AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems™

CrossMark
← click for updates

Multiplexed Metagenomic Deep Sequencing To Analyze the Composition of High-Priority Pathogen Reagents

Michael R. Wilson,[a,b] Greg Fedewa,[c] Mark D. Stenglein,[d] Judith Olejnik,[e,f]
Linda J. Rennick,[e,f] Sham Nambulli,[e,f] Friederike Feldmann,[g] W. Paul Duprex,[e,f]
John H. Connor,[e,f] Elke Mühlberger,[e,f] Joseph L. DeRisi[b,h]

Department of Neurology, University of California, San Francisco, San Francisco, California, USA[a]; Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California, USA[b]; Integrative Program in Quantitative Biology, Bioinformatics, University of California San Francisco, San Francisco, California, USA[c]; Department of Microbiology, Immunology, and Pathology, Colorado State University, Fort Collins, Colorado, USA[d]; Department of Microbiology, Boston University School of Medicine, Boston, Massachusetts, USA[e]; National Emerging Infectious Diseases Laboratories, Boston University, Boston, Massachusetts, USA[f]; Rocky Mountain Veterinary Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, Montana, USA[g]; Howard Hughes Medical Institute, Chevy Chase, Maryland, USA[h]

ABSTRACT Laboratories studying high-priority pathogens need comprehensive methods to confirm microbial species and strains while also detecting contamination. Metagenomic deep sequencing (MDS) inventories nucleic acids present in laboratory stocks, providing an unbiased assessment of pathogen identity, the extent of genomic variation, and the presence of contaminants. Double-stranded cDNA MDS libraries were constructed from RNA extracted from *in vitro*-passaged stocks of six viruses (La Crosse virus, Ebola virus, canine distemper virus, measles virus, human respiratory syncytial virus, and vesicular stomatitis virus). Each library was dual indexed and pooled for sequencing. A custom bioinformatics pipeline determined the organisms present in each sample in a blinded fashion. Single nucleotide variant (SNV) analysis identified viral isolates. We confirmed that (i) each sample contained the expected microbe, (ii) dual indexing of the samples minimized false assignments of individual sequences, (iii) multiple viral and bacterial contaminants were present, and (iv) SNV analysis of the viral genomes allowed precise identification of the viral isolates. MDS can be multiplexed to allow simultaneous and unbiased interrogation of mixed microbial cultures and (i) confirm pathogen identity, (ii) characterize the extent of genomic variation, (iii) confirm the cell line used for virus propagation, and (iv) assess for contaminating microbes. These assessments ensure the true composition of these high-priority reagents and generate a comprehensive database of microbial genomes studied in each facility. MDS can serve as an integral part of a pathogen-tracking program which in turn will enhance sample security and increase experimental rigor and precision.

IMPORTANCE Both the integrity and reproducibility of experiments using select agents depend in large part on unbiased validation to ensure the correct identity and purity of the species in question. Metagenomic deep sequencing (MDS) provides the required level of validation by allowing for an unbiased and comprehensive assessment of all the microbes in a laboratory stock.

KEYWORDS: metagenomics, pathogen tracking, phylogenetic analysis

Metagenomic deep sequencing can be multiplexed to ensure high priority virus stock integrity and tracking

Virology laboratories must maintain constant vigilance regarding the provenance of strains in their collections. In addition to confirming that experiments are being performed on the intended virus, unbiased methods that can detect viral coinfections

and/or other microbial contaminants are critical for ensuring that experiments are faithfully reporting on the biology of a particular virus and not on a polymicrobial exposure. Stock assurance is exceptionally important when working with viruses which require a high level of biocontainment, given the increased security requirements and higher costs and higher stakes of the research.

The validity and reproducibility of experiments, in addition to security and tracking concerns, demand an approach that can provide not only simple viral identification but also the entire spectrum of single nucleotide variants (SNVs) and their respective frequencies within a given stock. Molecular subtyping was critical in determining the source of the 2001 bioterrorism-associated anthrax outbreak (1). Furthermore, the recent missteps in the handling of *Bacillus anthracis*, influenza virus, and smallpox virus highlight the potential for inadvertent laboratory contamination and the need to improve isolate-tracking methods (2, 3).

In this study, we sequenced virus stocks of six distinct negative-sense RNA viruses, each of them with a unique passage history: (i) mycoplasma-contaminated Ebola virus (EBOV), (ii) mycoplasma-contaminated La Crosse virus (LACV) grown on cells with a known lymphocytic choriomeningitis virus (LCMV) contamination, (iii) vesicular stomatitis virus (VSV) grown on LCMV-contaminated cells, (iv) recombinant human respiratory syncytial virus (rHRSV) expressing enhanced green fluorescent protein (EGFP) from an additional transcription unit (ATU) (4), (v) recombinant canine distemper virus (rCDV) expressing Venus fluorescent protein from an ATU (5), and recombinant measles virus (rMV) expressing EGFP from an ATU (6). The investigators performing the sequence analysis were blinded to the virus isolates and the origins and known contaminations of the virus stocks. Here, we demonstrate the utility of metagenomic deep sequencing (MDS) for comprehensively assessing multiple viral isolates. We show that MDS and custom bioinformatics pipelines utilizing publicly available and free software packages can detect a wide range of bacterial and viral contaminants in high-priority virus stocks and that SNV analysis can successfully discriminate between two closely related EBOV isolates. We also demonstrate how dual-indexed barcodes dramatically decrease the false-positive assignment of sequencing reads to an input sample in multiplexed high-titer virus MDS libraries.

