Paul-Ehrlich-Institut
Federal Institute for Vaccines and Biomedicines

BLOODVIR

Surveillance system for novel viruses based on next generation sequencing and artificial intelligence

Martin Machyna, Markus Braun

FoG3 Host-Pathogen Interactions

Paul-Ehrlich-Institut

The Paul-Ehrlich-Institut is an Agency of the German Federal Ministry of Health.

Continuous viral surveillance is essential for public safety

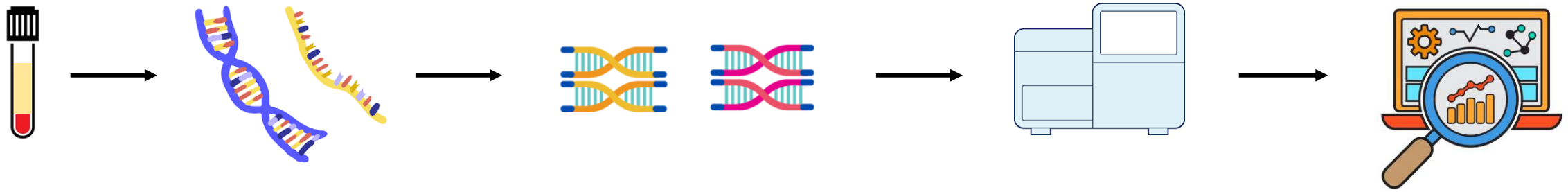
- 6,5 Million blood donations per year (Germany)
- 1:10,000 to 1:100,000 positive donations for HBV, HCV, HEV, HIV, Parvovirus B19
- Thanks to PCR testing only 11 transmission events in 4 years (2016-2020)

Identification of porcine circovirus type 1 (PCV-1) in Rotarix vaccine	*(Victoria et al. 2010)*
Identification of novel rhabdovirus in sf9 cells	*(Ma et al. 2014)*
Discovery of human circovirus 1 (HCirV-1) in human blood	*(Pérot et al. 2023)*

BLOODVIR surveillance system

Viral nucleic acid extraction

Sequencing

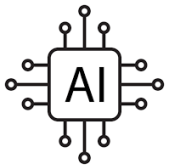


Blood plasma

Library preparation

Data Analysis

1. Known virus identification
2. Novel virus detection



BLOODVIR analysis pipeline

- Implemented in Snakemake
- Runs locally or on HPC cluster
- Read classification against a reduced RVDB
- Novel virus prediction with trained ML model
- Analysis results summarized in an interactive dashboard

Strategies for viral sequencing libraries

Non-targeted metagenomics

Host sequence depletion

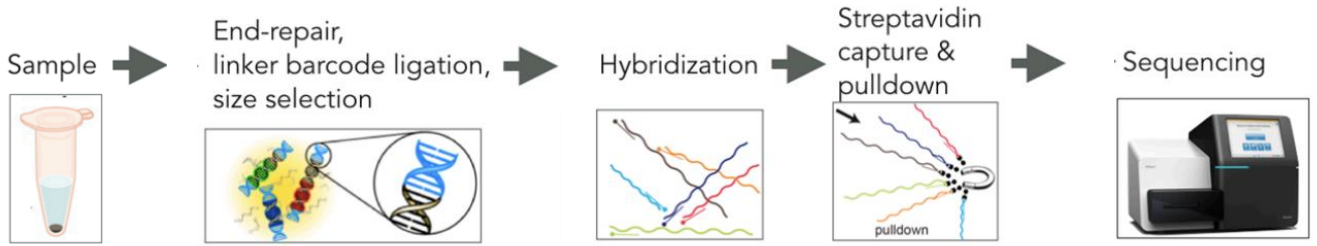
- Agnostic to virus type or genomic sequence
- Less virus material enrichment

Sample → Nuclease treatment → NEBNext Ultra II Library prep → Sequencing

Targeted virus enrichment

Virus enrichment with VirCapSeq-VERT (Ian Lipkin)

- 900,000 probes against 207 known vertebrate viral taxons
- Enables sequencing of genomes with as little as 75% sequence identity



Gibran Horemheb-Rubio
Csaba Miskey

Data analysis in a nutshell

Host sequence removal

- Removal of contaminating sequences
- Sequential alignment with bowtie2
- Common contaminants: human, mouse, bovine, bacteria

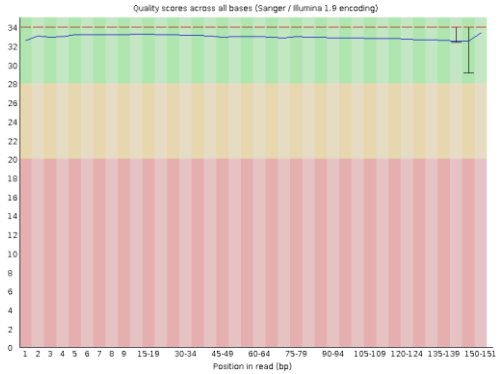
Data analysis in a nutshell

Host sequence removal

↓

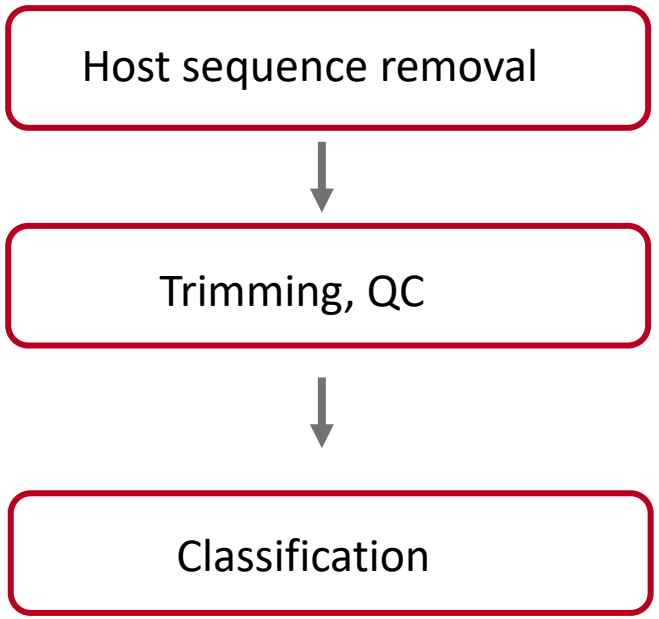
Trimming, QC

Before trimming



After trimming

Data analysis in a nutshell

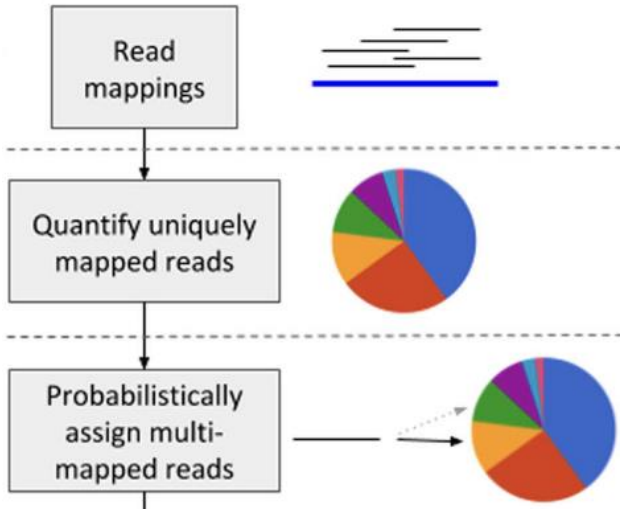


Classification methods benchmarking

Alignment-based

MiCoP *(LaPierre et al. 2019)*

TaxMaps *(Corvelo et al. 2018)*

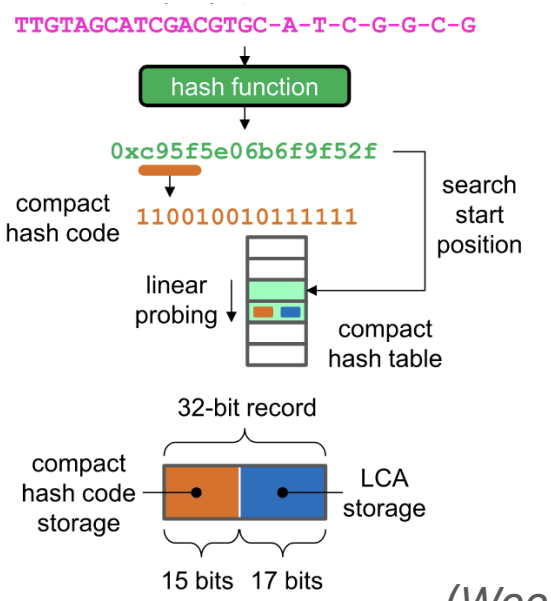


(LaPierre et al. 2019)

