Viral metagenomic analysis to complement the viral risk assessment and adventitious agent testing of live virus vaccines

Vanessa V. Sarathy, Jack Baker, Julia Maritz, Connor Geraghty

Analytical Research and Development

MSD

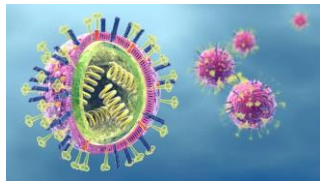
2024

Outline

1. Viral Metagenomic Analysis (VMA) to de-risk adventitious agents in live virus vaccines
2. HIVE has publicly-available tools for metagenomic data analysis
3. ViruScreen is a streamlined internal application for metagenomic data analysis
4. Benchmarking ViruScreen with HIVE
5. Comparison of different alignment workflows

Viral metagenomic analysis (VMA) for adventitious virus detection in products

- **Viral metagenomics** uses high throughput sequencing (HTS) or next generation sequencing (NGS) to identify contaminating viruses in biological products.
- NGS is unbiased and has shown increased sensitivity and detection capabilities
- The ICH Q5A Revision 2 includes recognition and support of NGS virus testing for inclusion in regulatory submissions for biotechnology products in scope.
- Several guidelines encourage NGS testing to replace *in vivo* testing.
- NGS testing can supplement gaps in traditional adventitious agent testing of live virus vaccines or as part of an early de-risking strategy for clinical studies.



Basic steps of viral vaccine metagenomic analysis study workflow

Sample processing → Nucleic acid library preparation → Nucleic acid sequencing → Data analysis using bioinformatics

- Study samples: Harvested virus fluids and control cell fluids as well as cells and media as controls.
- Viromic and/or transcriptomic analysis is performed based on risk assessment.
- Specificity and sensitivity are determined by spiking with a virus panel to determine recovery.
- High throughput sequencing: Illumina short reads, paired end, 2x150 bases.
- Bioinformatic application results are followed-up with manual analyses.

Spike viruses used to determine recovery of representative virus species

- Well-characterized virus stocks prepared by FDA/CBER available from ATCC for spiking into VMA samples
- Demonstrate broad virus detection in different matrices.
- Represent virus families of potential safety concern and with various physical and chemical properties.
- Have also used human rhinovirus 14 (+ssRNA, 7.2 kb, unenveloped) and porcine circovirus 2 *in silico* for analysis.

Description	Abbreviation	Genome type	Genome length	Virus particle size	Virus envelope	Chemical resistance
Porcine circovirus 1 (ATCC- FDA)	PCV1	DNA, ss, circular	1.8 kb	16-18 nm	No	High
Mammalian orthoreovirus type 1, strain Lang (ATCC- FDA)	Reo1 ("REO")	RNA, ds, linear (segmented)	23.6 kb (1196-3915 nt)	80 nm	No	Medium-high
Feline leukemia virus, strain Thielen (ATCC- FDA)	FeLV	RNA, ss, linear (dimeric)	8.5 kb	80-100 nm	Yes	Low
Human respiratory syncytial virus, strain A2 (ATCC- FDA)	RSVA2 ("HRSV")	RNA, ss, linear	15 kb	150-200 nm	Yes	Low-medium
Epstein-Barr virus (HHV-4), strain B95-8 (ATCC- FDA)	EBV ("HHV4")	DNA, ds, linear	172 kb	122-180 nm	Yes	Low-medium

Proposed 1st International Virus Reference Standards for Adventitious Virus Detection in Biological Products by Next-Generation Sequencing (NGS) Technologies (CBER-5). Arifa S. Khan and Study Group Participants. WHO Expert Committee on Biological Standardization. Geneva, 19-23 October 2020.

MSD

Overview of the short read alignment (SRA) workflow for Illumina sequences

Sample preparation and extraction.

Illumina high-throughput sequencing.

Trim adapter and low-quality sequences.

Virus harvests only: Stringent alignment to vaccine virus

Align to curated viral reference database and vaccine virus

Best hits within taxa reviewed across samples

Counterscreen 'Putative' positives

Identify bioinformatically-confirmed hits of interest

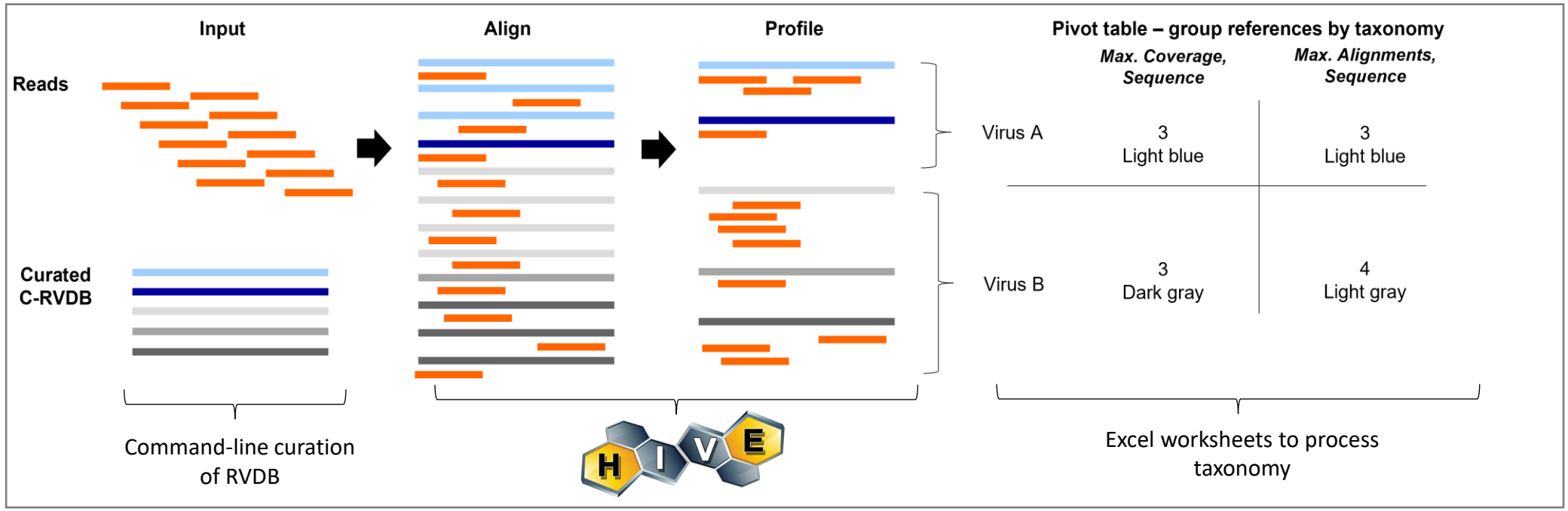
Launch investigation to confirm virus presence

MSD

Bioinformatic analysis environment used for characterization tests in regulatory filings

- HIVE:
  - High-performance Integrated Virtual Environment
  - HIVE Hexagon: HTS read aligner
  - HIVE Heptagon: Alignment profiler
- Highlights of HIVE:
  - Instances: US FDA, public domain
  - GWU maintenance, expert development, secure access
  - Well-known and published tools available in public instance
  - Generates BioCompute objects for pipeline communication

Analysis workflow using HIVE and manual processing: short read alignment to reference database and best hits filtering



- Several publications on HIVE: hive.biochemistry.gwu.edu/home
- FDA reference viral database: https://rvdb.dbi.udel.edu/

