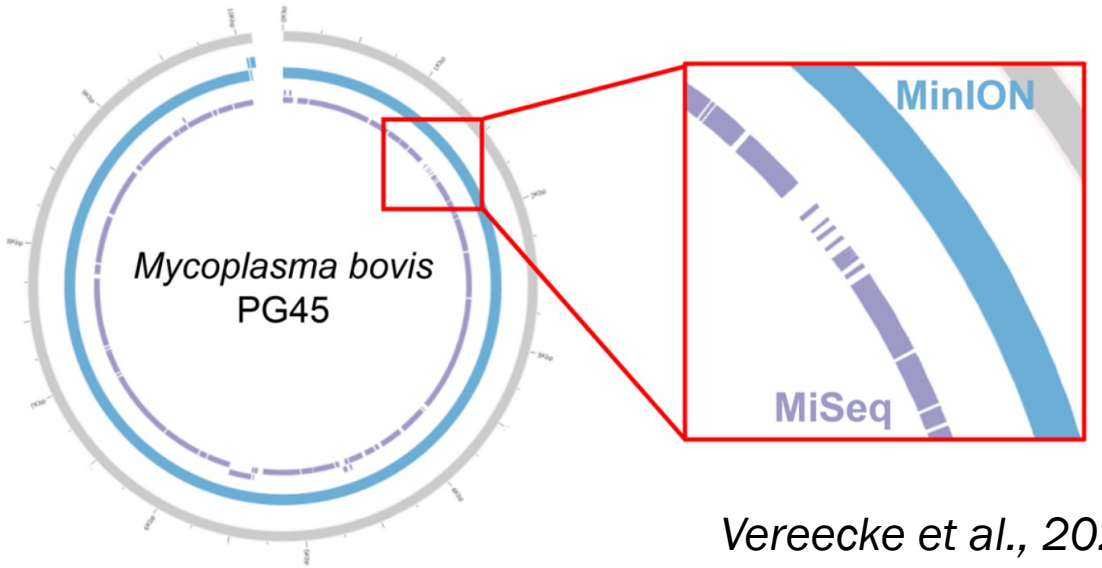
Bacterial Whole Genome Sequencing

- Long-read sequencing data
  - *easily resolves **complete** bacterial genomes & plasmids*
  - *Applicable to **diverse** bacterial species*
  - ***No PCR bias** due to GC content*
  - *Resolving **repetitive** regions*
  - *Advantages in **Metagenomics***

Mycoplasma sp.?