RESULTS

Importance of dual-index barcoding. In many sequencing applications, including MDS, it is critical to be able to distinguish low-level true-positive samples from false positives that result from assignment of sequencing reads to the wrong index ("index bleed-through"). Chimeric molecules produced during amplification steps in library preparation or on the sequencer during cluster generation are particularly susceptible to such misassignment. Creating library molecules with barcode sequences on both ends of the molecule (dual indexing) has been shown to reduce the rate at which this misassignment occurs by requiring concordance between the two barcode sequences (7).

To corroborate the benefit of minimizing misassignment of sequences by using dual indexing in the context of MDS, we analyzed the rate of index bleed-through in an independently generated sequencing data set that was demultiplexed using one or both index reads. We measured the rate of bleed-through in a set of 50 samples that included two samples that were positive for a recently discovered reptarenavirus. The other 48 samples were negative by quantitative reverse transcription PCR (qRT-PCR). We aligned reads from all 50 pooled samples to a virus genome segment present in the two positive samples (8). The data sets contained a median of 2.75×10^6 read pairs, and the two positive samples contained 65,284 and 203,565 virus-mapping reads (Fig. 1). Index bleed-through was determined to have occurred when reads identified as one of the 48 virus-negative samples by barcode actually aligned to this particular reptarenavirus genome segment. The median rate of bleed-through in the data sets that were demultiplexed using a single index was 17.9 per million filtered reads (Fig. 1). The median rate in dual-index demultiplexed data sets was 0.48 per million reads. Thus, in
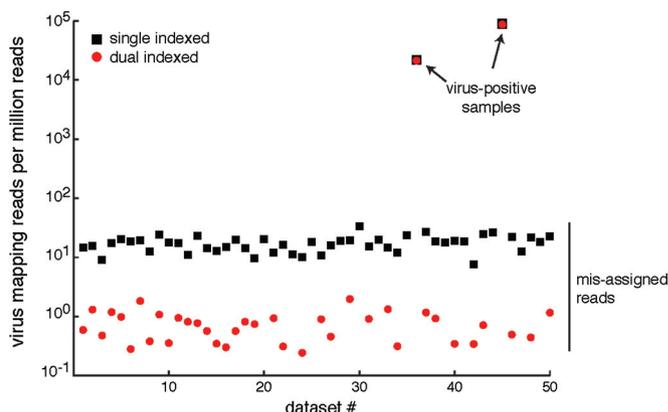
FIG 1 Dual indexing decreases the median rate of read misassignment by nearly 40-fold. Libraries from 50 samples were pooled and sequenced together. Two of the samples (indicated by arrows) were positive for snake arenavirus, and the other 48 were negative. Data sets were demultiplexed using a single index sequence (black squares) or dual index sequences (red circles), and reads from each data set were mapped with high stringency to the virus sequence. The number of virus mapping reads per million quality-filtered reads is indicated. Some dual-index-demultiplexed data sets had no misassigned reads. In these cases, red circles are not shown.

our data set, dual indexing reduced the median rate of read misassignment by 37-fold. The absolute number of mismapping reads decreased from a median of 50 reads to one read per data set.

Identification of microbes. The MDS of total RNA from each sample yielded 150-nucleotide (nt) paired-end sequences (Table 1). The data were analyzed using a rapid computational pipeline developed at the University of California, San Francisco (UCSF), to classify MDS reads and identify potential pathogens by comparison to the entire NCBI nucleotide reference database as described in Materials and Methods (9).

The results are summarized in Table 1. The EBOV and LACV samples were found to have significant contamination with *Mycoplasma* spp., with 3.65 and 0.02 *Mycoplasma* reads per viral read, respectively. The rMV sample contained more than 900 nonredundant, paired-end Epstein-Barr virus (EBV) sequences (0.013% of total reads) that were expected to come from the EBV-transformed human B-lymphoblastoid cell line (B-LCL) on which the virus was grown (10). Similarly, the rHRSV sample contained 121 paired-end sequences (0.0019% of total reads) that aligned to human papillomavirus 18 (HPV18). This was also unsurprising, since HEp-2 cells are the standard cell line used to culture HRSV and these are known to be a HeLa-contaminated cell line; HeLa cells were recently shown to have HPV18 DNA integrated into their genome (11). The LACV sample had evidence of Syrian hamster retroviruses, and it was confirmed that the virus had been grown on baby hamster kidney (BHK) cells. The rCDV sample contained sequence reads mapping to fowlpox virus, which had been used as part of the process

TABLE 1 Metagenomic deep sequencing results[a]

Sample	Total no. of sequencing reads	No. of unique nonhuman reads	Target virus	Other viral sequence(s)	Bacterial sequence
EBOV	47,328,387	17,777,406	EBOV (3,342,624)	None	*Mycoplasma hyorhinis* (12,215,655)
rMV	7,129,497	802,549	MV (159,184)	HHV4 (905)	Negative
rHRSV	6,274,384	1,165,688	HRSV (673,096)	HPV18 (121)	Negative
LACV	6,217,368	1,389,113	LACV (1,122,903)	LCMV (5,388), Syrian hamster IAP H10 (45), hamster gammaretrovirus (11)	*Mycoplasma arginini* (23,919)
rCDV	6,165,559	900,609	CDV (582,074)	Fowlpox (1,138), LACV (7)	Negative
VSV	10,164,946	2,130,521	VSV (1,481,456)	LCMV (3,315)	Negative

[a]Number of sequences aligning to each microbe are in parentheses. Abbreviations: EBOV, Ebola virus; rMV, recombinant measles virus; HHV4, human herpesvirus 4; rHRSV, recombinant human respiratory syncytial virus; LACV, La Crosse virus; HPV18, human papillomavirus 18; IAP, intracisternal A particle; LCMV, lymphocytic choriomeningitis virus; rCDV, recombinant canine distemper virus; VSV, vesicular stomatitis virus.