Automation and streamlining of data analysis for GMP or release testing

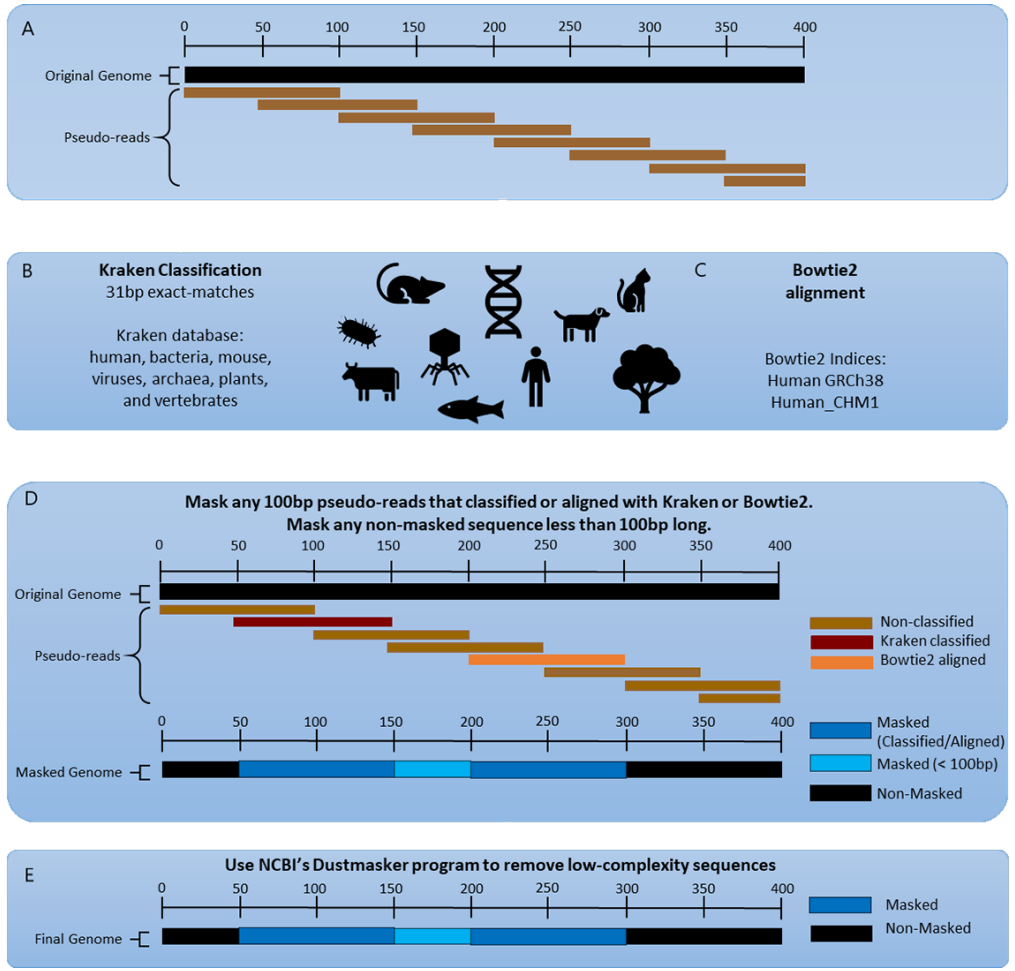
ViruScreen* is an internally-developed bioinformatic analysis application for VMA

- ✓ No computer programming skills required – user friendly
- ✓ Within company firewall – fewer copies of data
- ✓ Dedicated IT and Bioinformatics - customizable workflows
- ✓ Non-GMP and GMP environment for HTS data set analysis
- ✓ Downloadable report is generated

- Multiple workflows are available in ViruScreen for Illumina paired-end reads
  - FastQC
  - **Database curation** – Kraken, Dustmasker, custom
  - **Short read alignment (SRA)** – Bowtie 2, Magic-BLAST
  - ***De novo* assembly alignment –** Megahit, Magic-BLAST
  - Protein alignment – DIAMOND
  - Single nucleotide variant analysis – Lofreq2

*Patent application submitted 2023.

Curation of FDA clustered RVDB in UI to reduce false positive hits



- Clustered Reference Viral Database (cRVDB) generated by FDA
  - Diverse virus sequences; also includes retro-elements, etc
- Bulk vaccine virus harvest
  - Contains sequences from host cells
- Database curation is based on masking undesired regions
  - Described in Kenney, *et al*., adapted from Lu and Salzberg.
  - Split sequences into pseudoreads,
  - Taxonomically classify and align to custom database of human, bacteria, vectors, etc
  - Dustmasker used for low complexity regions
  - Undesired nucleotides converted to "N"
  - Retains same number of records

Lu J, Salzberg SL (2018) Removing contaminants from databases of draft genomes. *PLoS Comput Biol* 14(6): e1006277.

Kenney JG, et al. (2024) Communicating computational workflows in a regulatory environment. *Drug Discov Today*. 2024 Mar;29(3):103884

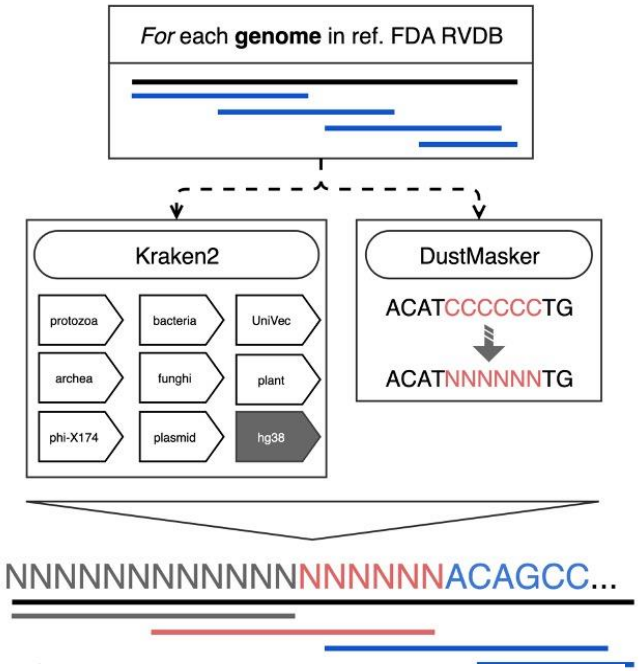
Curation statistics depicted in the user interface: cRVDB example

Communicating computational workflows in a regulatory environment

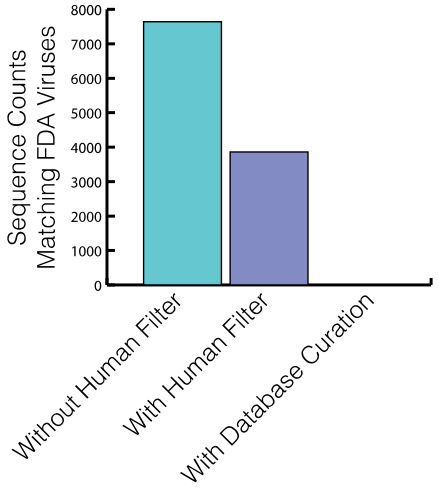
Jonathon G. Keeney [1,5,*], Naila Gulzar [1,5], Jack B. Baker [2], Ondrej Klempir [3], Geoffrey D. Hannigan [4], Danny A. Bitton [3], Julia M. Maritz [4], Charles H.S. King IV [1], Janisha A. Patel [1], Paul Duncan [2], Raja Mazumder [1]

[1] Department of Biochemistry and Molecular Medicine, School of Medicine and Health Sciences, The George Washington University, Washington, DC, USA
[2] Merck & Co., Kenilworth, NJ, USA
[3] R&D Informatics Solutions, MSD Czech Republic, Prague, Czech Republic
[4] Exploratory Science Center, Merck & Co., Cambridge, MA, USA