K-mer-based

Kraken2 *(Wood et al. 2019)*

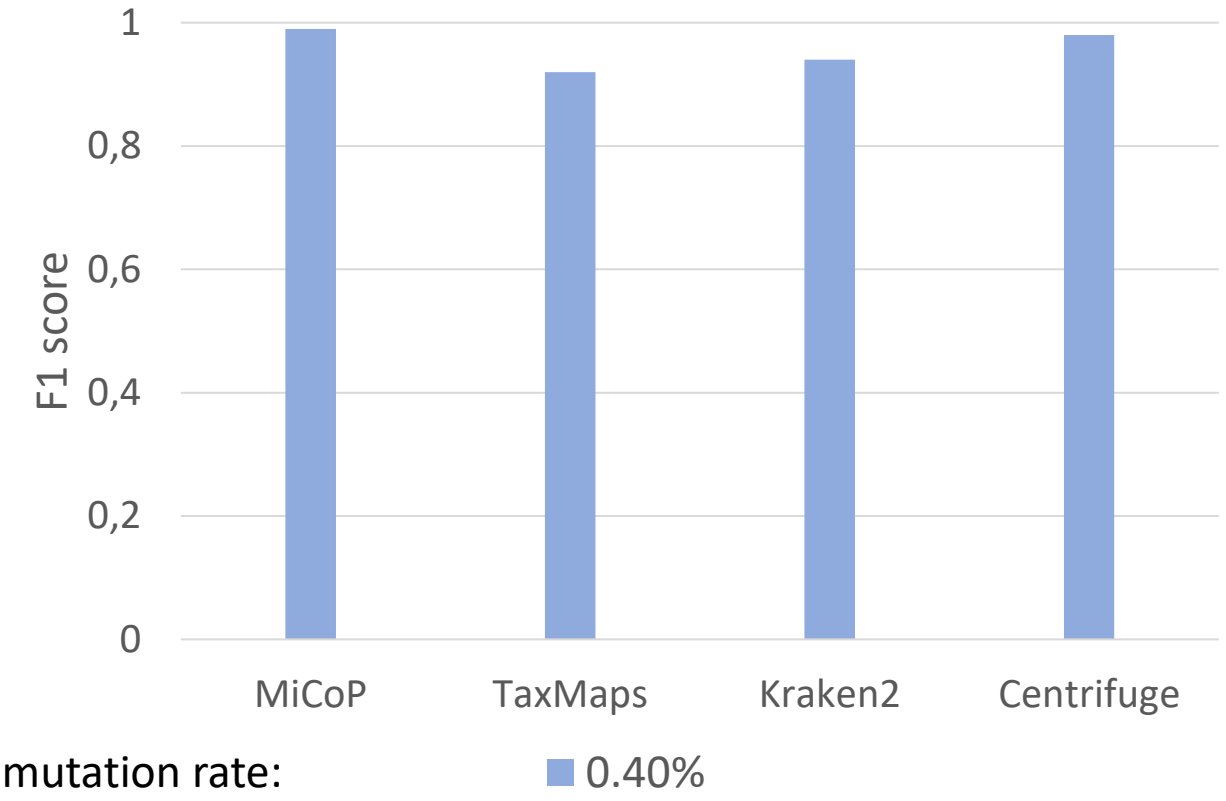
Centrifuge *(Kim et al. 2016)*



(Wood et al. 2019)

MiCoP provides best classification accuracy across a range of editing distances

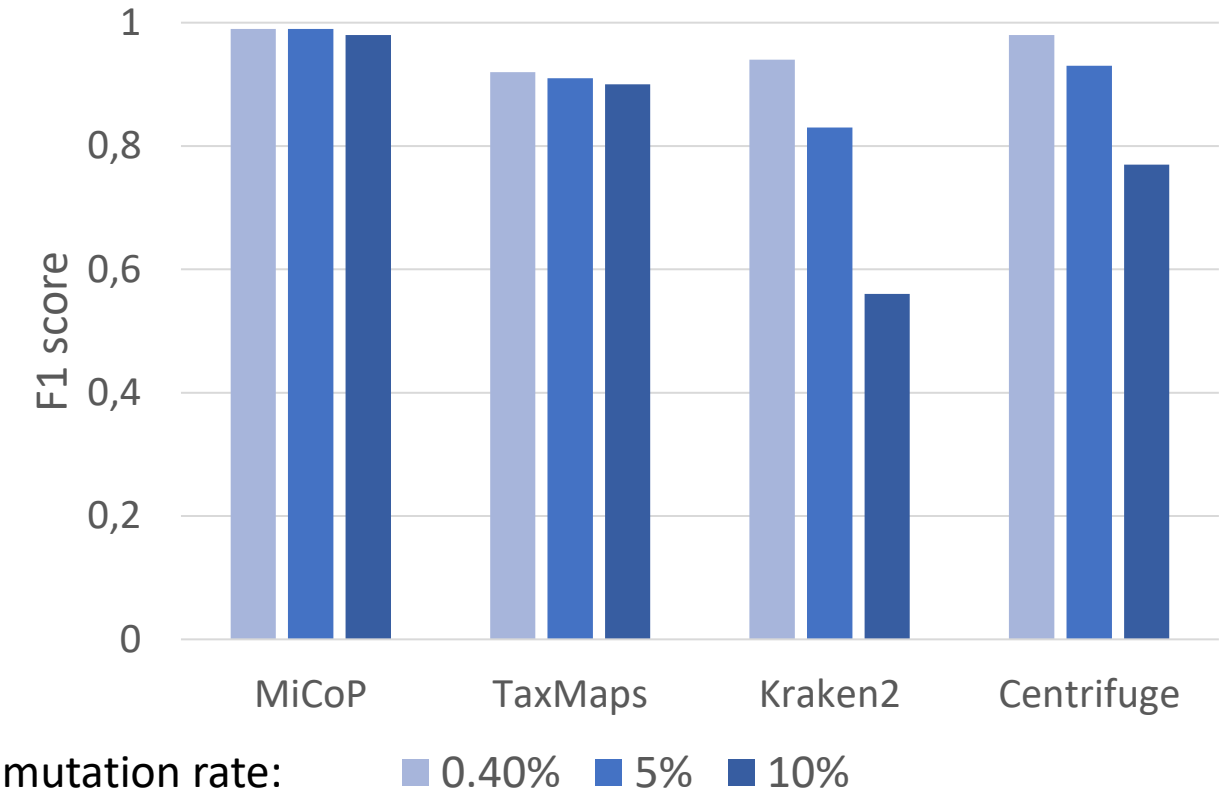
25,000,000	Random sequence
25,000,000	Human T2T
45,000,000	Influenza B
4,500,000	Hepatitis E
450,000	HIV
45,000	HPgV
4,500	Enterovirus C
450	Ebola
45	HTLV-1