29% GC

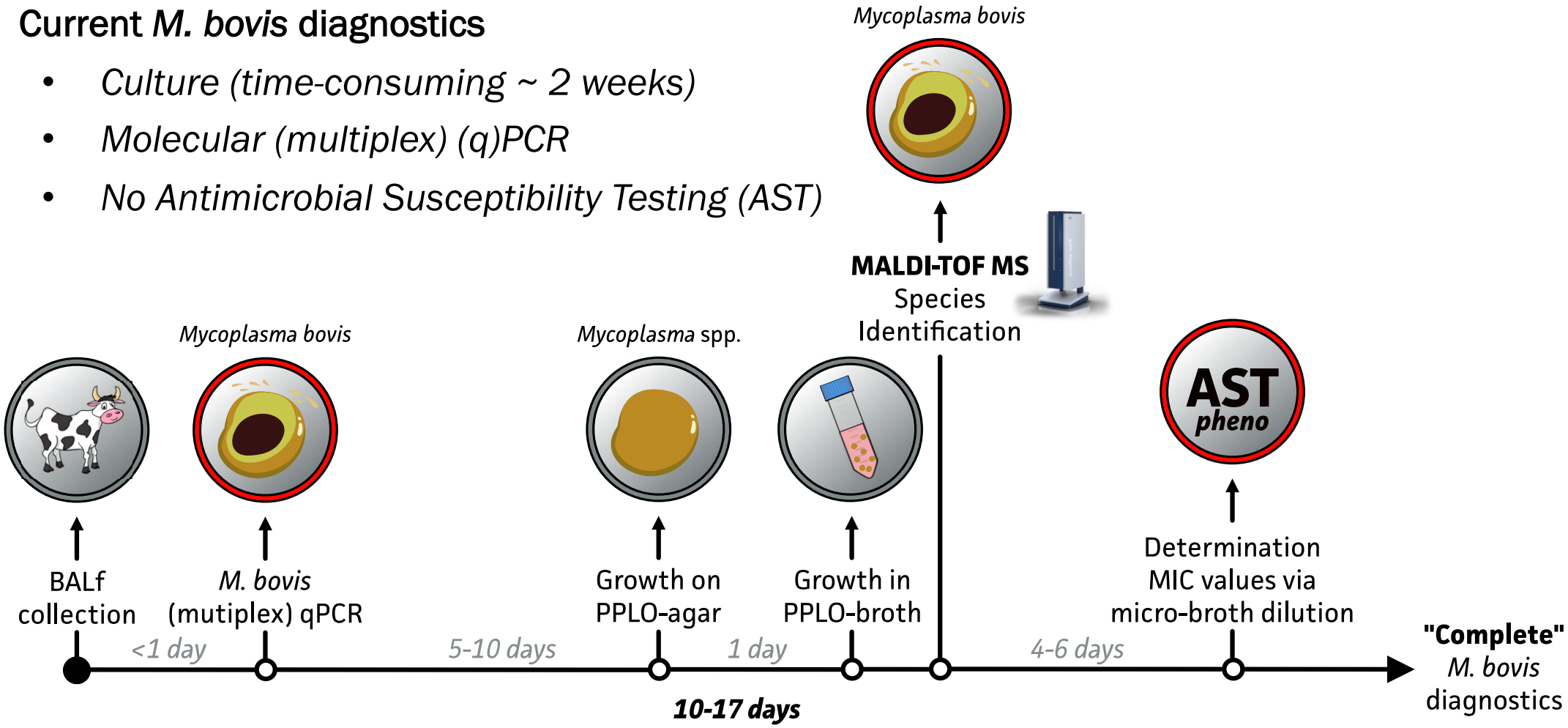
Highly repetitive



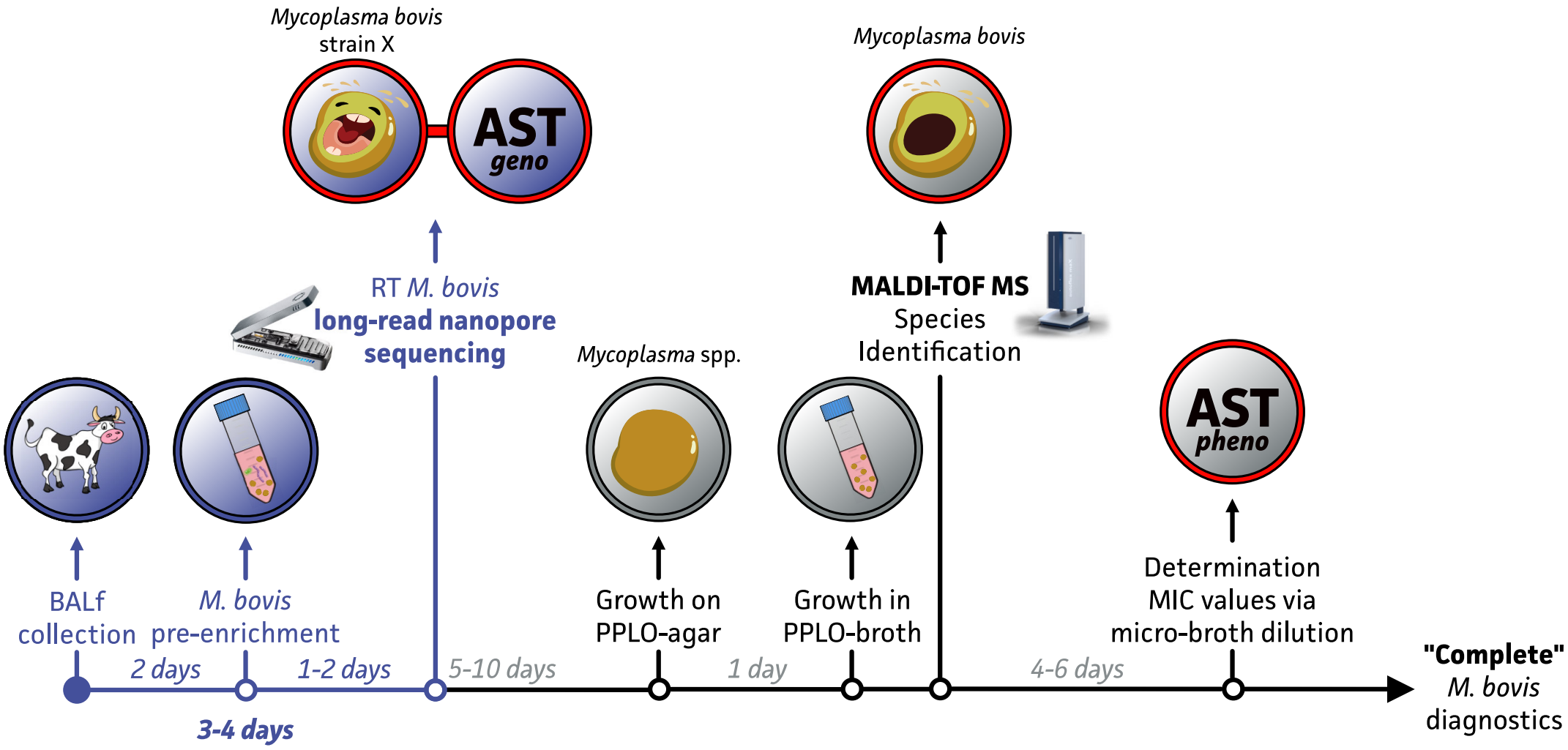
Vereecke et al., 2020 BMC Bioinformatics