TABLE 2 SNVs in each virus genome segment that are at least 0.5% of population[a]

| Viral genome	Mean coverage	No. of SNVs:				
		Total	≥1%	≥10%	≥50%	≥90%
EBOV	8,334	143 (70)	49 (19)	3 (1)	2 (1)	1 (1)
MV	1,622	122 (80)	72 (42)	13 (2)	11 (1)	11 (1)
HRSV	5,429	115 (71)	54 (27)	2 (1)	2 (1)	2 (1)
LACV L segment	5,539	61 (45)	27 (18)	4 (1)	3 (0)	2 (0)
LACV M segment	10,077	31 (21)	11 (8)	6 (4)	0 (0)	0 (0)
LACV S segment	14,906	6 (4)	4 (2)	1 (0)	1 (0)	1 (0)
CDV	4,741	427 (170)	297 (91)	30 (6)	3 (0)	1 (0)
VSV	13,401	109 (72)	53 (31)	24 (10)	24 (10)	24 (10)

[a]Mean coverage is the mean number of non-PCR-duplicate reads that mapped to each base of the virus genome segment. Each SNV column is listed as the number of total SNVs, followed by the number of nonsynonymous SNVs in parentheses, which are at least the percentage of the population listed in the header. Abbreviations: EBOV, Ebola virus; MV, measles virus; HRSV, human respiratory syncytial virus; LACV, La Crosse virus; CDV, canine distemper virus; VSV, vesicular stomatitis virus.

to generate rCDV from plasmid. Because these RNA extractions were not DNase treated, we cannot rule out that some of the EBV, HPV18, and fowlpox virus sequences were remnants of genomic DNA. Last, the LACV and VSV samples were found to be contaminated with LCMV. None of the isolates had evidence of fungal contamination.

As evidence of the minimal index bleed-through in these dual-indexed samples, only seven viral read pairs unique to LACV were found in the rCDV sample, and no other sample cross-contamination was present in any of the other samples in this set, despite being processed in parallel. Regardless, the fact that low-level index bleed-through may still occur, despite dual indexing, highlights the need for additional layers of error correction. Future iterations of this technology will likely include added features that will further reduce bleed-through of multiplexed samples, such as unique molecular identifier (UMI) barcodes, in which each cDNA molecule is uniquely indexed at the time of first-strand synthesis (12).

SNV experiment. All of the viruses contained several SNVs compared to their reference sequences that ranged in frequency, including some consensus-level SNVs, as listed in Table 2. At a frequency of greater than 0.005, rCDV had 427 SNVs (170 of those being nonsynonymous), EBOV had 143 SNVs (70 nonsynonymous), the LACV L segment had 61 SNVs (45 nonsynonymous), the LACV M segment had 31 SNVs (21 nonsynonymous), the LACV S segment had 6 SNVs (4 nonsynonymous), rMV had 122 SNVs (80 nonsynonymous), rHRSV had 115 SNVs (71 nonsynonymous), and VSV strain Indiana (VSV[IN]) had 109 SNVs (72 nonsynonymous). A majority of SNVs are rare variants, having a frequency of less than 0.10. All six viruses had nonsynonymous and synonymous SNVs at frequencies of greater than 0.10. Four viruses had nonsynonymous SNVs at a frequency of at least 0.90, while five viruses had synonymous SNVs at a frequency of at least 0.90. Figure 2A displays the distribution of the SNV frequency on each genomic segment.

