AFS ATCC FEDERAL SOLUTIONS

A Transcriptomic Workflow for Adventitious Agent Detection

ATCC™
Credible leads to Incredible™

David Yarmosh[1], Joseph Petrone[1], Ana Fernandes[1], Sophia Huff[1], Kumari Meenu Karna[1], Jeffrey I. Tokman[1], Anjan Panthee[1], John Bagnoli[1], Heather Couch[1]

[1]American Type Culture Collection

INTRODUCTION

Adventitious agents represent one of several key challenges facing the manufacture of any biological product. Current best practices in the detection of adventitious agents such as qPCR techniques or application of live-animal models are becoming recognized as inadequate. PCR and comparable techniques are highly accurate but target fixed sets of organisms; they cannot reliably cover the scope of all agents. *In vivo* analyses are expensive, time consuming, and conflicts with the three Rs (replacement, reduction, and refinement) principles.

Next Generation Sequencing (NGS) has been proposed as an alternative to both methods but suffers from difficulty in the determination of a true detection.

To better improve the position of NGS in adventitious agent detection, ATCC's Sequencing and Bioinformatics Center developed an NGS approach to mitigate the weaknesses of NGS by focusing on the strengths of traditional methods. Our method is multi-tiered to account for sequencing and analysis biases, including:

- Metagenomic classification with a minimally redundant database
- Multi-organism assembly
- Annotation
- Gene quantification targeted to the metagenomic classification hits

Each component is considered as a separate factor in determining the presence of an adventitious agent. This approach begins with RNA-Seq to reduce the sequencing of material that is not actively transcribing, thereby increasing the signal available of both the host organism and any adventitious agent. Post-sequencing metagenomic classification is performed, targeting sample for differential expression analysis relative to a control sample, amplifying adventitious agent signal and diminishing erroneous assignments. Assembly and annotation methods reduce the impact that isolated, but high-depth reads confound classification by genome-wide detection. Finally, results from each stage are collated into a single report for review. ATCC presents our method for NGS triage of biological products for adventitious agent detection.

MATERIALS AND METHODS

Virus	Strain	Genome Type	Genome Structure	Genome Size	Envelope
Porcine circovirus (PCV1)	Type 1	ssDNA	Circular Genome	~1.7 kb	Non-enveloped
Respiratory syncytial virus (RSV)	A2	ssRNA (-)	Linear Genome	~15.2 kb	Enveloped
Epstein-Barr virus (EBV)	B95-8	dsDNA	Linear Genome	~180kb	Enveloped
Influenza Virus (Flu)	A/New Caledonia/71/2014 H3N2	ssRNA (-)	Linear Genome	~13.5kb	Enveloped

- Cell Line: HeLa cells were used as a background matrix for all spike studies currently tested
- Spike Experiment: Hela cells at 1x10^6 cells were spiked with a purified virus in triplicate at varying concentrations, with HeLa cells and media with no added virus as a control. Nucleic acid from the spiked samples were then extracted using before sequencing and analysis.
- Analysis: mRNA reads are first subject to kraken2 classification. Kraken2 is run with default parameters, but with a custom database that comprises a minimally redundant selection of viral, bacterial, fungal, and archaeal genomes from RefSeq that is supplemented with common host organisms (human, mouse, etc.) and reference-grade genomes from the ATCC Genome Portal. This initial classification serves two purposes: 1) Determine which reads are likely viral in origin for assembly. 2) Identify representative genomes for differential gene expression analysis. This creates a fork in the workflow. Viral reads are assembled using SPAdes v3.15.5 in its metaviral mode. The assembled contigs are subject to the same kraken2 analysis for confirmation. These contigs are further annotated using the VIGA viral annotation pipeline. Algorithmic choices within predictive annotation tools prompt a BLAST search on annotated genes to identify the degree of homology between the annotated gene and similar genes in a process to identify possible false positive hits. Finally, the number of assembled genes is compared to the number of known genes for each virus assembled. Meanwhile, all viral species identified in the mRNA read kraken2 analysis are downloaded as well as the top 1% of species from non-viral domains. These are each used as reference genomes for Salmon v1.10.1 transcriptomic analysis. This approach provides Transcripts per Kilobase Mapped (TPKM) values for each gene in an organism's transcriptome. At this stage, the Salmon results for the control samples are compared to those of the query samples in differential expression, performed by custom python and R scripts. This allowed differential expression to separate the adventitious agent signal from the NGS noise.

