BioReliance®
Contract Testing Services

Rapid alignment-free detection of adventitious agents using next-generation sequencing

Tom J.B. de Man
Frankfurt, Germany, 5 December 2024

MERCK

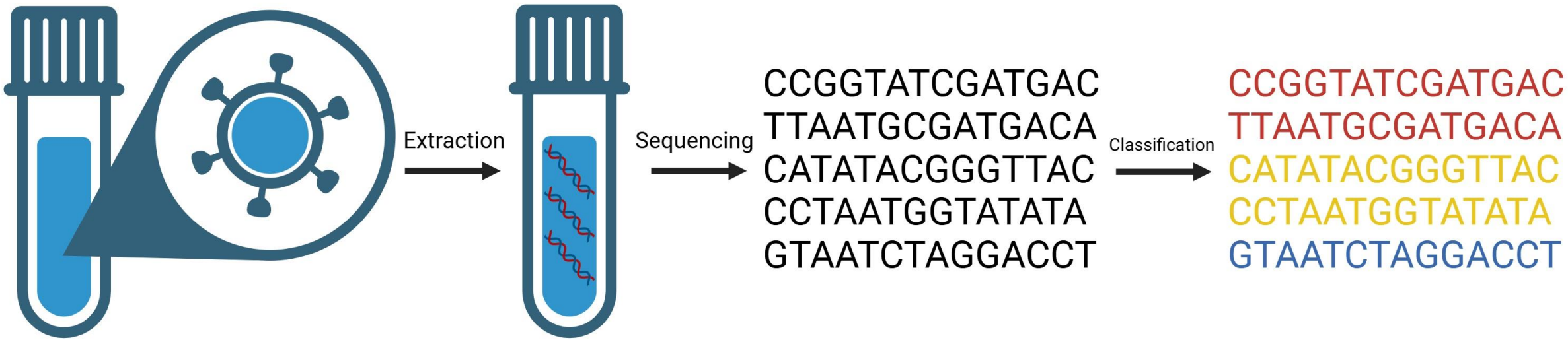
BioReliance®

Table of Contents

MERCK

Metagenomics Next Generation Sequencing (mNGS) flowchart

BioReliance®

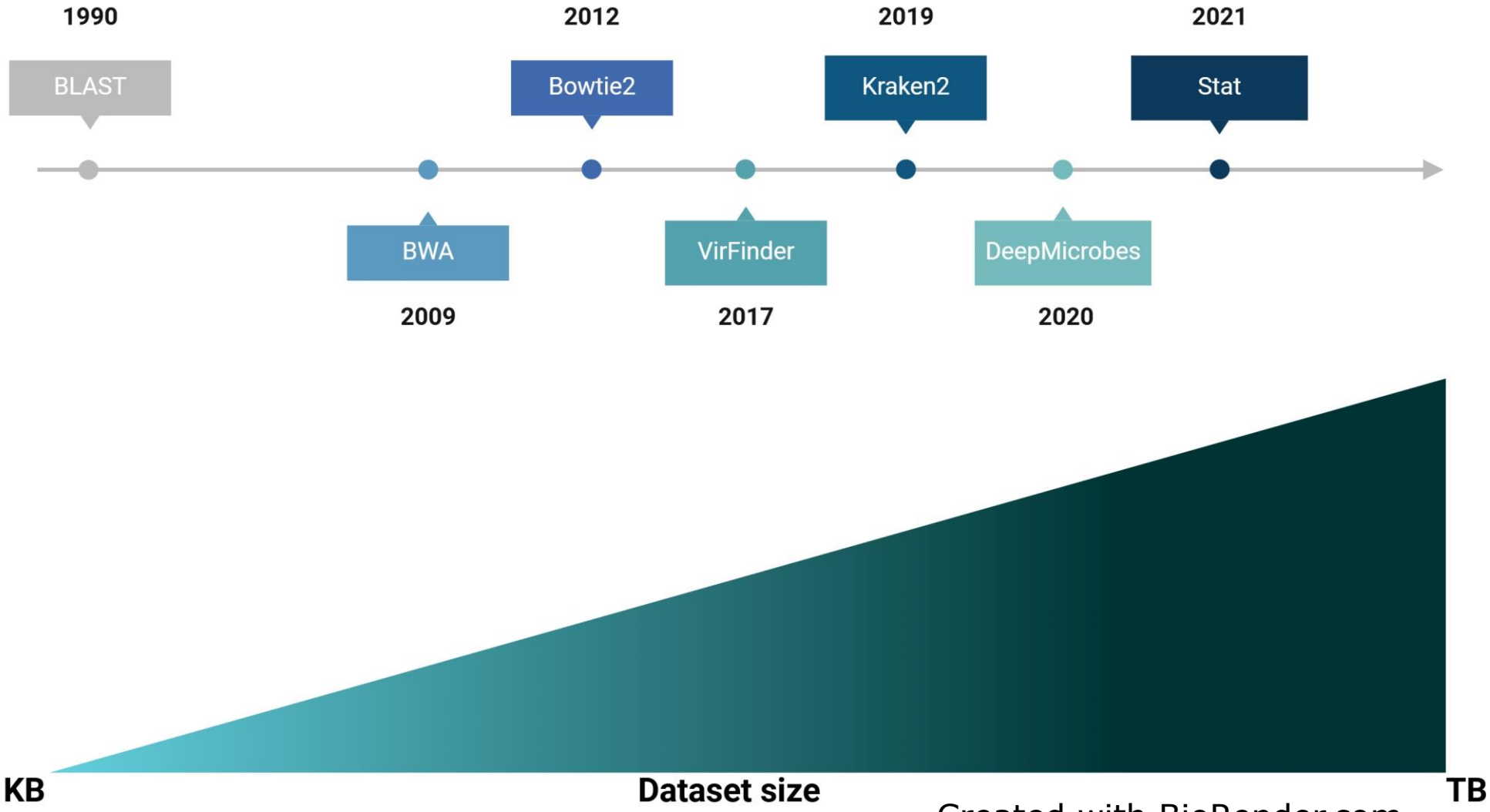


Created with BioRender.com

MERCK

More genomes and sequencing depth = a new type of algorithm

BioReliance®



Created with BioRender.com

MERCK

Evolution of data and bioinformatics tools at Merck

BioReliance®

- 454 data
- Average reads per run: 1M

- Illumina MiSeq
- Average reads per run: 25M

- Illumina NextSeq2000
- Average reads per run: 400M

MERCK

From read mapping to metagenomic read classification

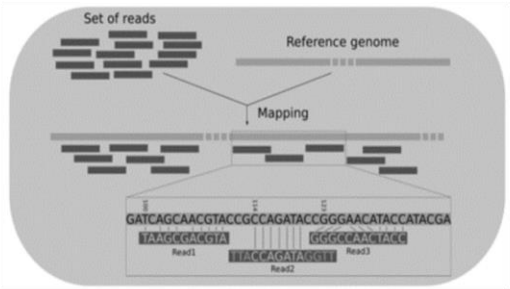
BioReliance®

Pipeline improvements

Current AAT algorithm

Capacity

1 TB database



Constraints

- Large reference databases
- Lengthy database build-time
- Produces redundant results
- Computationally intensive
- Prolonged analysis runtime
- Limited to small genomes

