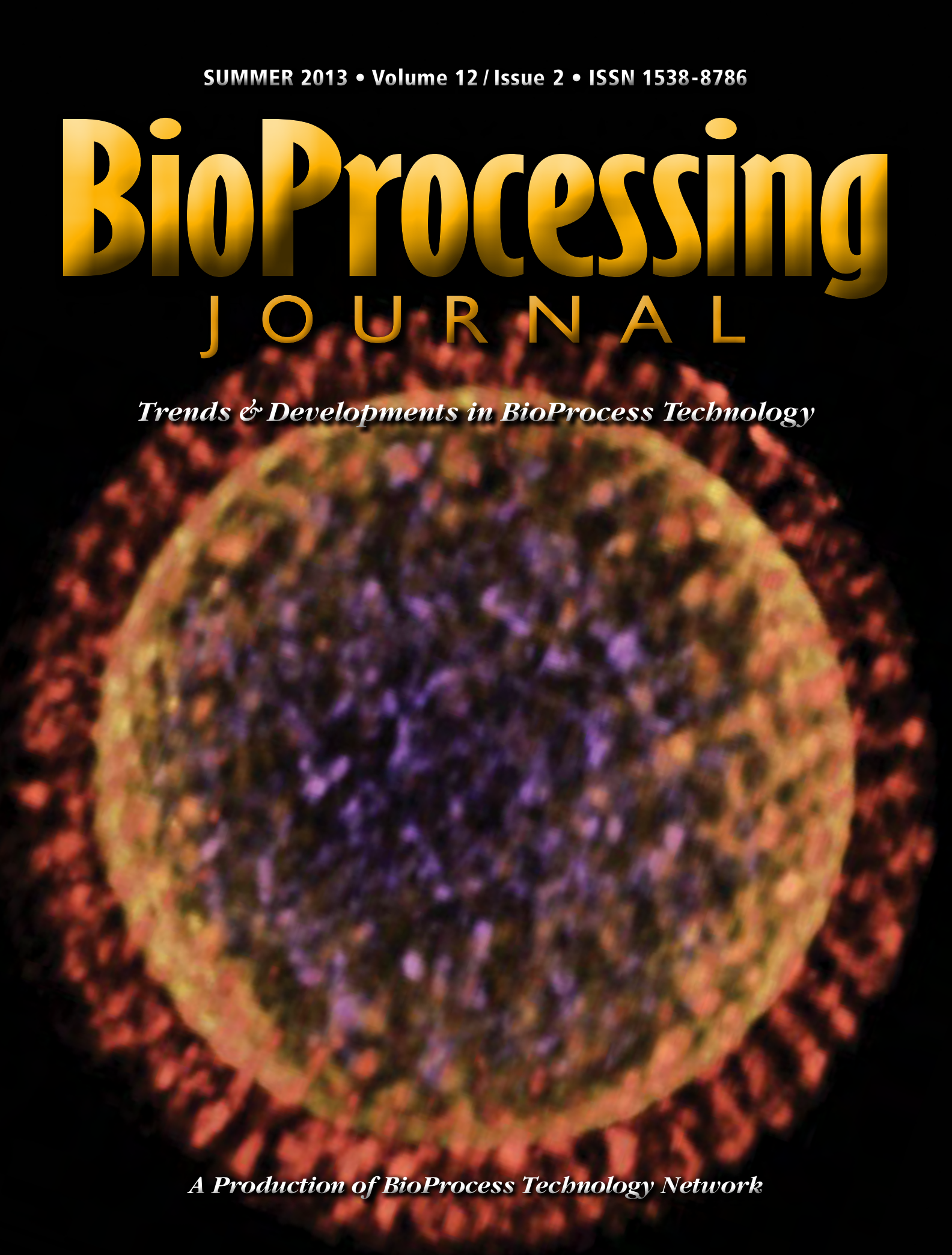
SUMMER 2013 • Volume 12 / Issue 2 • ISSN 1538-8786