F1 = (2 x precision x recall) / (precision + recall)

MiCoP provides best classification accuracy across a range of editing distances

25,000,000	Random sequence
25,000,000	Human T2T
45,000,000	Influenza B
4,500,000	Hepatitis E
450,000	HIV
45,000	HPgV
4,500	Enterovirus C
450	Ebola
45	HTLV-1



F1 = (2 x precision x recall) / (precision + recall)

Composition of genome database dramatically influences read classification

RVDB *(Goodacre et al. 2018)*

- Curated compilation of GenBank and RefSeq
- 850,309 records for 19,149 species (v23)
- Dominated by handful of viral species
- Many <1000 nt sequences

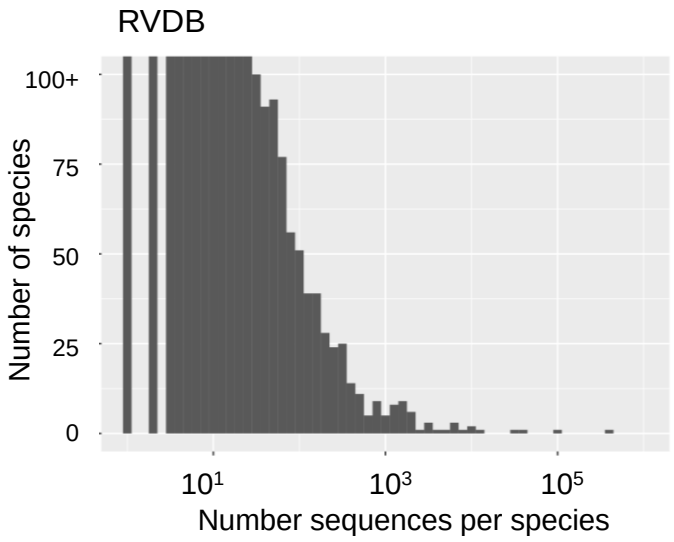
	RefSeq	RVDB
PCV1	✓	✓
Feline leukemia vrius	✓	✓
RSV	✓	✗
Gammaherpesvirus 4	✓	✓
Squirrel monkey virus	✓	✓
Viral reads	**3,395**	**1,569**

Composition of genome database dramatically influences read classification

RVDB *(Goodacre et al. 2018)*

- Curated compilation of GenBank and RefSeq
- 850,309 records for 19,149 species (v23)
- Dominated by handful of viral species
- Many <1000 nt sequences

	RefSeq	RVDB
PCV1	✓	✓
Feline leukemia vrius	✓	✓
RSV	✓	✗
Gammaherpesvirus 4	✓	✓
Squirrel monkey virus	✓	✓
Viral reads	**3,395**	**1,569**



Markus Braun

Composition of genome database dramatically influences read classification

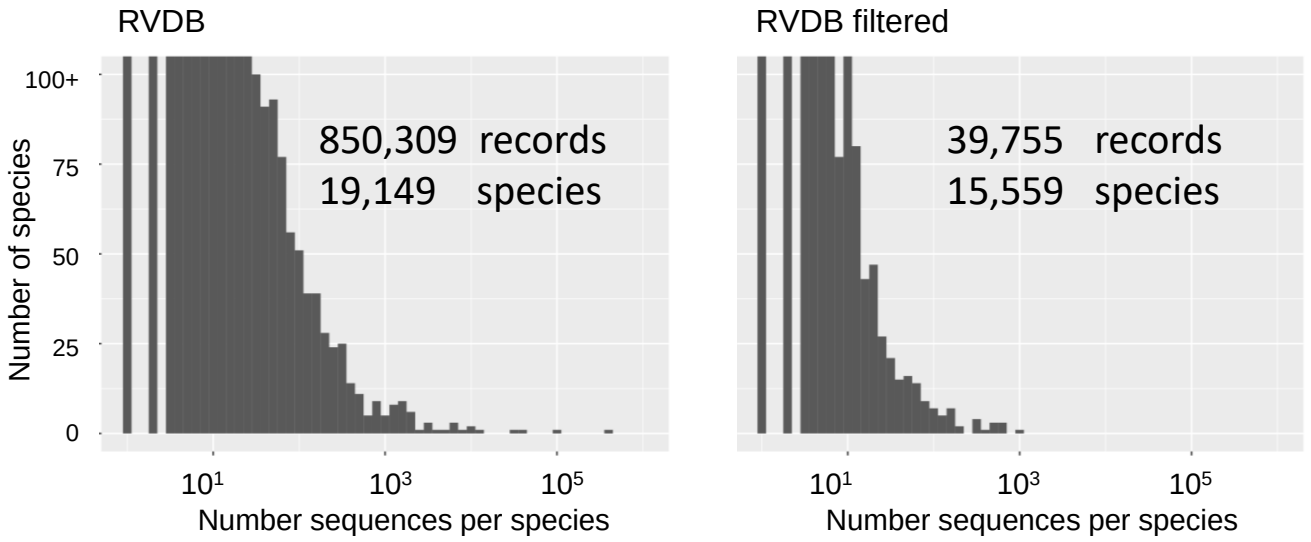
RVDB *(Goodacre et al. 2018)*

- Curated compilation of GenBank and RefSeq
- 850,309 records for 19,149 species (v23)
- Dominated by handful of viral species
- Many <1000 nt sequences

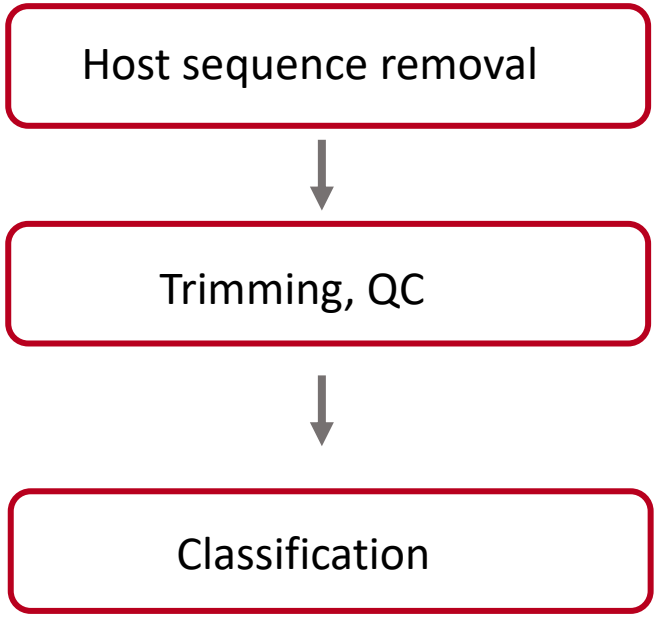
RVDB filtering

- Kept RefSeq records
- Removed genomes with Mash (MinHash) score < 0.15
- Removed sequences < 1000 nt

	RefSeq	RVDB	RVDB Filtered
PCV1	✓	✓	✓
Feline leukemia vrius	✓	✓	✓
RSV	✓	✗	✓
Gammaherpesvirus 4	✓	✓	✓
Squirrel monkey virus	✓	✓	✓
Viral reads	**3,395**	**1,569**	**6,917**