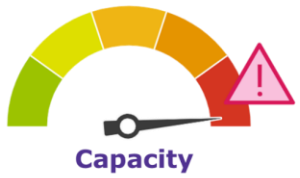
MERCK

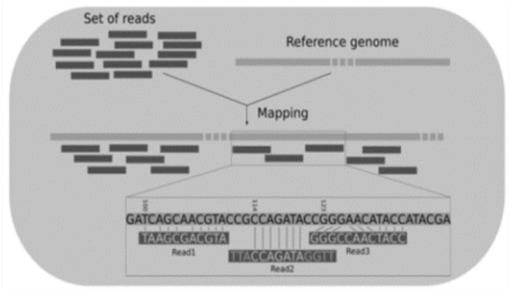
From read mapping to metagenomic read classification

BioReliance®

Pipeline improvements

Current AAT algorithm

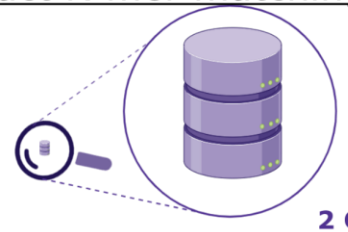




Constraints

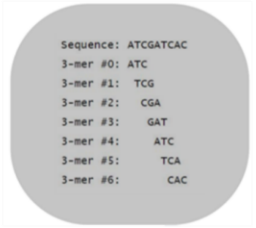
- Large reference databases
- Lengthy database build-time
- Produces redundant results
- Computationally intensive
- Prolonged analysis runtime
- Limited to small genomes

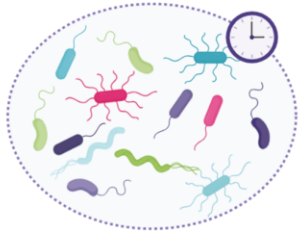
Revised pipeline with exact K-mer matching

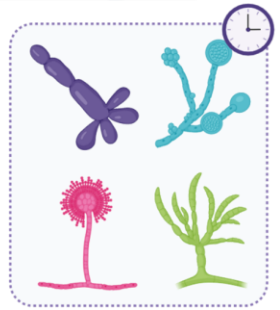


Improvements

- Smaller reference databases
- Reduced database build-time
- Avoids duplicate results
- More binary decision making for Study Management
- Reduced compute resource needs
- Minimized analysis runtime
- Ability to analyze larger genomes







Created with BioRender.com

MERCK

What is a k-mer

BioReliance®

- A k-mer is a subsequence of length k derived from a longer nucleic acid or protein sequence.
- K-mers are utilized in bioinformatics for tasks such as genome assembly, sequence alignment, and identifying genetic variations.
- DNA sequence: ATCGTAGC
  - 1st k-mer: ATC
  - 2nd k-mer: TCG
  - 3rd k-mer: CGT
  - 4th k-mer: GTA
  - 5th k-mer: TAG
  - 6th k-mer: AGC
- Each k-mer is derived by moving one nucleotide to the right.

MERCK

Why are mNGS k-mer-based algorithms so fast?

BioReliance®

- Speed advantage derives in large part from the use of **exact-match** database queries of *k*-mers, rather than inexact alignment of sequences

AAAAAAAAATGCGCGTAGCTGACGTGCAACGTGCACGTC

AAAATTAAATGCGCGTAGCTGACGTGCGGGGTGCACATC

Metagenomics classifier: **"No exact match, let's move to the next k-mer"**

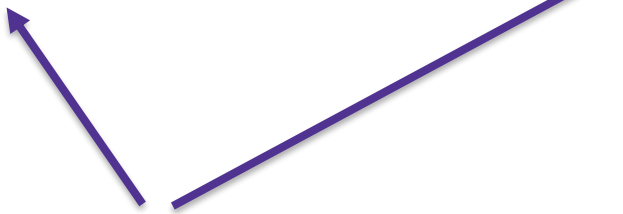
Read mapper: **"hold on, I can maybe align this and check against my gap open penalty, gap extension penalty, my mismatch penalty, my overall matching score. No worries, I only have to do all these steps for 250 million sequences"**

MERCK

Compression of data: Lowest Common Ancestor (LCA)

BioReliance®

• **Domain:**	Eukaryota	Virus	Virus
• **Kingdom:**	Animals	Orthornavirae	Orthornavirae
• **Phylum:**	Chordata	Kitrinoviricota	Kitrinoviricota
• **Class:**	Mammalia	Flasuviricetes	Flasuviricetes
• **Order:**	Primates	Amarillovirales	Amarillovirales
• **Family:**	Hominidae	Flaviviridae	Flaviviridae
• **Genus:**	Homo	Flavivirus	Hepacivirus
• **Species:**	H. sapiens	Zika virus	Hepatitis C virus



AAAAAAAAATGCGCGTAGCTGACGTGCAACGTGCACGTC

MERCK

Benchmarking the new k-mer algorithm – Materials & Methods

BioReliance®

- Data from "**Rebecca Bova *et al*. (2024). *Biologicals*, 86, 101771**"

- **Four cell lines commonly used in biomanufacturing:**
  - HeLa cells (human)
  - CHO (Chinese hamster)
  - Vero (non-human primate)
  - Sf9 (insect)
- **Three representative viruses:**
  - Mammalian Orthoreovirus 1 (REO1)
  - Feline Leukemia Virus (FeLV)
  - Human Respiratory Syncytial Virus (RSV)

- Four spike levels per virus; 1E+06, 1E+05, 1E+04, or 1E+03 Viral Genome Copies (VGC)/1E+06 cell-line cells + unspiked control, all in triplicate.
- Operate the new mNGS algorithm + data qualification