BioProcessing JOURNAL

Trends & Developments in BioProcess Technology



A Production of BioProcess Technology Network

BioProcessing JOURNAL
TRENDS & DEVELOPMENTS IN BIOPROCESS TECHNOLOGY

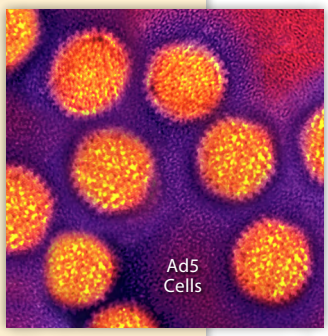
Adventitious Virus Contamination Testing:

Massively Parallel Sequencing in a Multimodal Solution for Biopharmaceutical Safety Testing

By JACK X. YU, STEPHEN KARAKASIDIS, GINGER ZHOU, WENYING HUANG, YANKAI JIA, MICHAEL J. HANTMAN, and JEFFREY A. SHAMAN

Abstract

Virus contamination of commercially valued cell cultures can be a health risk to the general population and imposes financial burdens on manufacturing and biopharmaceutical companies. We investigated the use of massively parallel sequencing/next-generation sequencing (MPS/NGS) and bioinformatics for the detection and subsequent identification of adventitious contamination. Specifically, the Illumina® HiSeq 2000 instrument, in the 2×100 base pair (bp) paired end (PE) run configuration, was used to determine the identity and limit of detection of a DNA virus spiked into a virus vaccine, and an RNA virus spiked into a mammalian master cell bank (MCB). This configuration provided sufficient sequence read lengths and depth of coverage to detect and identify spiked-in SV40 and measles viruses in a background of vaccine, MCB, and host nucleic acid that make up the bulk of these samples. Furthermore, the detection of 30 SV40 reads within one tested sample suggested a < 1 plaque forming unit (pfu) sensitivity in a background of 2.8×10^7 infectious adenovirus serotype 5 (Ad5) particles.



Results from subsequent virus vaccine testing suggested the presence of non-viable virus DNA contamination in the sample. Therefore, we propose that a multimodal approach, in which broad-range screening for known or unknown adventitious agents by MPS/NGS is complemented by targeted virus detection assays (*e.g.*, PCR-based and infectivity assays), should provide the most useful safety monitoring information. We suggest that MPS/NGS and the accompanying bioinformatics is a sensitive, broad-range, and long-lasting tool with the ability to improve upon existing biosafety testing within a larger testing program.

Introduction

Biosafety testing of biologics produced in cell substrates is necessary to determine the presence of adventitious virus contamination. Cell-based infectivity assays, targeted nucleic acid amplification techniques, and broad-range hybridization-based microarray tests are the current industry standards utilized to detect these contaminants. Over time, these assays have been developed to detect contaminants of pre-established interest. In some instances, however, these tests may not be sufficient to detect very low levels of viral contamination, or to identify the species of the contaminant.

Biosafety testing can be more effective by introducing an additional step that can provide a high degree of sensitivity, detect and identify virus contaminants not directly targeted—and at low levels—and identify the species and variants of the contaminant. MPS/NGS offers the ability to detect all existing nucleotides, RNA, and DNA from adventitious virus contamination, thus fine-tuning the degree of sensitivity by increasing or decreasing the depth of coverage. With a high depth of coverage, the sensitivity is sufficient to detect very low levels of DNA contamination. In addition, keen sensitivity enables the potential for identification of contaminants

by cross-referencing unmapped sequence reads to a database. While efforts have been made to improve biosafety testing, few studies have examined MPS/NGS as a complementary tool for the detection and identification of low levels of viral contamination.

A compelling example of applying MPS/NGS to biosafety testing was made by screening a selection of live attenuated viral vaccines for sequence variations as well as for the presence of adventitious viruses that may have contaminated the vaccines during production.[1] In this study, one of six orally administered rotavirus vaccines was found to contain porcine circovirus-1. As a result, the vaccine was pulled off the market. Furthermore, the contamination reinforced the distinct need for a standard practice in biosafety testing utilizing highly sensitive technologies, such as MPS/NGS. An advantage of MPS/NGS is the capability to sequence the whole genome[2] and the transcriptome,[3] as well as target against specific genomic regions such as the exome[4] or a gene panel focused on a particular disease.[5] Thus, the test can be customized for researchers' specific needs.