The investigators were blinded to the exact isolate of EBOV since the major goal of this study was to identify a state-of-the-art approach which could precisely characterize a virus reagent with isolate specificity in an unbiased fashion. As described in Materials and Methods, SNV analysis was performed to determine the EBOV isolate. Reads were mapped to two different reference genomes: (i) Zaire ebolavirus isolate Ebola virus/*Homo sapiens*-tc/COD/1976/Yambuku-Mayinga, complete genome strain (GenBank accession no. NC_002549), and (ii) Zaire ebolavirus strain Kikwit, complete genome (GenBank accession no. JQ352763). Comparing rates of correctly mapped reads did not reveal isolate identity. The same 2,501,8050 filtered read-pairs mapped at virtually identical percentages, 21.03% to EBOV Mayinga and 21.05% to EBOV Kikwit. However, SNV analysis revealed 223 SNVs in the reads that mapped to EBOV Mayinga with a frequency of ≥0.90, while there was only one SNV with a frequency of ≥0.90 in the reads that mapped to EBOV Kikwit (Fig. 2B). Thus, we correctly concluded that the precise isolate sequenced in this experiment was EBOV Kikwit.
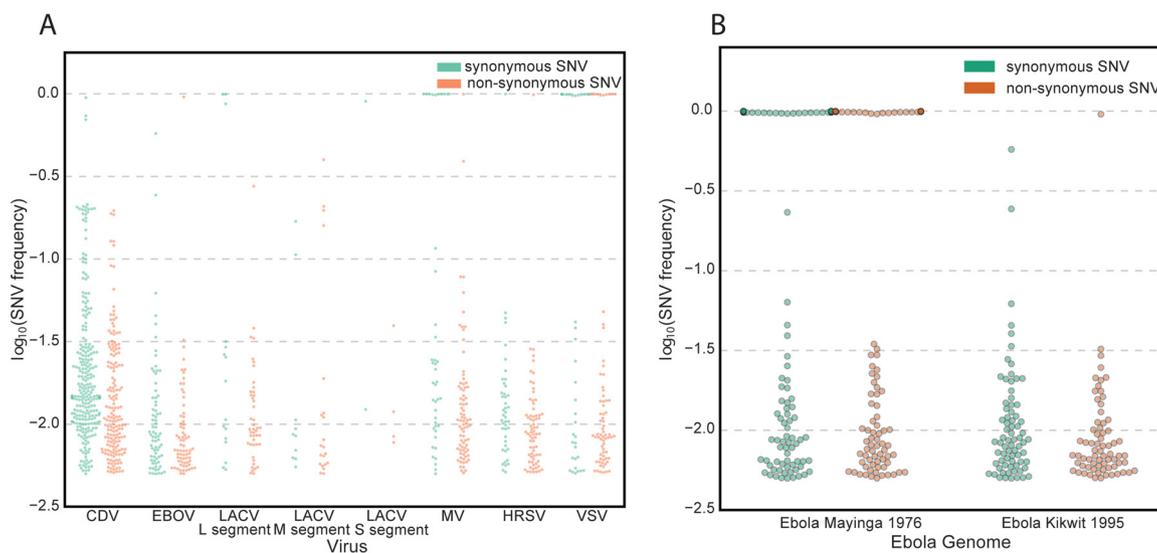
FIG 2 Distribution of single nucleotide variants (SNVs) for each virus and for two Ebola virus strains. For each plot, the *y* axis is the log$_{10}$(SNV frequency) in order to display the range of low-frequency SNVs more accurately. (A) SNV analysis for each virus sample using its reference genome. Abbreviations: CDV, canine distemper virus; EBOV, Ebola virus; LACV, La Crosse virus; MV, measles virus; HRSV, human respiratory syncytial virus; VSV, vesicular stomatitis virus. (B) SNV analysis of 2,501,8050 filtered read-pairs revealed 223 SNVs in the reads that mapped to Ebola virus Mayinga with a frequency of ≥0.90, while there was only one SNV with a frequency of ≥0.90 in the reads that mapped to Ebola virus Kikwit.

DISCUSSION

We describe the successful application of a multiplexed MDS assay and bioinformatics pipelines for comprehensively interrogating the provenance of six virus isolates in a blinded fashion. With regard to multiplexing, the frequency of reads that are assigned an incorrect barcode is a function of the abundance of those reads in the correct barcode library and the degree of multiplexing. So, for high-titer viral stock-derived data sets, where viral reads are abundant, bleed-through occurs at a higher frequency. We demonstrated that dual indexing reduces the misassignment of sequencing reads by 37-fold in a pilot experiment and saw essentially no intersample cross-contamination in our multiplexed assay despite very-high-titer virus isolates. It is important to note that most commercially available dual-indexing kits do not provide this benefit because the reduction in misassignment depends on samples being identified by unique index pairs. Most commercially available kits do not actually provide unique index pairs. Instead, they mix and match indexes to create pairs with repeated index sequences, for instance, by permuting eight i5 indexes with 12 i7 indexes to achieve 96 pairs. We also found that using dual indexing decreased the total number of reads per data set by ~6% due to the removal of unassigned read pairs. However, this is a relatively small tradeoff for the large decrease in misassigned reads.

Contaminating sequences were accurately identified, including bacteria (*Mycoplasma* spp.) and six viruses across the six virus stocks. Four of these viral contaminating sequences (i.e., EBV, HPV18, hamster gammaretrovirus, and LCMV [*n* = 2]) reflected the cells on which the viruses were cultured (10, 11, 13, 14). The fowlpox virus present in the rCDV sample was a remnant of the procedure used to generate rCDV, which uses a recombinant fowlpox virus to express T7 RNA polymerase (5). This is unsurprising given the low passage number of the rCDV.

The results from MDS of different virus stocks highlight the utility of this approach for in-depth analysis of virus sequences. There were different contaminating pathogens in each of the stocks, and these biological "fingerprints" provided precise information about their origin and passage history. Such information is invaluable for virologists embarking on time-consuming and expensive studies in biocontainment but is also useful for a broad range of microbiologists who would benefit from sample assurance. For rMV and rHRSV, there was evidence of the cell lines used to propagate the viruses,

and for rCDV, there was evidence of the helper virus used for reverse genetics. For viruses with an extensive propagation history, the cell type in which the virus had been passaged could be inferred based on contaminating sequences (i.e., hamster retrovirus sequences) as well as the additional, nonmicrobial sequences that can be assembled to characterize aspects of the host cell genome. This suggests that MDS is useful for virus forensic analysis, helping to identify the manner and cell type in which the virus was cultured or recovered.