MERCK

Benchmarking the new k-mer algorithm – Results

BioReliance®

- Specificity of the AAT assay was evaluated by analysis of the unspiked cells and virus spiked cells at spike level 1E+05 VGC

Spike Conc. VGC/1E6 cells	Cell Line	# Replicates with Sequence Match by Virus		
		RSV	FeLV	REO
Unspiked	CHO	0/3	0/3	0/3
	HeLa	0/3	0/3	0/3
	VERO	0/3	0/3	0/3
	SF9	0/3	0/3	0/3
1.00E+05	CHO	3/3	3/3	3/3
	HeLa	3/3	3/3	3/3
	VERO	3/3	3/3	3/3
	SF9	3/3	3/3	3/3

MERCK

Benchmarking the new k-mer algorithm – Results

BioReliance®

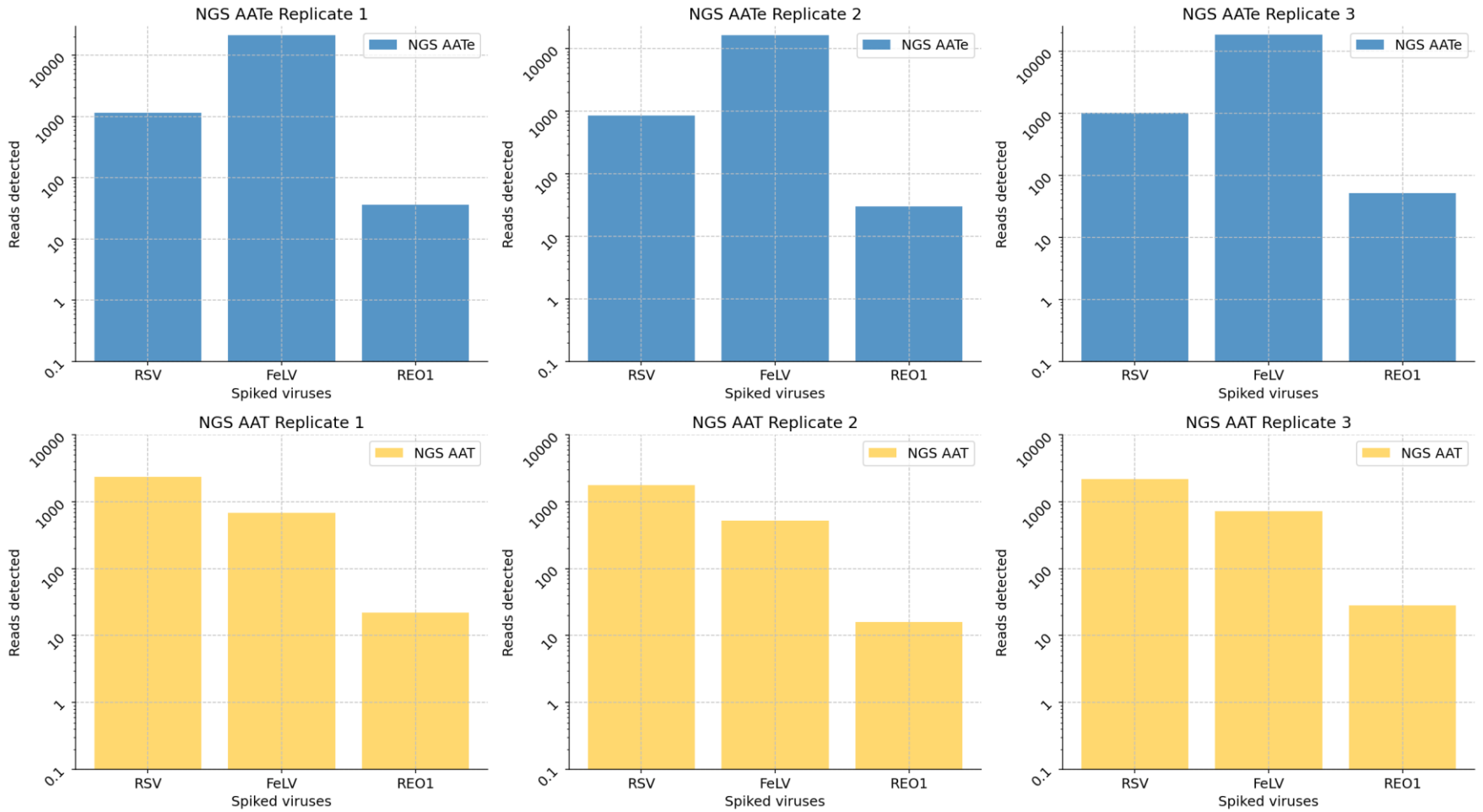
- The limit of detection (LOD) of the AAT assay was determined to be as follows for all four cell lines spiked with the three representative viruses.
  - RSV: 1E+03 VGC/1E6 cells
  - FeLV: 1E+04 VGC/1E6 cells
  - REO 1: 1E+04 VGC/1E6 cells
- The overall LOD of the assay was determined to be **1E+04 Viral Genome Copies per 1E6 cells** across all four cell lines tested.

Spike Conc. VGC/1E6 cells	Cell Line	# Replicates with Sequence Match by Virus		
		RSV	FeLV	REO
1.00E+05	CHO	3/3	3/3	3/3
	HeLa	3/3	3/3	3/3
	VERO	3/3	3/3	3/3
	SF9	3/3	3/3	3/3
1.00E+04	CHO	3/3	**2/3**	3/3
	HeLa	3/3	3/3	**2/3**
	VERO	3/3	3/3	3/3
	SF9	3/3	3/3	**2/3**
1.00E+03	CHO	3/3	0/3	**2/3**
	HeLa	3/3	**2/3**	0/3
	VERO	3/3	**2/3**	1/3
	SF9	3/3	3/3	1/3