Database curation

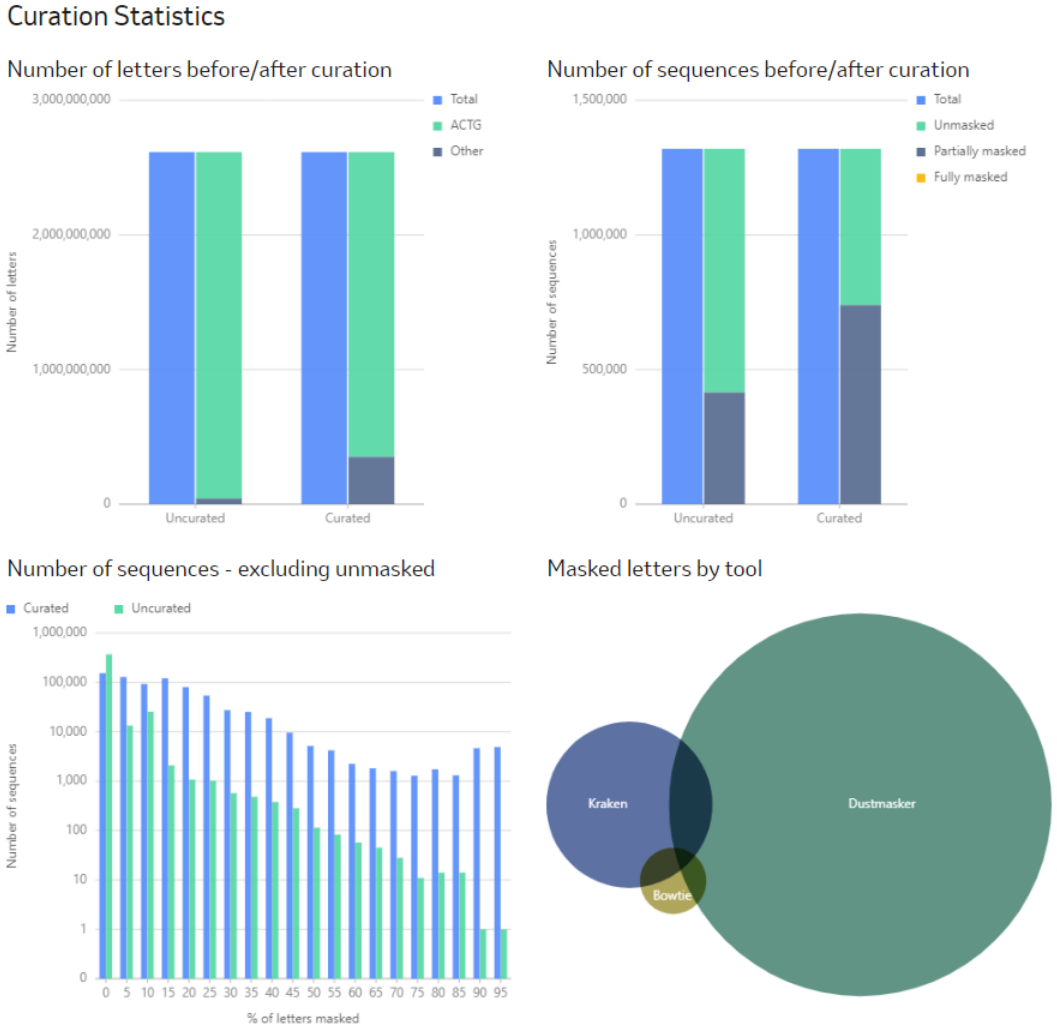


Simulated dataset results



Drug Discovery Today • Volume 29, Number 3 • March 2024

Output example from recent cRVDB version: ViruScreen UI and report



Overview of the short read alignment (SRA) workflow for Illumina sequences

Sample preparation and extraction

Illumina high-throughput sequencing

Trim adapter and low-quality sequences

<u>Virus harvests only</u>: Stringent alignment to vaccine virus

Align to curated viral reference database and vaccine virus

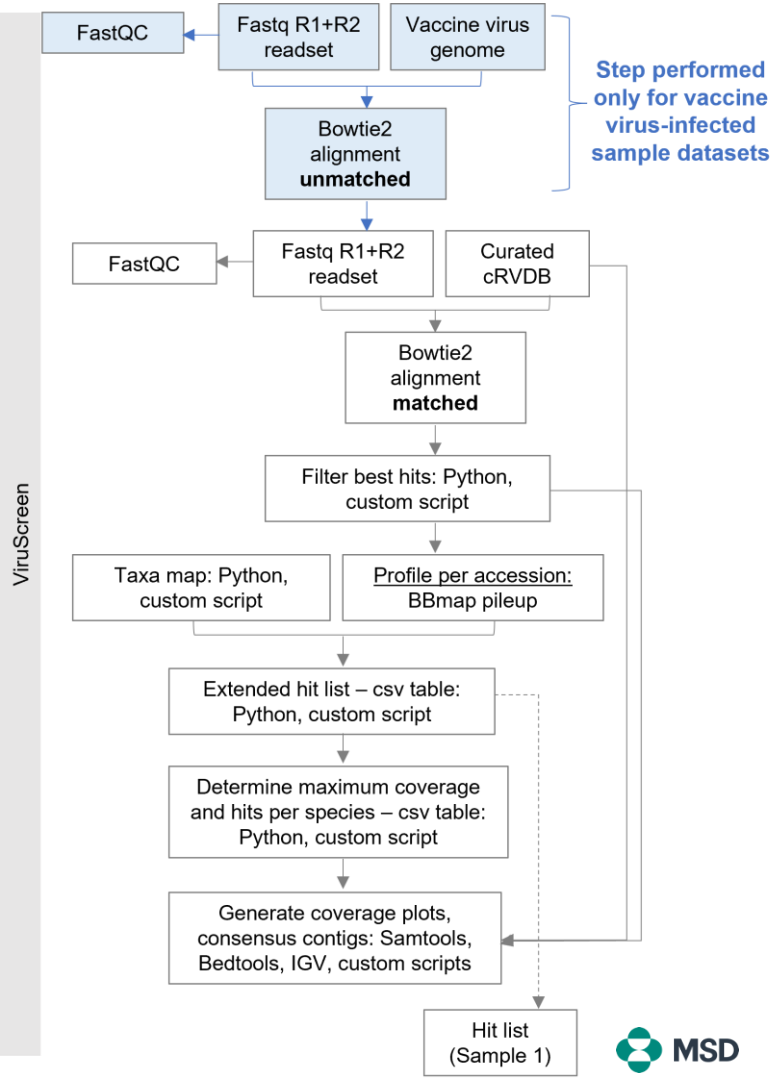
Best hits within taxa reviewed across samples

Counterscreen 'Putative' positives

Identify bioinformatically-confirmed hits of interest

Launch investigation to confirm virus presence

Automated SRA workflow in ViruScreen



MSD

SRA case-study: Benchmarking ViruScreen GMP against HIVE

- **Purpose**: Analyze data sets representing different types of data and sample preparations to inform on the sensitivity and specificity across HIVE and ViruScreen workflows.
- **Data**: Research datasets spiked with known virus amounts or virus sequences for comparison.
- **Analysis**: Perform short read alignments (SRA) of different spiked datasets to the cRVDBv20 with added curation.
  - HIVE alignments performed with Hexagon
  - ViruScreen alignments performed with Bowtie2
- **Results**: Determine the recovery of the spiked viruses / virus sequences by maximum total hits and coverage length

Datasets used in the HIVE–ViruScreen comparison study

In silico-spiked data

- Concentrated uninfected media
- Background sequence reads
- Spiked *in silico* with 20 virus sequences from five virus genomes.
- ~22 M reads dataset

Internal data

- Research samples corresponding to harvest control fluids **(HCF)** and harvest virus fluids **(HVF)** from an early-phase SARS-CoV2 live viral vaccine generated in **Vero cells**
- Spiked with ~1e5 genome copies / mL of 5 virus panel.
- ~13 M HCF and ~50 M HVF reads

Published data

- **HeLa cells**
- Spiked with four mixed viruses at three levels of viral particles per cell.
- AVDTIG collaborative study. (Lab-C, RNA datasets).
- ~300 M reads each dataset

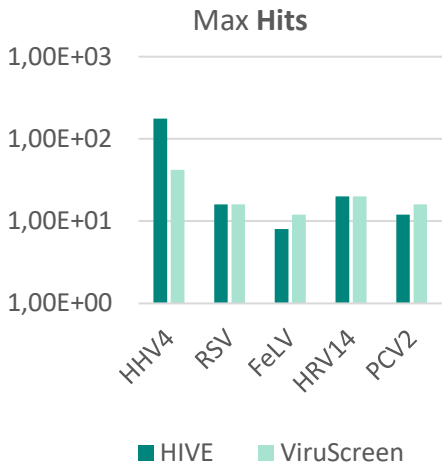
Khan AS and Study Group Participants. Proposed 1st International Virus Reference Standards for Adventitious Virus Detection in Biological Products by Next-Generation Sequencing (NGS) Technologies (CBER-5). 7 Sep 2020. WHO/BS/2020.2394.