Bacterial Whole Genome Sequencing

- **Current *M. bovis* diagnostics**
  - *Culture (time-consuming ~ 2 weeks)*
  - *Molecular (multiplex) (q)PCR*
  - *No Antimicrobial Susceptibility Testing (AST)*



Bacterial Whole Genome Sequencing



Importance of HPC

- Bigger genomes = More data = More **computational power** required
- HPC implementation
  - *Basecaller training*
  - *Raw data basecalling (**GPU version** available)*
  - *Bacterial genome construction*
  - *Phylogenetic analyses (bootstrapping = repeating for significance)*
  - *Genome-Wide Association Studies (GWAS)*

Importance of HPC – Basecaller training

- **Teaching** software to more accurately translate raw data into bases.
- **Bonito** Research Basecaller (ONT)
- Multi-GPU support = increased speed of training!
- Generate genomes with Consensus Quality = Illumina data

METHODOLOGY ARTICLE Open Access

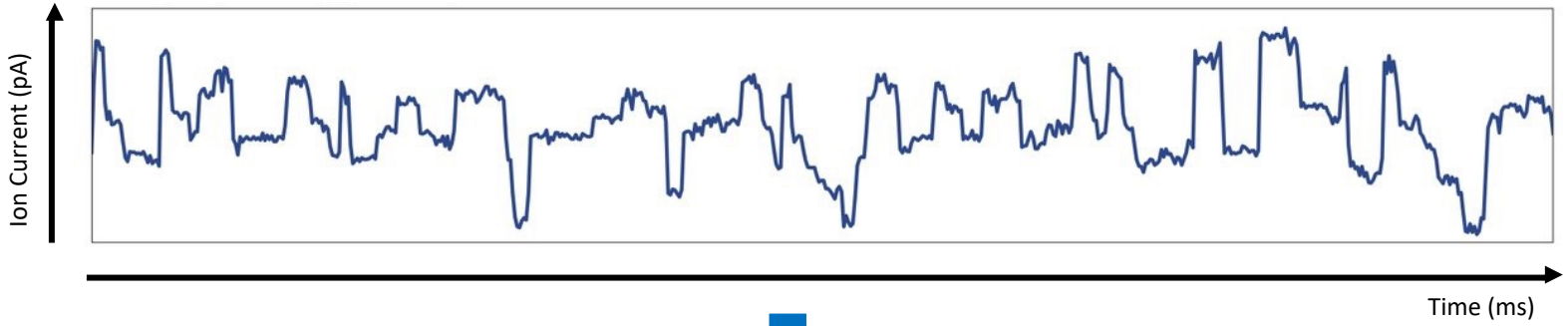
Check for updates

High quality genome assemblies of *Mycoplasma bovis* using a taxon-specific Bonito basecaller for MinION and Flongle long-read nanopore sequencing