As a technology, MPS/NGS can be studied systematically for use in biosafety testing by introducing an unknown contaminating virus into a culture of known DNA. Importantly, the limit of detection of MPS/NGS can be determined by decreasing the concentration of the spiked-in contaminant to a point where the nucleic acid is not detected. The resulting sequence data of these contamination studies can be mapped to reference genomes to determine the origin of nucleic acid present within a sample. The presence of a contaminant can be confirmed if any sequence exists outside of the genomes known to be in the sample. The identity of any virus contaminant can be determined by mapping the remaining sequence reads to a virus database.

In this study, we tested the use of MPS/NGS as a biosafety tool by contaminating a virus vaccine stock with a DNA virus and a mammalian cell bank with an RNA virus. SV40, the DNA virus, was spiked into three Ad5 virus vaccine stocks in concentrations decreasing ten-fold in each sample. A fourth sample contained Ad5 alone as a control. The RNA virus, measles, was spiked into a mammalian MRC-5 cell bank as a model for virus-contaminated mammalian cell culture. The extracted DNA and RNA from the measles spike-in experiment were used to create two samples. All samples were sequenced using MPS/NGS and then mapped to reference genomes known to be present in the sample. Any remaining reads were mapped against a virus database to determine the identity of any unknown, contaminating virus nucleic acid.

Materials and Methods

Sample Preparation

A titered stock of Ad5 (2.7×10^7 tissue culture infectious dose 50 [TCID50] per mL prepared at Charles River Laboratories from ATCC® VR-5™ grown in human A549 cells) was spiked with known amounts of SV40 (prepared at Charles River Laboratories) at ratios of 1 SV40 pfu to 10^2, 10^3, or 10^4 infectious Ad5 particles. Nucleic acid was purified from the mixtures as well as an unspiked control Ad5 sample using QIAamp DNA Blood Mini Kit (QIAGEN) at Charles River Laboratories.

Measles virus (Edmonston strain, ATCC VR-24™) with a titer of 8.0×10^4 TCID 50/mL was infected onto MRC-5 cells, seeded into T-225 cm^2 flasks at 50% confluency, at a multiplicity of infection of 0.05. Each flask was inoculated with virus in non-complete 1× MEM growth medium. Following adsorption for 1 h at 36.5 ± 1°C, complete growth medium (1× MEM + 5% FBS) was added and cells were incubated for seven days (36.5 ± 1°C, 5% CO_2) with an observed 60% cytopathic effect (CPE) . Then one infected flask was mixed with six uninfected MRC-5 cell culture flasks to dilute the amount of virus present. Cells were washed with PBS, pelleted by centrifugation, and then frozen. One DNA preparation and one RNA preparation were derived from a measles-infected MRC-5 cell pellet using the QIAamp DNA Blood Mini Kit and the QIAamp RNA Viral Mini Kit (QIAGEN), respectively.

DNA Library Preparation and HiSeq Sequencing

Genomic DNA (gDNA) samples were quantified using a Qubit® 2.0 Fluorometer (Invitrogen, Life Technologies), and the DNA integrity was checked on a 1% agarose gel. DNA library preparations, sequencing reactions, and bioinformatics analyses were conducted at GENEWIZ, Inc.

Illumina TruSeq™ DNA library preparation and clustering was performed and sequencing reagents were used according to the manufacturer's recommendations. Briefly, the gDNA was fragmented and subsequently end-repaired. Adapters were ligated after adenylation of the 3' ends, followed by limited cycle PCR enrichment to add barcodes for multiplexing. The sizes of the resulting DNA fragments were then validated using a DNA 1000 Chip on the Agilent 2100 Bioanalyzer (Agilent Technologies). The nucleic acid of the DNA sequencing libraries were quantified using the Qubit 2.0 Fluorometer and by real-time qPCR (Applied Biosystems® 7500 Real-Time PCR System, Life Technologies).

