

**COMPARISON OF HIGH-THROUGHPUT SEQUENCING METHODS
FOR MONITORING GENETIC CHANGES IN EBOLAVIRUS POPULATIONS**

by

Elyse R. Nagle

B.S. (Towson University) 2009

THESIS

Submitted in partial satisfaction of the requirements

for the degree of

MASTER OF SCIENCE

in

BIOMEDICAL SCIENCE

in the

GRADUATE SCHOOL

of

HOOD COLLEGE

May 2017

Accepted:

Oney Smith, Ph.D.
Committee Member

Rachel K. Bagni, Ph.D.
Director, Biomedical Science Program

Ann Boyd, Ph.D.
Committee Member

Gustavo Palacios, Ph.D.
Thesis Adviser

April M. Boulton, Ph.D.
Dean of the Graduate School

STATEMENT OF USE AND COPYRIGHT WAIVER

I do authorize Hood College to lend this thesis, or reproductions of it, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

ACKNOWLEDGEMENTS

All of the work described in this thesis was performed at the Center for Genome Sciences at USAMRIID, Fort Detrick, Maryland. First, I would like to thank Dr. Gustavo Palacios for acting as my thesis advisor, for his sharing his time and knowledge in support of this project, and for allowing me to complete this milestone in addition to my normal duties in the CGS. I give many thanks to Bradley Pfeffer of CGS for his time and knowledge of the analysis pipelines, and for his patience with my frequent analysis changes. I would also like to thank Dr. Michael Wiley, Dr. Jason Ladner, CPT Jeffrey Kugelman, Ph. D., and Dr. Mariano Sanchez-Lockhart, all of CGS, for their expertise, input, and edits on this thesis. Additionally, I would like to thank Dr. Oney Smith and Dr. Ann Boyd, both of Hood College, for serving on my thesis committee and for their advice, edits, and input.

Finally, I would like to thank my family for their unwavering support of this journey. My parents, Kimberly and Keith Nagle provided countless hours of babysitting and help so that I could attend classes, study, and write. They, along with my partner Matthew Abbott and his family, also provided extraordinary moral support and love. This work is dedicated to them and my young daughter, Julianne.

Research was conducted under an IACUC approved protocol in compliance with the Animal Welfare Act, PHS Policy, and other Federal statutes and regulations relating to animals and experiments involving animals. The facility where this research was conducted is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care, International and adheres to principles stated in the Guide for the Care and Use of Laboratory Animals, National Research Council, 2011.

The research described herein was sponsored by the Defense Threat Reduction Agency, Project No. CB10246.

Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

TABLE OF CONTENTS

| | Page |
|--|------|
| ABSTRACT | v |
| LIST OF TABLES | vi |
| LIST OF FIGURES | vii |
| LIST OF ABBREVIATIONS | viii |
| INTRODUCTION | 1 |
| MATERIALS AND METHODS | 13 |
| Amplicon-Based Sample Preparation | 13 |
| RNA Access Sample Preparation | 17 |
| Data Analysis | 21 |
| RESULTS | 24 |
| RNA Access Results | 24 |
| Comparison of Amplicon Sequencing and RNA Access | 30 |
| DISCUSSION | 39 |
| REFERENCES | 45 |

ABSTRACT

Amplicon-based sequencing of Ebolavirus is a powerful tool to monitor the genetic changes in the viral population during a drug study. Short amplicons are generated covering the whole virus genome and a library is generated and sequenced. This method can detect complete Ebolavirus in samples with only 10^5 genome copies/mL, but it is time-consuming and requires extensive PCR, potentially producing errors. A new Ebolavirus-targeted capture and enrichment method, RNA Access was used to increase the sensitivity of detection and reduce PCR error. The two methods were compared with the same set of samples. Surprisingly, RNA Access is not as sensitive as the amplicon-based method; complete Ebolavirus genomes were sequenced from 10^6 genome copies/mL and required more sequencing reads than the amplicon-based method. Despite these shortcomings, the protocol speed and minimal PCR make RNA Access a viable method to monitor genetic variation in an Ebolavirus population.

LIST OF TABLES

| Table | | Page |
|-------|--|------|
| 1 | A) Coverage and depth of RNA Access samples using standard VSALIGN inputs. B) Coverage and depth of RNA Access samples using all reads in VSALIGN. | 25 |
| 2 | Coverage and depth of low-titer RNA Access samples. | 27 |
| 3 | Coverage and depth of RNA Access samples after duplicate removals with Picard. | 29 |
| 4 | Comparison of bases sequenced above the target depth and detected SNPs in amplicon and RNA Access samples. | 32 |
| 5 | A) Shared Ebolavirus SNP frequency in amplicon data and B) Shared Ebolavirus SNP frequency in RNA Access data. | 33 |
| 6 | Comparison of sequencing read duplicate removal with the Picard analysis of amplicon and RNA Access data. | 35 |
| 7 | Comparison of the percentage of reads removed through duplicate analysis and cleaning in VSALIGN of amplicon and RNA Access data | 36 |
| 8 | Comparison of frequencies of known mutations in amplicon and RNA Access data. to compare data sensitivity | 38 |

LIST OF FIGURES

| Figure | | Page |
|--------|--|------|
| 1 | A) Ebolavirus genome amplification. B) Ebolavirus replication cycle. | 3 |
| 2 | Schematic for amplicon primer design. | 8 |
| 3 | Schematic for RNA Access probe design and target capture. | 11 |
| 4 | Amplicon-based sample preparation workflow. | 14 |
| 5 | RNA Access sample preparation workflow. | 18 |

LIST OF ABBREVIATIONS

| | |
|-------------------|--|
| BDBV | Bundibugyo virus |
| BSAT | biological select agents and toxins |
| BSL-4 | biosafety level 4 |
| CGS | Center for Genome Sciences |
| DV ₂₀₀ | percentage of RNA fragments below 200 nt |
| EBOV | Zaire Ebolavirus |
| EDTA | ethylenediaminetetraacetic acid |
| FDA | Federal Drug Administration |
| FFPE | formalin-fixed paraffin-embedded |
| HPLC | high-performance liquid chromatography |
| JCVI | J. Craig Venter Institute |
| LLOD | lower limit of detection |
| NHP | non-human primate |
| PAGE | polyacrylamide gel electrophoresis |
| PCR | polymerase chain reaction |
| PMO | phosphorodiamidate morpholino oligomer |
| qPCR | quantitative polymerase chain reaction |
| RESTV | Reston virus |
| RIN | RNA integrity number |
| rVSV | recombinant vesicular stomatitis virus |
| siRNA | small interfering RNA |
| SNP | single nucleotide polymorphism |

| | |
|----------|--|
| SUDV | Sudan virus |
| TAFV | Tai Forest virus |
| USAMRIID | United States Army Medical Research Institute of Infectious Disease |
| WHO | World Health Organization |

INTRODUCTION

Ebolavirus initially emerged in Zaire, present day Democratic Republic of the Congo, in 1976 (CDC 2015) causing severe viral hemorrhagic fever. Since then, there have been a total of five different ebolavirus species discovered: Tai Forest virus (TAFV), Sudan virus (SUDV), Zaire virus (EBOV), Reston virus (RESTV), and Bundibugyo virus (BDBV) (Kuhn *et al.* 2014). Four of the five species are known to cause severe disease in humans, while the Reston species has only been documented in non-human primates (NHPs). Until recently, Ebola outbreaks have been relatively small and occurring in remote, isolated areas with a range of mortality from 36 - 89% and the largest total case load recorded for an Ebola outbreak prior to 2013 was 318 cases in Zaire in 1976 (CDC 2015). Ebola causes fever, malaise, severe diarrhea, vomiting, muscle and joint pain, bleeding, and bruising.

The virus is a linear, non-segmented, negative-sense, single-stranded RNA virus about 19-kb in length (Figure 1A). The genome contains seven genes encoding nine proteins: nucleoprotein (NP), viral protein 35 (VP35), VP40, glycoprotein (sGP, GP_{1,2}, ssGP), VP30, VP24, and the RNA-dependent RNA polymerase (L) (Kuhn 2008). A brief outline of the viral replication cycle is contained in Figure 1B. Figures 1A and 1B are reprinted with permission of Nature Publishing Group. The Ebola glycoprotein resides on the surface of the virion and interacts with host cell receptors; fusion between the host and virion causes signal transduction in the host cell for viral entry via macropinocytosis. Inside the endosome, GP1 on the virion interacts with a host receptor to allow fusion and release of the viral nucleocapsid into the host cytoplasm. The RNA-dependent RNA polymerase transcribes the Ebola genes in order, releasing from the template strand after

each gene is transcribed, the transcripts are then capped and polyadenylated in the cytoplasm. The polymerase reinitiates transcription at the next gene, but the downstream activity is attenuated. For this reason, NP exists in the highest levels in the cytoplasm. Replication of the virus is triggered when there is enough NP to package new virus copies and budding of new virions from the host cell is directed by VP40 (Messaoudi *et al.* 2015). Because GP is responsible for entry into the host cell and the polymerase for replication, both GP and L are often therapeutic targets.