Nick Vereecke[1,4*], Jade Bokma[2], Freddy Haesebrouck[3], Hans Nauwynck[1,4], Filip Boyen[3], Bart Pardon[2] and Sebastiaan Theuns[1,4]

GPU

Importance of HPC – Basecaller training



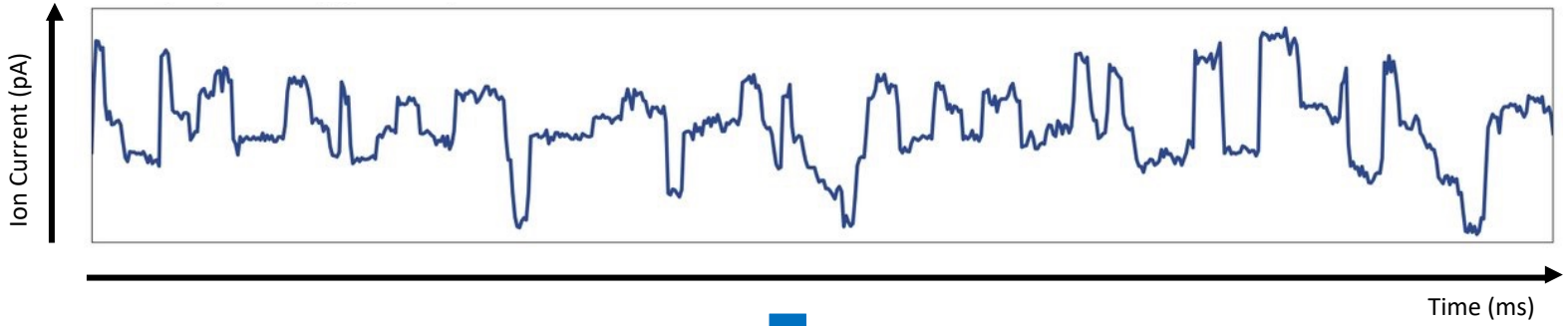
Default Basecalling

A C T T A C T T A A G C G G G C G A T C T A A C C G A A G T C A C C C C T *Sequenced read*

A C T T A C T T T A G C G G G C G A T C T A A A C G A A G T C A C C C G T *"True" Reference*

GPU

Importance of HPC – Basecaller training



Default Basecalling

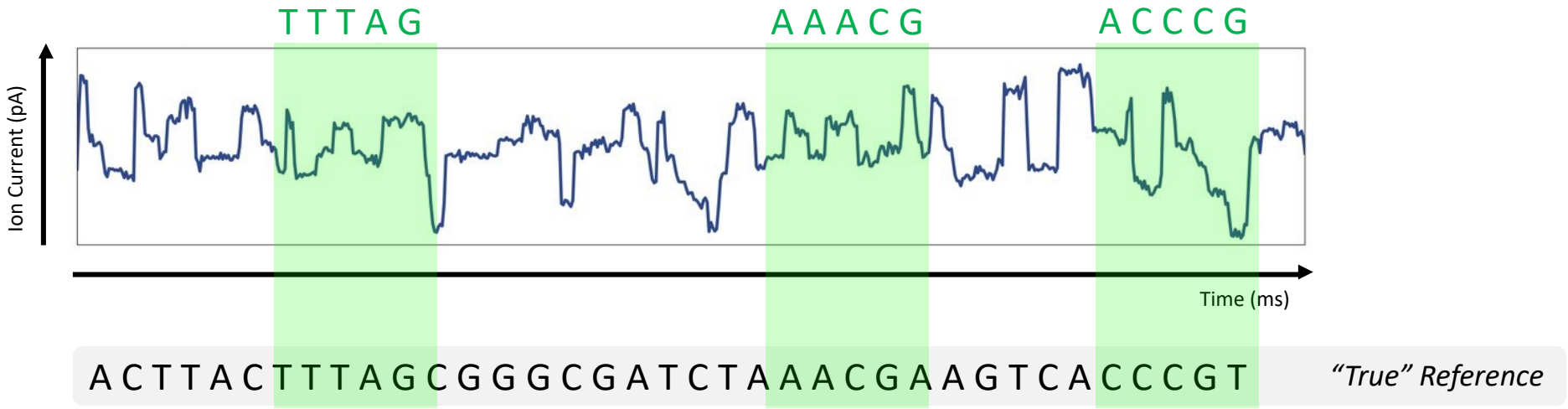
A C T T A C T T A A G C G G G C G A T C T A A C C G A A G T C A C C C C T *Sequenced read*