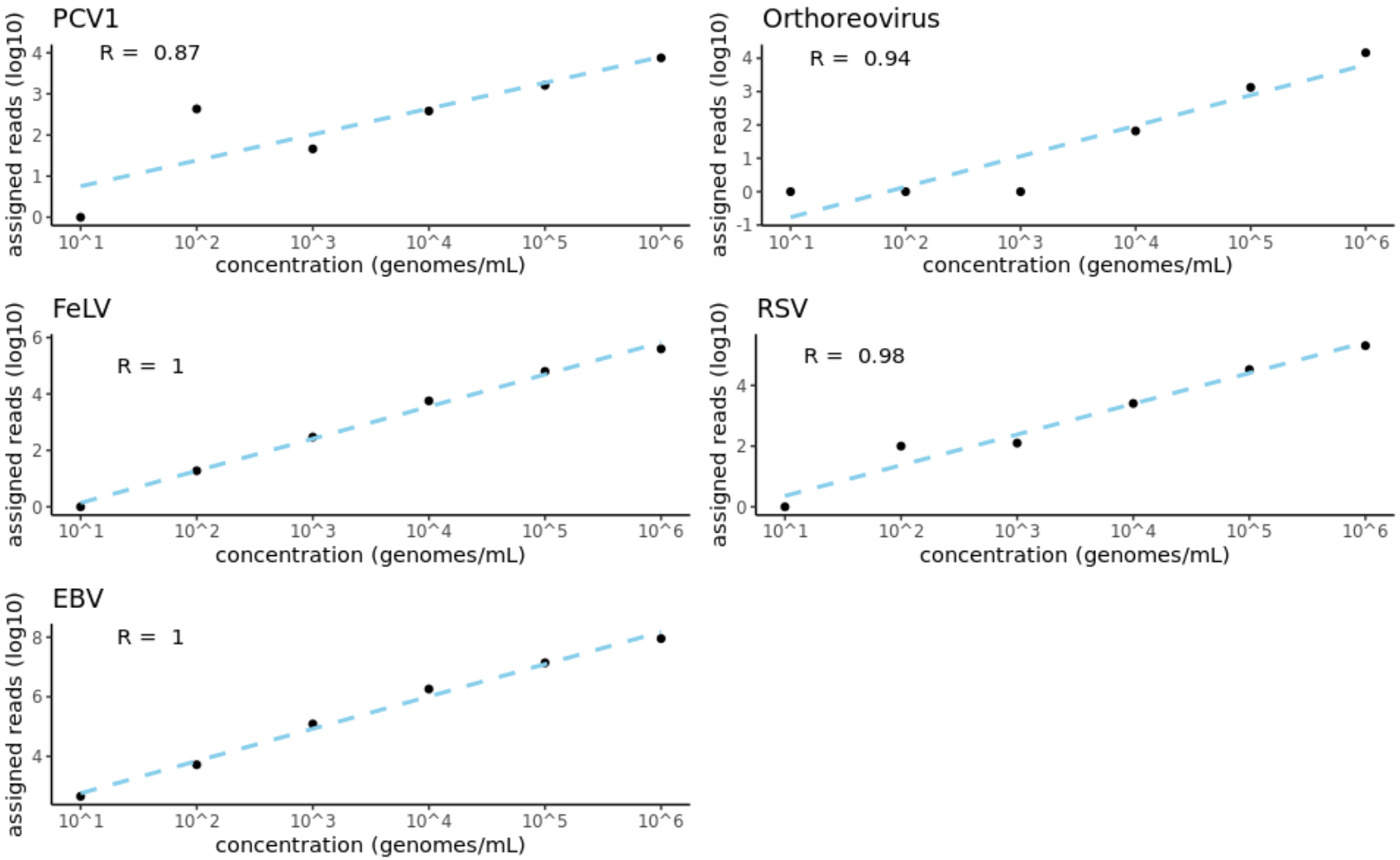
Data analysis in a nutshell



MiCoP (bwa) using reduced RVDB

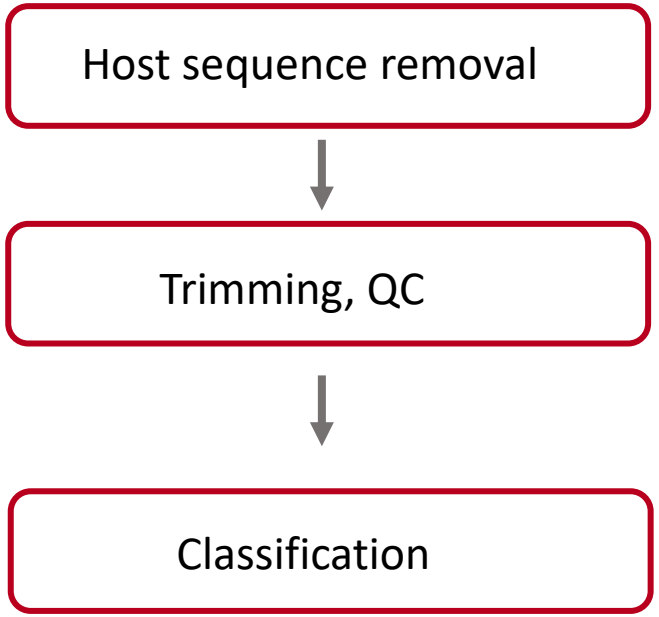
+ Kaiju secondary classification

BLOODVIR can detect 100 viral copies per mL

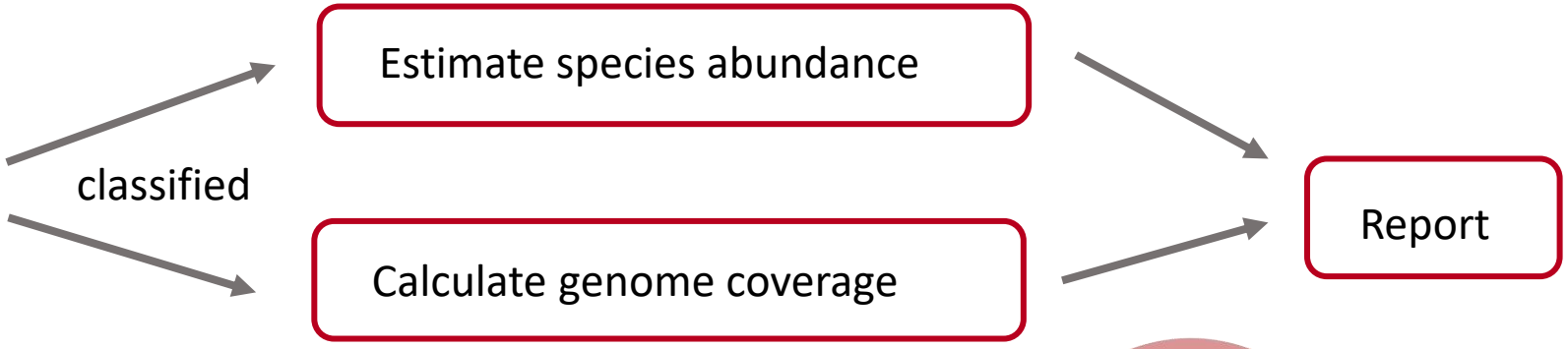


Markus Braun

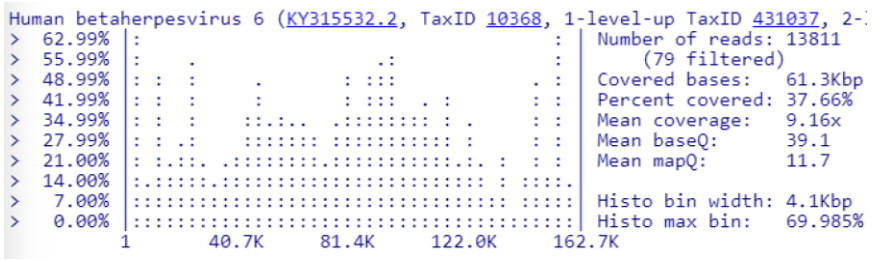
Data analysis in a nutshell

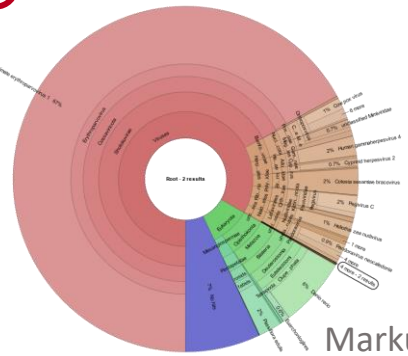


- Minimum coverage filtering
- Length filtering



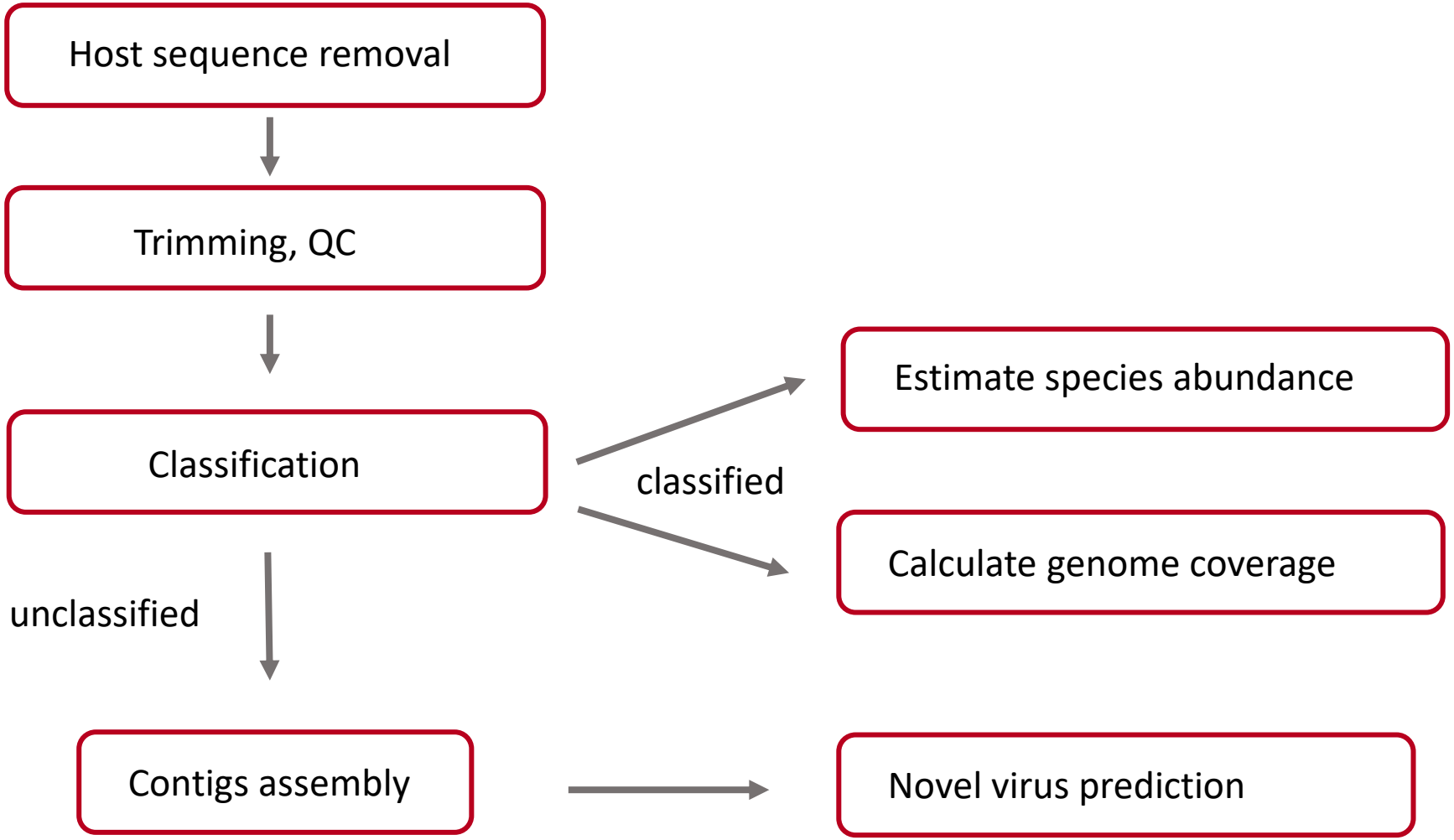
TaxID	Taxon name	Read number
10376	Human gammaherpesvirus 4	19,403
11856	Squirrel monkey retrovirus	4,741
68887	Torque teno virus	258
351073	Mammalian orthoreovirus	217



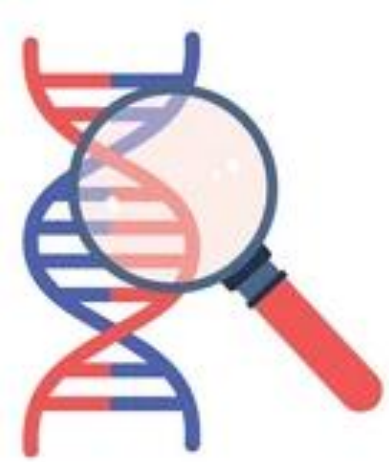