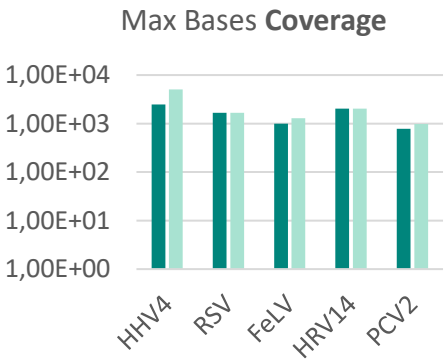
Khan AS, Ng SHS, Vandeputte O, Aljanahi A, Deyati A, Cassart JP, Charlebois RL, Taliaferro LP. A Multicenter Study To Evaluate the Performance of High-Throughput Sequencing for Virus Detection. mSphere. 2017 Sep 13;2(5):e00307-17. PMID: 28932815.

MSD

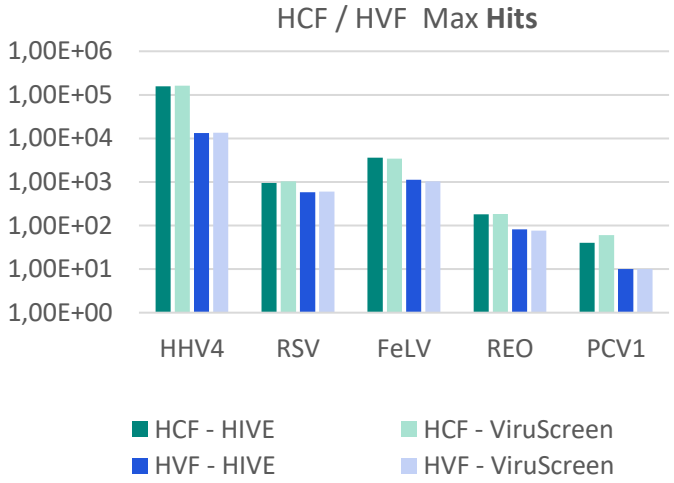
ViruScreen and HIVE workflows result in similar spike virus recovery

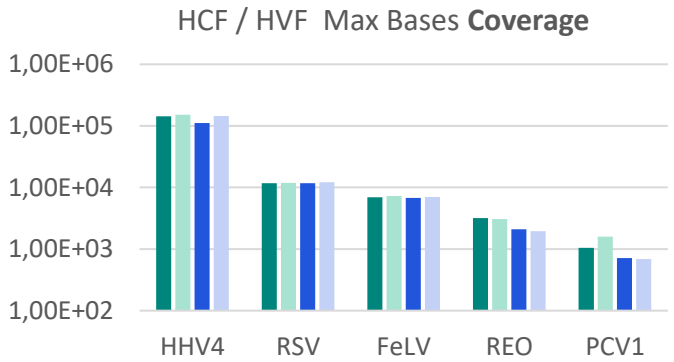
Results of *in silico* virus sequences



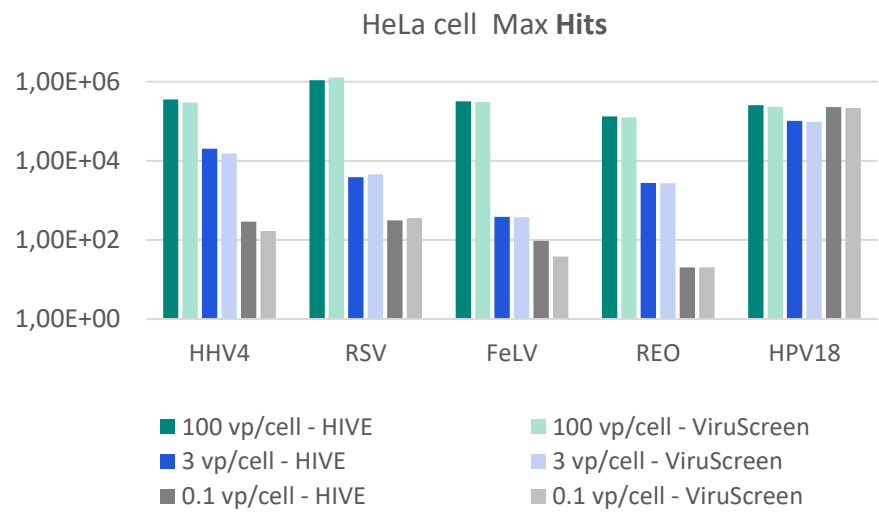


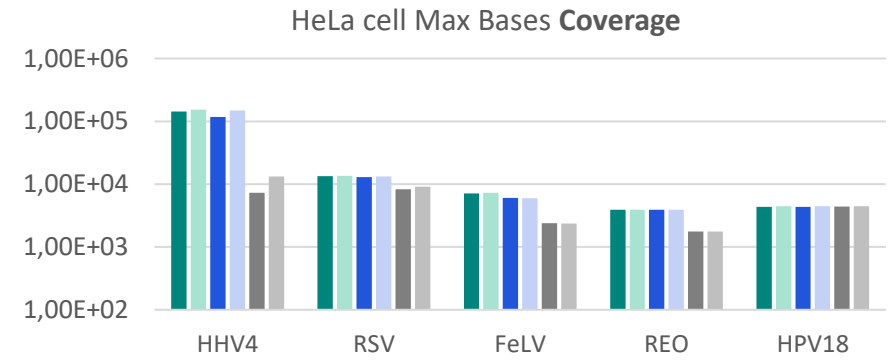
Virus panel spiked into CoV2 live virus vaccine samples





AVDTIG Lab-C datasets (virus spiked into HeLa cells)

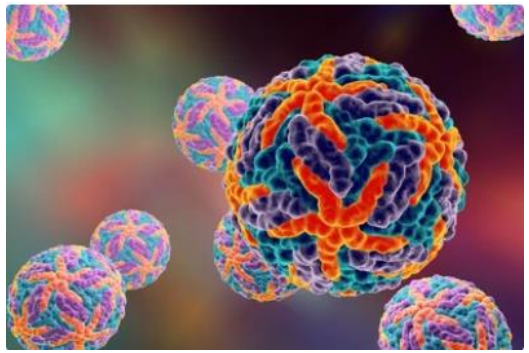




- **Both workflows led to similar recovery (~10^3 bases) from *in silico* reads corresponding to different viruses.**
- **Viruses spiked into samples and processed internally or those from a published study also exhibited similar recovery.**

HIVE-ViruScreen research study using dengue live virus vaccine samples

- V181 is an investigational live attenuated quadrivalent dengue vaccine produced in Vero cells
- Non-GMP samples of V181 dengue 3 vaccine and controls used for
  - A spike recovery study using the WHO/FDA/CBER virus panel
  - Unspiked samples used to generate research results to understand the workflows
- Short read alignments (SRA) of readsets to a curated, clustered RVDBv25 using HIVE and ViruScreen
- *De novo* assembly alignments performed in ViruScreen to compare with SRA
  - Assemble short reads into contigs with Megahit
  - Align contigs to the reference using Magic-BLAST