A C T T A C T T T A G C G G G C G A T C T A A A C G A A G T C A C C C G T *"True" Reference*

Teach the original model how to translate = **model training**

GPU

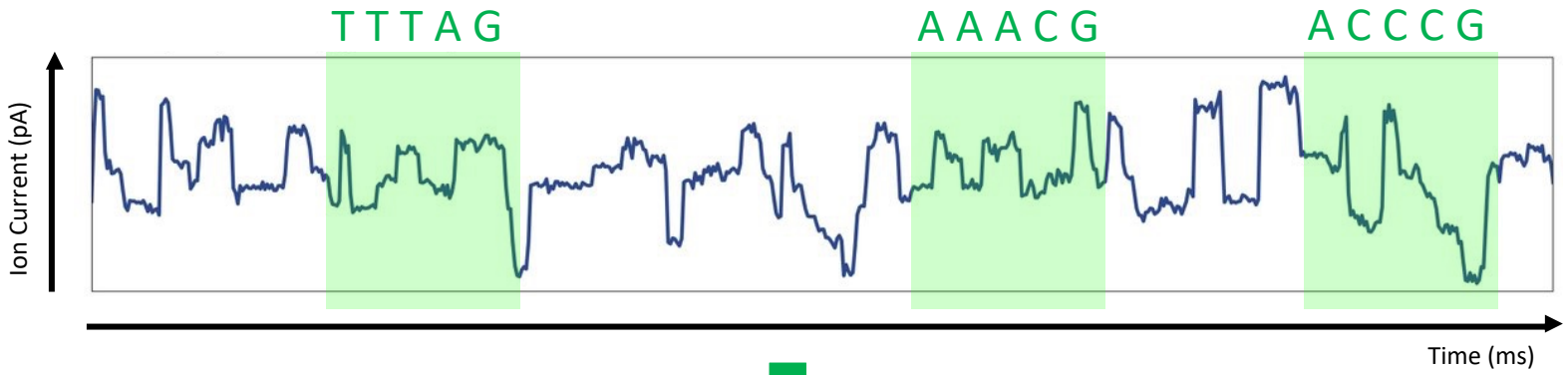
Importance of HPC – Basecaller training



Teach the original model how to translate = **model training**

GPU

Importance of HPC – Basecaller training



Trained Basecalling

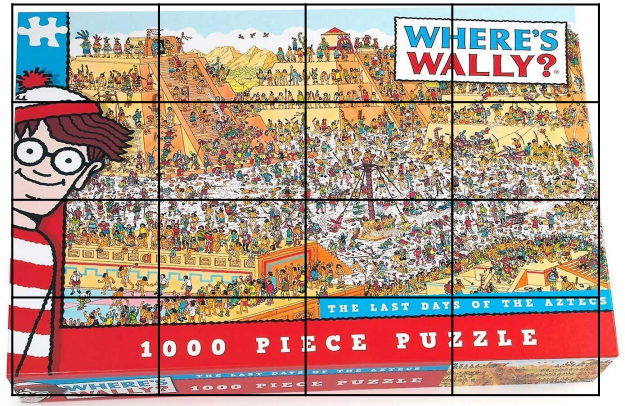
A C T T A C T T T A G C G G G C G A T C T A A A C G A A G T C A C C C G T — ***Trained*** *Sequenced read*

