VGEA: A snakemake pipeline for RNA virus genome assembly from next generation sequencing data

Paul E. Oluniyi [1, 2], **Fehintola V. Ajogbasile** [1, 2], **Judith U. Oguzie** [1, 2], **Jessica N. Uwanibe** [1, 2], **Adeyemi T. Kayode** [1, 2], **Anise N. Happi** [3], **Chinedu A. Ugwu** [1, 2], **Testimony J. Olumade** [1, 2], **Olusola Ogunsanya** [3], **Philomena E. Eromon** [2], **Onikepe A. Folarin** [1, 2], **Simon D.W. Frost** [4, 5], **Jonathan L. Heeney** [6], **Christian T. Happi** Corresp. [1, 2]

[1] Department of Biological Sciences, Faculty of Natural Sciences, Redeemer's University, Ede, Osun, Nigeria

[2] African Centre of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University, Ede, Osun, Nigeria

[3] Department of Veterinary Pathology, Faculty of Veterinary Medicine, University of Ibadan, Ibadan, Oyo, Nigeria

[4] Microsoft Research, Redmond, 98052, Washington, United States of America

[5] London School of Hygiene & Tropical Medicine, London, United Kingdom

[6] Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom

Corresponding Author: Christian T. Happi
Email address: happic@run.edu.ng

Background. Next generation sequencing (NGS)-based studies have vastly increased our understanding of viral diversity. Viral sequence data obtained from NGS experiments are a rich source of information, these data can be used to study their epidemiology, evolution, transmission patterns, and can also inform drug and vaccine design. Viral genomes however represent a great challenge to bioinformatics due to their high mutation rate and forming quasispecies in the same infected host. This has therefore brought about the need to develop/implement advanced bioinformatics tools to assemble genomes well-representative of the viral population circulating in individual patients.

Results. Here we present VGEA (Viral Genomes Easily Assembled), a snakemake workflow for advanced assembly of RNA viral genomes from NGS data. VGEA enables users to split bam files into forward and reverse reads, carry out de novo assembly of forward and reverse reads to generate contigs, pre-process reads for quality and contamination, and map reads to a reference tailored to the sample using corrected contigs supplemented by the user's choice of reference sequences.

Conclusion. VGEA is freely available on GitHub at: https://github.com/pauloluniyi/VGEA under the GNU General Public License and also on Zenodo (doi: 10.5281/zenodo.3702287).

VGEA: A snakemake pipeline for RNA virus genome assembly from next generation sequencing data

Paul E. Oluniyi[1,2], Fehintola V. Ajogbasile[1,2], Judith U. Oguzie[1,2], Jessica N. Uwanibe[1,2], Adeyemi T. Kayode[1,2], Anise N. Happi[3], Chinedu A. Ugwu[1,2], Testimony J. Olumade[1,2], Olusola Ogunsanya[3], Philomena E. Eromon[2], Onikepe A. Folarin[1,2], Simon D.W. Frost[4,5], Jonathan L. Heeney[6], Christian T. Happi[1,2*]

1. Department of Biological Sciences, Faculty of Natural Sciences, Redeemer's University, Ede, Osun State, Nigeria.
2. African Centre of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University, Ede, Osun State, Nigeria.
3. Department of Veterinary Pathology, Faculty of Veterinary Medicine, University of Ibadan, Ibadan, Nigeria.
4. Microsoft Research, Redmond, 98052, WA, USA.
5. London School of Hygiene & Tropical Medicine, London, UK.
6. Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom.

*Corresponding Author:
Christian Happi[1,2]
Email address: happic@run.edu.ng

ABSTRACT

Background. Next generation sequencing (NGS)-based studies have vastly increased our understanding of viral diversity. Viral sequence data obtained from NGS experiments are a rich source of information, these data can be used to study their epidemiology, evolution, transmission patterns, and can also inform drug and vaccine design. Viral genomes however represent a great challenge to bioinformatics due to their high mutation rate and forming quasispecies in the same infected host. This has therefore brought about the need to develop/implement advanced bioinformatics tools to assemble genomes well-representative of the viral population circulating in individual patients.

Results. Here we present **VGEA** (**V**iral **G**enomes **E**asily **A**ssembled), a snakemake workflow for advanced assembly of RNA viral genomes from NGS data. VGEA enables users to split bam files into forward and reverse reads, carry out *de novo* assembly of forward and reverse reads to generate contigs, pre-process reads for quality and contamination, and map reads to a reference tailored to the sample using corrected contigs supplemented by the user's choice of reference sequences.