The SNV analysis identified numerous synonymous and nonsynonymous mutations present in each of the isolates. The presence and frequency of such mutations in longitudinal samples may be used to monitor inevitable virus adaptation rigorously in the laboratory setting and the quality of seed stocks (15) as well as to define particular isolates (16). Here, SNV analysis allowed the precise determination of the particular EBOV isolate (Fig. 2B). Last, detailed cataloging of SNVs for each isolate allows rapid identification and tracking of high-priority pathogens in the event of an accidental or nonaccidental release of virus into the environment and could be an integral part of any microbial source-tracking program. Viral genome sequencing including SNV analysis has become an important tool to monitor viral spread, viral evolution, and routes of transmission in an outbreak situation as exemplified by the recent EBOV disease outbreak in West Africa (17–26). If a validated reference sequence does not exist, we recommend that the first step be a *de novo* assembly. However, this should be followed with a remapping of all the reads back to the assembled reference and a LoFreq* or similar SNV analysis, as we have demonstrated here. The latter analysis will reveal the presence of variations or even a mixture of strains that would be obscured or lost through an assemble-and-BLAST strategy alone.

In summary, we show that a single multiplex MDS assay can comprehensively assess (i) virus and isolate identity, (ii) SNVs in virus populations, (iii) the presence of viral coinfection, and (iv) the presence of bacterial contamination and (v) can provide information about the cell line used to propagate the virus.

MATERIALS AND METHODS

Viruses and RNA purification. All work with infectious EBOV was performed under biosafety level 4 (BSL-4) conditions at the Integrated Research Facility, Rocky Mountain Laboratories (RML), Division of Intramural Research, NIAID, NIH, Hamilton, MT. Vero E6 cells (ATCC CRL 1586) were infected with EBOV at a multiplicity of infection (MOI) of 1, and virus-containing supernatants were clarified by low-speed centrifugation at 4 days postinfection (p.i.). Total RNA from supernatant (140 μl) was purified using the Qiagen viral RNA minikit, according to the RML standard operating procedures for virus inactivation. Purified RNA was eluted in H_2O. The EBOV stock used for this analysis contained a known mycoplasma contamination.

To generate rMV Khartoum, Sudan (rMV^{KS}) stocks, B-LCL cells were infected at an MOI of 0.01. After 4 days, when viral cytopathic effect was maximal, the stock was subjected to one freeze-thaw cycle to release the highly cell-associated measles virus into the medium. Cell debris was removed by low-speed centrifugation, and the supernatant was stored at −80°C as virus stock.

To generate rCDV Rhode Island (rCDV^{RI}) stocks, Vero cells expressing the CDV receptor canine CD150 were infected at an MOI of 0.01. After 2 days, when viral cytopathic effect was maximal, the stock was subjected to one freeze-thaw cycle to release cell-associated virus into the medium. Cell debris was removed by low-speed centrifugation, and the supernatant was stored at −80°C as virus stock.

To generate rHRSV^{B05} stocks, HEp-2 cells were infected at an MOI of 0.01. After 3 days, when viral cytopathic effect was maximal, the medium was transferred to 50-ml tubes, and cell debris was removed by low-speed centrifugation. The supernatant was stored at −80°C as virus stock.

The LACV H78 strain was grown on BHK cells (27). BHK cells were infected with passage 3 LACV H78 at an MOI of 0.01. Virus was propagated for 48 h before medium was removed. Cell debris was removed by low-speed centrifugation, and individual aliquots of virus were frozen for future analysis. The LACV stock contained a known mycoplasma contamination.

For growth of the VSV Indiana (VSV^{IN}) strain, BHK cells were infected with VSV^{IN} at an MOI of 0.01. Virus was propagated for 24 h before medium was removed. Cell debris was removed by low-speed centrifugation, and individual aliquots of virus were frozen for future analysis.

For rMV, rCDV, rHRSV, LACV, and VSV, 250 μl of virus-containing supernatant was mixed with 750 μl of Trizol-LS (Ambion), and RNA was isolated according to the manufacturer's protocol. Purified RNA was resuspended in H_2O.

Comparison of single and dual indexing. A library containing a pool of 50 dual-indexed libraries was created as previously described (8). The library was sequenced on an Illumina HiSeq 2500 instrument at the UCSF Center for Advanced Technology. The 150-nt read-length, paired-end, dual-indexed data set was manually demultiplexed using either one or both index reads. To measure the rate of read

misassignment, reads were first processed to remove low-quality and adapter sequences as described below. Trimmed reads were then aligned using the Bowtie2 tool to the sequence of a reptarenavirus genome segment (GenBank accession no. KP071661.1) actually present in two of the 50 samples (8, 28). Bowtie2 was run with parameters –local – qc-filter –score-min C,120,1. The number of aligning reads for each data set was counted.

Confirming virus identity and detecting contamination. RNA of six individual viral isolates extracted from supernatants of infected cells was provided for MDS. The investigators preparing the MDS libraries and performing the bioinformatics analysis were blinded to the presence or absence of any known microbial contaminants and also to the strain and isolate identity of each virus. The six viral isolates were LACV strain H78, EBOV, rCDV[RI], rMV[KS], rHRSV[B05], and the VSV[IN] strain. Samples were processed for MDS analysis as previously described (9). Samples were randomly amplified to double-stranded cDNA using the NuGEN Ovation v.2 kit (NuGEN, San Carlos, CA), and MDS libraries were constructed using the Nextera protocol (Illumina, San Diego, CA). Each library was dual indexed in the same manner as described above. Samples were pooled into a single library before library size and concentration were determined using the Blue Pippin (Sage Science, Beverly, MA) and Kapa universal quantitative PCR (qPCR) kit (Kapa Biosystems, Woburn, MA), respectively. Samples were sequenced on an Illumina HiSeq 2500 instrument using 150-nt paired-end sequencing. The paired-end sequences were analyzed for microbes using a rapid computational pipeline for microbial detection.