Markus Braun

Data analysis in a nutshell



How to detect viruses that have not been observed yet?



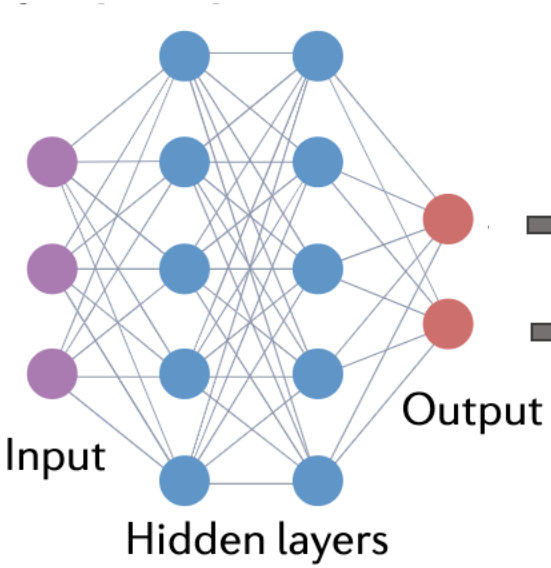
Virus ?

Machine Learning – model training

Input data

```
CACCGATANATCTTTCCCAGCGCGTTGCA
TTGTTACAGCGAGTTGCCCTCGCATCCGA
CCGGCCCCCCGCCCCACCTCTGGAGATCT
TCAAACCCAGGATCTGTGCCTTGGTCGTC
TTTGAAGCCCCGGCCCCTGGCACAGCGCC
TGGCAC
```

Model



Results

75% Virus

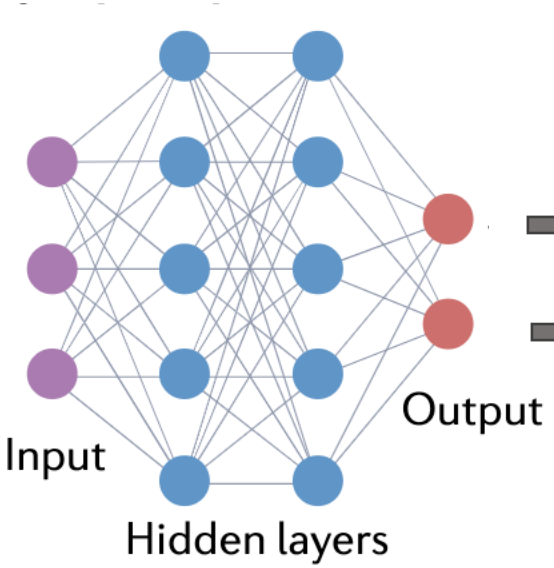
25% Not a virus

Machine Learning – model training

Input data

```
CACCGATANATCTTTCCCAGCGCGTTGCA
TTGTTACAGCGAGTTGCCCTCGCATCCGA
CCGGCCCCCCGCCCCACCTCTGGAGATCT
TCAAACCCAGGATCTGTGCCTTGGTCGTC
TTTGAAGCCCCGGCCCCTGGCACAGCGCC
TGGCAC
```