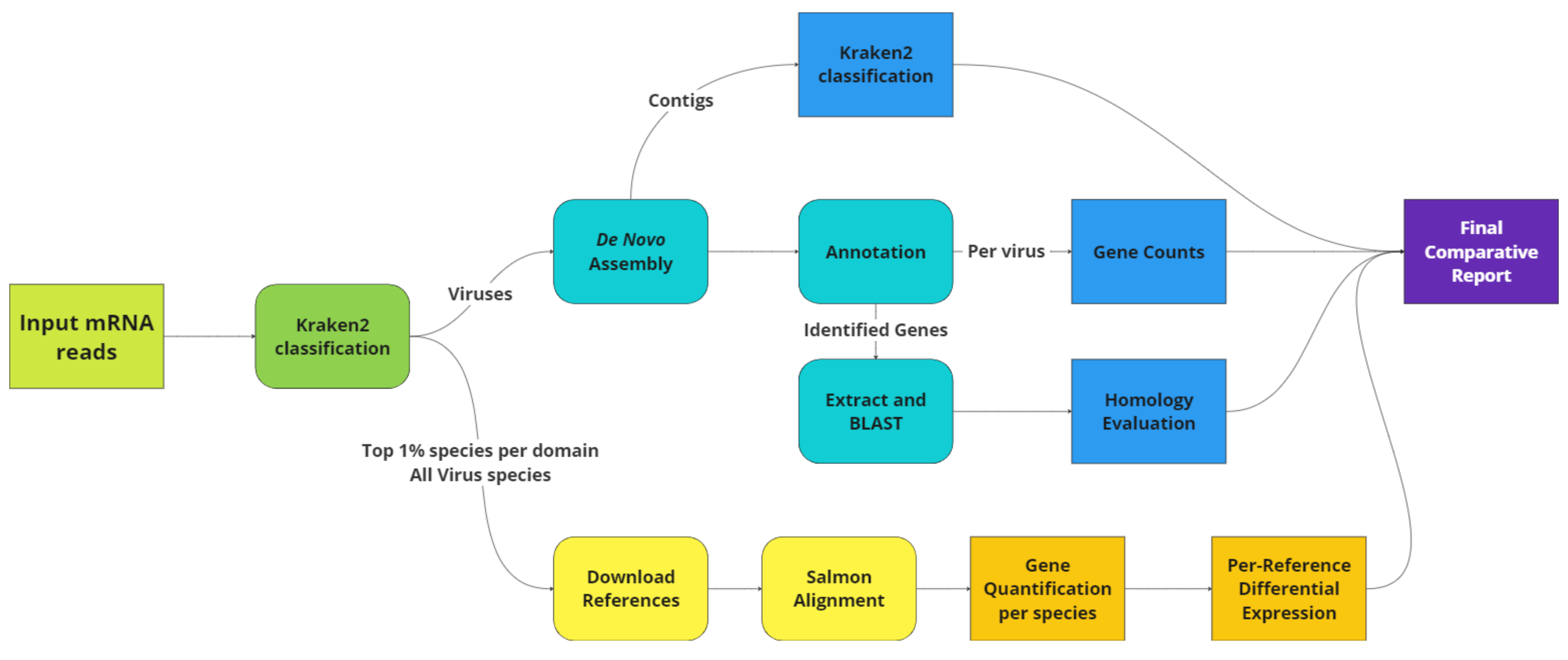
RESULTS

Number of Samples	Spike Target	Host Cell	Library Type	Titers	Quantification Method
15	RSV (VR-1540)	HeLa	WGS	0.5,1,2.5,5,10 ng/µL	Qubit
9	RSV (VR-1540)	HeLa	WGS	0.1,0.25,0.5 ng/µL	Qubit
12	PCV1	HeLa	WGS	0.1,0.25,0.5,1,2.5,5 ng/µL	qPCR
4	PCV1	HeLa	WGS	0.1,0.25,0.5,1 ng/µL	qPCR
9	RSV (VR-1540)	HeLa	WGS	0.1,3,100 copies/cell	qPCR
9	RSV (VR-1540)	HeLa	WGS	0.1,0.5,5 ng/µL	qPCR
12	RSV (VR-1540)	HeLa	RNA-Seq	0.1,3,100,1000 copies/cell	qPCR
13	RSV (VR-1540)	HeLa	RNA-Seq	(hostX1),0.01,0.1,3,25 copies/cell	qPCR
13	PCV1	HeLa	RNA-Seq	(hostX1),0.01,0.1,3,25 copies/cell	qPCR
13	PCV1	HeLa	RNA-Seq	(hostX1),3,25,50,100 copies/cell	qPCR
16	EBV	HeLa	RNA-Seq	(hostX1),0.001,0.1,1,3,1 copies/cell	qPCR
16	Flu	HeLa	RNA-Seq	(hostX1),0.001,0.1,1,3,1 copies/cell	qPCR
16	Mixture	HeLa	RNA-Seq	Mixture	qPCR

Table 1 Summary of spike experiments performed. Initial runs were performed with WGS, but later switched to mRNA-Seq resulting in a greater signal/noise ratio.

Replicate	PCV Titer	Virus	ATCC Read Pairs	CLC Read Count	ATCC Genes Detected	Total Genes
1	3	PCV	1	0	2	2
2	3	PCV	0	0	0	2
3	3	PCV	1	0	2	2
1	25	PCV	15	0	2	2
2	25	PCV	14	16	2	2
3	25	PCV	17	0	2	2
1	50	PCV	32	10	2	2
2	50	PCV	33	8	2	2
3	50	PCV	39	12	2	2
1	100	PCV	76	12	2	2
2	100	PCV	69	10	2	2
3	100	PCV	84	22	2	2
Host	0	N/A	0	0	0	N/A

Table 2 LOD determination of PCV Read pairs column indicates how many forward and reverse read pairs were classified using Kraken2 database for porcine circovirus, forward and reverse reads are counted independently. In the CLC read count column it can be seen at 25 gc/cell 2/3 samples were unable to make any identifications and no reads were identified in the 3gc/cell samples.


Input mRNA reads

Kraken2 classification

Viruses

De Novo Assembly

Contigs

Kraken2 classification

Annotation

Per virus

Gene Counts

Identified Genes

Extract and BLAST

Homology Evaluation

Top 1% species per domain
All Virus species

Download References

Salmon Alignment

Gene Quantification per species

Per-Reference Differential Expression

Final Comparative Report

Figure 1 Overview of bioinformatics workflow. Reference selection is based on identifying non-host organisms, and downloading sequence and annotation information for organisms that represent at least 1% of their respective domain-level classification. Differential expression was then performed on a reference by reference basis.