Three separate proteins are translated from the GP gene due to stuttering by the RNA-dependent RNA polymerase. There is a section of seven uridylyls in GP at genome positions 6,918 – 6,924 before the poly-U editing site at position 6,925 (Kugelman *et al.* 2012). When all seven adenylyls are synthesized, sGP is the primary product and the virus population is called a 7U variant. The product GP_{1,2} is generated when the polymerase stutters and adds an extra adenylyl. The resultant virus population is called an 8U variant. Occasionally the polymerase will either skip one uridylyl or add two adenylyls producing ssGP and the resultant viral populations are called 6U or 9U variants, respectively (Kugelman *et al.* 2012; Volchkova *et al.* 2011). The functions of sGP and ssGP are unknown, but GP_{1,2} is the structural protein that resides on the virus envelope and is responsible for viral entry into the host cell. Interestingly, the 7U variant is predominant in *in vivo* populations, while the 8U variant is predominant *in vitro* (Trefry *et al.* 2015; Volchkova *et al.* 2011). This is true even when a 7U variant is used as the challenge agent in cell culture; the shift to 8U occurs within the first few passages. When compared with an 8U reference genome, there are four genome positions within the GP gene that are known to be SNPs in the majority of 7U populations (Kugelman *et*

al. 2012; Kugelman et al. 2016; Trefry et al. 2015). These SNPs will be used as control points to compare the sensitivity between sequencing methods.



Figure 1A. Ebola virus genome organization. Reprinted by permission of Nature Publishing Group.

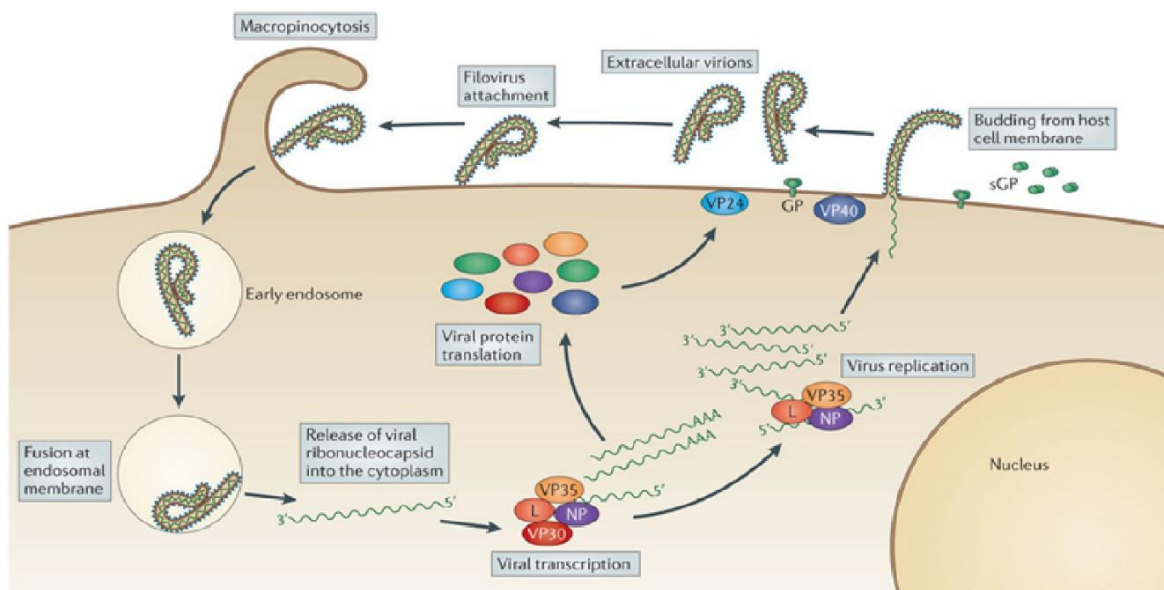


Figure 1B. Ebola virus replication cycle. Reprinted by permission of Nature Publishing Group.

There is no officially approved drug to treat Ebola. Several different categories of countermeasures are under development: (1) Phosphorodiamidate morpholino oligomers (PMOs) and small interfering RNAs (siRNAs) target specific sequences in the viral genome which block translation or target transcripts for degradation, respectively (Kugelman *et al.* 2015b); (2), nucleoside analogs that block the activity of the RNA-dependent RNA polymerase have been recently demonstrated to be useful for the treatment of filovirus infections. Nucleoside analog GS-5734 (Gilead Sciences, Inc., Foster City, CA) is the first to protect NHPs against Ebola post-exposure (Warren *et al.* 2016). BCX4430 (BioCryst Pharmaceuticals, Inc., Durham, NC) (Warren *et al.* 2014) and Favipiravir (Toyama Chemicals, Tokyo, Japan) (Oestereich *et al.* 2014) have been shown to protect against Ebola in rodent models; (3) Passive immunotherapy treatments based on monoclonal antibody cocktails neutralize the virus long enough for the host immune system to act (Qiu *et al.* 2014); (4) Finally, a potential vaccine candidate composed of a recombinant vesicular stomatitis virus (rVSV) expressing Ebola GP instead of the VSV GP has shown some efficacy when used post-exposure (Geisbert and Feldmann 2011). All of these therapeutic categories have been tested in NHPs and many have entered clinical trials in West Africa (WHO 2015). Of these four types of therapeutics; two of them are sequence-based.

A major concern, especially with the first three classes of countermeasures, is the possibility of resistance due to the emergence of viral escape mutants. RNA viruses can mutate very quickly under selective pressure (Kugelman *et al.* 2015a; Moya *et al.* 2004) and if mutations occur in the target area of a sequence-based drug, the virus may beat the

treatment increasing the window of transmission, initiating a new node of infection in an outbreak where current therapeutics would be ineffective.

Population genetics is the study of the natural selection, genetic drift, mutations, recombination, and migration that build the genetic landscape within a population (Moya *et al.* 2004). At the United States Army Medical Research Institute for Infectious Disease Center for Genome Sciences (USAMRIID CGS), we are using population genetics methods to monitor any changes within the Ebolavirus population of an infected group, *in vivo* or *in vitro*, that decrease the efficacy of a particular vaccine or therapeutic drug. This has become extremely important while evaluating countermeasures against agents classified as Biological Select Agent and Toxin (BSAT) where, due to regulations and policies, as well as ethical, logistical, and financial issues, the advancement to regulatory approval will be made under the “Animal Rule”. Essentially, this means that approval of the drug for human use by the Federal Drug Administration (FDA) will be made in the absence of human clinical studies through studies in animal models that resemble human disease. Given that many animal studies involving NHPs are limited in size, the analysis of the resulting inconsistent cases (e.g. treated animals with no response) for the presence of escape mutants or reduced efficacy become crucial (Kugelman *et al.* 2015a). Interestingly, in emergency situations, viral population genetics can be used to monitor an infected group treated with an experimental drug in near real-time.

As an example of the above, we have recently monitored the appearance of viral escape mutants during treatment with ZMapp, a synthetic neutralizing antiserum (Kugelman *et al.* 2015a). Passive Immunotherapy was thrust into the global spotlight when ZMapp was given fast track designation by the FDA and approved for

compassionate use to treat Ebola by the World Health Organization (WHO). ZMapp (Mapp Biopharmaceutical, Inc., San Diego, CA) is a monoclonal antibody cocktail consisting of specifically chosen antibodies from two predecessor drugs: MB-003 (Pettit *et al.* 2013) and ZMAb (Qiu *et al.* 2014). Using population genetics methods, we have detected a case of viral escape in each of two MB-003 NHP studies (Kugelman *et al.* 2015a). In one study, two of six animals were not protected by MB-003. One animal died in the normal expected range, while the other had an atypical, delayed time of death. Mutations were found in the antibody-binding regions that later demonstrated decreased binding in a follow-up study; the same mutations were confirmed in a virus isolated samples collected at the time of death. Moreover, the isolated virus was not neutralized in vitro by the cocktail, pointing to viral escape (Kugelman *et al.* 2015a). Later, in a retrospective analysis, similar mutations were found in a separate NHP MB-003 study. Nevertheless, ZMapp has been used in NHPs and seven human Ebola-infected patients as of September 2015 without any observed escape mutants. However, our study demonstrates the importance of viral population genetics analysis as a tool to monitor potential viral escape.

At the USAMRIID Center for Genome Sciences, we routinely use a method based on amplicon-based whole genome amplification for studies of viral population genetics. Total RNA is extracted from Trizol, and cDNA (complementary DNA) synthesis is performed using random hexamers. The cDNA template is used to generate amplicons that tile across a particular viral genome in an overlapping manner. Specifically, the primer set for Ebola Zaire was designed using the Kikwit variant (EBOV Kikwit) and consists of 39 primer pairs generating approximately 1,500-bp amplicons with a 500-bp

overlap that provide double-coverage of the whole genome (Figure 2). Amplicons are generated in individual polymerase chain reactions (PCR), and then screened for concentration and correct size. The successful amplicons are normalized for concentration and pooled on a per-sample basis. Libraries are created from the pools and sequenced on an Illumina Next-Generation Sequencing platform. Using these methods, whole viral genomes can be generated from only 500,000 (5×10^5) copies of virus (unpublished data).