Libraries were clustered on four lanes of one flow cell using the cBOT clonal cluster generation instrument (Illumina) following the manufacturer's instructions. After clustering, the libraries were loaded onto the Illumina HiSeq 2000 instrument using a 2×100 bp PE configuration. Image analysis and base calling were conducted with Illumina HiSeq Control Software (HCS) on the HiSeq 2000 instrument. The resulting data was quality tested using Illumina CASAVA 1.8.2 informatics software (Table 1).

RNA Library Preparation and HiSeq Sequencing

RNA samples were quantified using the Qubit fluorometer. The RNA integrity was checked with the Agilent bioanalyzer. Illumina TruSeq RNA library preparations, sequencing reactions, and initial bioinformatics analysis were conducted at GENEWIZ, Inc. RNA library preparation, clustering, and sequencing reagents were used throughout the process following the manufacturer's recommendations. Briefly, 2 µg of RNA was used as starting material for library preparation. Host ribosomal RNA was removed using the RiboMinus™ Eukaryote Kit (Life Technologies). Subsequently, first-strand and second-strand DNA were synthesized. Adapters were ligated after adenylation of the 3′ ends, followed by enrichment and barcode addition for multiplexing by limited-cycle PCR. DNA libraries were validated using the Agilent bioanalyzer with the DNA 1000 chip. DNA libraries were quantified using the Qubit fluorometer and by real-time PCR with the Applied Biosystems system. The samples were clustered on four lanes of a flow cell using the cBOT. After clustering, the samples were loaded on the HiSeq instrument according to manufacturer's instructions. The samples were sequenced using a 2×100 bp PE configuration. Image analysis and base calling were conducted by the Illumina HCS on the HiSeq instrument. The resulting data was quality tested using the CASAVA program (Table 2).

Data Analysis for Virus Detection

Raw sequence data generated from the HiSeq was converted into FASTQ files utilizing the CASAVA program. Then sequence reads for the four samples were aligned to the human genome GRCh37 (Genome Reference Consortium build 37). Sequence reads that could not be mapped to the human genome were collected and blasted against a viral genome database containing 4,193 reference genomes, which were downloaded from the National Center for Biotechnology Information (NCBI). Sequence reads that were mapped to a viral genome and had an E value $< 10^{-20}$ with cutoff values of 99% identity and > 99 nucleotide (nt) read lengths were retained. The BLAST (a rapid search algorithm) results were parsed and further analyzed. A similar workflow was used for the DNA and RNA data from the mammalian cell bank. Further investigations into sequence mapping were performed by changing cutoff criteria while retaining E values $< 10^{-20}$.

TABLE 1. Summary of MPS/NGS data quality from DNA virus contamination experiment.

Lane	Sample ID	Index	Description	Project	Yield (Mbases)	# Reads	% of Q Score ≥ 30	Mean Q Score
1	Sample 1	CGATGT	PE100	AVT_Pilot1	29,126	288,372,530	91.23	35.48
2	Sample 2	TGACCA	PE100	AVT_Pilot1	26,547	262,838,866	93.29	36.46
3	Sample 3	ACAGTG	PE100	AVT_Pilot1	25,541	252,877,696	90.65	35.32
4	Sample 4	GCCAAT	PE100	AVT_Pilot1	32,911	325,854,480	90.40	35.34

TABLE 2. Summary of MPS/NGS data quality from RNA virus contamination experiment.

Lane	Sample ID	Index	Description	Project	Yield (Mbases)	# Reads	% of Q Score ≥ 30	Mean Q Score
1	CRO1RNA	GTGGCC	PE100	AVT_Pilot2	43,157	427,300,002	89.57	34.68
2	CRO2RNA	ATCACG	PE100	AVT_Pilot2	30,191	298,918,694	93.29	36.59

Results

DNA Virus Detection and Identification

Isolated DNA from the Ad5 virus vaccine was mixed with SV40 DNA in a single-blind study. To determine the limit of detection, four samples were prepared. Sample 1 contained SV40 and Ad5 in a 1:100 ratio, Sample 2 at 1:1000, Sample 3 at 1:10,000, and the fourth, "Control," contained Ad5 only as a negative control. The sequencing and bioinformatics were performed blind of the spike-in DNA identity and concentrations. Illumina HiSeq 2000 MPS/NGS reactions yielded an average of 282.4 million sequencing reads for the four samples (summarized in Table 3). Between 52.1–95.4% of the reads mapped to the human genome since these are remnants of the virus vaccine host cell DNA. 4.2–47.0% of the reads mapped to the Ad5 genome. The majority of the remaining reads mapped to the contamination, which bioinformatics accurately identified as SV40.