Bioinformatic pipeline. Paired-end reads were quality filtered using PriceSeqFilter (version 1.2, parameters –rqf 95 0.98), a component of the paired-read iterative contig extension (PRICE) assembler (29), followed by removal of human sequences by alignment to a combined reference genome including human genome build 38 (hg38) and chimpanzee (*Pan troglodytes*) using the Spliced Transcripts Alignment to a Reference (STAR) aligner (30). Unaligned reads that were at least 95% identical were compressed by cd-hit-dup (v4.6.1) (31, 32). After a second Bowtie2 alignment to remove residual human sequences (using an hg38 reference database), these reads were then used as queries to search the NCBI nt database (July 2015) using gsnapl (Genentech, v2015-09-29) (33).

SNV analysis. Illumina data from all the samples were processed in the same way for variant analysis. First reads were quality filtered using PriceSeqFilter (version 1.2) set to remove any read with less than 95% of nucleotides having a 0.98 probability of being correct (-rqf 95 0.98), any read with less than 90% called nucleotides, and any read that matched to the Illumina adapter sequences (29). Filtered high-quality reads were then aligned to reference genomes using GSNAP (version 2015-09-29) using default settings (33). Reference genome accession numbers were as follows: EBOV, JQ352763; LACV long (L) segment, NC_004108; LACV medium (M) segment, NC_004109; LACV short (S) segment, NC_004110; MV, HM439386; rHRSV, KF640637; VSV[IN], J02428. The rCDV[RI] sequence (unpublished) was supplied by W. Paul Duprex (Boston University). According to the recommended guidelines for the variant caller, PCR-duplicate reads were removed using Picard tools (version 2.2.4). Variants were called using LoFreq* (version 2.1.2) (34) using default settings with the exception of a conservative lower cutoff of ≥0.005 in frequency. LoFreq* is a sequencing read quality aware variant caller that models each location in the genome as a Poisson-binomial distribution of the number of nonreference bases. It also tests for common sequencing errors that may result in false positives, such as testing for the strand bias of a variant. It is also designed to take advantage of high-coverage ($>500\times$ coverage) data sets, such as those that we have (Table 2). It uses high-coverage data sets to call true-positive low-frequency variants while maintaining a low false-positive rate but still shows reasonable sensitivity at low coverage levels ($50\times$ coverage) (34). Variants were determined to be synonymous or nonsynonymous using a custom Python script.

Accession number(s). The 150-nt paired-end sequences are located in the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA accession number SRP076690).

REFERENCES

1. **Rasko DA, Worsham PL, Abshire TG, Stanley ST, Bannan JD, Wilson MR, Langham RJ, Decker RS, Jiang L, Read TD, Phillippy AM, Salzberg SL, Pop M, Van Ert MN, Kenefic LJ, Keim PS, Fraser-Liggett CM, Ravel J**. 2011. Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation. Proc Natl Acad Sci U S A **108:** 5027–5032. http://dx.doi.org/10.1073/pnas.1016657108.

2. **Hoffmaster AR, Fitzgerald CC, Ribot E, Mayer LW, Popovic T**. 2002. Molecular subtyping of Bacillus anthracis and the 2001 bioterrorism-associated anthrax outbreak, United States. Emerg Infect Dis **8:**1111–1116. http://dx.doi.org/10.3201/eid0810.020394.

3. **Cohen J**. 2014. Lab safety. Alarm over biosafety blunders. Science **345:**247–248. http://dx.doi.org/10.1126/science.345.6194.247.

4. **Lemon K, Nguyen DT, Ludlow M, Rennick LJ, Yüksel S, van Amerongen G, McQuaid S, Rima BK, de Swart RL, Duprex WP**. 2015. Recombinant subgroup B human respiratory syncytial virus expressing enhanced green fluorescent protein efficiently replicates in primary human cells and is virulent in cotton rats. J Virol **89:**2849–2856. http://dx.doi.org/10.1128/JVI.03587-14.

5. **Ludlow M, Nguyen DT, Silin D, Lyubomska O, de Vries RD, von Messling V, McQuaid S, De Swart RL, Duprex WP**. 2012. Recombinant canine distemper virus strain Snyder Hill expressing green or red fluorescent proteins causes meningoencephalitis in the ferret. J Virol **86:** 7508–7519. http://dx.doi.org/10.1128/JVI.06725-11.

6. **Lemon K, de Vries RD, Mesman AW, McQuaid S, van Amerongen G, Yüksel S, Ludlow M, Rennick LJ, Kuiken T, Rima BK, Geijtenbeek TB, Osterhaus AD, Duprex WP, de Swart RL**. 2011. Early target cells of measles virus after aerosol infection of non-human primates. PLoS Pathog **7:**e1001263. http://dx.doi.org/10.1371/journal.ppat.1001263.

7. **Kircher M, Sawyer S, Meyer M**. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res **40:**e3. http://dx.doi.org/10.1093/nar/gkr771.

8. **Stenglein MD, Jacobson ER, Chang LW, Sanders C, Hawkins MG, Guzman DS, Drazenovich T, Dunker F, Kamaka EK, Fisher D, Reavill DR, Meola LF, Levens G, DeRisi JL**. 2015. Widespread recombination, reassortment, and transmission of unbalanced compound viral genotypes in natural arenavirus infections. PLoS Pathog **11:**e1004900. http://dx.doi.org/10.1371/journal.ppat.1004900.