Figure 2. Schematic for amplicon primer design. Primers are designed to create 1,500-bp amplicons that tile across the genome with 500-bp overlaps. The amplicons provide at least double coverage of every genome position.

These methods are well tested and will work for any virus being studied by using a virus-specific primer set. Furthermore, this exact method was used to detect the MB-003 viral escape mutants (Kugelman *et al.* 2015a). Still, there is plenty of room for improvement, partially driven by the extreme need of a swift response to an unprecedented outbreak. Many parts of the protocol are automated with liquid handling robots, but the system is not truly high-throughput and still requires a large amount of hands-on work which is vulnerable to error. Most population genetics studies in NHPs involve up to 32 animals, each of which generally has a blood draw for analysis on day 0, 3, 5, 7, 9, 12, 14, 21, and 28. While we have been able to recover whole genomes from viral titer as little as 5×10^5 copies detected by qPCR with GP-specific primers and probes, there are still many samples, especially early in the infection, that have a detectable viral titer for which we are not able to recover using current methods. Recovering virus from low-titer samples is a priority to increase the population study and help confirm rare mutations. Viral RNA in low titer samples may be fragmented or simply too rare to be amplified with large amplicons, so using smaller amplicons would most likely produce better results at the expense of efficiency. Sensitivity must be balanced with physically feasible and reasonable protocols. Functionally, amplicon generation involves a large amount of PCR which could introduce errors and bias. Error management is very important; it must be mitigated to be able to determine whether we are seeing a true mutation.

In search of an improved method to optimize the current methods, several technologies were evaluated. One of them, developed by Illumina, Inc. (San Diego, CA), was considered for further refinement given its potential improvements over the

amplicon-based methods. The Illumina TruSeq RNA Access Library Prep kit is marketed to sequence RNA from formalin-fixed paraffin-embedded (FFPE) tissue or tumor samples for cancer research. The system requires very little RNA input and it can accommodate degraded RNA. Libraries are prepared from RNA and then enriched by using biotinylated probes targeted to a human reference exome to extract all coding sequences by binding to streptavidin-coated magnetic beads (Figure 3). Illumina has demonstrated that using this enrichment method to sequence rare transcripts requires only a fraction of the reads and depth normally needed (Illumina 2014). In addition, the protocol is automation friendly and completed libraries can be sequenced on any Illumina platform. We designed a pool of 80-mer biotinylated probes that provide at least triple-coverage of the Ebolavirus genomes to be used for enrichment (Figure 3). The Illumina TruSeq RNA Access kit coupled with Ebola-specific probes covering the entire genome will replace traditional population genetics methods as a fast and efficient, truly high-throughput protocol. The improvements over the current protocol include: less input RNA to save precious sample, reduction of the error rates by avoiding amplification steps that will artificially add variation to the viral population, more sensitive and efficient detection of whole viral genomes, and a significant reduction in the amount of PCR required.

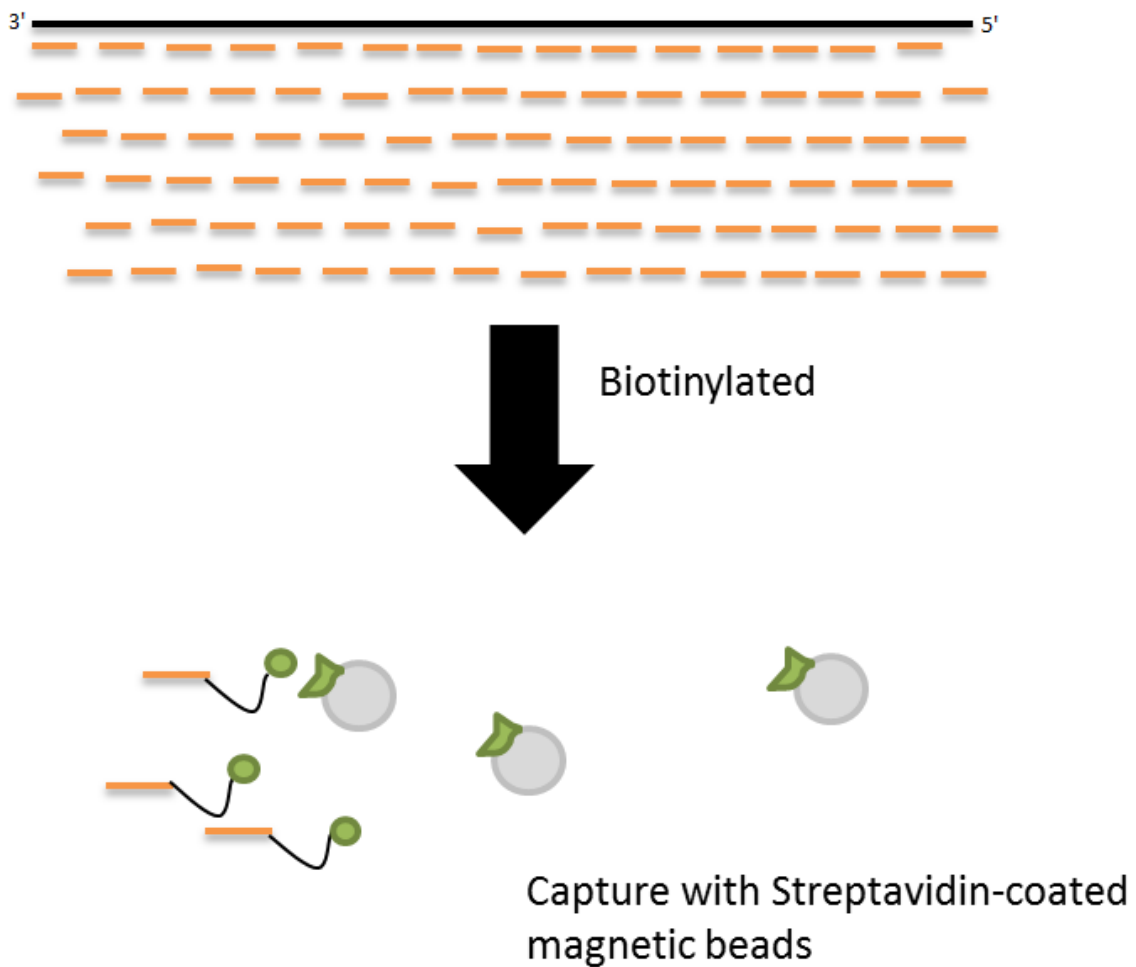


Figure 3. Schematic for RNA Access probe design and target capture. A pool of 80-mer probes was designed to provide at least triple-coverage of the virus genome. The probes were biotinylated to allow capture of virus-specific targets with Streptavidin coated magnetic beads.

The objectives of this study are to 1) determine the sensitivity of RNA Access compared to amplicon generation and establish a new LLOD and 2) determine the amount of sequencing required to achieve at least the current depth and coverage produced by amplicon generation. Additionally, the RNA Access automated protocol will be customized to accommodate changes used in CGS.

MATERIALS AND METHODS

Eight whole blood samples from a previous population genetics study in cynomolgus macaques (*Macaca fascicularis*) infected with either 7U EBOV Kikwit (GenBank ID: KT762962) or 8U EBOV Kikwit (GenBank ID: KT582109) were chosen to include a range of titers from 3.52×10^3 to 1.23×10^{11} genome copies/mL (gc/mL). The viral titers were supplied with the samples and determined by a qPCR (quantitative PCR) assay that uses probes for the Ebolavirus GP gene. A second set of eight samples from the same project were chosen to confirm the lower limit of detection (LLOD). The viral titers in the samples ranged from 3.23×10^6 to 3.52×10^3 genome copies/mL. The whole blood samples were mixed with TRIzol LS (ThermoFisher Scientific, Waltham, MA) at a ratio of 1:3 for virus inactivation prior to exiting the biosafety level 4 (BSL-4) suite.

Amplicon-Based Sample Preparation

Figure 4 provides an overview of amplicon-based sample preparation. RNA was extracted from whole blood in Trizol LS and first-strand cDNA was produced. Individual amplicons were generated tiling the viral genome, then screened for size and concentration and pooled. After purification, the pools were sheared and libraries were processed, amplified, and screened for concentration. The completed libraries were sequenced and analyzed.