Interestingly, 30 reads that mapped to SV40 were identified within the 288.3 million total reads of the negative control (Table 3, "Control"). These 30 reads were sufficient to identify the SV40 virus DNA but only accounted for 39.5% of the SV40 genome. The remaining reads, ranging from 0.4–0.7% of the total reads, contained information that was unmapped to the reference sequences.

SV40 Infectivity and qPCR

The Ad5 stock that generated the 30 SV40 sequence reads was tested using an infectivity assay followed by a targeted qPCR assay.

In the infectivity assay, Ad5 stock was first treated with an Ad5-neutralizing antibody (abcam®), then mixed with a culture of VERO-76 cells (ATCC CRL-1587™), and finally tested for the presence of SV40 by immunofluorescence (Anti-SV40 T Antigen, US Biological). The neutralized-Ad5-plus-SV40 positive control showed both a CPE and a positive immunofluorescence signal on each day. The neutralized-Ad5 sample did not generate CPE, and could not be detected by immunofluorescence on either day.

In the SV40-specific qPCR assay, Ad5 stock was analyzed for the presence of SV40 nucleic acid. In one out of 11 qPCR reactions, fewer than ten target copies of SV40 DNA were detected. This represents < 10 SV40 sequences in a background Ad5 titer of 1.54×10^6 TCID50. No SV40 sequences were detectable by qPCR in the A549 cell stock, which was used to propagate the Ad5 stock.

RNA Virus Detection and Identification

In a single-blind experiment, nucleic acid was extracted from a human cell line infected with measles virus to model a virus-contaminated mammalian cell culture. The blinded sample, split into a DNA or RNA fraction, was sequenced using MPS/NGS. The sequencing and bioinformatics steps revealed no DNA virus contamination accurately—99.32% of the reads mapped to the human cell line (Table 4). However, 0.21% of the sequencing reads from the RNA fraction mapped

TABLE 3. Results from MPS/NGS of DNA extracted from an Ad5 virus vaccine that was spiked with SV40 virus DNA to mimic contamination.

Blinded Sample Name	Sample 1	Sample 2	Sample 3	Control
Input DNA	"1:100"	"1:1,000"	"1:10,000"	"Control"
Total Sequence Reads (M)	262.8	252.8	325.8	288.3
Reads Mapped to Human Genome (M)	202.4 (77.1%)	131.8 (52.1%)	194.4 (59.7%)	275.1 (95.4%)
Reads Mapped to Ad5 Genome (M)	50.2 (19.1%)	118.9 (47.0%)	129.7 (39.8%)	12.0 (4.2%)
Reads Mapped to Simian Virus 40 (M)	8.2 (3.1%)	1.1 (0.4%)	0.21 (0.06%)	0.00003 (0.0%)
Remaining Reads (M)	2.00 (0.7%)	0.96 (0.4%)	1.50 (0.5%)	1.20 (0.4%)

TABLE 4. Results from MPS/NGS of nucleic acid extracted from a mammalian cell bank that was spiked with RNA to mimic a contaminated master cell bank.

Blinded Sample Name	CRO1RNA	CRO1DNA
Total Sequence Reads (M)	427.3	298.9
Reads Mapped to Human Genome (M)	421.1 (98.54%)	296.9 (99.32%)
Measles Virus, Complete Genome (M)	0.9 (0.21%)	0.0 (0.0%)
Remaining Reads (M)	5.3 (1.25%)	2.0 (0.68%)

to the measles virus, yielding the correct identification of the spiked-in virus. Further sequence analysis revealed that 99.0% of the 15,894 bp measles genome was covered, with a ≥ 95% identity to the reference genome. The remaining reads, ranging from 0.68–1.25% of the total reads, contained information that was unmatched to reference sequences.