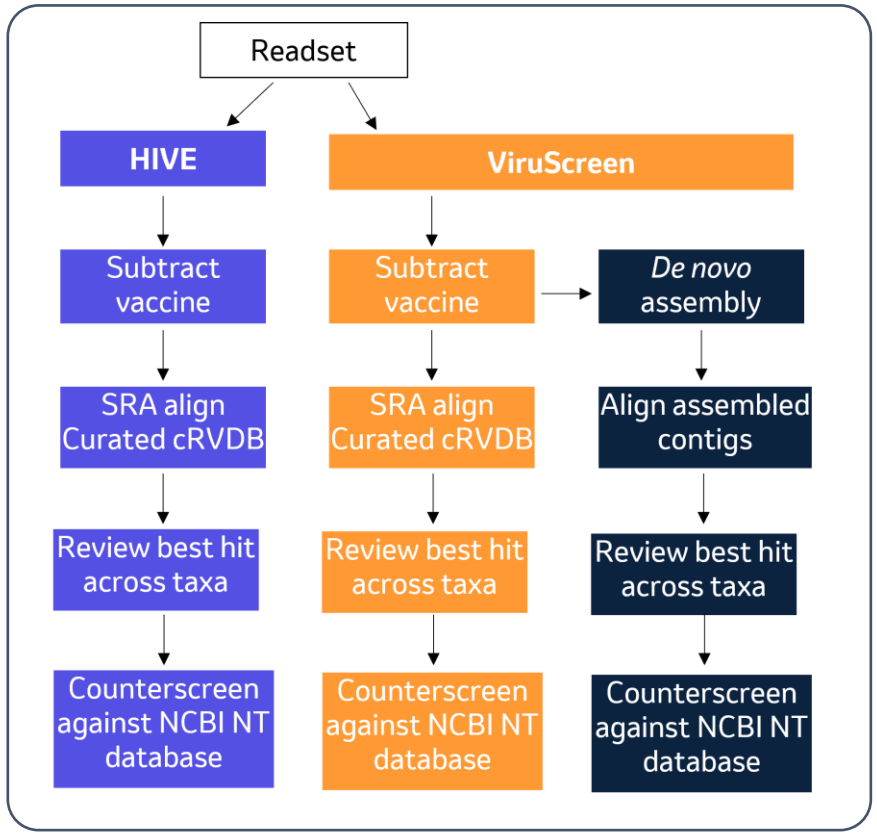

Dengue virus vaccine VMA research study design

Samples and controls to be analyzed in research study

Sample name	Condition*	Purpose	Sample treatment	Extraction method	Library
Harvest virus fluids (HVF)	Unspiked	Secreted or released viruses	+/- Virus spikes Benzonase digestion	Phenol:Chloroform: Isoamyl alcohol (total nucleic acids)	**NexteraXT**
	Spiked ~10^3 gc/mL				
	Spiked ~10^4 gc/mL				
	Spiked ~10^5 gc/mL				
Harvest control fluids (HCF)	Unspiked				
	Spiked ~10^3 gc/mL				
	Spiked ~10^4 gc/mL				
	Spiked ~10^5 gc/mL				
Control cell pellet	Unspiked	Transcriptome analysis	Trypsinization Wash 2x	Qiagen RNeasy mini kit (RNA) + Ribosomal depletion	**TruSeq mRNA Stranded**
	Spiked (pre-extraction HRSV RNA ~1 gc/cell)				
Medium control	Unspiked	Reagent control	+/- Virus spikes Benzonase digestion	Phenol:Chloroform: Isoamyl alcohol (total nucleic acids)	**NexteraXT**
Buffer control	Unspiked				

*gc/mL: genome copies per mL of fluid sample

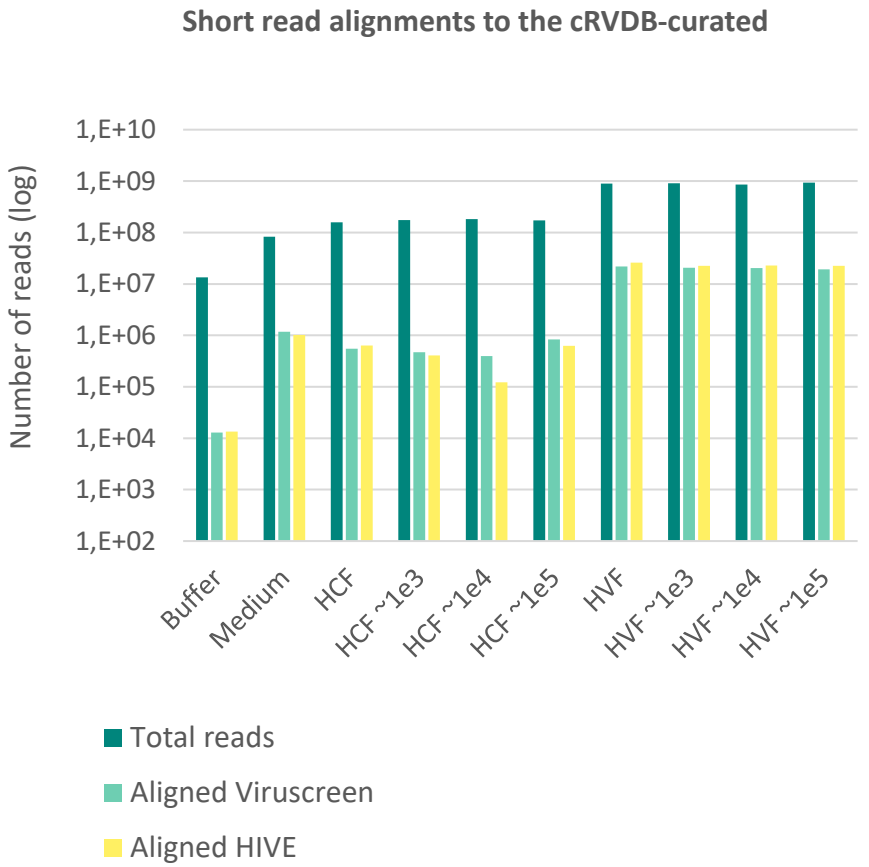
Study workflows



- **Evaluate readsets in SRA: run in parallel in both HIVE and ViruScreen.**
- **Assess *de novo* assembled contig analysis in ViruScreen.**

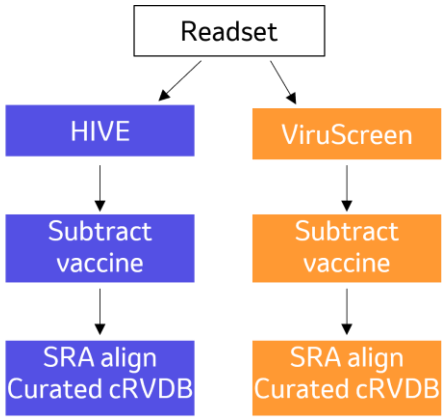
Sequencing yield and short read alignments

Sample name	Condition	Total Reads	ViruScreen vaccine-subtracted unaligned	ViruScreen aligned to cRVDB	HIVE vaccine-subtracted unaligned	HIVE aligned to cRVDB
Harvest virus fluids (HVF)	Unspiked	897,683,136	350,189,438	21,816,410	315,625,265	26,075,925
	Spiked ~10^3 gc/mL	902,878,748	400,753,954	20,561,490	365,580,953	22,570,006
	Spiked ~10^4 gc/mL	862,157,904	320,467,052	20,276,366	286,466,278	22,799,702
	Spiked ~10^5 gc/mL	931,119,204	396,219,060	19,321,400	366,443,715	22,406,014
Harvest control fluids (HCF)	Unspiked	157,024,866	n/a	554,054	n/a	638,168
	Spiked ~10^3 gc/mL	174,846,764	n/a	469,026	n/a	406,406
	Spiked ~10^4 gc/mL	182,711,880	n/a	394,130	n/a	122,296
	Spiked ~10^5 gc/mL	173,280,070	n/a	831,432	n/a	625,580
Control cell pellet	Unspiked	835,769,224	n/a	18,514	n/a	N.P.
	Spiked (pre-extraction HRSV RNA ~1 gc/cell)	843,169,168	n/a	21,004	n/a	N.P.
Medium control	Unspiked	82,745,728	n/a	1,182,448	n/a	1,013,066
Buffer control	Unspiked	13,405,896	n/a	12,966	n/a	13,480

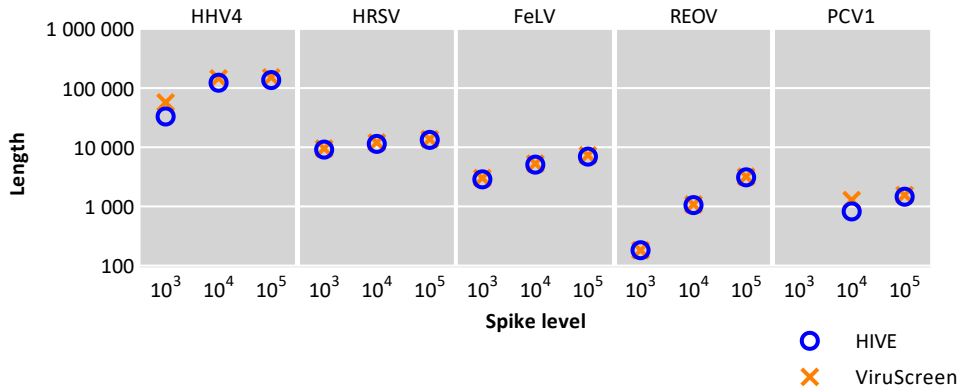


- **The readsets generated for the benchmarking study exhibited similar alignment to both the vaccine virus, where applicable, and to the curated cRVDB**