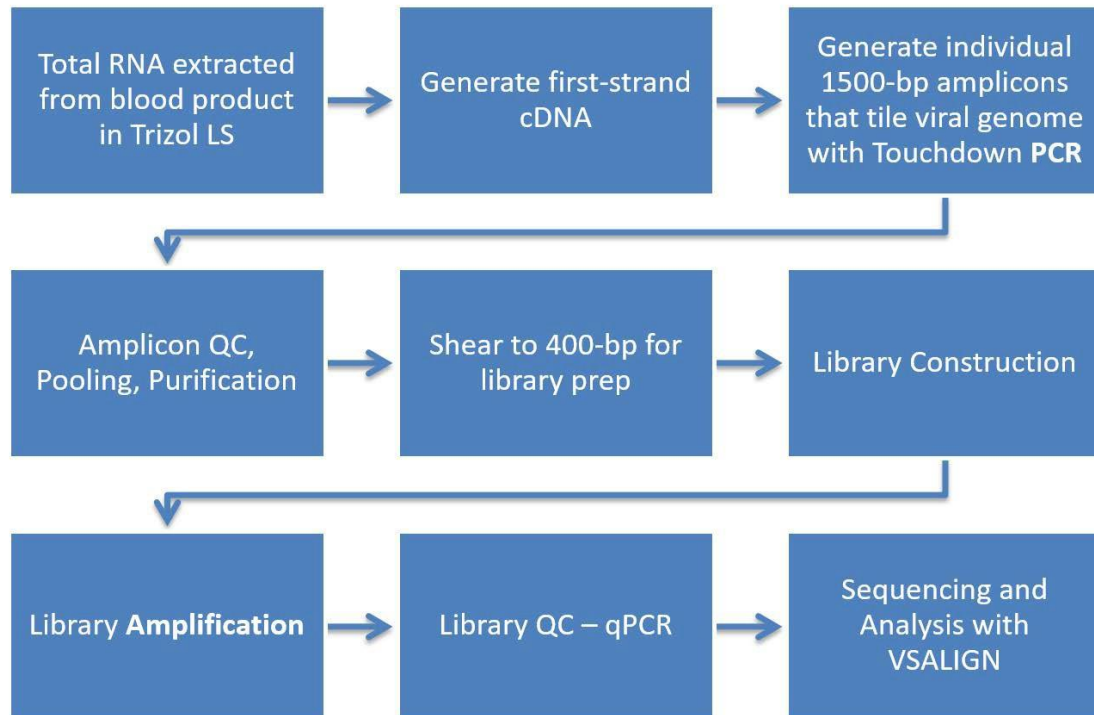


Figure 4. Amplicon-based sample preparation workflow.

RNA Extraction and cDNA Synthesis

Total RNA was extracted from 250 μ L of whole blood in TRIzol LS using the PureLink RNA Mini kit (ThermoFisher Scientific, Waltham, MA) per the manufacturer's instructions and eluted in 30 μ L RNase-free water. First-Strand cDNA synthesis was performed using the Superscript III First-Strand Synthesis System (ThermoFisher Scientific, Waltham, MA) using 9 μ L of RNA and random hexamers supplied with the kit.

Primer Design for Amplicon Generation

The JCVI Primer Design software (Li *et al.* 2008) was used to design primers that tile the complete genome of Ebola Zaire Kikwit (GenBank ID: AY354458). The input parameters were set to generate amplicons about 1,500-bp in length with a 500-bp overlap to provide redundant coverage of the genome. The primers were ordered from Eurofins MWG Operon (Louisville, KY) with high-performance liquid chromatography polyacrylamide gel electrophoresis (HPLC-PAGE) purification and lyophilized. Upon arrival, the primers were rehydrated to 100 μ M with 8.0 pH Tris-EDTA (TE) buffer. The forward and reverse primers were paired for each amplicon and diluted to 10 μ M for use in PCR.

Amplicon Generation, QC, and Pooling

A custom application specifically designed for the CGS population genetics amplicon PCR protocol was used on the Zephyr Liquid Handling System (PerkinElmer, Waltham, MA) to automate individual amplicon generation. The forward and reverse primers were paired for each amplicon at 10 μ M in 150 μ L 8.0 pH TE buffer in each well

of a 96-well stock plate to be added separately to the PCR reaction. The amplicon generation reaction was conducted using Phusion Hot Start Flex (New England BioLabs, Inc., Ipswich, MA) in a final volume of 20 μ L consisting of 5X Phusion Hot Start Flex HF Buffer, 10X dNTPs, and 100X Phusion Hot Start Flex HF polymerase. A master mix was prepared for each sample and 5 μ L first-strand cDNA was added to the master mix. The Zephyr aliquoted master mix and primers to individual plates which were sealed and amplified using a touchdown PCR method with the following conditions: 30 sec at 98°C, 20 cycles of 10 sec denaturation at 98 °C, 10 sec annealing at 64 °C which is decreased by 1 °C/cycle, and 30 sec extension at 72 °C, followed by 30 cycles of 10 sec at 98 °C, 10 sec at 58 °C, and 30 sec at 72 °C, followed by 10 min final extension at 72 °C, and hold at 12 °C.

Amplicons were screened individually for concentration and correct size on the LabChip GX Nucleic Acid Separation System (PerkinElmer, Waltham, MA) using the DNA 5K assay. Amplicons within +/- 10% of the expected size were pooled at an equimolar concentration on a per sample basis. The samples were purified with a 0.6X concentration of AMPure XP Reagent (Beckman Coulter, Brea, CA) to remove any excess primer and erroneous product up to 500 bp. The concentration of the pooled and purified samples was measured on the Nanodrop 2000 (ThermoScientific, Waltham, MA).

Library Preparation

The Covaris LE220 Focused-Ultrasonicator (Covaris, Inc., Woburn, MA) was used to fragment 500 ng of each amplicon pool to an average size for 400 bp in a volume of 50 μ L using the manufacturer's protocol. Libraries were produced using the KAPA

Library Preparation Kit for Illumina (KAPA Biosystems, Inc., Wilmington, MA) with dual-indexes for multiplexing on the PerkinElmer Sciclone G3 Liquid Handling Workstation and amplified per the manufacturer's protocol.

The purified libraries were screened for concentration and size on the LabChipGX system using the DNA 1K assay. Pools of eight libraries were generated for qPCR based on library concentration and diluted to 10 nM. The library pools were quantified for sequencing using the KAPA Library Quantification Complete (Universal) kit for Illumina platforms. The library pools were diluted to 2 nM then pooled into one final library pool for sequencing.

Amplicon Sequencing

The final library was denatured and diluted to 10 pM for cluster generation on the Illumina cBOT (San Diego, CA) following the manufacturer's protocol. The final library pool was sequenced in one lane of the Illumina HiSeq 2500 using the 101-bp paired-end protocol.

RNA Access Sample Preparation

Figure 5 provides an overview of RNA Access sample preparation. Total RNA was extracted from whole blood in Trizol LS. Libraries were prepared directly from RNA template, and then enriched with Ebolavirus-specific biotinylated probes. The final libraries were amplified and screened for concentration and quality, then sequenced and analyzed.

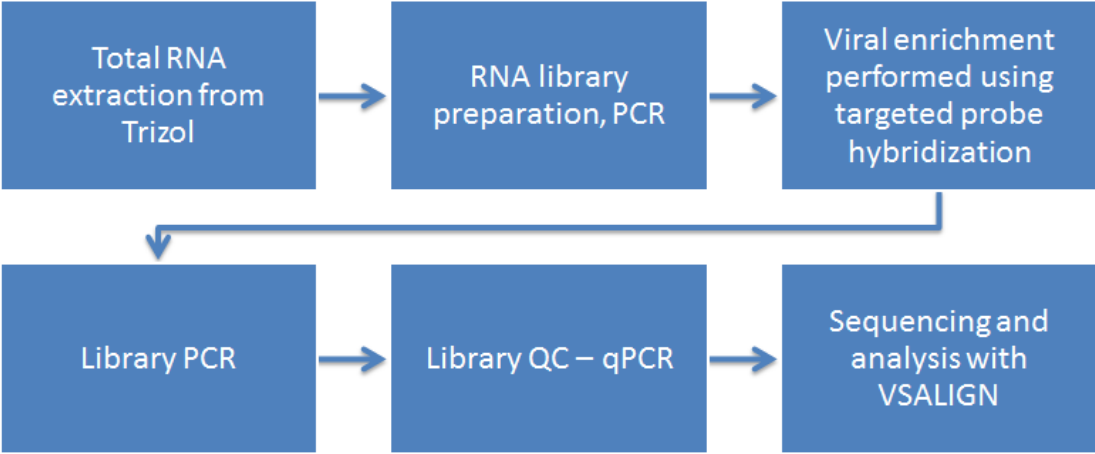


Figure 5. RNA Access sample preparation workflow.

RNA Extraction and QC

RNA was extracted from samples using the PureLink RNA Mini kit (ThermoFisher Scientific, Waltham, MA) per the manufacturer's protocol and eluted in 30 μ L RNase-free water. The RNA was screened for quality and concentration using the RNA ScreenTape on the Agilent TapeStation 2200 (Santa Clara, CA). The percentage of RNA fragments below 200 nucleotides (DV_{200}) was calculated by using smear analysis in the TapeStation software to determine the percentage of RNA fragments above 200 nucleotides. Illumina uses the DV_{200} value as a more specific measure of RNA quality than the standard RNA Integrity Number (RIN). The DV_{200} of all samples was above 90%, so the minimum RNA input of 20 ng was used.

Filovirus Capture Oligo Design

The capture oligo pool is made up of 1,897 80-mer biotinylated oligos that provide 3X coverage of nine filovirus genomes including Tai Forest virus (GenBank ID: NC014372), Sudan virus (GenBank ID: NC006432), Reston virus (GenBank ID: NC004161), Bundibugyo virus (GenBank ID: NC014372), Zaire virus Makona isolate (GenBank ID: KJ660348), Zaire virus Mayinga isolate (GenBank ID: NC002549), Marburg virus Musoke isolate (GenBank ID: NC001608), RAVN virus (GenBank ID: NC024781), and Lloviu virus (GenBank ID: NC016144). The pool was designed with and provided by Illumina, Inc. (San Diego, CA).