Conclusion. VGEA is freely available on GitHub at: https://github.com/pauloluniyi/VGEA under the GNU General Public License and also on Zenodo (doi: 10.5281/zenodo.3702287).

Keywords: VGEA, NGS, Genome, Assembly

INTRODUCTION

The most abundant biological entities on Earth are viruses as they can be found among all cellular forms of life. So far, over four thousand five hundred viral species have been discovered, from which a huge amount of sequence information has been collected by researchers and scientists all over the world (*Pickett et al., 2012; Sharma et al., 2015; Brister et al., 2015*). In recent times (past two decades), a number of these viruses have emerged in the human population causing outbreaks sometimes pandemics., These viruses include mainly: Influenza virus, Severe Acute Respiratory Syndrome (SARS) coronavirus, Middle East Respiratory Syndrome (MERS) coronavirus, Ebola virus, Yellow fever virus, Lassa virus (LASV), Zika virus (*Chan, 2002; Bean et al., 2013; Folarin et al., 2016; Grubaugh et al., 2017; Metsky et al., 2017; Siddle et al., 2018; Ajogbasile et al., 2020*) and SARS-CoV-2 (*Chen et al., 2020; Holshue et al., 2020; Sohrabi et al., 2020*). During these outbreaks and pandemics, identification of the causative agents and carrying out sequencing to obtain the genomes of the viruses have proved to be critical in helping inform disease surveillance and epidemiology.

NGS platforms have been widely accepted as high-throughput, unbiased technologies that have many attractive features compared to conventional diagnostic methods for virus detection and assembly (*Tang & Chiu, 2010; Mokili et al., 2012*). NGS-based studies have vastly increased our understanding of viral diversity (*Reyes et al., 2010; Cantalupo et al., 2011*). Pathogen sequence data obtained from NGS experiments are a rich source of information, these data can be used to study their epidemiology, evolution, transmission patterns, and can also inform drug and vaccine design. The field of genomics, especially pathogen genomics has been transformed by NGS, with costs constantly decreasing, equipment becoming more portable/field deployable during

outbreaks and remarkable increase in data availability. The huge amount of data being generated has brought about the need to develop simple and user friendly bioinformatics tools to assemble pathogen genomes well-representative of the pathogen population circulating in individual patients. Here we present **VGEA**, a pipeline for assembly of RNA viral genomes from next generation sequencing data**.**

MATERIALS AND METHODS

The VGEA pipeline is built on the snakemake workflow management system (*Köster & Rahmann, 2012*), a workflow management system that allows the effortless deployment and execution of complex distributed computational workflows in any UNIX-based system, from local machines to high-performance computing clusters. In order to guarantee reproducibility of the results obtained, the VGEA pipeline integrates fixed versions of the tools implemented in the pipeline from conda (https://docs.conda.io/en/latest/). Several tools are used to perform different tasks within the pipeline: **Samtools** (*Li et al., 2009*) for splitting of bam files into forward and reverse reads; **IVA** (*Hunt et al., 2015*) for *de novo* assembly to generate contigs; **Shiver** (*Wymant et al., 2018*) to pre-process reads for quality and contamination, then map to a reference tailored to the sample using corrected contigs supplemented with the user's choice of existing reference sequences.

The **VGEA** pipeline requires the following dependencies:

- Python 3 (www.python.org).
- Samtools (*Li et al., 2009*).
- IVA (*Hunt et al., 2015*).
- Shiver (*Wymant et al., 2018*).

- Fastp (*Chen et al., 2018*).
- Trimmommatic, optional but highly recommended (*Bolger et al., 2014*).
- KMC (*Kokot et al., 2017*).
- MUMmer (*Marçais et al., 2018*).
- SMALT (https://www.sanger.ac.uk/science/tools/smalt-0) or BWA (*Li & Durbin, 2009*) or BOWTIE (*Langmead, 2010*).
- Fastaq (https://github.com/sanger-pathogens/Fastaq).
- Biopython (*Cock et al., 2009)*.
- MAFFT (*Katoh et al., 2002*).
- BLAST version 2.2.28 (*Altschul et al., 1990*).
- SPAdes (*Bankevich et al., 2012*).