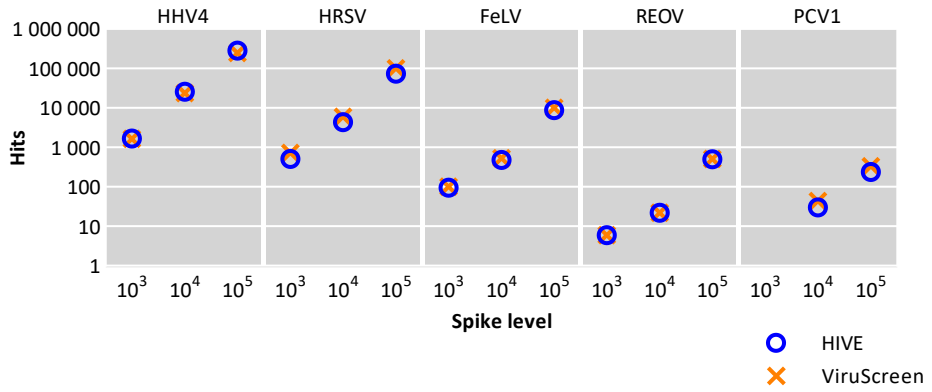
Spike virus analysis of dengue HVF and HCF in HIVE and ViruScreen



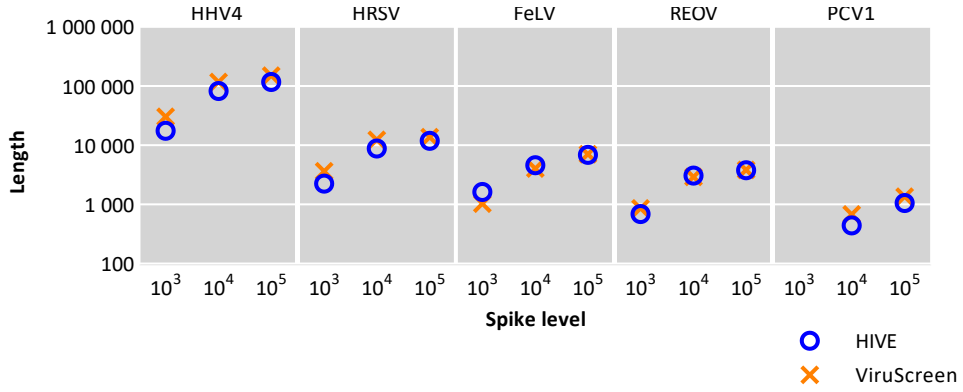
Harvested control fluids **max bases covered**



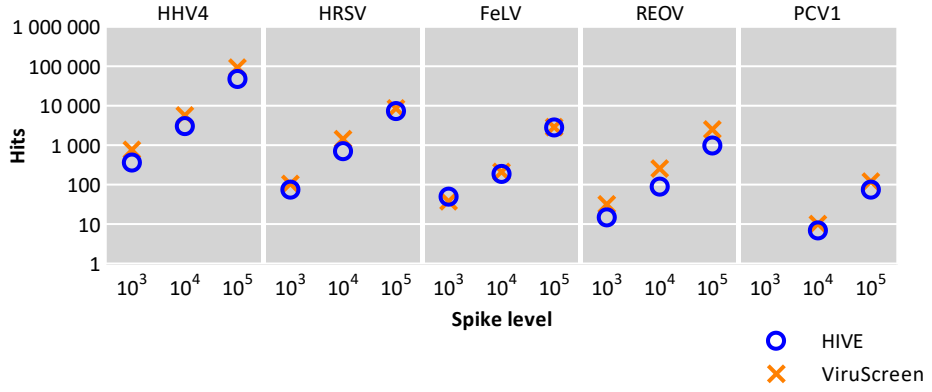
Harvested control fluids **max hits**



Harvested Virus fluids **max bases covered**

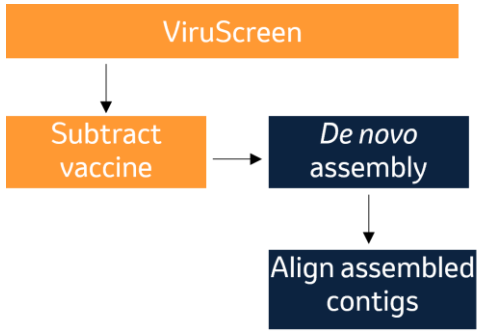


Harvested virus fluids **max hits**



- **Limit of detection is ~10^3 spike level for all viruses except PCV1 (~10^4 genome copies / ml) across both HIVE and ViruScreen for both uninfected and infected cell fluids.**

Spike virus analysis of *de novo* assembled contig alignments in ViruScreen



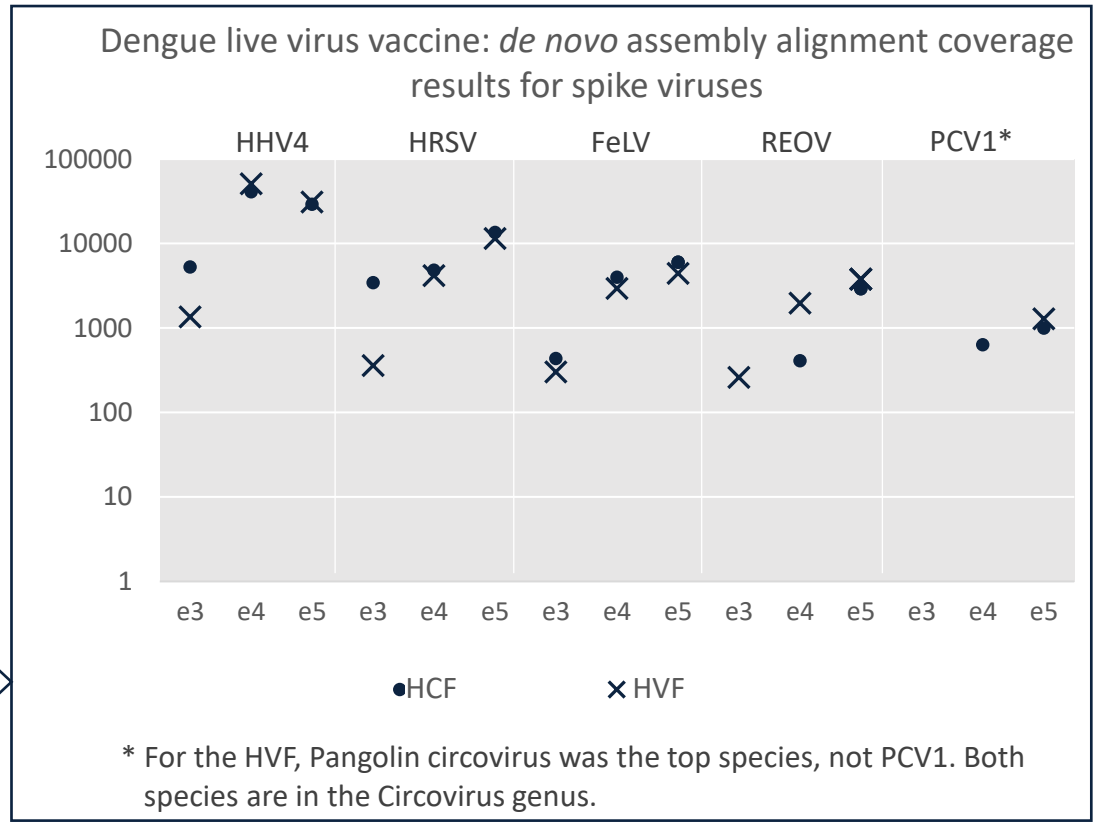
Assembled alignments

Limit of detection:

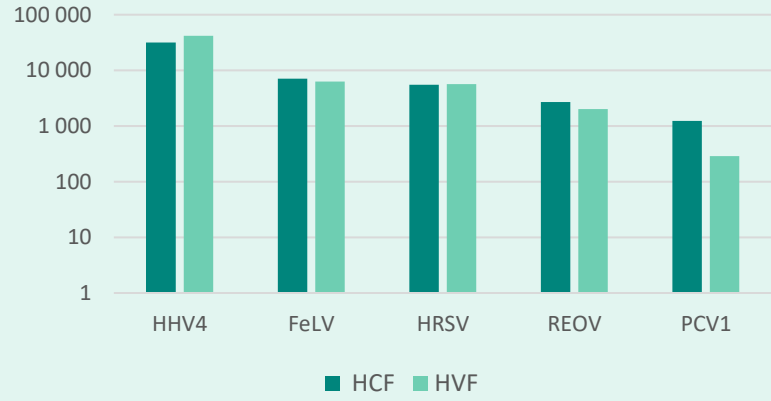
~10^3 gc/mL: HHV4, HRSV, FeLV

~ 10^4 gc/mL: REOV

~ 10^5 gc/mL: Circovirus



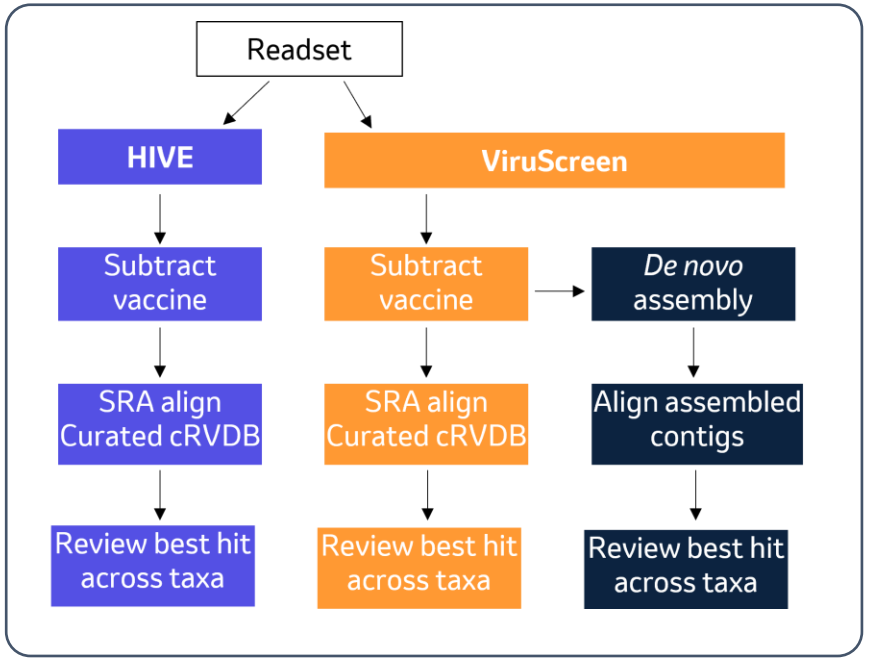
* For the HVF, Pangolin circovirus was the top species, not PCV1. Both species are in the Circovirus genus.



Analysis of the CoV2 live vaccine readsets (~10^5) used in the SRA benchmarking study indicates that PCV1 species is recovered following *de novo* contig alignment to the curated cRVDB, in addition to other Circovirus spp.

- **Across sample readsets, spike recovery following *de novo* assembly is not as sensitive as short read alignments.**
- **For *de novo*-assembled CoV2 live vaccine sample readsets, PCV1 was detected at ~ 10^5 gc/mL**

Analysis across unspiked samples for different applications and workflows

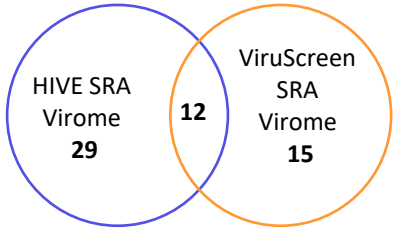


Number of best hits across taxa: species and genus

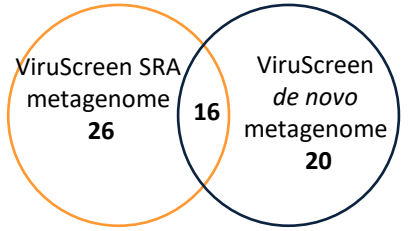
	HIVE SRA	ViruScreen SRA	ViruScreen *de novo* alignments
Number of *species* – virome	29	15	6
Number of genera – virome	11	8	6
Number of *species* –metagenome	n.d.	26	20
Number of genera – metagenome	n.d.	11	10