Model



Results

75% Virus

25% Not a virus

Expected results:

100% Virus

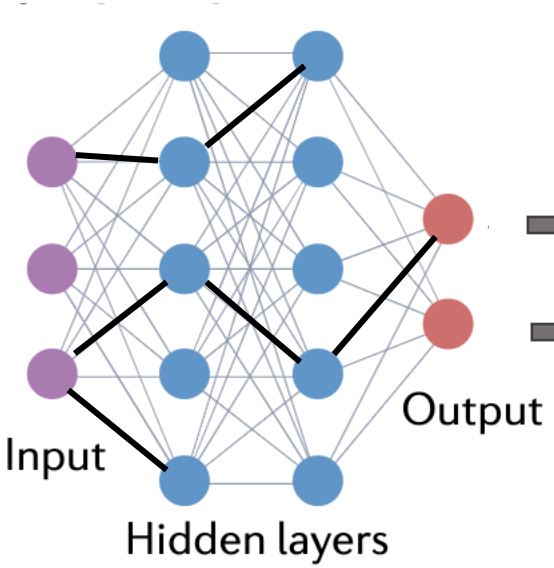
0% Not a virus

Machine Learning – model training

Input data

```
CACCGATANATCTTTCCCAGCGCGTTGCA
TTGTTACAGCGAGTTGCCCTCGCATCCGA
CCGGCCCCCCGCCCCACCTCTGGAGATCT
TCAAACCCAGGATCTGTGCCTTGGTCGTC
TTTGAAGCCCCGGCCCCTGGCACAGCGCC
TGGCAC
```

Model



Results

75% Virus

25% Not a virus

Expected results:

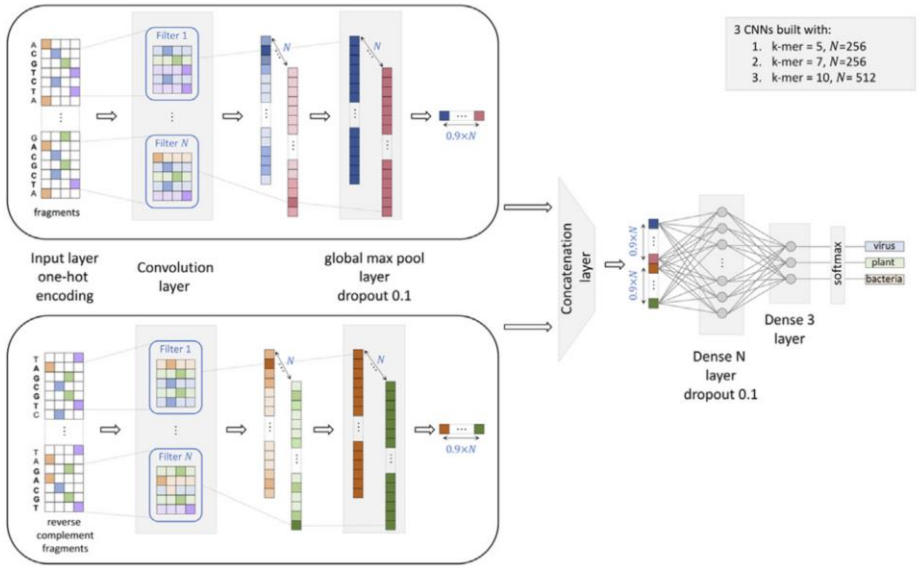
100% Virus

0% Not a virus

ML architectures

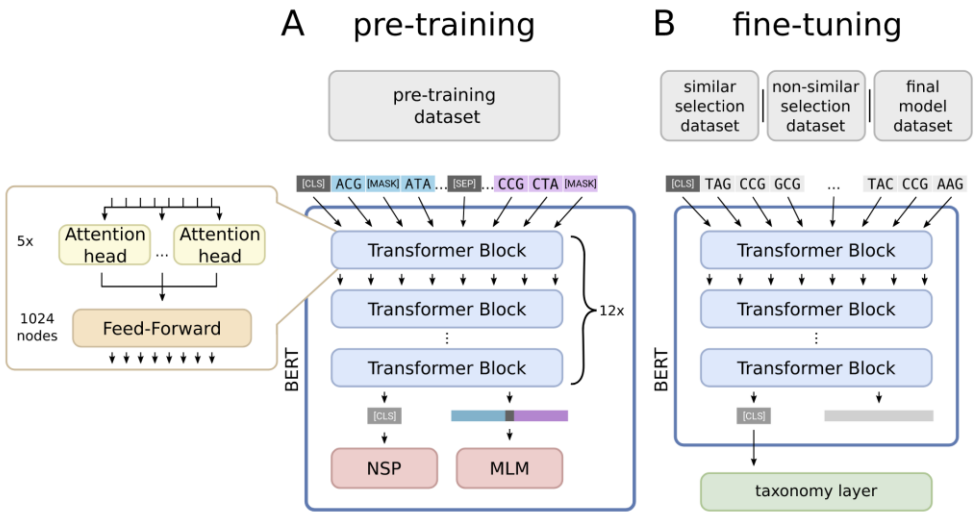
VirHunter *(Sukhorukov et al. 2022)*

- 3x CNN + random forest classifier



BERTax *(Mock, Kretschmer et al. 2022)*

- BERT transformer model



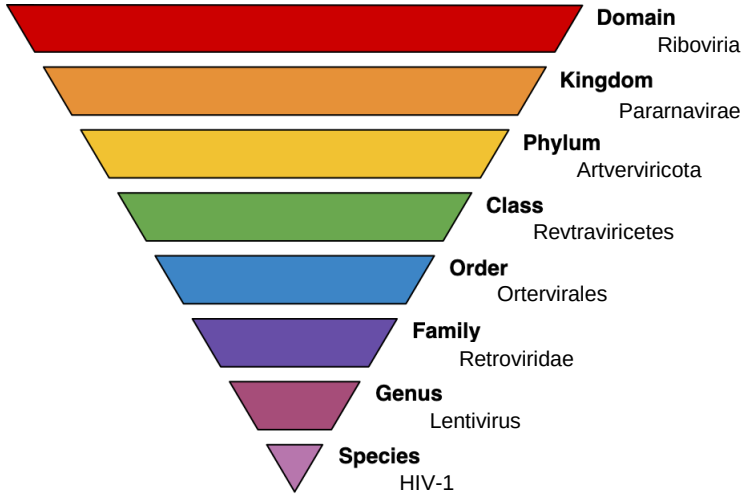
Selected candidate ML architectures

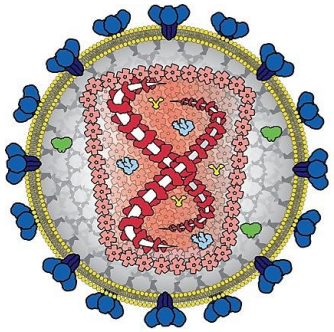
Model Name	Year Published	Architecture	Code Availability
VirHunter	2022	3x CNN + Random Forest	https://github.com/cbib/virhunter
BERTax	2022	Transformer BERT	https://github.com/f-kretschmer/bertax
Deep6	2023	CNN	https://github.com/janfelix/Deep6
ViraMiner	2019	2x CNN	https://github.com/NeuroCSUT/ViraMiner
Virtifier (Seq2Vec)	2021	LSTM	https://github.com/crazyinter/Seq2Vec

Evaluating ML architectures with Leave-One-Out strategy

Training:

- Human
- Bacteria
- Viruses - Leave Out one Family of viruses

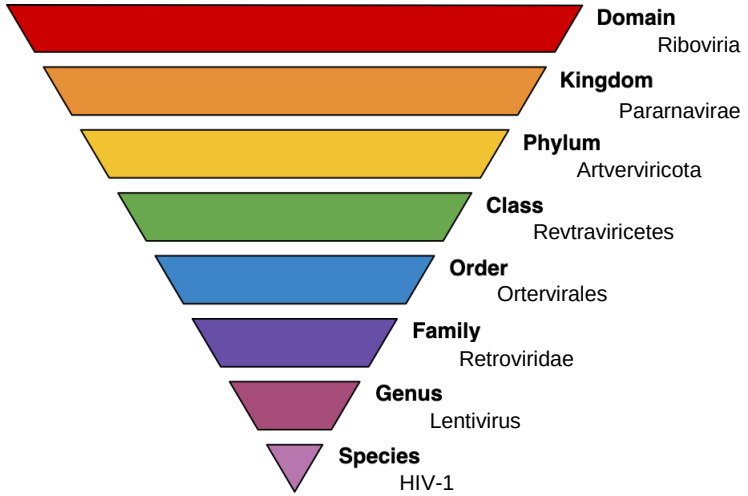


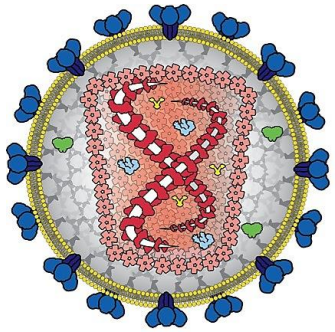


Evaluating ML architectures with Leave-One-Out strategy

Training:

- Human
- Bacteria
- Viruses - Leave Out one Family of viruses





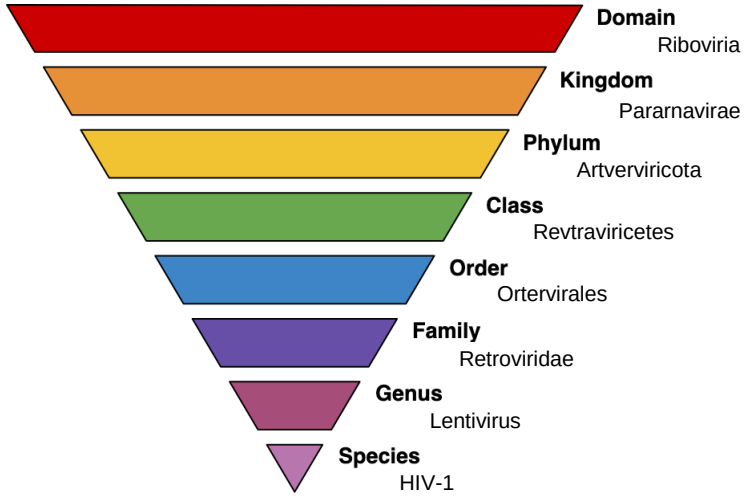
1	Chrysoviridae	Riboviria
2	Geminiviridae	Monodnaviria
3	Fimoviridae	Riboviria
4	Botourmiaviridae	Riboviria
5	Picornaviridae	Riboviria

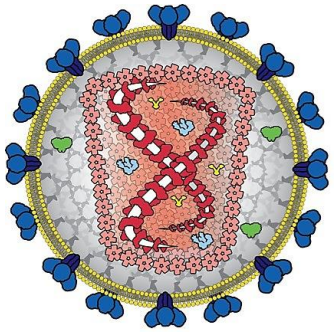
6	Picobirnaviridae	Riboviria
7	Tospoviridae	Riboviria
8	Potyviridae	Riboviria
9	Cruciviridae	DNA viruses
10	Secoviridae	Riboviria

Evaluating ML architectures with Leave-One-Out strategy

Training:

- Human
- Bacteria
- Viruses - Leave Out one Family of viruses



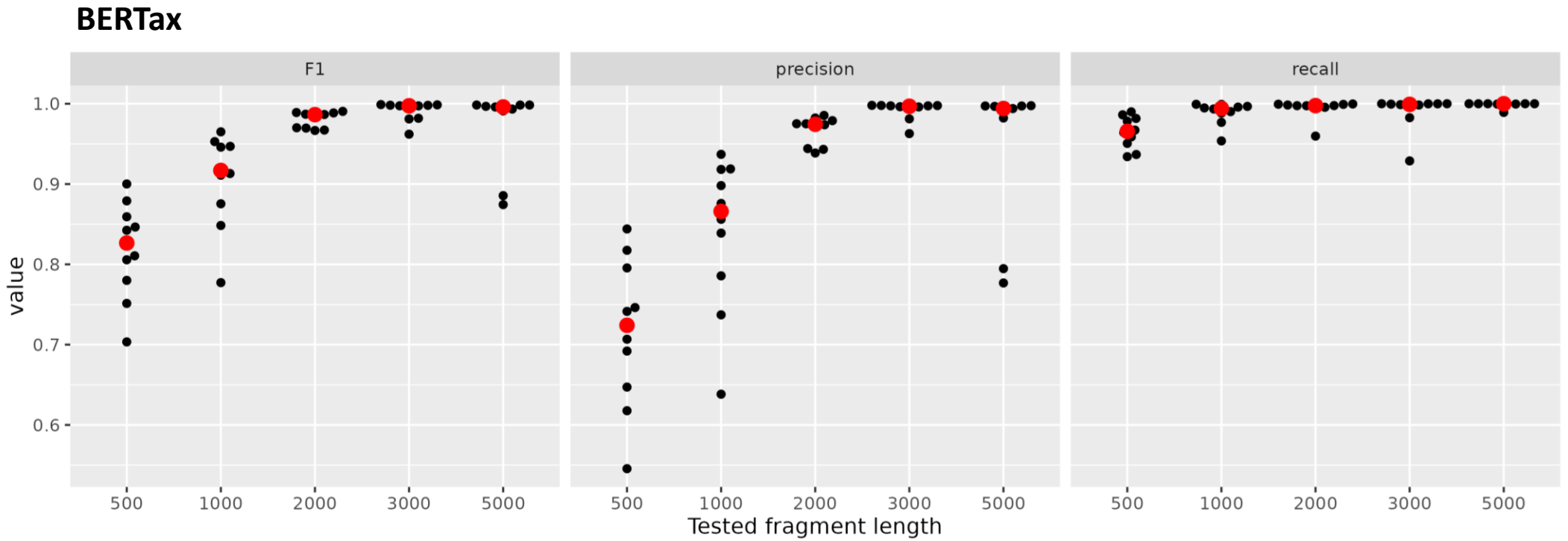


Evaluation:

- Contigs: Human, Bacteria, left out Viral Family
- Number of contigs: 10,000 each
- Contigs length: 0.5 - 5kb

- Evaluation metrics: Precision, Recall, F1

Accuracy of novel virus prediction is contig length dependent



Precision = How many false positive virus contigs?
Recall = How many false negative virus contigs?
F1 = (2 x precision x recall) / (precision + recall)

VirHunter shows best performance in novel virus prediction

Precision	500	1000	2000	3000	5000
VirHunter	0.942119	0.9629739	0.9817503	0.9885331	0.989903
BERTax	0.7240722	0.8659623	0.97442	0.9968432	0.9940233
Deep6	0.994781	0.9963495	0.9969462	0.9942555	0.9862932
Virtifier	0.931027	0.9562499	0.9654454	0.9418713	0.9350113
ViraMiner	0.8770039	0.9315166	0.9485944	0.9521701	0.9453604

Recall	500	1000	2000	3000	5000
VirHunter	0.9876357	0.9944948	0.9998499	1	1
BERTax	0.96565	0.99415	0.99765	0.9992	1
Deep6	0.8875513	0.9434528	0.9838685	0.9959537	1
Virtifier	0.8141	0.89565	0.92635	0.8999	0.9299
ViraMiner	0.8428	0.9116	0.9581	0.9768442	0.9989