We have also made available a singularity recipe file (*Kurtzer et al., 2017*) on the GitHub page: https://github.com/paululuniyi/VGEA. With the provision of the singularity recipe file, users can easily build a local image of the VGEA container that includes all necessary tools, in their fixed versions, and their dependencies using the command below:

sudo singularity build vgea.simg Singularity

Users can then proceed to run the entire VGEA pipeline with all the dependencies installed from the singularity container. This approach ensures the reproducibility and the tracking of both software code and version, regardless of the operating system used. With the provision of the singularity container, users can easily deploy VGEA to run in the cloud or on high performance computing (HPC) clusters.

The VGEA pipeline consists of three major steps:

115

116 **Splitting of BAM files**

117 BAM (and SAM and CRAM) are file formats that contain sequencing reads: either aligned,
118 unaligned, or a combination of the two. One initial step after carrying out next-generation
119 sequencing is often to get rid of host contamination, e.g., by mapping the FASTQ reads obtained
120 from the sequencing machine against the human genome. This will ultimately yield a BAM file.
121 The BAM file obtained from host contaminant removal can then be used as input for the VGEA
122 pipeline. The pipeline has been developed to facilitate splitting of bam files into fastq files of
123 forward and reverse reads using **Samtools** (*Li et al., 2009*). Another reason for having BAM files
124 as starting input for the VGEA pipeline is that scientists, especially in resource-limited settings,
125 usually have BAM files handy. BAM files are usually smaller in size than their corresponding
126 FASTQ files making them easily transferable or uploadable to the cloud.
127 During the splitting step, the pipeline checks in the current directory or within the container (if
128 the user is making use of the singularity container provided) for any file with a '.bam' extension,
129 if it finds any, it splits into FASTQ files of forward and reverse reads.

130

131 **Assembly**

132 Following splitting of the BAM files into FASTQ files of forward and reverse reads, the VGEA
133 pipeline carries out *de novo* assembly to generate contigs using **IVA** (*Hunt et al., 2015*). **IVA** is
134 used as our default assembler because it was designed specifically for read pairs sequenced at
135 highly variable depth from RNA virus samples and has been demonstrated to outperform all
136 other virus de novo assemblers (*Hunt et al., 2015*).

137

- **Mapping**

The VGEA pipeline then uses the **Shiver** software (*Wymant et al., 2018*) for mapping of reads to a reference. First, the contigs generated in the assembly step are compared with reference(s) supplied by the user using BLASTN (*Altschul et al., 1990*), this is to remove contaminants and low-quality contigs. Using **MAFFT** (*Katoh et al., 2002*), the processed contigs are added to the alignment of existing references initially supplied by the user from which **Shiver** identifies the closest existing reference by comparison with all of the contigs. Using contig sequences and the closest existing reference to fill in gaps between contigs, if any exists, **Shiver** creates a reference for mapping. User-supplied raw reads are then mapped to the shiver-created reference to generate a consensus sequence well-representative of the viral population in the patient sample. Before mapping however, the reads are trimmed using Trimmomatic (*Bolger et al., 2014*) and Fastaq (https://github.com/sanger-pathogens/Fastaq) in order to remove low-quality bases, adapters and primer sequences. Adapter and Primer sequences are provided by the user.

RESULTS

We demonstrated the usage and performance of the VGEA pipeline by applying it to generate consensus whole genomes from Lassa virus and SARS-COV-2 datasets sequenced on the illumina MiSeq and illumina FGx sequencing machines in our lab at the African Centre of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University, Nigeria. We also applied the pipeline to generate whole genomes from HIV-1 datasets sequenced on the illumina HiSeq 2500 obtained from NCBI Sequence Read Archive (SRA). We made use of 60 test datasets (Lassa Virus (20), SARS-CoV-2 (20) and HIV-1 (20)) for the validation of the

VGEA pipeline. All our test datasets are available on figshare (https://doi.org/10.6084/m9.figshare.13009997). The performance of our pipeline was consistent irrespective of whether the samples were sequenced with the illumina MiSeq, FGx or HiSeq 2500. We compared our genomes obtained with **VGEA** to genomes obtained with **viral-ngs** (https://github.com/broadinstitute/viral-ngs) (*Park et al., 2015*) which is a suite of genomic analysis pipelines for viral sequencing and is one of the most widely used resources for whole genome assembly of viruses (Tables 1,2,3&4).