MERCK

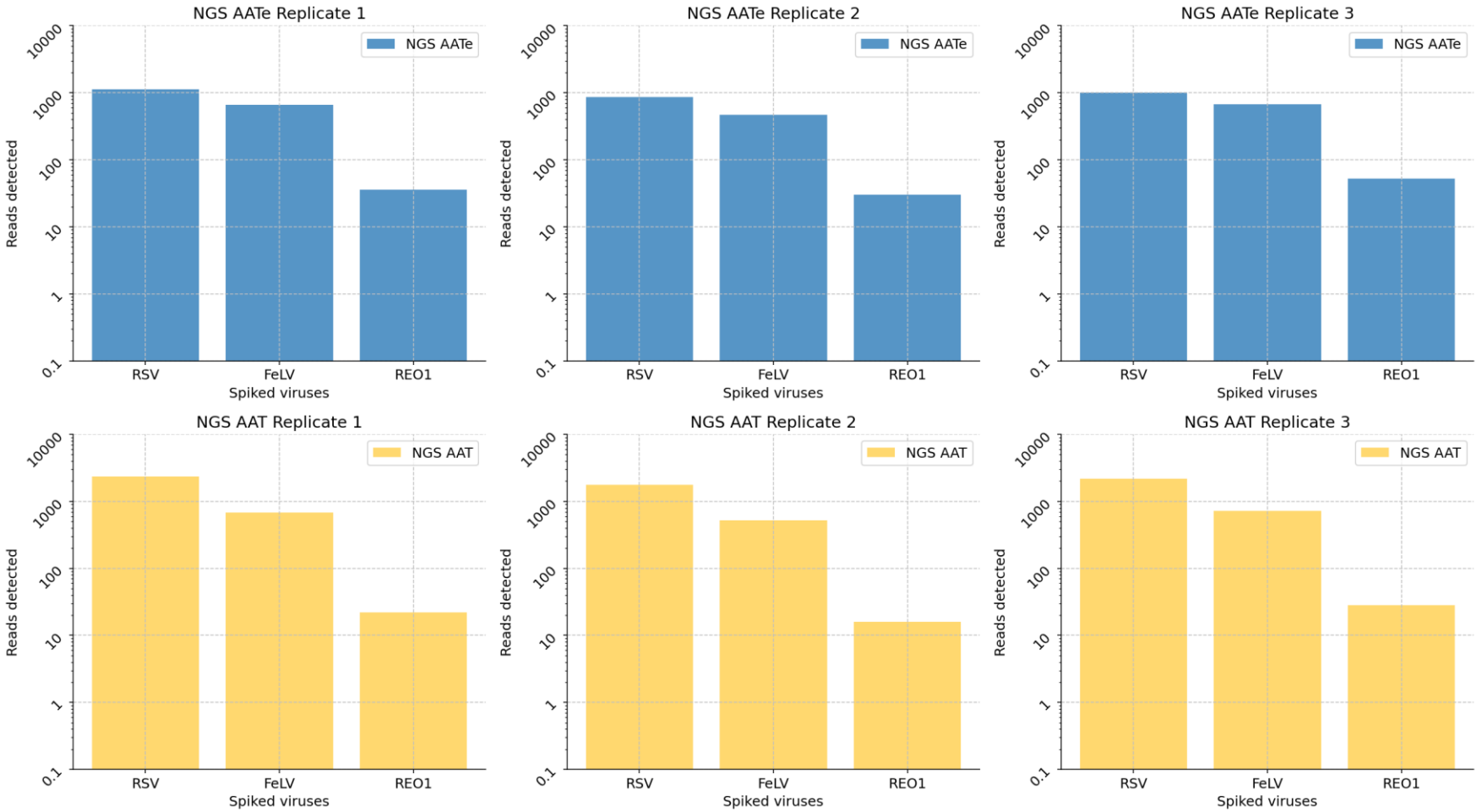
Deep dive 1: 1E+05 VGC/1E6 cells: HeLa

BioReliance®



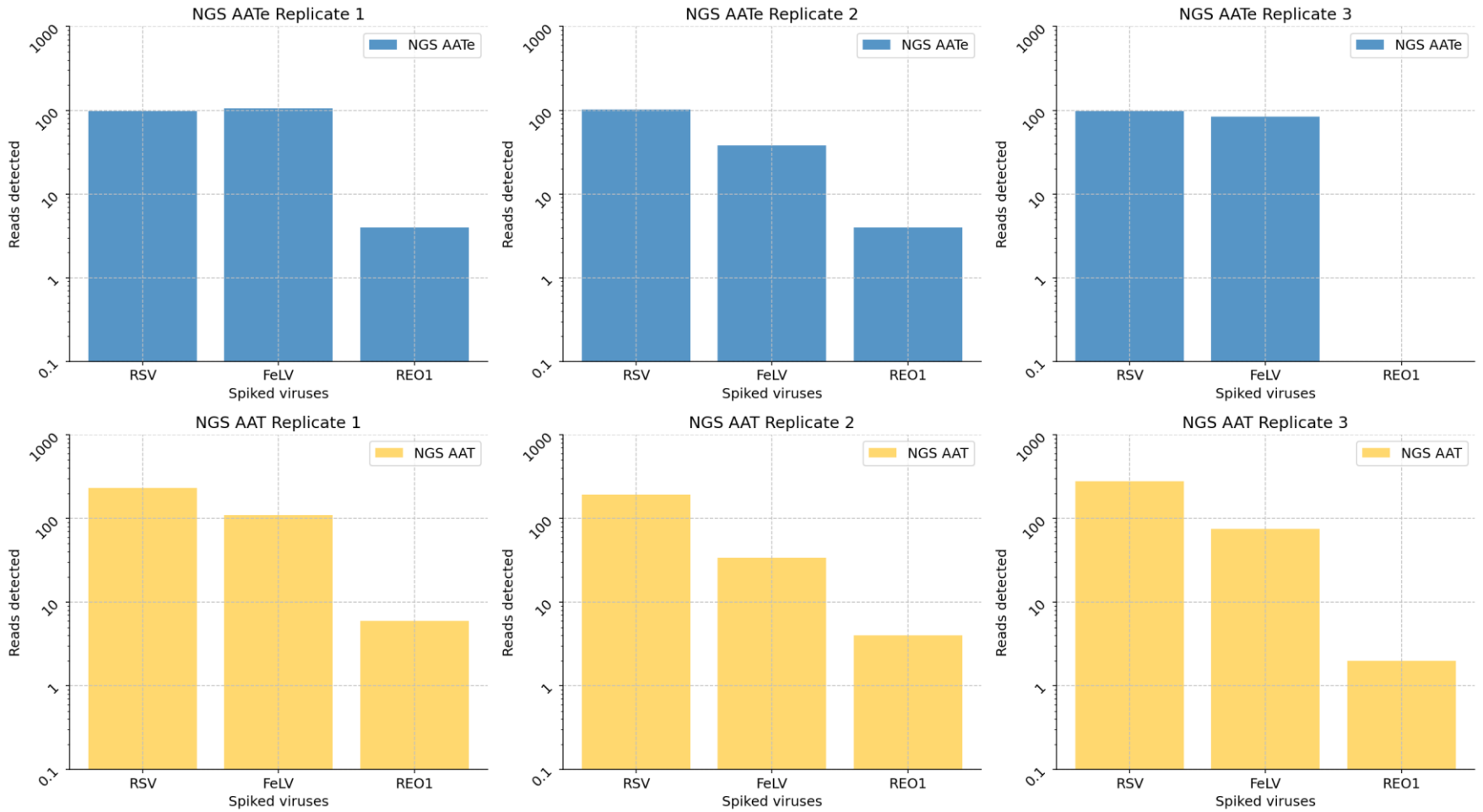
MERCK

Deep dive 2: 1E+05 VGC/1E6 cells: HeLa + human genome filter



Deep dive 3: 1E+04 VGC/1E6 cells: SF9

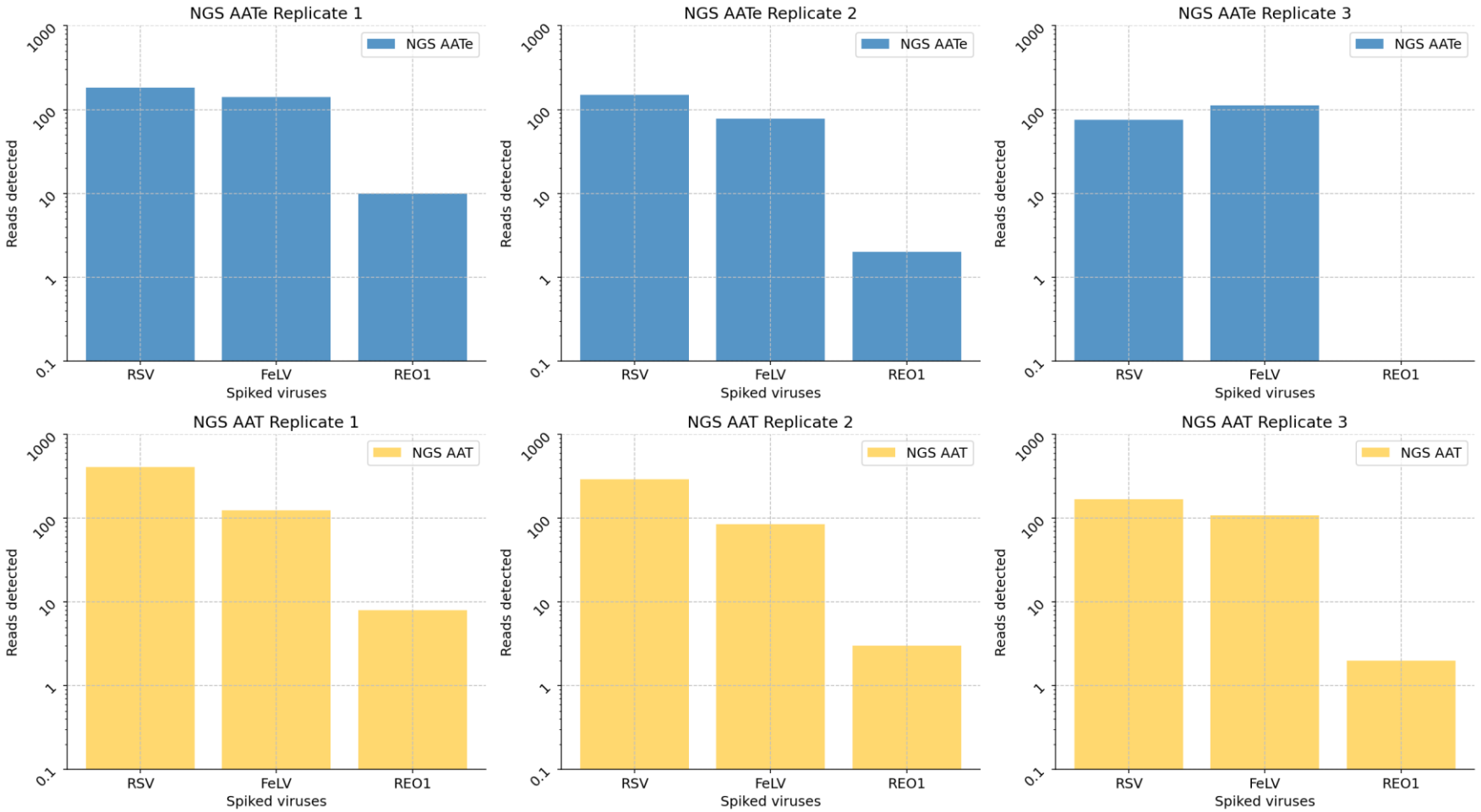
BioReliance®



MERCK

Deep dive 4: 1E+04 VGC/1E6 cells: HeLa + human genome filter

BioReliance®



MERCK