9. **Wilson MR, Shanbhag NM, Reid MJ, Singhal NS, Gelfand JM, Sample HA, Benkli B, O'Donovan BD, Ali IK, Keating MK, Dunnebacke TH, Wood MD, Bollen A, DeRisi JL**. 2015. Diagnosing Balamuthia mandrillaris encephalitis with metagenomic deep sequencing. Ann Neurol **78:** 722–730. http://dx.doi.org/10.1002/ana.24499.

10. **Van Binnendijk RS, Poelen MC, de Vries P, Voorma HO, Osterhaus AD, Uytdehaag FG**. 1989. Measles virus-specific human T cell clones. Characterization of specificity and function of CD4+ helper/cytotoxic and CD8+ cytotoxic T cell clones. J Immunol **142:**2847–2854.

11. **Ogura H, Yoshinouchi M, Kudo T, Imura M, Fujiwara T, Yabe Y**. 1993. Human papillomavirus type 18 DNA in so-called HEP-2, KB and FL cells—further evidence that these cells are HeLa cell derivatives. Cell Mol Biol (Noisy-le-grand) **39:**463–467.

12. **Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S**. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods **11:**163–166. http://dx.doi.org/10.1038/nmeth.2772.

13. **Switzer WM, Zheng H, Simmons G, Zhou Y, Tang S, Shankar A, Kapusinszky B, Delwart EL, Heneine W**. 2011. No evidence of murine leukemia virus-related viruses in live attenuated human vaccines. PLoS One **6:**e29223. http://dx.doi.org/10.1371/journal.pone.0029223.

14. **Van der Zeijst BA, Bleumink N, Crawford LV, Swyryd EA, Stark GR**. 1983. Viral proteins and RNAs in BHK cells persistently infected by lymphocytic choriomeningitis virus. J Virol **48:**262–270.

15. **Kugelman JR, Rossi CA, Wiley MR, Ladner JT, Nagle ER, Pfeffer BP, Garcia K, Prieto K, Wada U, Kuhn JH, Palacios G**. 2016. Informing the historical record of experimental nonhuman primate infections with Ebola virus: genomic characterization of USAMRIID Ebola virus/H.sapiens-tc/COD/1995/Kikwit-9510621 challenge stock "R4368" and its replacement "R4415." PLoS One **11:**e0150919. http://dx.doi.org/10.1371/journal.pone.0150919.

16. **Kuhn JH, Lofts LL, Kugelman JR, Smither SJ, Lever MS, van der Groen G, Johnson KM, Radoshitzky SR, Bavari S, Jahrling PB, Towner JS, Nichol ST, Palacios G**. 2014. Reidentification of Ebola virus E718 and ME as Ebola virus/H.sapiens-tc/COD/1976/Yambuku-Ecran. Genome Announc **2:**e01178-14. http://dx.doi.org/10.1128/genomeA.01178-14.

17. **Ladner JT, Wiley MR, Mate S, Dudas G, Prieto K, Lovett S, Nagle ER, Beitzel B, Gilbert ML, Fakoli L, Diclaro JW, II, Schoepp RJ, Fair J, Kuhn JH, Hensley LE, Park DJ, Sabeti PC, Rambaut A, Sanchez-Lockhart M, Bolay FK, Kugelman JR, Palacios G**. 2015. Evolution and spread of Ebola virus in Liberia, 2014-2015. Cell Host Microbe **18:** 659–669. http://dx.doi.org/10.1016/j.chom.2015.11.008.

18. **Hoenen T, Groseth A, Rosenke K, Fischer RJ, Hoenen A, Judson SD, Martellaro C, Falzarano D, Marzi A, Squires RB, Wollenberg KR, de Wit E, Prescott J, Safronetz D, van Doremalen N, Bushmaker T, Feldmann F, McNally K, Bolay FK, Fields B, Sealy T, Rayfield M, Nichol ST, Zoon KC, Massaquoi M, Munster VJ, Feldmann H**. 2016. Nanopore sequencing as a rapidly deployable Ebola outbreak Tool. Emerg Infect Dis **22:**331–334. http://dx.doi.org/10.3201/eid2202.151796.

19. **Moon SJ, Kim TH, Yoon SY, Chung JH, Hwang HJ**. 2015. Relationship between stage of chronic kidney disease and sarcopenia in Korean aged 40 years and older using the Korea National health and nutrition examination surveys (KNHANES IV-2, 3, and V-1, 2), 2008–2011. PLoS One **10:**e0130740. http://dx.doi.org/10.1371/journal.pone.0130740.

20. **Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, Soropogui B, Sow MS, Keïta S, De Clerck H, Tiffany A, Dominguez G, Loua M, Traoré A, Kolié M, Malano ER, Heleze E, Bocquin A, Mély S, Raoul H, Caro V, Cadar D, Gabriel M, Pahlmann M, Tappe D, Schmidt-Chanasit J, Impouma B, Diallo AK, Formenty P, Van Herp M, Gunther S**. 2014. Emergence of Zaire Ebola virus disease in Guinea. N Engl J Med **371:**1418–1425. http://dx.doi.org/10.1056/NEJMoa1404505.

21. **Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang PP, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A, Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally JL, Chapman SB, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheiffelin JS, Lander ES, Happi C, Gevao SM, Gnirke A, Rambaut A, Garry RF, Khan SH, Sabeti PC**. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science **345:**1369–1372. http://dx.doi.org/10.1126/science.1259657.