We compared the means of our genome lengths using **VGEA** and **viral-ngs** to determine the significance of the differences in genome lengths from the two pipelines; using the **R programming language**, we carried out a student's t-test and found that genome length with **VGEA** are significantly longer than genome length with **viral-ngs** (p-value = 0.002962, C.I = 95% for Lassa virus (S) dataset; p-value = 0.01748, C.I = 95% for Lassa virus (L) dataset; p-value = 0.04283, C.I = 95% for SARS-CoV-2 dataset; p-value = 0.001286, C.I = 95% for HIV-1 datasets).

Using the **VGEA** pipeline, we also obtained Lassa virus partial genomes from three rodent samples, however running the same samples through **viral-ngs** we couldn't obtain any genome despite making the pipeline as less stringent as possible.

DISCUSSION

Using the workflow management system, **Snakemake**, we have developed a user-friendly pipeline for advanced assembly of viral genomes from NGS data. Its main features are: (i) splitting bam files into forward and reverse reads, (ii) *de novo* assembly to generate contigs, (iii) pre-processing of reads for quality and contamination, (iv) mapping reads to a sample-tailored reference. This significantly improves the quality of the genomes obtained from NGS data and generates genomes well-representative of the pathogen population circulating in individual patients and communities. VGEA is freely available on GitHub at: https://github.com/paululuniyi/VGEA under the GNU General Public License and also on Zenodo (doi: 10.5281/zenodo.3702287). All test datasets used for the validation of the pipeline are available on NCBI and also on figshare (https://doi.org/10.6084/m9.figshare.13009997).

CONCLUSION

VGEA was built primarily by biologists and in a manner that is easy to be employed by users without significant computational background. As new and innovative tools for viral genome analysis and assembly are increasingly being developed, these can easily be incorporated into the VGEA pipeline. We hope that other scientists can build upon and improve VGEA as a tool to extract more qualitative and quantitative information from viral genomes.

ACKNOWLEDGEMENTS

We appreciate the continuous support of ACEGID staff and the management of Redeemer's University.

Abbreviations

VGEA Viral Genomes Easily Assembled

NGS Next generation sequencing

RNA Ribonucleic acid

SARS Severe Acute Respiratory Syndrome

MERS Middle East Respiratory Syndrome

IVA Iterative Virus Assembler

SHIVER Sequences from HIV Easily Reconstructed

HPC High Performance Computing

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by grants from the National Institute of Allergy and Infectious Diseases (https://www.niaid.nih.gov), NIH-H3Africa (https://h3africa.org) (U01HG007480 and U54HG007480 to C.T.H), the World Bank grant (worldbank.org) (project ACE019 to C.T.H.) and the Wellcome Trust grant (https://wellcome.ac.uk) (216619/Z/19/Z to C.T.H and J.L.H). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

National Institute of Allergy and Infectious Diseases, NIH-H3Africa: U01HG007480 and U54HG007480.

The World Bank: ACE019.

The Wellcome Trust: 216619/Z/19/Z.

Competing Interests

Simon D.W. Frost is employed by Microsoft Research. All other authors have declared that no competing interests exist.

Author Contributions

- Paul E. Oluniyi developed/implemented the pipeline and wrote the manuscript.
- Paul E. Oluniyi, Fehintola V. Ajogbasile, Judith U. Oguzie, Jessica N. Uwanibe, Adeyemi T. Kayode, Philomena E. Eromon, Anise N. Happi, Testimony J. Olumade and Olusola Ogunsanya performed sequencing and metagenomics analysis.
- Anise N. Happi, Chinedu A. Ugwu, Testimony J. Olumade and Olusola Ogunsanya collected and processed rodent samples for sequencing.
- Onikepe A. Folarin, Simon D.W. Frost, Jonathan L. Heeney and Christian T. Happi supervised the project.
- Christian T. Happi reviewed and corrected the manuscript and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

VGEA is freely available on GitHub at: https://github.com/paululuniyi/VGEA under the GNU General Public License.

All primary test datasets used for the validation of the VGEA pipeline are available on figshare (https://doi.org/10.6084/m9.figshare.13009997). All SARS-CoV-2 and Lassa virus test datasets have been submitted to NCBI SRA (BioProject accession numbers, PRJNA666685 and PRJNA666664). All HIV-1 test datasets are available on NCBI SRA (accession numbers: ERR3953696, ERR3953853, ERR3953893, ERR3953891, ERR3953866, ERR3953846, ERR3953756, ERR3953877, ERR3953876, ERR3953750, ERR3953741, ERR3953697, ERR3953699, ERR3953706, ERR3953708, ERR3953710, ERR3953712, ERR3953716, ERR3953295, ERR3953693).