Library Preparation

Twenty nanograms of RNA were used as input for all samples in the RNA Access Library Prep Kit (Illumina, San Diego, CA). Libraries were prepared with dual-indexes for multiplexing per the manufacturer's protocol prior to hybridization except that the

First PCR was amplified for 20 cycles. The libraries were validated prior to hybridization using High Sensitivity D1000 ScreenTape on the Agilent TapeStation. The Illumina protocol recommends pooling groups of up to 4 samples, but in order to preserve sensitivity, samples were processed individually. Additionally, all reagent volumes in the Illumina hybridization protocol were quartered because the samples were not pooled. All incubation steps were followed as written in the manufacturer's protocol. The enriched libraries were validated using the High Sensitivity D1000 ScreenTape on the Agilent TapeStation, diluted to 2nM and quantified for sequencing with qPCR using the KAPA Library Quantification Complete (Universal) kit for Illumina platforms.

Sequencing

Samples R1-R8 were pooled into one final library at 2 nM and denatured and diluted to 10 pM for cluster generation and sequencing per the manufacturer's protocol on the Illumina MiSeq using the 150-bp paired-end protocol. Samples R9 - R16 were pooled into one final library pool at 1.5 nM and denatured and diluted to 10 pM for cluster generation and sequencing per the manufacturer's protocol on the Illumina MiSeq using the 150-bp paired-end protocol.

Automation Development

The RNA Access protocol was automated on the Sciclone G3 Liquid Handling System by PerkinElmer. The Hybridization steps in the original application were adjusted with the assistance of PerkinElmer to accommodate the custom reaction volumes used in Hybridization in the CGS.

Data Analysis

VSALIGN

VSALIGN (USAMRIID, Ft. Detrick, MD) is an in-house CGS software pipeline used to clean, align, and analyze raw deep-sequencing data for population genetics (Kugelman *et al.*, 2017). Briefly, VSALIGN first processes raw sequencing data to exclude low quality reads < 20 Phred, removes adaptor sequences used for multiplexing, removes non-viral sequences, and trims the identified sequencing primer sequences at the end of the read. The reads are trimmed until there is a base with > 15 Phred. Additionally, VSALIGN will remove chimeric reads and read pairs, remove paired reads with an index quality < 30 Phred, paired reads that do not have opposing directionality, and exact duplicate pairs that are more than three standard deviations from the average number of replicates in the sample. Alignment follows cleaning and is done in DNASTar Lasergene nGen (Madison, WI) within VSALIGN (Kugelman *et al.*, 2017). The user can input the parameters for the number of reads used in the alignment and the minimum and maximum depth. An extensive explanation of the procedures is detailed by Kugelman *et al.* (2017).

In this study, all available reads were used for alignment to Ebola virus/H.sapiens-tc/COD/1995/Kikwit-9510621 (GenBank ID: AY354458) which is an 8U reference genome (Kugelman *et al.* 2016). The maximum depth was set to 1,500. Typically, the minimum read depth parameter used for amplicon sequencing is 200 because of the error rates associated with primer generation, however, if a single nucleotide polymorphism (SNP) is detected below 200 it may still be real and require further analysis. A coverage depth of 20 is considered the absolute minimum to call a position by the CGS. The

amplicon data was originally analyzed with a minimum depth of 200 and the RNA Access data was analyzed with a minimum depth of 20. The amplicon data was analyzed a second time with a minimum depth of 20 for direct comparison of known SNPs in the RNA Access data.

Studies have shown that Ebolavirus passaged *in vitro* primarily contain the 8U GP gene editing site, while Ebolavirus population *in vivo* primarily contains the 7U GP gene editing site (Trefry *et al.* 2015; Volchkova *et al.* 2011). Ebolavirus passages *in vitro* and *in vivo* were characterized to identify markers of passage history by Kugelman, *et al.* (2012, 2016). There are four genome positions within the GP gene that are SNPs occurring in the majority of the 7U viral population when compared with the 8U reference (Kugelman *et al.* 2012; Kugelman *et al.* 2016; Trefry *et al.* 2015). These four SNPs, the known SNPs, were used as control points to compare sensitivity between amplicon and RNA Access methods.

Picard

Picard is an informatics toolset from the Broad Institute (Cambridge, MA) for sequencing data analysis that is available publicly. The Picard tool MarkDuplicates is a more strict method of duplicate removal than the VSALIGN. MarkDuplicates is a part of a pipeline containing in-house and other open-source tools for alignment and duplicate removal of sequencing data. Fastq files were input with the Ebola virus/H.sapiens-tc/COD/1995/Kikwit-9510621 (GenBank ID: AY354458) reference fasta for alignment in Bowtie (Langmead *et al.* 2009). The results were compiled in a sam file and Samtools (Li *et al.* 2009) was used to convert to a bam file for input in MarkDuplicates. The bam

output file was converted to individual fastq files with bamtofastq, a part of bedtools (Quinlan and Hall 2010).

RESULTS

RNA Access Results

RNA Access

Eight samples from a previous population genetics study were chosen to include a range of titers from 3.52×10^3 to 1.23×10^{11} Ebolavirus genome copies/mL. This sample set represents viral titers well below and well above the previously determined limit of detection of 5×10^5 virus copies. It will be harder to detect variants in low-titer viral samples because less viral RNA will be available for capture and enrichment. The viral titers were supplied with the samples and determined by a qPCR assay that uses probes for the Ebola GP gene. RNA was extracted from the Trizol samples, denoted R1 - R8, and processed with the RNA Access kit and sequenced on the MiSeq in the 2 x 150-bp format. The raw sequencing reads generated on the MiSeq were analyzed with VSALIGN using a random input of 200,000 reads and a target depth of 200, which is the CGS standard for amplicon-based population genetics studies. The coverage and average depth of R1 - R8 are contained in Table 1a. Coverage is above 99% for R1 - R4 and depth is well above the 200 target with the exception of R4 at 300.21. Samples R5 - R8 do not meet the target depth and do not have full coverage of the genome.

All of the available raw reads for R4 - R8 were analyzed in VSALIGN in an attempt to increase coverage and depth (Table 1b). Using all reads resulted in 99.81% coverage and 870.85 average depth for R4. Depth and coverage started to decrease significantly at 10^5 genome copies/mL. However, the depth and coverage for R8 were unexpectedly high as the lowest-titer sample.

TABLE 1a. Coverage and depth of RNA Access samples using standard VSALIGN input parameters.

| R1 – R8 VSALIGN 200,000 Reads | | | |
|-------------------------------|---------------|----------|---------|
| Sample | Titer (ge/mL) | Coverage | Depth |
| R1 | 1.23E+11 | 99.74 | 1154.74 |
| R2 | 1.85E+09 | 99.83 | 1189.98 |
| R3 | 7.01E+08 | 99.82 | 1121.2 |
| R4 | 2.62E+07 | 99.73 | 300.21 |
| R5 | 1.50E+06 | 60.62 | 58.49 |
| R6 | 6.56E+05 | 39.18 | 39.71 |
| R7 | 1.41E+04 | 0.68 | 35.75 |
| R8 | 3.52E+03 | 67.79 | 43.42 |

TABLE 1b. Coverage and depth of RNA Access samples using all available reads in VSALIGN.

| R4 – R8 VSALIGN All Reads | | | |
|---------------------------|---------------|----------|--------|
| Sample | Titer (ge/mL) | Coverage | Depth |
| R4 | 2.62E+07 | 99.81 | 870.85 |
| R5 | 1.50E+06 | 97.44 | 288.25 |
| R6 | 6.56E+05 | 74.13 | 80.61 |
| R7 | 1.41E+04 | 3.51 | 50.52 |
| R8 | 3.52E+03 | 94.59 | 124.25 |

RNA Access for Low Titer Samples

In order to confirm the possible lower limit of detection determined in the analysis of R1 - R8, a second set of eight low titer samples were chosen from the same study. Samples R9 - R16 includes two samples at each 10^6 , 10^5 , 10^4 , and 10^3 genome copies/mL. RNA was extracted from the Trizol samples and processed and sequenced using the same methods as R1 - R8. The results in Table 2 summarize the coverage and depth using all of the raw reads used as input in VSALIGN. Though there is slight variation between samples of similar viral titer, the depth and coverage decrease with viral titer.

Table 2. Coverage and depth of low-titer RNA Access samples using all available reads.

| Sample | Titer (ge/mL) | Coverage | Depth |
|--------|---------------|----------|--------|
| R9 | 3.23E+06 | 89.69 | 451.1 |
| R10 | 1.31E+06 | 40.61 | 274.78 |
| R11 | 6.78E+05 | 64.02 | 127.05 |
| R12 | 2.73E+05 | 34.88 | 271.82 |
| R13 | 6.65E+04 | 11.06 | 353.67 |
| R14 | 1.05E+04 | 0.01 | 20 |
| R15 | 8.77E+03 | 0 | 0 |
| R16 | 6.22E+03 | 0 | 0 |

RNA Access Duplicate Analysis

The Picard pipeline developed by the Broad Institute (Cambridge, MA) has a stricter method of duplicate removal than VSALIGN called MarkDuplicates. The raw data for R1 - R16 was run through the Picard pipeline prior to alignment with VSALIGN to determine whether the coverage would change. The results outlined in Table 3 indicate little change in coverage and depth for R1 - R4, except for a drop in depth for R2 and R4, still within target. There is an extreme drop in coverage and depth starting with R5 at 10^6 . This was confirmed with R9 - R16; there was not enough data left after Picard duplicate removal to run VSALIGN.