F1	500	1000	2000	3000	5000
VirHunter	0.9636378	0.9784936	0.9900745	0.9936862	0.9949259
BERTax	0.826529	0.9168926	0.9864239	0.9974045	0.9959579
Deep6	0.9342183	0.9695196	0.9909379	0.992236	0.992008
Virtifier	0.8633674	0.9283334	0.9480911	0.9197223	0.8666775
ViraMiner	0.8541662	0.9214337	0.9534818	0.950492	0.9241737

	precision	recall	F1
VirHunter	0.97305586	0.99639608	0.9841636
BERTax	0.9110642	0.99133	0.94464158
Deep6	0.99372508	0.96216526	0.97578396
Virtifier	0.94592098	0.89318	0.90523834
ViraMiner	0.93092908	0.93764884	0.92074948

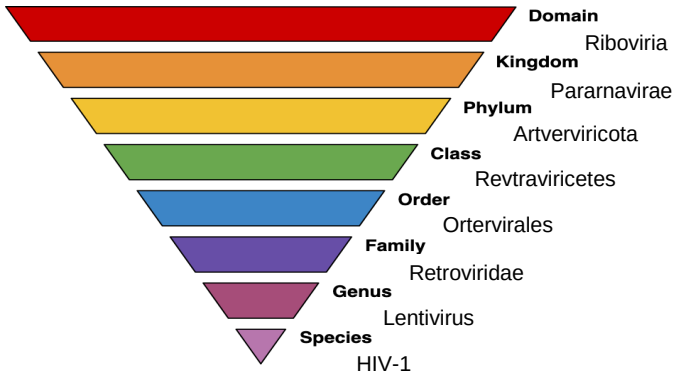
VirHunter shows best performance in novel virus prediction

Family-level testing

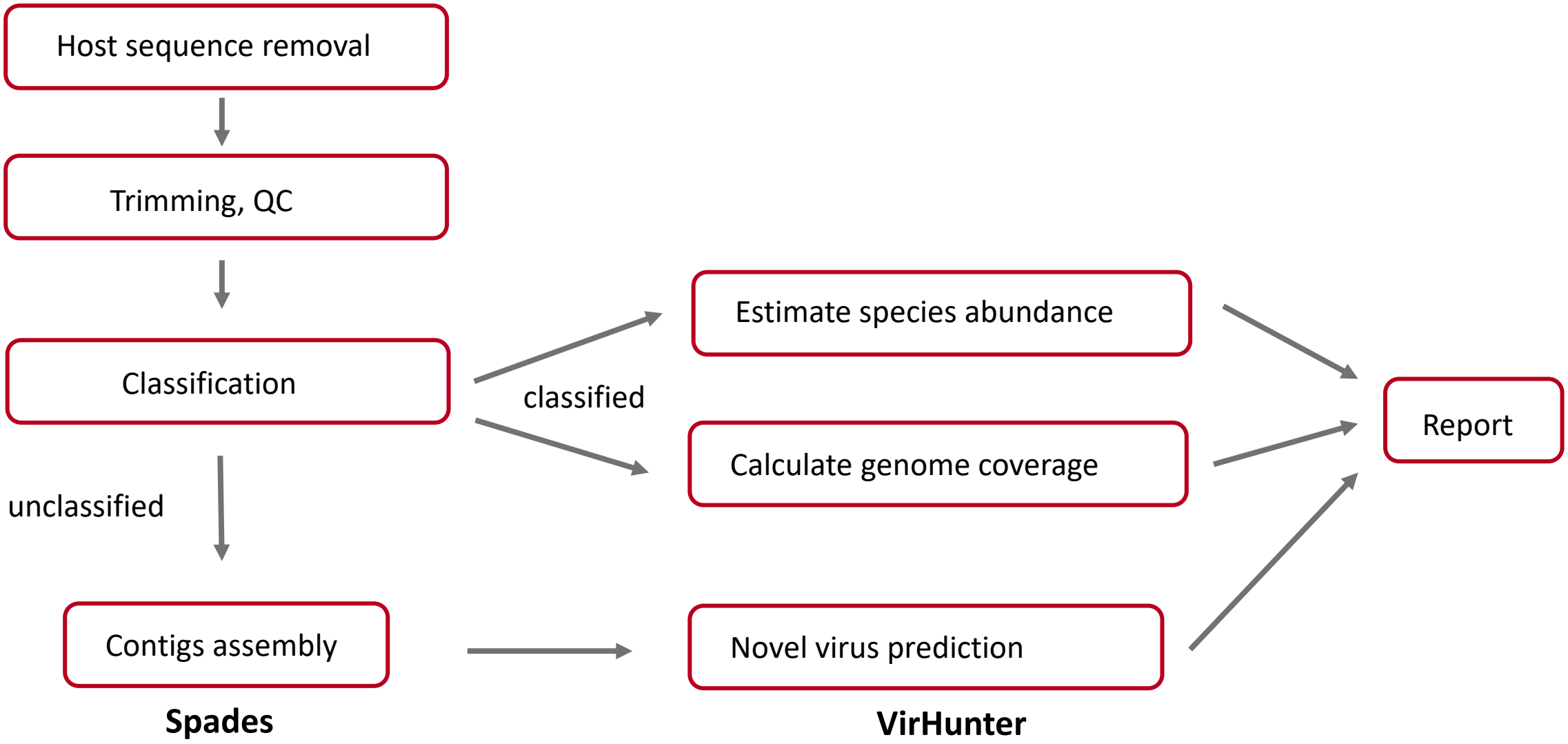
	precision	recall	F1
VirHunter	0.97305586	0.99639608	0.9841636
BERTax	0.9110642	0.99133	0.94464158
Deep6	0.99372508	0.96216526	0.97578396
Virtifier	0.94592098	0.89318	0.90523834
ViraMiner	0.93092908	0.93764884	0.92074948

Class-level testing

	precision	recall	F1
VirHunter	0.9711604	0.994151	0.9802635
BERTax	0.92200458	0.98743516	0.9492529
Deep6	0.99402326	0.9430469	0.96282532
Virtifier	0.94146532	0.83847706	0.8743416
ViraMiner	0.9320866	0.95767938	0.92969558



Data analysis in a nutshell



Conclusions

- BLOODVIR pipeline implemented in snakemake for reproducible execution
- Sequence classification with MiCoP (bwa) is more accurate than k-mer based solutions
- Linear detection response with 100 virus copies per mL detection limit
- Prediction of novel viruses using custom trained VirHunter ML model

BLOODVIR team

Paul-Ehrlich-Institut

Renate König

Markus Braun

Martin Machyna

Liam Childs

Leona Enke

Arezoo Jamali

Zsófia Nacsa

Erica Margiotta

Jana Schatzl

Johannes Blümel

Gibran Horemheb Rubio Quintanares

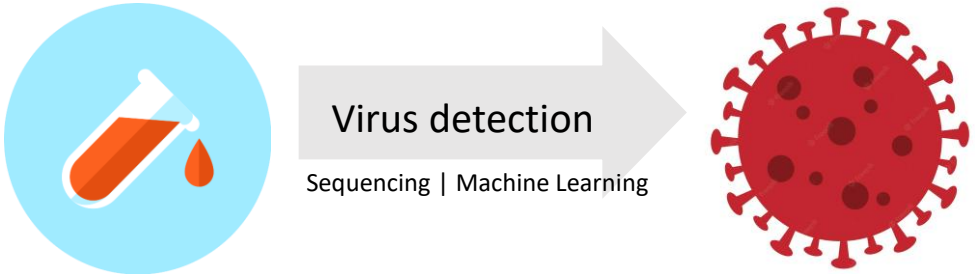
Janice Brückmann

Csaba Miskey

Dóra Spekhardt

Pauline Santos

Bundesministerium für Gesundheit



This work is supported by the Federal Ministry of Health of Germany