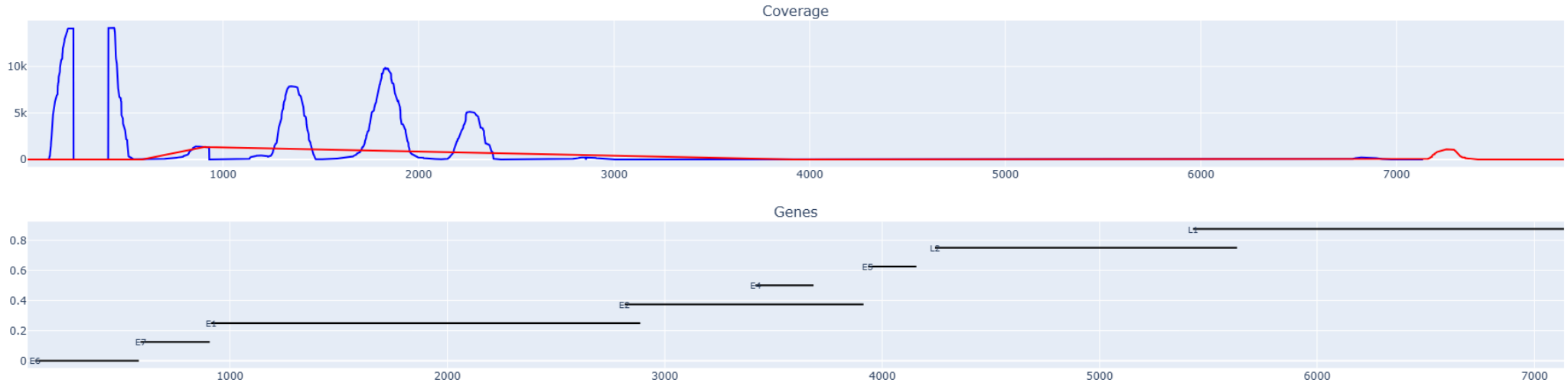
mNGS AAT data qualification

BioReliance®

- Each positive result, of a virus potentially infectious to the test sample and/or capable of harming human health, would need to be confirmed through additional comparative analyses
  - **The results contain read classifications but not their aligned positions in the genomes**

- Do sequencing reads originate from across the pathogen's genome?