Table 3. Coverage and depth of RNA Access samples after duplicate removal with the Picard analysis pipeline.

| Sample | Titer (ge/mL) | Coverage | Depth |
|--------|---------------|----------|---------|
| R1 | 1.23E+11 | 99.83 | 1357.42 |
| R2 | 1.85E+09 | 99.74 | 663.48 |
| R3 | 7.01E+08 | 99.84 | 1283.63 |
| R4 | 2.62E+07 | 99.69 | 208.49 |
| R5 | 1.50E+06 | 3.91 | 22.91 |
| R6 | 6.56E+05 | 0 | 0 |
| R7 | 1.41E+04 | 0 | 0 |
| R8 | 3.52E+03 | 1.86 | 22.8 |

Comparison of Amplicon Sequencing and RNA Access

VSALIGN Results

The data produced via traditional amplicon sequencing and via RNA Access were aligned to the reference Zaire ebolavirus 1995, complete genome (GenBank reference: AY354458.1). VSALIGN generates a report that calculates the percent of the population that has a SNP at each position in the genome compared to the reference. Only data for the GP gene was compared between amplicon sequencing and RNA Access. Table 4 contains the count of the number of bases sequenced at or above the target depth of 200 and the number of SNPs detected $\geq 2\%$ frequency. Sample R8 persistently produced data that indicates sample contamination and was removed from this comparison analysis. The amplicon sequencing data meets or exceeds the target depth for the full length of GP to 6.56×10^5 genome copies/mL while the RNA Access data produced full length GP to 7.01×10^6 genome copies/mL. Cells reading 'nan' were not sequenced above the minimum depth and cells reading '0' were sequenced above the target depth, but no change was detected. RNA Access detects more SNPs in many samples than amplicon sequencing, especially in high-titer samples, even without full coverage of the GP gene. To determine which of these SNPs can be considered 'true' SNPs, genome positions that had a SNP frequency $\geq 2\%$ in both amplicon and RNA Access data in at least any one sample were compared across all samples in Tables 5a and 5b. There were just two samples that share SNPs $\geq 2\%$ in the same position in both preparation methods, highlighted in green. Many more SNPs were detected that are not shared between

methods. These SNPs are highlighted in red and mostly appear in the amplicon data. The unshared SNPs may be considered erroneous due to the excessive PCR used in the amplicon-generation.

Table 4. Comparison of the number of bases sequenced in the Ebolavirus GP gene above the target depth and the number of detected SNPs in amplicon and RNA Access samples.

| Sample | Titer (ge/mL) | Amplicon | | RNA Access | |
|--------|---------------|---------------------|----------------|---------------------|----------------|
| | | All Reads - GP Gene | | All reads - GP Gene | |
| | | Count | SNPs \geq 2% | Count | SNPs \geq 2% |
| R1 | 1.23E+11 | 2,031 | 4 | 2,031 | 6 |
| R2 | 1.85E+09 | 2,031 | 8 | 2,031 | 9 |
| R3 | 7.01E+08 | 2,031 | 6 | 2,031 | 7 |
| R4 | 2.62E+07 | 2,031 | 4 | 2,020 | 4 |
| R9 | 3.23E+06 | 2,031 | 5 | 1,720 | 9 |
| R5 | 1.50E+06 | 2,031 | 9 | 1,765 | 7 |
| R10 | 1.31E+06 | 2,031 | 10 | 844 | 4 |
| R11 | 6.78E+05 | 2,031 | 5 | 119 | 0 |
| R6 | 6.56E+05 | 2,031 | 6 | 154 | 1 |
| R12 | 2.73E+05 | 1,040 | 2 | 381 | 0 |
| R13 | 6.65E+04 | 1,335 | 3 | 0 | 0 |
| R7 | 1.41E+04 | 517 | 0 | 0 | 0 |
| R14 | 1.05E+04 | 0 | 0 | 0 | 0 |
| R15 | 8.77E+03 | 1,329 | 4 | 0 | 0 |
| R16 | 6.22E+03 | 967 | 2 | 0 | 0 |

Duplicate Comparison

All of the raw reads for R1 - R16 amplicon and RNA Access samples were processed with MarkDuplicates from Picard to determine the number of paired, mapped reads considered duplicates. Table 6 contains the duplicate data. The percentage of duplicates was calculated by dividing the aligned read pair duplicates by the total aligned read pairs. At the highest viral titer, a smaller percentage of the RNA Access read pairs are duplicates compared to amplicon samples. However, in line with the sharp drop in coverage at 10^6 ge/mL, there is a sharp increase in the number of duplicates in the RNA Access data. In most samples, more than half of the aligned reads are marked as duplicates, and then the total number of aligned read pairs is compared between the amplicon and RNA Access samples.

Duplicates in VSALIGN are removed with reads during the many cleaning steps. Table 7 summarizes the percent of amplicon and RNA Access reads removed. Similar to the Picard duplicate removal in RNA Access samples, those with highest titers had a significantly smaller percentage of reads removed than any others.

Table 6. Comparison of aligned sequencing read duplicate removal with the Picard analysis pipeline in amplicon and RNA Access data.

| Sample | Titer (ge/mL) | Amplicon | | | RNA Access | | |
|--------|---------------|----------------------|------------------------------|---------------------|----------------------|------------------------------|---------------------|
| | | Aligned Paired Reads | Aligned Read Pair Duplicates | Percent Duplication | Aligned Paired Reads | Aligned Read Pair Duplicates | Percent Duplication |
| R1 | 1.23E+11 | 1,854,936 | 1,180,607 | 63.65 | 1,830,381 | 713,309 | 38.97 |
| R2 | 1.85E+09 | 2,044,973 | 1,259,020 | 61.57 | 265,611 | 68,032 | 25.61 |
| R3 | 7.01E+08 | 2,314,232 | 1,562,612 | 67.52 | 279,045 | 83,017 | 29.75 |
| R4 | 2.62E+07 | 1,978,442 | 1,256,489 | 63.51 | 37,042 | 22,142 | 59.78 |
| R9 | 3.23E+06 | 549,816 | 212,372 | 38.63 | 8,514 | 8,277 | 97.22 |
| R5 | 1.50E+06 | 1,198,331 | 684,354 | 57.11 | 6,001 | 5,270 | 87.82 |
| R10 | 1.31E+06 | 1,519,526 | 913,397 | 60.11 | 1,745 | 1,676 | 96.05 |
| R11 | 6.78E+05 | 3,126,256 | 2,334,976 | 74.69 | 1,161 | 1,049 | 90.35 |
| R6 | 6.56E+05 | 1,371,059 | 824,648 | 60.15 | 1,370 | 1,064 | 77.66 |
| R12 | 2.73E+05 | 648,818 | 419,302 | 64.63 | 1,865 | 1,815 | 97.32 |
| R13 | 6.65E+04 | 368,755 | 154,550 | 41.91 | 648 | 632 | 97.53 |
| R7 | 1.41E+04 | 2,574 | 81 | 3.15 | 78 | 60 | 76.92 |
| R14 | 1.05E+04 | 824 | 13 | 1.58 | 9 | 8 | 88.89 |
| R15 | 8.77E+03 | 101,397 | 2,7136 | 26.76 | 0 | 0 | 0.00 |
| R16 | 6.22E+03 | 15876 | 1,401 | 8.82 | 2 | 1 | 50.00 |

Table 7. Comparison of the reads removed in VSALIGN through duplicate analysis and cleaning of amplicon and RNA Access data.

| Sample | Titer (ge/mL) | Amplicon | | | RNA Access | | |
|--------|------------------|----------------|------------------|--------------|----------------|------------------|--------------|
| | | Input Reads | Reads Removed | % Removed | Input Reads | Reads Removed | % Removed |
| R1 | 1.23E+11 | 3,179,050 | 3,141,032 | 98.98 | 3,491,663 | 896,526 | 25.9 |
| R2 | 1.85E+09 | 3,462,350 | 3,423,465 | 99.06 | 553,386 | 192,652 | 34.81 |
| R3 | 7.01E+08 | 3,678,866 | 3,639,322 | 99.11 | 584,393 | 187,884 | 32.15 |
| R4 | 2.62E+07 | 3,597,800 | 3,558,269 | 99.08 | 716,871 | 567,176 | 79.12 |
| R9 | 3.23E+06 | 912,342 | 877,718 | 96.38 | 2,163,734 | 2,136,355 | 98.73 |
| R5 | 1.50E+06 | 2,930,119 | 2,893,188 | 99.08 | 1,535,310 | 1,349,353 | 87.89 |
| R10 | 1.31E+06 | 2,925,671 | 2,888,946 | 98.93 | 1,327,614 | 1,320,066 | 99.43 |
| R11 | 6.78E+05 | 6,957,631 | 6,913,013 | 99.54 | 1,310,095 | 1,304,598 | 99.58 |
| R6 | 6.56E+05 | 3,172,866 | 3,135,568 | 98.92 | 628,483 | 548,969 | 87.35 |
| R12 | 2.73E+05 | 8,688,577 | 8,663,065 | 99.89 | 2,524,924 | 2,518,439 | 99.74 |
| R13 | 6.65E+04 | 2,294,162 | 2,272,667 | 99.24 | 1,731,047 | 1,728,401 | 99.85 |
| R7 | 1.41E+04 | 1,005,391 | 1,001,848 | 99 | 883,170 | 772,873 | 87.51 |
| R14 | 1.05E+04 | 1,168,084 | 1,165,519 | 99.96 | 1,656,035 | 1,479,855 | 89.36 |
| R15 | 8.77E+03 | 1,193,613 | 1,173,765 | 98.51 | 1,652,888 | 1,475,837 | 89.29 |
| R6 | 6.22E+03 | 2,701,931 | 2,686,331 | 99.6 | 1,364,034 | 1,364,028 | 100 |
| R8 | 3.52E+03 | 3,378,181 | 3,340,332 | 99.06 | 712,332 | 613,373 | 86.11 |