22. **Park DJ, Dudas G, Wohl S, Goba A, Whitmer SL, Andersen KG, Sealfon RS, Ladner JT, Kugelman JR, Matranga CB, Winnicki SM, Qu J, Gire SK, Gladden-Young A, Jalloh S, Nosamiefan D, Yozwiak NL, Moses LM, Jiang PP, Lin AE, Schaffner SF, Bird B, Towner J, Mamoh M, Gbakie M, Kanneh L, Kargbo D, Massally JL, Kamara FK, Konuwa E, Sellu J, Jalloh AA, Mustapha I, Foday M, Yillah M, Erickson BR, Sealy T, Blau D, Paddock C, Brault A, Amman B, Basile J, Bearden S, Belser J, Bergeron E, Campbell S, Chakrabarti A, Dodd K, Flint M, Gibbons A, Goodman C, Klena J, McMullan L, Morgan L, Russell B, Salzer J, Sanchez A, Wang D, Jungreis I, Tomkins-Tinch C, Kislyuk A, Lin MF, Chapman S, MacInnis B, Matthews A, Bochicchio J, Hensley LE, Kuhn JH, Nusbaum C, Schieffelin JS, Birren BW, Forget M, Nichol ST, Palacios GF, Ndiaye D, Happi C, Gevao SM, Vandi MA, Kargbo B, Holmes EC, Bedford T, Gnirke A, Ströher U, Rambaut A, Garry RF, Sabeti PC**. 2015. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. Cell **161:**1516–1526. http://dx.doi.org/10.1016/j.cell.2015.06.007.

23. **Kugelman JR, Wiley MR, Mate S, Ladner JT, Beitzel B, Fakoli L, Taweh F, Prieto K, Diclaro JW, Minogue T, Schoepp RJ, Schaecher KE, Pettitt J, Bateman S, Fair J, Kuhn JH, Hensley L, Park DJ, Sabeti PC, Sanchez-Lockhart M, Bolay FK, Palacios G, US Army Medical Research Institute of Infectious Diseases, National Institutes of Health, Integrated Research Facility—Frederick Ebola Response Team 2014–2015**. 2015. Monitoring of Ebola virus Makona evolution through establishment of advanced genomic capability in Liberia. Emerg Infect Dis **21:**1135–1143. http://dx.doi.org/10.3201/eid2107.150522.

24. **Simon-Loriere E, Faye O, Faye O, Koivogui L, Magassouba N, Keita S, Thiberge JM, Diancourt L, Bouchier C, Vandenbogaert M, Caro V, Fall G, Buchmann JP, Matranga CB, Sabeti PC, Manuguerra JC,**

Holmes EC, Sall AA. 2015. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. Nature **524:**102–104. http://dx.doi.org/10.1038/nature14612.

25. **Tong YG, Shi WF, Liu D, Qian J, Liang L, Bo XC, Liu J, Ren HG, Fan H, Ni M, Sun Y, Jin Y, Teng Y, Li Z, Kargbo D, Dafae F, Kanu A, Chen CC, Lan ZH, Jiang H, Luo Y, Lu HJ, Zhang XG, Yang F, Hu Y, Cao YX, Deng YQ, Su HX, Sun Y, Liu WS, Wang Z, Wang CY, Bu ZY, Guo ZD, Zhang LB, Nie WM, Bai CQ, Sun CH, An XP, Xu PS, Zhang XL, Huang Y, Mi ZQ, Yu D, Yao HW, Feng Y, Xia ZP, Zheng XX, Yang ST, Lu B, Jiang JF, Kargbo B, He FC, Gao GF, Cao WC, China Mobile Laboratory Testing Team in Sierra Leone**. 2015. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. Nature **524:**93–96. http://dx.doi.org/10.1038/nature14490.

26. **Hoenen T, Safronetz D, Groseth A, Wollenberg KR, Koita OA, Diarra B, Fall IS, Haidara FC, Diallo F, Sanogo M, Sarro YS, Kone A, Togo AC, Traore A, Kodio M, Dosseh A, Rosenke K, de Wit E, Feldmann F, Ebihara H, Munster VJ, Zoon KC, Feldmann H, Sow S**. 2015. Virology. Mutation rate and genotype variation of Ebola virus from Mali case sequences. Science **348:**117–119. http://dx.doi.org/10.1126/science.aaa5646.

27. **Chandler LJ, Borucki MK, Dobie DK, Wasieloski LP, Thompson WH, Gundersen CB, Case K, Beaty BJ**. 1998. Characterization of La Crosse virus RNA in autopsied central nervous system tissues. J Clin Microbiol **36:**3332–3336.

28. **Langmead B, Salzberg SL**. 2012. Fast gapped-read alignment with bowtie 2. Nat Methods **9:**357–359. http://dx.doi.org/10.1038/nmeth.1923.

29. **Ruby JG, Bellare P, Derisi JL**. 2013. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 (Bethesda) **3:**865–880. http://dx.doi.org/10.1534/g3.113.005967.

30. **Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR**. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29:**15–21. http://dx.doi.org/10.1093/bioinformatics/bts635.

31. **Fu L, Niu B, Zhu Z, Wu S, Li W**. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics **28:**3150–3152. http://dx.doi.org/10.1093/bioinformatics/bts565.

32. **Li W, Godzik A**. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22:**1658–1659. http://dx.doi.org/10.1093/bioinformatics/btl158.

33. **Wu TD, Nacu S**. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics **26:**873–881. http://dx.doi.org/10.1093/bioinformatics/btq057.

34. **Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N**. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res **40:**11189–11201. http://dx.doi.org/10.1093/nar/gks918.