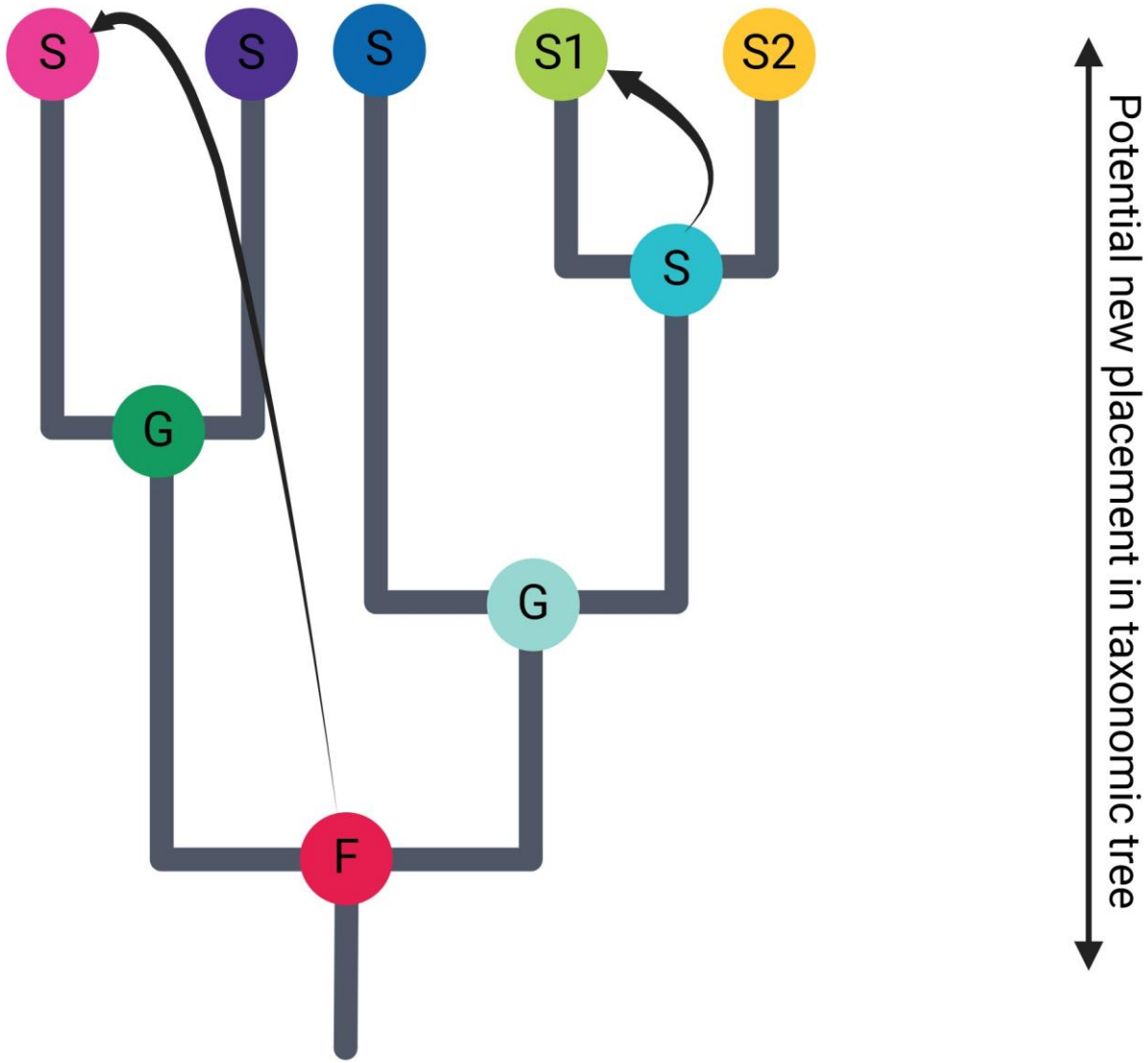
- Human papillomavirus 18 (E6, E7, and E1 genes) is integrated into the HeLa cell line, but it can also infect HeLa cells