Known Mutations Data Comparison

There are four genome positions within the GP gene that are known to be SNPs occurring in the majority of the 7U viral population when compared with the 8U reference (Kugelman et al. 2012; Kugelman et al. 2016; Trefry et al. 2015). These four SNPs were used as control points to compare sensitivity between amplicon and RNA Access methods. Table 8 summarizes the SNP frequencies of known mutations at positions 6,179, 6,925, 7,327, and 7,669 for amplicon sequencing and RNA Access, respectively. All of these SNPs are expected to be $\geq 99\%$ of the viral population. That many of these SNPs were detected at levels well below 99% was unexpected, however the comparison between methods is similar. The SNP frequencies in red text were detected below the target depth of 200. The control SNPs were detected completely at 10^5 genome copies/mL in the amplicon method and 10^8 in the RNA Access method.

Table 8. Frequencies of the known SNP positions in amplicon and RNA Access data to compare method sensitivity.

| | | Amplicon | | | | RNA Access | | | |
|------------|----------|----------|-------|-------|-------|------------|-------|-------|-------|
| | | 6179 | 6925 | 7327 | 7669 | 6179 | 6925 | 7327 | 7669 |
| R1 | 1.23E+11 | 99.89 | 99.25 | 100 | 95.08 | 99.96 | 96.86 | 99.89 | 95.7 |
| R2 | 1.85E+09 | 20.23 | 88.56 | 100 | 5.5 | 25.39 | 87.73 | 99.85 | 8.86 |
| R3 | 7.01E+08 | 9.94 | 93.11 | 100 | 5.81 | 9.29 | 88.81 | 100 | 7.81 |
| R4 | 2.62E+07 | 99.69 | 97.22 | 98.44 | 98.43 | 100 | 96.13 | 100 | 99.85 |
| R9 | 3.23E+06 | 99.17 | 99.37 | 100 | 85.62 | 100 | 100 | 100 | nan |
| R5 | 1.50E+06 | 99.71 | 100 | 100 | 98.89 | 100 | 83.13 | 100 | 100 |
| R10 | 1.31E+06 | 61.27 | 79.13 | 100 | 22.24 | nan | 100 | nan | 0 |
| R11 | 6.78E+05 | 100 | 99.61 | 99.77 | 100 | nan | nan | nan | nan |
| R6 | 6.56E+05 | 99.89 | 100 | 99.79 | 0.14 | nan | 77.65 | nan | 0 |
| R12 | 2.73E+05 | 33 | 100 | nan | nan | nan | 100 | nan | nan |
| R13 | 6.65E+04 | 100 | 99.61 | nan | nan | nan | nan | nan | nan |
| R7 | 1.41E+04 | nan | nan | nan | nan | nan | nan | nan | nan |
| R14 | 1.05E+04 | nan | nan | nan | nan | nan | nan | nan | nan |
| R15 | 8.77E+03 | nan | 99.77 | 99.8 | 0.88 | nan | nan | nan | nan |
| R16 | 6.22E+03 | 88.55 | 99.09 | nan | nan | nan | nan | nan | nan |

DISCUSSION

The current amplicon-based whole genome amplification method is well-characterized as the standard, however, the method is quite labor-intensive and the excessive amplification can introduce errors (Kugelman *et al.* 2017). The RNA Access method from Illumina was designed to specifically enrich and separate a target, in this case Ebolavirus, from a cDNA library with minimal amplification steps. The objectives of this study were to 1) determine the sensitivity of RNA Access compared to amplicon generation and establish a new lower limit of detection (LLOD) and 2) determine the amount of sequencing required to achieve at least the current depth and coverage produced by amplicon generation.

Ultimately, 16 whole blood samples in TRIzol from cynomolgus macaques infected with either a 7U or 8U variant of Ebola Kikwit were used in this study. The first eight samples, R1 – R8, ranged in viral titer from 1.23×10^{11} to 3.52×10^3 genome copies/mL, which included samples both above and below the current limit of detection for amplicon generation of 10^5 genome copies/mL. In order to confirm the LLOD results established in the first set of samples, a second set of samples, R9 - R16, processed with two samples each at 10^6 , 10^5 , 10^4 , and 10^3 genome copies/mL. After sequencing, the raw reads for samples R1 – R8 were cleaned and aligned in VSALIGN first using the standard input of 200,000 random reads and a minimum target depth of 200 which are the standard VSALIGN input parameters for amplicon sequencing in the USAMRIID CGS. Using these input parameters, full genomes were recovered only to a minimum of 10^7 genome copies/mL, though partial genomes were recovered at lower viral titers (Table 1a). Samples R4 – R8 were processed through the VSALIGN pipeline a second time

using all available reads as the input. Using all available reads increased the RNA Access sensitivity to 10^6 genome copies/mL (Table 1b). Sample R4 was used as a control to compare using 200,000 reads and all reads. While the coverage was consistent, the depth increased significantly. VSALIGN uses a random assortment of 200,000 reads per run and will produce slightly different results for each instance. All reads could be used for every sample, but the analysis time will be significantly increased from a few days to a week or more. Therefore, If RNA Access is the method of choice for population genetics, using all of the reads for VSALIGN should be considered for samples with lower viral titer. Samples R9 - R16 were sequenced and processed with VSALIGN using all available reads. The results in Table 2 confirm the lower limit of detection for full genomes at 10^6 genome copies/mL with partial genome recovery to 10^4 genome copies/mL.

The Picard pipeline is another tool used at CGS that is a combination of in-house and open-source tools for cleaning, duplicate removal, and alignment. Picard method MarkDuplicates is a more strict method of duplicate removal that also incorporates the flowcell coordinate data for fragments and removes those that are too close together called optical duplicates. Prinseq (Schmieder and Edwards 2011), used in VSALIGN does not remove optical duplicates. The RNA Access reads were also used in the Picard pipeline to determine if there were any changes in the LLOD. When processed in the Picard pipeline, full genomes were recovered to 10^7 genome copies/mL and there was a sharp cutoff in coverage at 10^6 with only 3.91% genome coverage. Acknowledging that Picard decreases the sensitivity because the cleaning is more complete, VSALIGN will be used with a caveat.

With a clear lower limit of detection defined for RNA Access, the amplicon data was compared to the RNA Access data. Only the data for the GP gene was compared. The glycoprotein of Ebolavirus is essential for entry into host cells, and is most often the target of therapeutics. Additionally, there are several well-characterized mutations within the GP gene that could be used as controls when comparing detection (Kugelman *et al.* 2012, 2016; Trefry *et al.* 2015; Volchkova *et al.* 2011). First, the count of the number of bases sequenced and the SNPs detected in $\geq 2\%$ of the viral population, which is expected to be 2031 bases (Table 4). The $\geq 2\%$ frequency is the standard set in the CGS for significant changes. The LLOD for amplicon generation was met with full GP coverage at 10^5 genome copies/mL, however the RNA Access LLOD is not quite met.

Interestingly, more SNPs were detected in four of the higher titer samples. The ‘true’ SNPs were considered to be those that were shared between both methods. Genome positions that had a SNP frequency $\geq 2\%$ in any one sample in both methods were compared (Tables 5a and 5b). Only four positions in two samples had similar frequency in both methods. Many more SNPs were detected in the amplicon data, supporting the hypothesis that the excess of PCR in amplicon generation incorporates errors. Further, the amplicon data was also processed with Picard. When compared with the Picard output for RNA Access (Table 6), over 60% of the reads are duplicates at the highest viral titers while the RNA Access duplicates are about half as much. RNA Access does show a significant increase in duplicates to $> 90\%$ at 10^6 genome copies/mL and amplicon duplicates stay about the same, however those RNA Access samples have 99% fewer input reads than the amplicon samples. The same comparison was performed for the percent of reads removed from the VSALIGN analysis (Table 7). The reads

removed from VSALIGN include all of the many cleaning steps in addition to the duplicate analysis. These results support the duplicate analysis from Picard, further supporting the hypothesis that amplicon generation is introducing more errors than RNA Access.