Discussion

Data obtained from the MPS/NGS of the virus vaccine and mammalian cell bank demonstrate the ability to detect spiked-in contaminating virus DNA and RNA at very low levels. MPS/NGS is also successful in detecting contaminants that are not part of targeted contaminant detection strategies. Furthermore, an updatable virus database and adjustable bioinformatics stringency thresholds ensure no loss of original sequence data while providing for future analysis of sequencing data for new viruses as they are discovered, sequenced, and added to the database. The workflow described here, in which biologics are tested for adventitious virus contamination, can provide valuable information when included as part of a safety monitoring program.

In total, 99.3% or more of sequence reads from all virus vaccine samples were identified as either human (host cell DNA), Ad5 (virus vaccine), or SV40 (spiked-in contaminant). While testing the level of sensitivity and limit of detection of this MPS/NGS process, 30 reads of SV40 were identified within the "Control" sample — sufficient to identify the SV40 virus and to suggest a limit of detection of ~ 1.04×10^{-7} in a total of 288.3×10^{6} reads. To determine the significance of the SV40 DNA in the presumed negative sample, the original lot of Ad5 was tested using a specific SV40 virus viability study and a targeted qPCR assay. These subsequent test results verified that the contaminating nucleic acid existed in the original stock and was not from a viable SV40 virus. Importantly, these results demonstrate the high sensitivity of the MPS/NGS assay. Specifically, the non-targeted MPS/NGS assay detected the contamination in the control sample in concentrations high enough to identify it. Due to the high sensitivity of the MPS/NGS assay, there is the potential to detect non-viable, environmental DNA. However, the benefit of detecting heretofore unassayed, public health-relevant, virus contamination is evident.

Targeted testing of biologics has the disadvantage of missing contaminants. Particularly, the standard cadre of PCR or qPCR assays designed for detecting DNA and RNA contamination is designed explicitly for known contaminations. By using MPS/NGS technologies, all nucleotides within a sample are observed—not just the targeted prokaryotic, eukaryotic, and virus sequences. Through the use of a well-designed bioinformatics pipeline, these nucleotides can be distinguished between acceptable components and those that may be contaminants. Furthermore, with a strong understanding of the host genome, bioinformatics can be used to detect endogenous retroviruses and nucleotide changes within small genomes.

In this study, virus contamination was identified through the use of a bioinformatics workflow and a database of 4,193 virus sequences. One advantage of this workflow (Figure 1) is the ease in which it can be updated from multiple