RESULTS

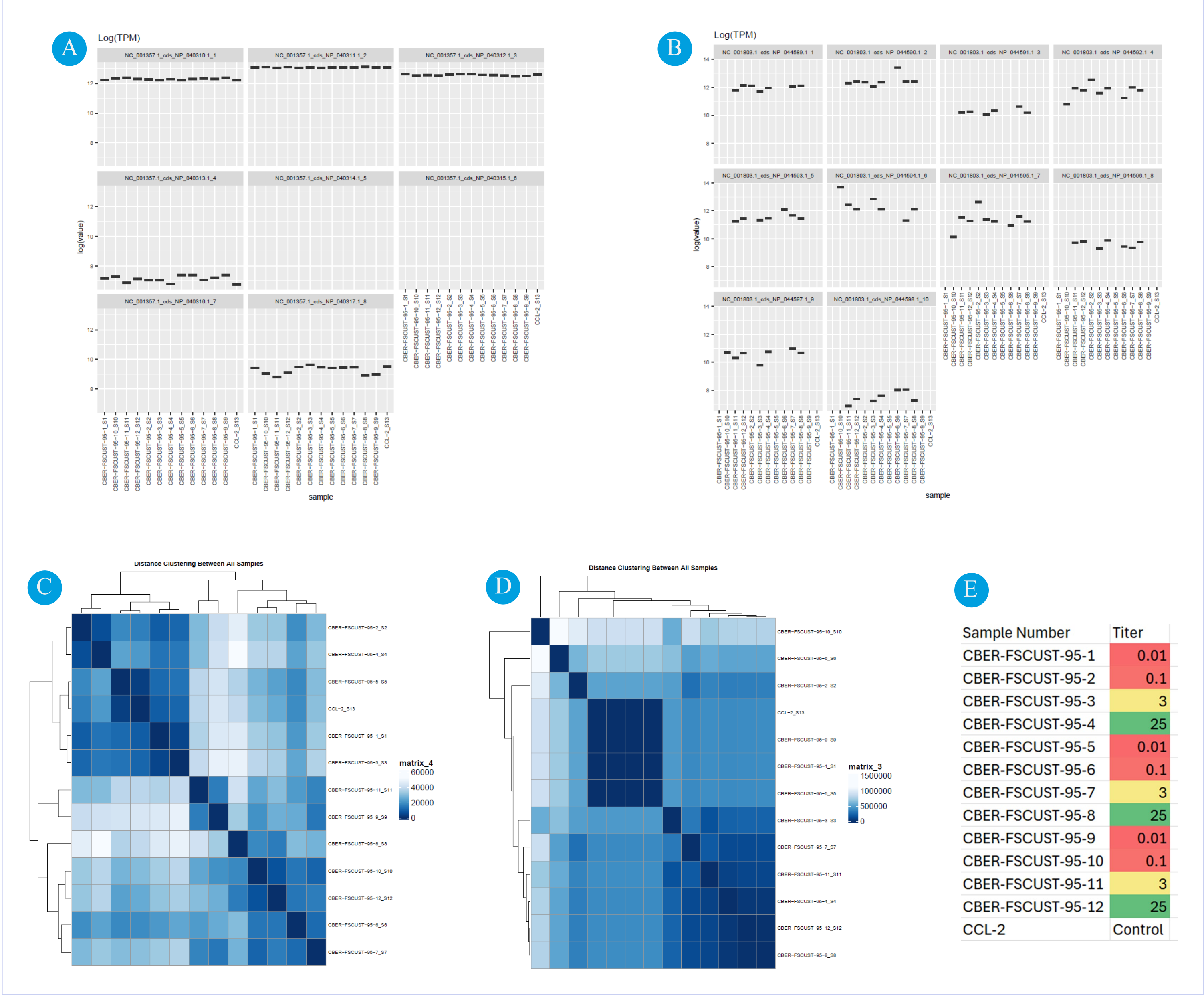


Figure 2 Differential expression illustrating the difference in terms of HPV (a. & c.) or RSV (b. & d.) expression. Presence of HPV in HeLa cells allows it to act as a viral control in all the samples tested and we can see in the heatmap (c.) the expression of HPV genes is not tied to the levels of RSV added to each sample compared to the heatmap for RSV genes (d.) you can see clustering based on titer of the virus added. It is also seen that the negative control sample clusters most closely with the samples that have the lowest titer of RSV spiking (0.01gc/cell). The box plots show how the genes for HPV (a.) is much more consistent among titers of spiked sample compared to the RSV spiked samples (b.). Scale e. is a simple legend showing the levels of spiked sample with accompanying titers.

CONCLUSION

This transcriptomic workflow represents an attempt to address the key issues facing modern adventitious agent detection: sensitivity to next generation signal noise, databases constraining specificity, and methodological bias confounding results. Our intention is supplementing or reducing the need for *in vivo* methods for detection.

Taking advantage of ATCC's extensive viral stocks, spike studies involving several viruses representing different viral groups showed the limit of detection of some viruses of interest which may be extrapolated to other similar viruses.

Utilizing the differential expression of viral genes between spiked samples and control samples allows for greater separation of the typical NGS noise and the potential adventitious agent signal while increasing the resolution to the gene-level, which further improves our ability to determine viral presence and viability, contrasted with less significant contamination or homology.

The standardization of these methods we employ is paramount for establishing a reliable and robust workflow from the laboratory through the analysis to ensure that the system provides the greatest degree of consistently valuable information regarding a sample.

Acknowledgements: We would like to thank Arifa Khan and the Advanced Virus Detection Technologies Working Group (AVDTWG) for the development of the spiking study.

We would also like to thank ATCC and AFS leadership for this development opportunity

AFS 10801 University Boulevard, Manassas, VA 20110-2209 **Phone:** 800.359.7370 **Email:** federalsolutions@atcc.org **Web:** https://www.atcc.org/federal-solutions