Finally, the ‘control’ SNPs were compared between methods. The GP gene can produce three proteins: sGP, GP_{1,2}, and ssGP (Kugelman *et al.* 2012; Kuhn 2008; Volchkova *et al.* 2011). The differences occur in the mRNA editing site which is a section of seven uridylys (Kugelman *et al.* 2012). When the RNA-dependent RNA polymerase synthesizes all seven adenylys it produces sGP and is called a 7U variant, when the polymerase stutters and adds an extra adenylyl it produces GP_{1,2} and is called an 8U variant, and when the polymerase skips one uridylyl or adds two adenylyl it produces ssGP and is either a 6U or 9U variant (Kugelman *et al.* 2012; Volchkova *et al.* 2011). Ebolavirus passaged *in vivo* is primarily 7U, even when the virus used for challenge is 8U (Volchkova *et al.* 2011). When compared to an 8U reference, we expect position 6,925 at the end of the mRNA editing site to be a deletion with $\geq 99\%$ frequency in *in vivo* studies. Three other positions in the GP gene generate mutation with near $\geq 99\%$ frequency *in vivo* when compared to the 8U reference (Kugelman *et al.* 2012; Kugelman *et al.* 2016). Some of the results were unexpected; not all of the control SNPs were detected at the expected frequency. Encouragingly, the SNPs that are lower than expected do match between methods, but further confirmation work will need to be performed. Nevertheless, the control SNPs confirms the amplicon sequencing lower limit of detection of 10^5 genome copies/mL. At least one position in each of the 10^6 genome copies/mL samples was either not detected at all or not detected above the target

depth of 200. Those SNPs may still be real, but will need to be confirmed with a second validation. When the SNPs below 200 depth are included, the proposed lower limit of detection for RNA Access of 10^6 genome copies/mL is also confirmed by the controls. However, it may be that RNA Access samples require more sequencing than previously thought. Samples R2 – R4 all had less than 1,000,000 reads. Far fewer reads were removed from R2 – R4 and they still had complete genome coverage well above the target depth. Possibly, lower titer samples require more sequencing to counter the reads removed. Still, RNA Access has been shown to recover Ebolavirus in a sample from a human patient when amplicon sequencing could not recover it (Mate *et al.* 2015).

In conclusion, RNA Access is a viable option to replace the current standard of amplicon sequencing. This method is not as sensitive in the typical samples received for amplicon sequencing. Furthermore, most samples $< 10^6$ genome copies/mL will need more than 1,000,000 reads (the standard target for amplicon sequencing) and all of those reads will be required for VSALIGN, increasing the analysis time. Nevertheless, these shortcomings are outweighed by the benefits gained. A large population genetics study processed with amplicon sequencing could take a month or more to process in the lab, plus sequencing and analysis time. RNA Access could be completed through analysis in less than 3 weeks, conservatively. Though many parts of amplicon sequencing have been automated, it is extremely labor intensive and there is plenty of room for human error in addition to the errors introduced to the sample by touchdown PCR. RNA Access was completely automated and customized for the small volumes used in the CGS. Automation improves both sample-to-sample consistency and reproducibility. Amplicon

sequencing will be kept as a backup method, but RNA Access should be the new method of choice for population genetics studies.

REFERENCES

- [CDC] Centers for Disease Control and Prevention. 2015. Outbreaks chronology: Ebola virus disease [Internet]. [cited 2015 December 1]. Available from: <http://www.cdc.gov/vhf/ebola/outbreaks/history/chronology.html>.
- Geisbert TW, Feldmann H. 2011. Recombinant vesicular stomatitis virus-based vaccines against ebola and marburg virus infections. *J Infect Dis.* 204 Suppl 3:S1075-1081.
- Illumina. 2014. Truseq rna access library prep kit [Internet]. [cited 2015 October 15]. <http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-truseq-rna-access.pdf>.
- Kugelman JR, Kugelman-Tonos J, Ladner JT, Pettit J, Keeton CM, Nagle ER, Garcia KY, Froude JW, Kuehne AI, Kuhn JH et al. 2015a. Emergence of ebola virus escape variants in infected nonhuman primates treated with the mb-003 antibody cocktail. *Cell Rep.* 12(12):2111-2120.
- Kugelman JR, Lee MS, Rossi CA, McCarthy SE, Radoshitzky SR, Dye JM, Hensley LE, Honko A, Kuhn JH, Jahrling PB et al. 2012. Ebola virus genome plasticity as a marker of its passaging history: A comparison of in vitro passaging to non-human primate infection. *PLOS One.* 7(11):e50316.
- Kugelman JR, Rossi CA, Wiley MR, Ladner JT, Nagle ER, Pfeffer BP, Garcia K, Prieto K, Wada J, Kuhn JH et al. 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(3):e0150919.
- Kugelman JR, Sanchez-Lockhart M, Andersen KG, Gire S, Park DJ, Sealfon R, Lin AE, Wohl S, Sabeti PC, Kuhn JH et al. 2015b. Evaluation of the potential impact of ebola virus genomic drift on the efficacy of sequence-based candidate therapeutics. *MBio.* 6(1).
- Kugelman JR, Wiley MR, Nagle ER, Reyes D, Pfeffer BP, Kuhn JH, et al. 2017. Error baseline rates of five sample preparation methods used to characterize RNA virus populations. *PLOS ONE* 12(2): e0171333.
- Kuhn JH. 2008. Chapter 11 molecular characteristics of filoviruses. In: Calisher CH, editor. *Filoviruses: A compendium of 40 years of epidemiological, clinical, and laboratory studies.* New York, NY: Springer-Verlag Wien. p. 175-263.
- Kuhn JH, Andersen KG, Baize S, Bao Y, Bavari S, Berthet N, Blinkova O, Brister JR, Clawson AN, Fair J *et al.* 2014. Nomenclature- and database-compatible names for the two ebola virus variants that emerged in guinea and the democratic republic of the congo in 2014. *Viruses.* 6(11):4760-4799.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10(3):R25-R25.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and samtools. *Bioinformatics.* 25(16):2078-2079.

Li K, Brownley A, Stockwell TB, Beeson K, McIntosh TC, Busam D, Ferriera S, Murphy S, Levy S. 2008. Novel computational methods for increasing pcr primer design effectiveness in directed sequencing. *BMC Bioinformatics.* 9:191.

Mate SE, Kugelman JR, Nyenswah TG, Ladner JT, Wiley MR, Cordier-Lassalle T, Christie A, Schroth GP, Gross SM, Davies-Wayne GJ et al. 2015. Molecular evidence of sexual transmission of ebola virus. *N Engl J Med.* 373(25):2448-2454.

Messaoudi I, Amarasinghe GK, Basler CF. 2015. Filovirus pathogenesis and immune evasion: insights from Ebola virus and Marburg virus. *Nat Rev Microbiol.* 13:663-676.

Moya A, Holmes EC, Gonzalez-Candelas F. 2004. The population genetics and evolutionary epidemiology of rna viruses. *Nat Rev Microbiol.* 2(4):279-288.

Oestereich L, Ludtke A, Wurr S, Rieger T, Munoz-Fontela C, Gunther S. 2014. Successful treatment of advanced ebola virus infection with t-705 (favipiravir) in a small animal model. *Antiviral Res.* 105:17-21.

Qiu X, Wong G, Audet J, Bello A, Fernando L, Alimonti JB, Fausther-Bovendo H, Wei H, Aviles J, Hiatt E et al. 2014. Reversion of advanced ebola virus disease in nonhuman primates with zmapp. *Nature.* 514(7520):47-53.

Quinlan AR, Hall IM. 2010. Bedtools: A flexible suite of utilities for comparing genomic features. *Bioinformatics.* 26(6):841-842.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics.* 27(6):863-864.

Trefry JC, Wollen SE, Nasar F, Shamblin JD, Kern SJ, Bearss JJ, Jefferson MA, Chance TB, Kugelman JR, Ladner JT et al. 2015. Ebola virus infections in nonhuman primates are temporally influenced by glycoprotein poly-u editing site populations in the exposure material. *Viruses.* 7(12):6739-6754.

Volchkova VA, Dolnik O, Martinez MJ, Reynard O, Volchkov VE. 2011. Genomic rna editing and its impact on ebola virus adaptation during serial passages in cell culture and infection of guinea pigs. *J Infect Dis.* 204(suppl_3):S941-S946.

Warren TK, Jordan R, Lo MK, Ray AS, Mackman RL, Soloveva V, Siegel D, Perron M, Bannister R, Hui HC et al. 2016. Therapeutic efficacy of the small molecule gs-5734 against ebola virus in rhesus monkeys. *Nature*. 531(7594):381-385.

Warren TK, Wells J, Panchal RG, Stuthman KS, Garza NL, Van Tongeren SA, Dong L, Retterer CJ, Eaton BP, Pegoraro G et al. 2014. Protection against filovirus diseases by a novel broad-spectrum nucleoside analogue bcx4430. *Nature*. 508(7496):402-405.

[WHO] World Health Organization. 2015. Ebola vaccines, therapies, and diagnostics [Internet]. [cited 2015 December 10]. Available from: http://www.who.int/medicines/emp_ebola_q_as/en/.