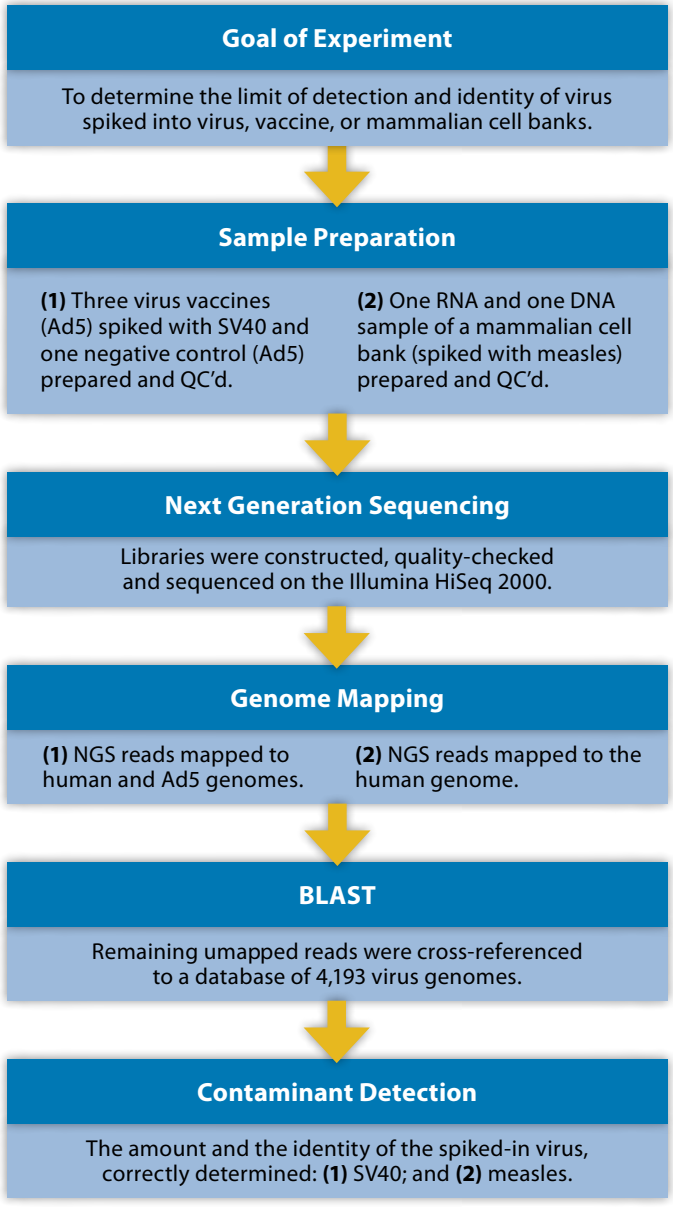
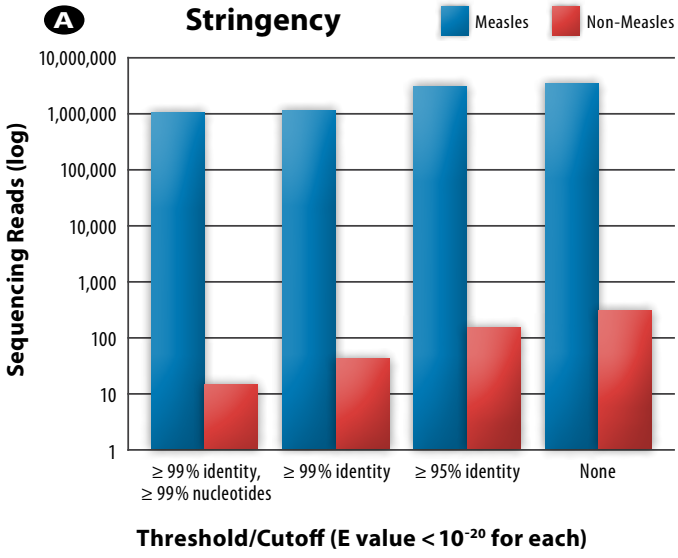
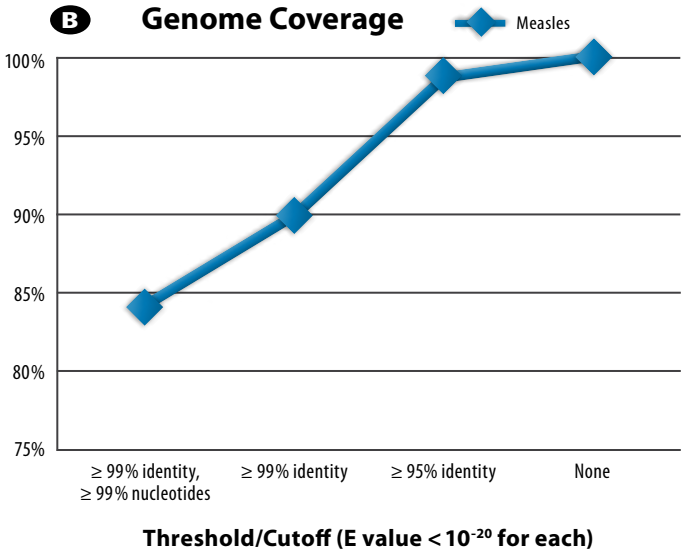


FIGURE 1. Workflow of experimental procedures: **(1)** SV40 spike-in; and **(2)** measles spike-in.