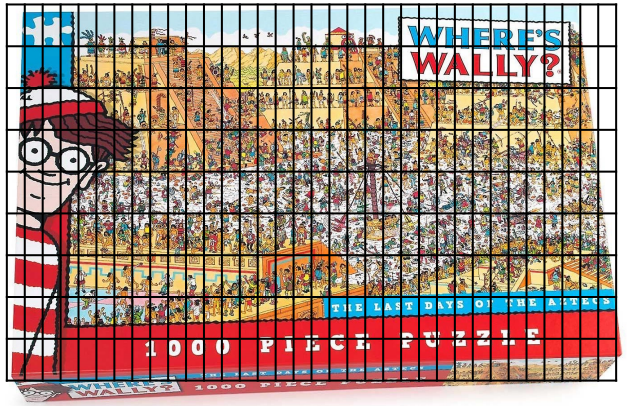
A C T T A C T T A A G C G G G C G A T C T A A C C G A A G T C A C C C C T — ***Default*** *Sequenced read*

A C T T A C T T T A G C G G G C G A T C T A A A C G A A G T C A C C C G T — *"True" Reference*

GPU

Importance of HPC – Genome Construction



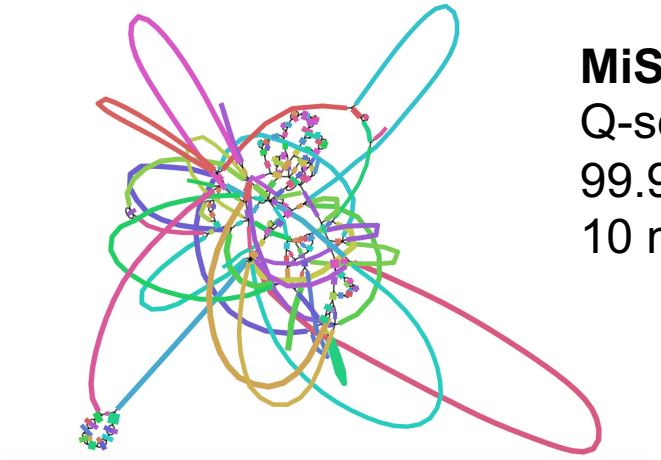


Oxford NANOPORE Technologies

illumina

ONT *default* (2020)
Q-score: 30
99.9%
1000 mistakes in genome

ONT *trained* (2021)
Q-score: 50
99.999%
10 mistakes in genome



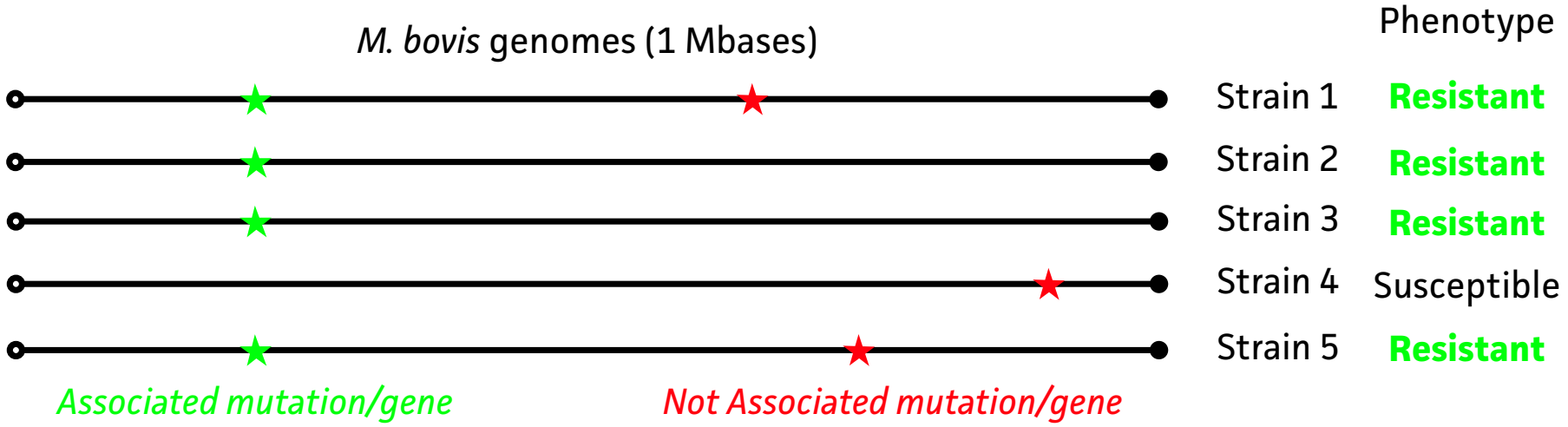
MiSeq (2020)
Q-score: 50
99.999%
10 mistakes in genome