MERCK

Data qualification confirms the taxon's place in the taxonomic tree

BioReliance®

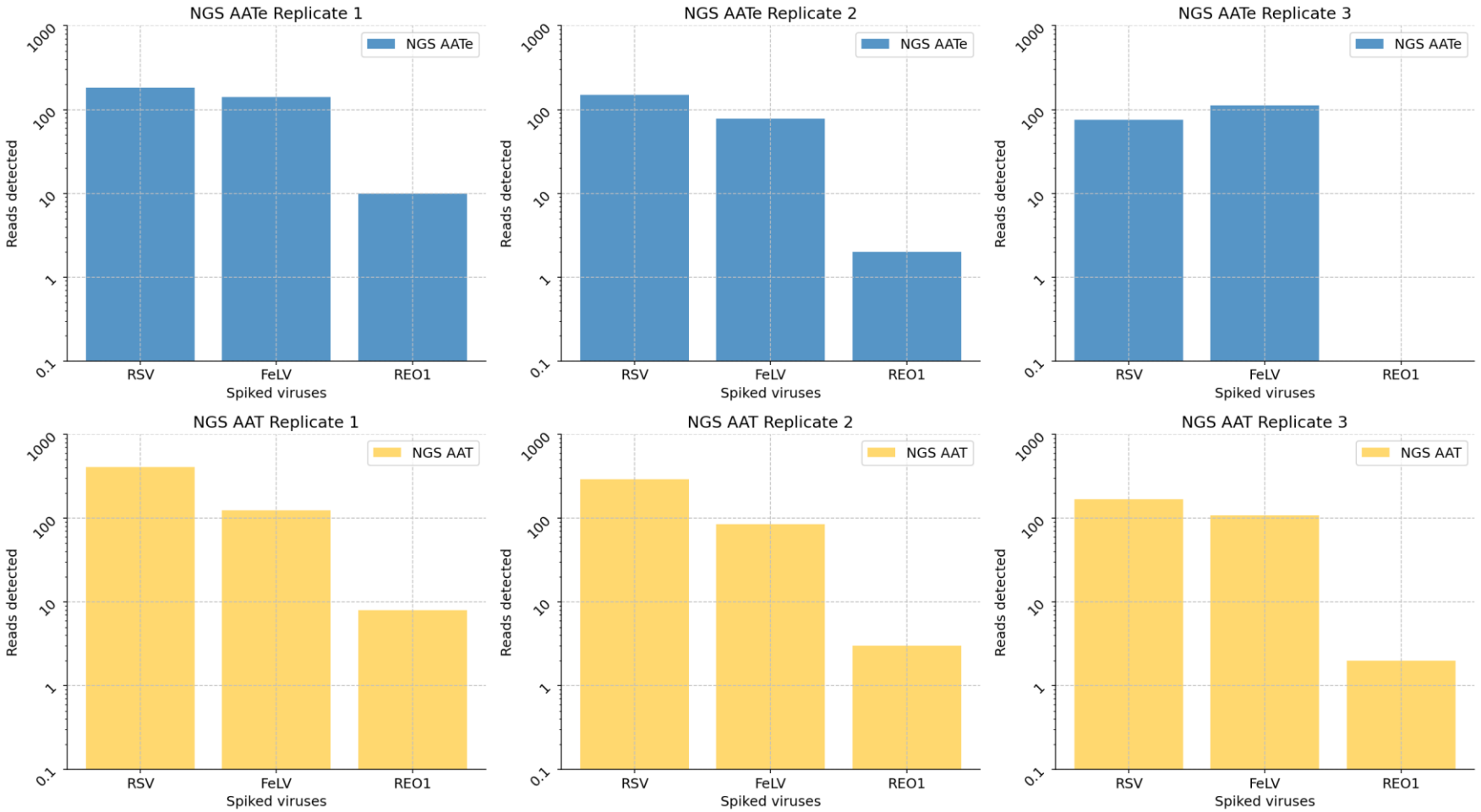


Created with BioRender.com

MERCK

Deep dive 4: 1E+04 VGC/1E6 cells: HeLa + human genome filter

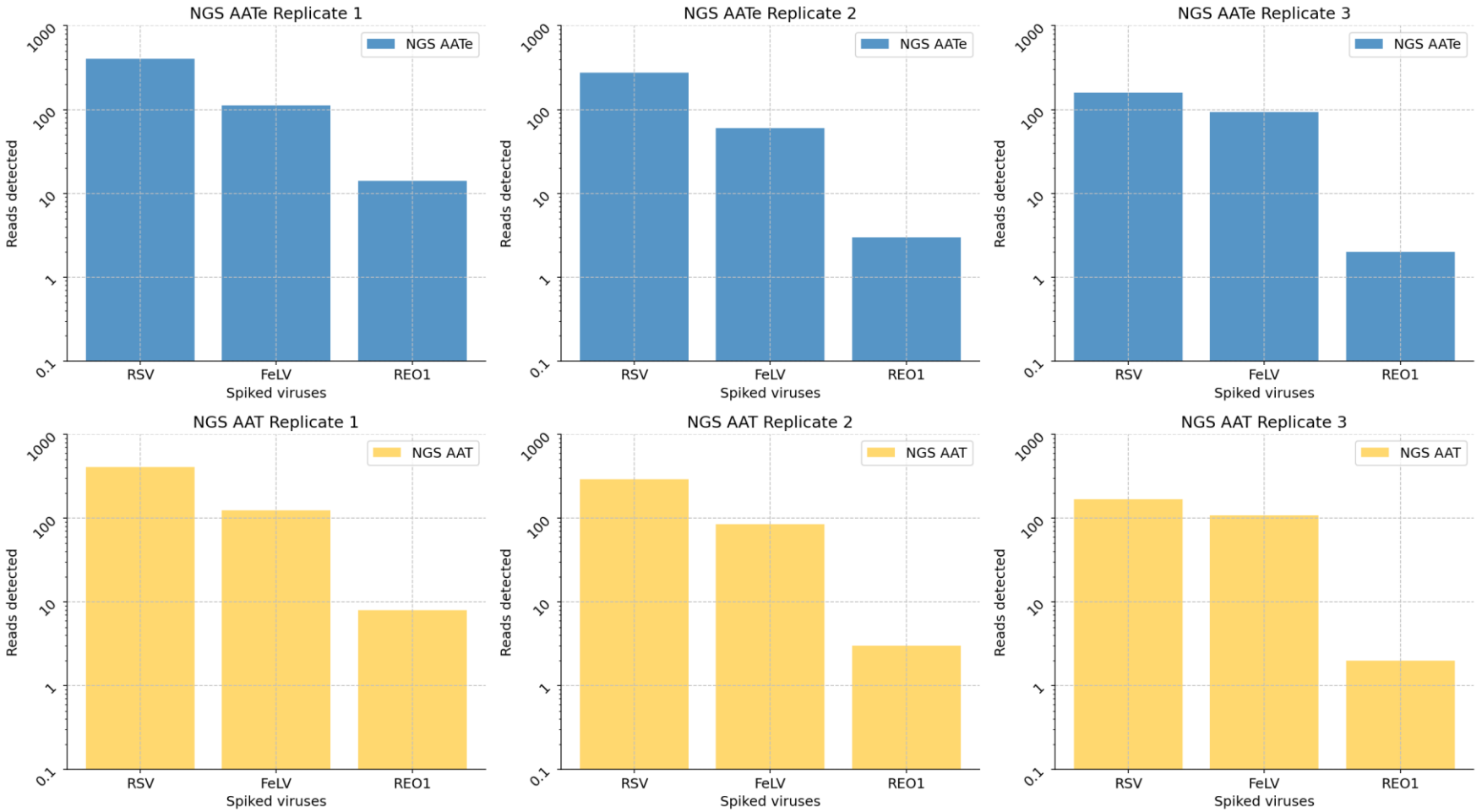
BioReliance®



MERCK

Deep dive 5: 1E+04 VGC/1E6 cells: HeLa + human genome filter + data qualification