public sources. As an example, in the measles spike-in experiment, there are 5.3×10^6 reads (Table 4) that do not map to the human, measles, or any other virus genome in the database. These remain unmapped because of the high stringency established for mapping to known sequences (human and virus), some unresolved base pairs and short reads, and reads not in the databases. Additional analysis can be performed on the MPS/NGS data as new virus sequence information is added to the database without sequencing the sample again. This is in contrast to current direct target technologies where, as new viruses are detected, new assays are required to be run. Keeping the database up to date will be essential for the identification of nucleic acid components of this and other samples over time. In addition, querying an updated database over time can assist in the characterization of a biologic. Another advantage of the bioinformatics workflow is the ability to easily change the stringency requirements without the need to sequence the sample again. This saves precious sample while providing detailed virus information down to a single base pair. For example, the reads mapped to the measles genome increase from 90.2% to approximately 100% (Figure 2) when the stringency is lowered. The number of reads that match other viruses in the database also increase from 15 to 157 as the stringency is lowered. All of those reads at the lowest stringency (E value $< 10^{-20}$) represent less than 15% coverage of any distinct virus genome. We do recommend follow-up studies (bioinformatic evaluation followed by other assays) depending on the value of the bank and the potential risk.

This study demonstrates the advantages of MPS/NGS in the detection of adventitious virus contamination. We propose that MPS/NGS should be used as a routine screening procedure to identify all potential virus contamination. Subsequent assays should be performed to provide information on any infectivity of the viruses detected by MPS/NGS. The proposed workflow should ensure valuable safety monitoring information and allow for limitless bioinformatic inquiry.





C

	CRO1RNA (total reads)	CRO1DNA (total reads)	Threshold/ Cutoff (E value < 10^{-20} for each)	Measles Genome Coverage
Measles	904776	0	≥ 99% identity, ≥ 99% nucleotides	84.0%
Non-Measles Virus	15	16	—	—
Measles	1094806	0	≥ 99% identity	90.2%
Non-Measles Virus	45	21	—	—
Measles	3112503	0	≥ 95% identity	99.0%
Non-Measles Virus	157	21	—	—
Measles	3238957	0	None	~100%; only the first 10 nt are not covered
Non-Measles Virus	322	22	—	—

FIGURE 2. Results of *in silico* experiments. The remaining unmapped reads from the measles spike-in experiment were compared with sequences in the virus database as the stringency was decreased. **(A)** MPS/NGS reads mapped to measles virus sequence increase as the bioinformatics stringency is decreased (blue). The reads mapped to other virus sequences in the database increase as stringency is decreased (red). These detected virus nucleic acids do not represent more than 15% coverage of any virus genome other than measles. **(B)** The sequence coverage of the measles genome approaches 100% as the stringency is decreased. **(C)** Summary of *in silico* results.

Acknowledgments

Our special thanks go to Ali Manzar, MS, and Genna Rubnitz, MA, for their critical review and editing of this paper, and Jason Hoopes for nucleic acid preparations.

References

[1] Victoria JG, Wang C, Jones MS, Jaing C, McLoughlin K, Gardner S, Delwart EL. Viral nucleic acids in live attenuated vaccines: detection of minority variants and an adventitious virus. *J Virol*, 2010; 84:6033. http://dx.doi.org/10.1128/JVI.02690-09. PMid:19028524.

[2] Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics*, 2012; 13:818. http://dx.doi.org/10.1038/nrg3226.

[3] Costa V, Angelini C, De Feis I, Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol*, 2010; 2010:853916. PMid:14172021, PMCid:441218.

[4] Bamsha MJ, Ng SB, Biham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 2010; 11:31.

[5] Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. *Nature Reviews Genetics*, 2013; 14:295. http://dx.doi.org/10.1038/nrg3463. PMid:16274824.

About the Authors

Jack X. Yu[1], Stephen Karakasidis[2], Ginger Zhou[1], Wenying Huang[1], Yankai Jia[1], Michael J. Hantman[2]*, and Jeffrey A. Shaman[1]*

1. GENEWIZ, Inc., 115 Corporate Boulevard, South Plainfield, New Jersey 07080 USA
2. Charles River Laboratories, 358 Technology Drive, Malvern, Pennsylvania 19355 USA

Drs. Shaman and Hantman are the corresponding authors:

Email: Jeffrey.Shaman@genewiz.com; Phone: 908-222-0711
Email: Michael.Hantman@crl.com; Phone: 610-407-1096