Vereecke et al., 2020 BMC Bioinformatics

CPU & GPU

Importance of HPC – GWAS

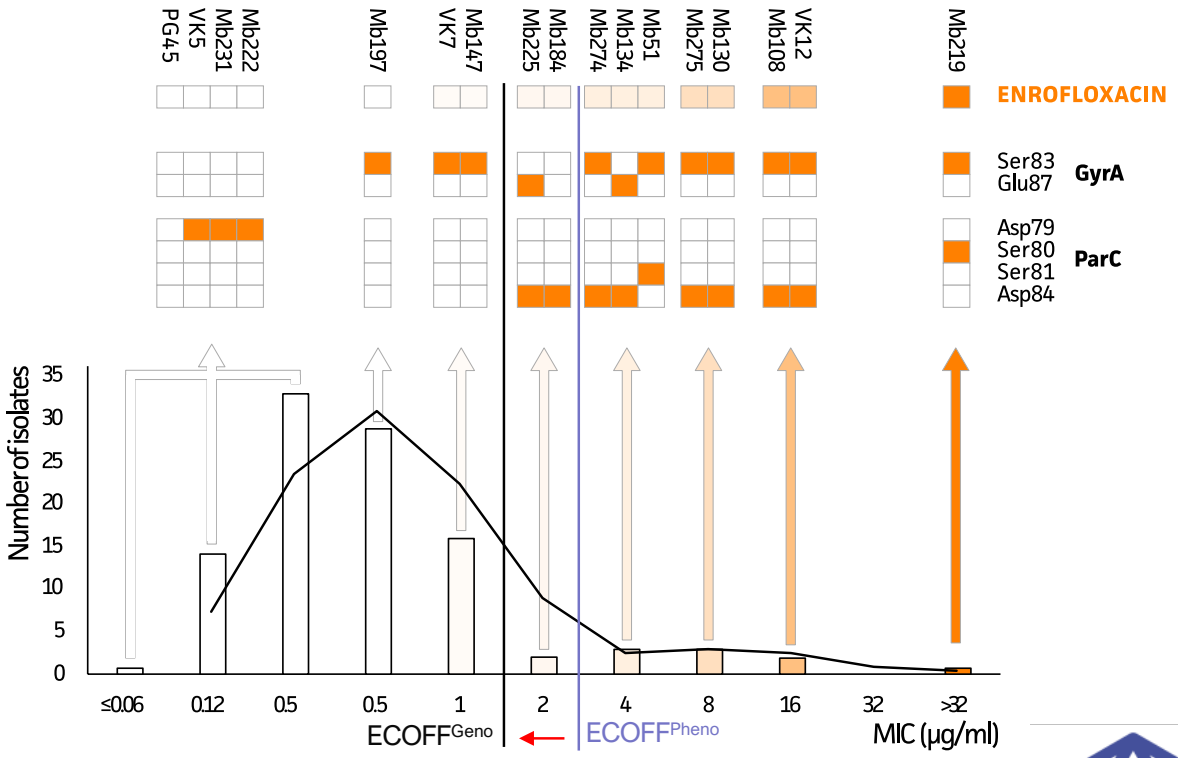
- Need for **highly accurate** & **complete genomes**
- Compare **ALL** genomes of 1 species (*e.g. M. bovis*)
- Identify genes & point mutations associated with phenotypes (*e.g. virulence or AMR*)
- More genomes = Higher resolution



CPU

Importance of HPC – GWAS

- 100 Belgian ***M. bovis*** genomes
- Resistance to Critical Antimicrobial **Enrofloxacin** (Fluoroquinolone)



Bokma & Vereecke et al., 2021; Under Review

CPU

Nick Vereecke
PhD Fellow
dr. Sebastiaan Theuns

E nick.vereecke@pathosense.com
T +32 (0)9 264 73 87
M +32 (0)467 03 78 04

DEPARTMENT VIROLOGY, PARASITOLOGY, AND IMMUNOLOGY
RESEARCH GROUP OF VIROLOGY

GHENT UNIVERSITY

FACULTY OF VETERINARY MEDICINE
accredited by EAEVE

PathoSense