n.d. not determined

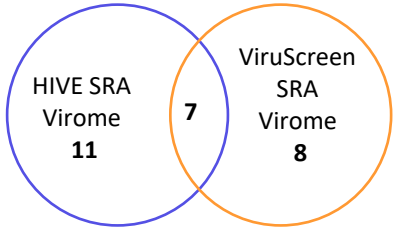
Number of species



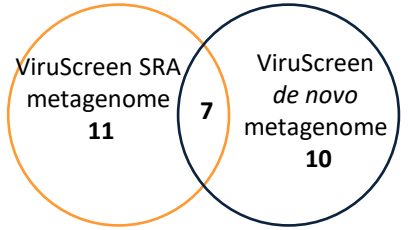
Number of species



Number of genera

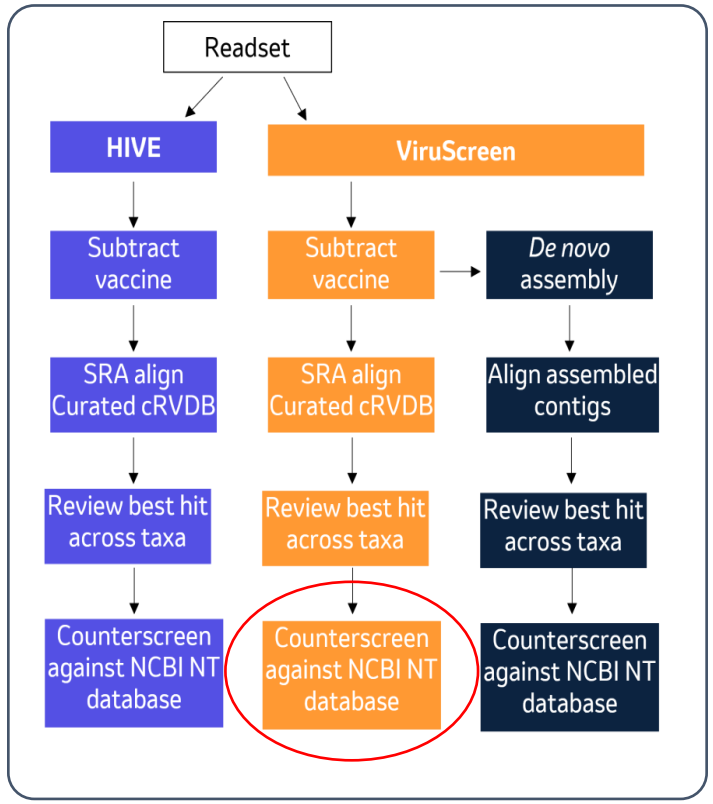


Number of genera



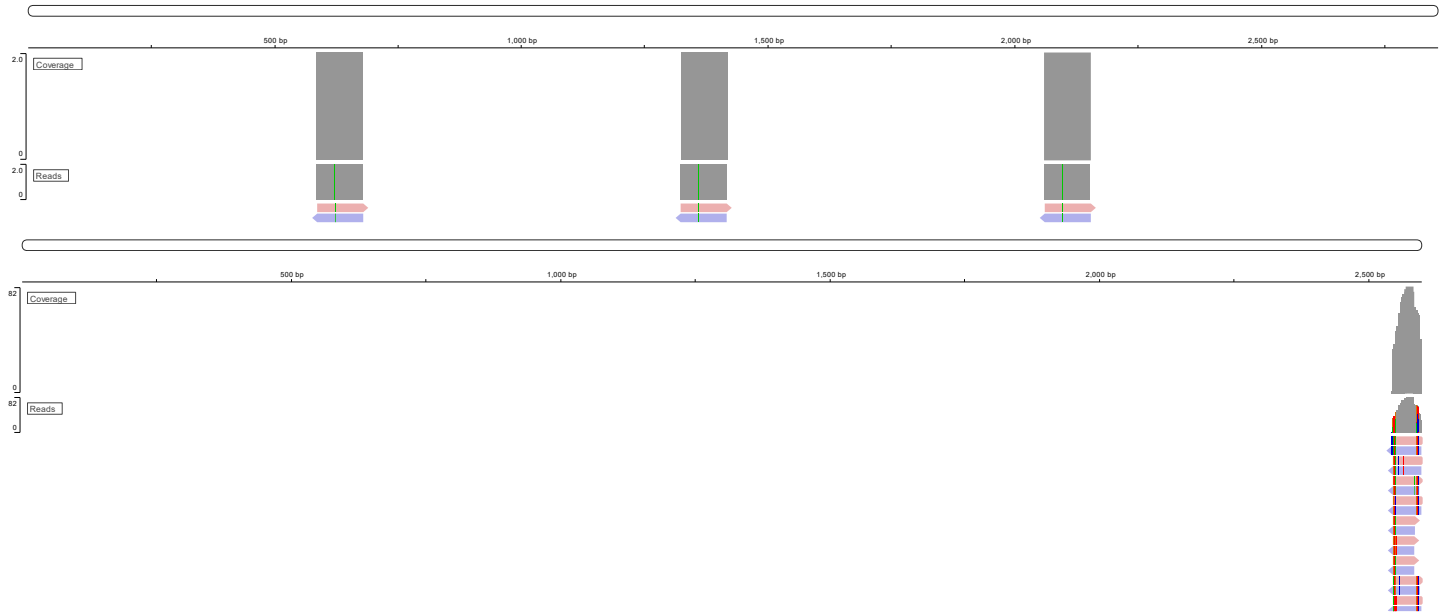
- SRA across unspiked samples HIVE returned the most hits
  - Presumably partially due to lack of paired read filtering in HIVE workflow.
- In ViruScreen, the SRA leads to higher number of putative hits to counterscreen than does the *de novo* alignment workflow.

Example I: Counterscreening of a hit that was *not* bioinformatically confirmed



Family	Genus	Best hit species within Genus	Maximum of total bases covered
			Harvest Virus Fluids
Retroviridae	Lentivirus	HIV-1	279

Example coverage plots (different accessions)



Arguments against specificity to HIV-1:

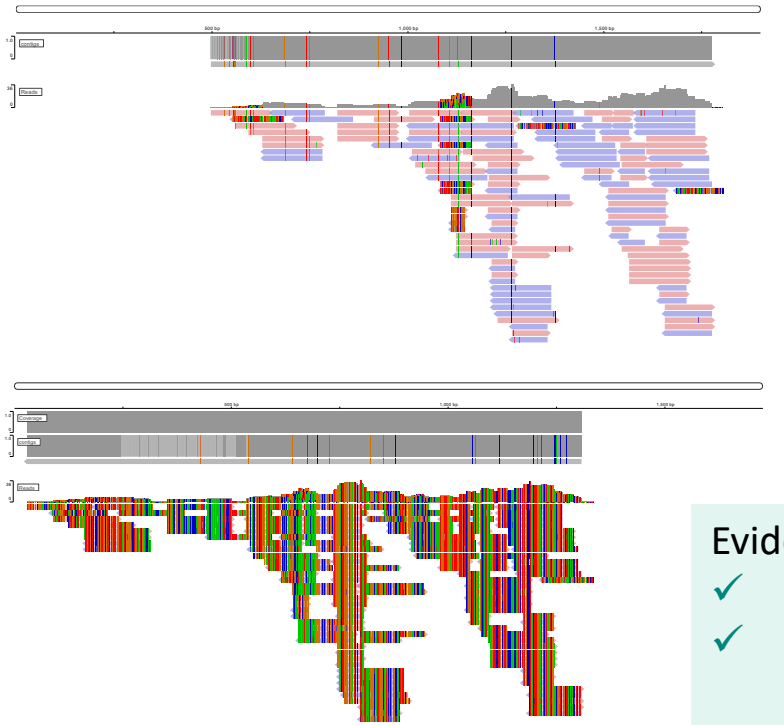
- ☒ Representative BLASTX alignments demonstrate best match to non-viral accessions.
- ☒ Representative BLASTn and BLASTx alignments to primate proteins or lentiviral cloning vector
- ☒ No coverage of uniquely-HIV-1 sequences

Example II: Counterscreening a hit of a true positive (HVF spiked sample)

Family	Genus	Best hit species within Genus	Maximum of total bases covered
			Harvest Virus Fluids spiked ~1e5 gc/mL
Circoviridae	Circovirus	Pangolin Circovirus B2	1,279
Circoviridae	Circovirus	Pangolin Circovirus T2	1,279

Spiked samples contain true positive hits – use as example to show counterscreening to confirm best match

Example coverage plots to ViruScreen results



- Perform BLASTx
- Perform BLASTn

BLASTn results taxonomy tree indicates several Circovirus species

Evidence to bioinformatically confirm the result

- ✓ Representative BLASTx alignments have best match to PCV1 replicase
- ✓ Representative BLASTn alignments demonstrate high match scores to several Circovirus species.

Conclusions

- Viral Metagenomic Analysis (VMA) studies have been undertaken to supplement the virus risk assessment for select live virus vaccines and cell banks.
- HIVE bioinformatic analyses have been successfully filed as characterization studies.
- An internal application for customized workflows was created to streamline bioinformatic analysis for potential future release of GMP materials.
- Comparison across workflows provides opportunity to benchmark bioinformatic methods.
- Large datasets from the dengue live vaccine samples were generated to complement the initial comparisons between HIVE and ViruScreen which exhibited similar performance for spiked virus recovery.
- Additionally, testing of the end-to-end workflow led to identification of successes and opportunities to leverage in studies for filing of advanced virus detection methods for biological products.
- Further automation of largely manual analysis such as counterscreening and comparative analysis in ViruScreen will facilitate future applicability for product-specific release or GMP assays.

- **Vanessa Sarathy**
- **(Paul Duncan)**
- **Jack Baker**
- **Connor Geraghty**
- **Julia Maritz**
- **Semina Arampatzi**
- **(Geof Hannigan)**
- **Christopher Wang**
- (Topher Woelk)
- **Danny Bitton**
- **Anna Gromek**
- (Ondrej Klempir)
- (Ondrej Tupa)
- **Ron Smeral**
- **Ales Vondra**
- (Jakob Goldmann)
- Bernice Westrek
- Luca Benetti
- Szi Fei Feng
- (Begüm Topçuoglu)
- Elek Dobos
- **Ram Ramaswamy**
- Matthew Balmer

- Kasia Bankowska
- Vern W. Henery
- Erik Talens
- Jos Weusten
- (Michelle de Groot)
- Megan Calafati
- Marc Sze
- (Sushmita Parajuli)
- **Alena Navratilova**
- Viktor Ivanov
- Shara Dellatore
- (Andrew Brown)
- (Sarah Dunaj)
- (Daria Hazuda)
- Olga Korpacheva
- Tereza Neuwirthova
- Luis Diego Gené
- Xudong Qiao
- Vincent Antonucci
- James Cass

- ❖ West Point, PA, USA
- ❖ Cambridge, MA, USA
- ❖ Prague, Czech Republic
- ❖ Oss, The Netherlands

- **Raja Mazumder**
- **Jonathon Keeney**
- Emily Pennington
- (Naila Gulzar)

MSD
Tel: 215-652-4198
E-mail: vanessa.sarathy@merck.com
Address: 770 Sumneytown Pike, West Point, PA, USA

Thank you