BioReliance®

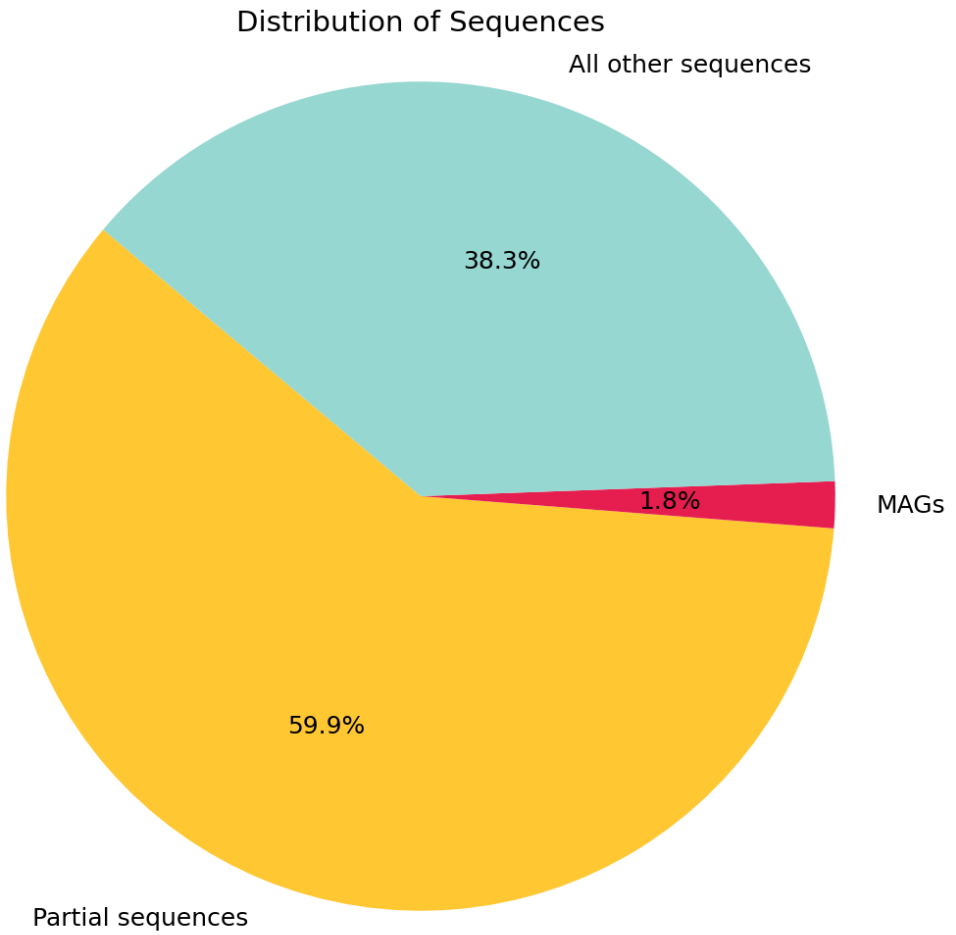


MERCK

Genomic reference database curation is key for mNGS analyses

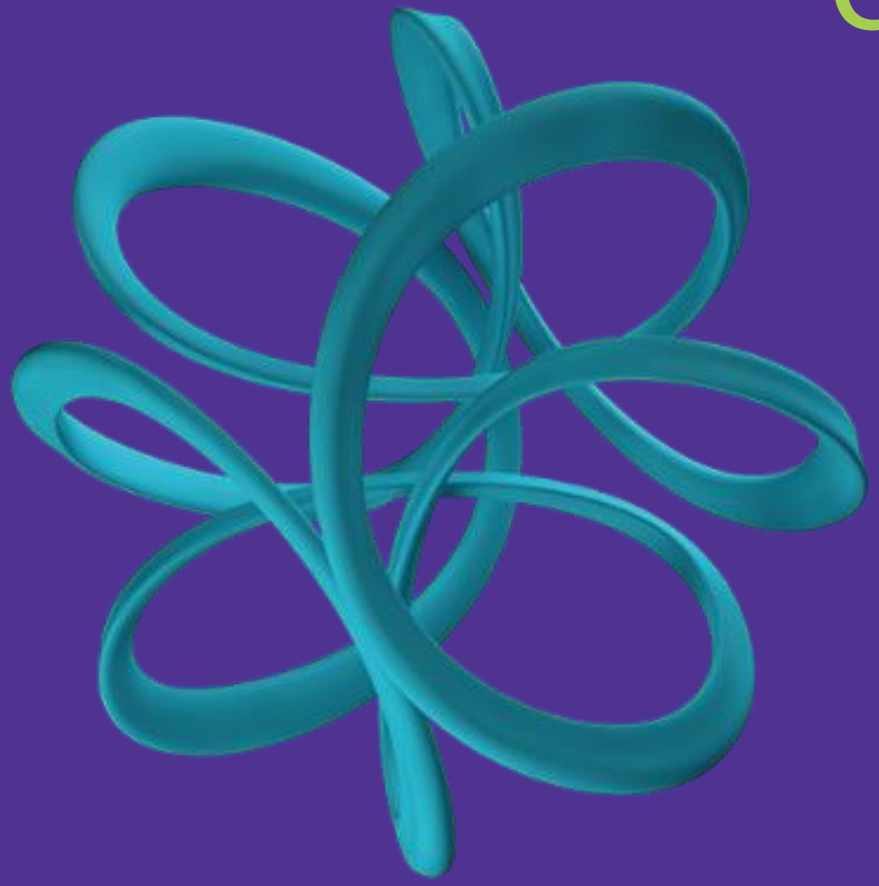
BioReliance®

- Metagenome assembled genomes (MAGs) are models
- MAGs are generated via numerous strategies, and it is not uncommon to pool data from hundreds of samples, using different NGS platforms, extraction and library preparation methods, and different quality control criteria
- MAGs are assembled from these data



MERCK

Conclusion



1. Employing a metagenomics classifier (k-mer) algorithm is future-proofing mNGS AAT operations
2. The overall LOD of the AAT assay was determined to be **1E+04 Viral Genome Copies per 1E6 cells** across four cell lines tested.
3. Cell-line genome filtering may reduce false positive detections
4. AAT data qualification can increase taxonomic resolution and relevancy of k-mer results

Tom J.B. de Man

tom.de-man@milliporesigma.com