REFERENCES

Ajogbasile FV, Oguzie JU, Oluniyi PE, Eromon PE, Uwanibe JN, Mehta SB, Siddle KJ, Odia I, Winnicki SM, Akpede N, Akpede G, Okogbenin S, Ogbaini-Emovon E, MacInnis BL, Folarin OA, Modjarrad K, Schaffner SF, Tomori O, Ihekweazu C, Sabeti PC, Happi CT. 2020. Real-time Metagenomic Analysis of Undiagnosed Fever Cases Unveils a Yellow Fever Outbreak in Edo State, Nigeria. *Scientific reports* 10:3180. DOI: 10.1038/s41598-020-59880-w.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* 215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology: a journal of computational molecular cell biology* 19:455–477. DOI: 10.1089/cmb.2012.0021.

Bean AGD, Baker ML, Stewart CR, Cowled C, Deffrasnes C, Wang L-F, Lowenthal JW. 2013. Studying immunity to zoonotic diseases in the natural host - keeping it real. *Nature reviews. Immunology* 13:851–861. DOI: 10.1038/nri3551.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.

Brister JR, Ako-Adjei D, Bao Y, Blinkova O. 2015. NCBI viral genomes resource. *Nucleic acids research* 43:D571–7. DOI: 10.1093/nar/gku1207.

Cantalupo PG, Calgua B, Zhao G, Hundesa A, Wier AD, Katz JP, Grabe M, Hendrix RW, Girones R, Wang D, Pipas JM. 2011. Raw sewage harbors diverse viral populations. *mBio* 2. DOI: 10.1128/mBio.00180-11.

Chan PKS. 2002. Outbreak of avian influenza A(H5N1) virus infection in Hong Kong in 1997. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* 34 Suppl 2:S58–64. DOI: 10.1086/338820.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890. DOI: 10.1093/bioinformatics/bty560.

Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J 'an,

Yu T, Zhang X, Zhang L. 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet* 395:507–513. DOI: 10.1016/S0140-6736(20)30211-7.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423. DOI: 10.1093/bioinformatics/btp163.

Folarin OA, Ehichioya D, Schaffner SF, Winnicki SM, Wohl S, Eromon P, West KL, Gladden-Young A, Oyejide NE, Matranga CB, Deme AB, James A, Tomkins-Tinch C, Onyewurunwa K, Ladner JT, Palacios G, Nosamiefan I, Andersen KG, Omilabu S, Park DJ, Yozwiak NL, Nasidi A, Garry RF, Tomori O, Sabeti PC, Happi CT. 2016. Ebola Virus Epidemiology and Evolution in Nigeria. *The Journal of infectious diseases* 214:S102–S109. DOI: 10.1093/infdis/jiw190.

Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K, Wiley MR, White S, Thézé J, Magnani DM, Prieto K, Reyes D, Bingham AM, Paul LM, Robles-Sikisaka R, Oliveira G, Pronty D, Barcellona CM, Metsky HC, Baniecki ML, Barnes KG, Chak B, Freije CA, Gladden-Young A, Gnirke A, Luo C, MacInnis B, Matranga CB, Park DJ, Qu J, Schaffner SF, Tomkins-Tinch C, West KL, Winnicki SM, Wohl S, Yozwiak NL, Quick J, Fauver JR, Khan K, Brent SE, Reiner RC Jr, Lichtenberger PN, Ricciardi MJ, Bailey VK, Watkins DI, Cone MR, Kopp EW 4th, Hogan KN, Cannons AC, Jean R, Monaghan AJ, Garry RF, Loman NJ, Faria NR, Porcelli MC, Vasquez C, Nagle ER, Cummings DAT, Stanek D, Rambaut A, Sanchez-Lockhart M, Sabeti PC, Gillis LD, Michael SF, Bedford T, Pybus OG, Isern S, Palacios G, Andersen KG. 2017. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* 546:401–405. DOI: 10.1038/nature22400.

Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Spitters C, Ericson K, Wilkerson S, Tural A, Diaz G, Cohn A, Fox L, Patel A, Gerber SI, Kim L, Tong S, Lu X, Lindstrom S, Pallansch MA, Weldon WC, Biggs HM, Uyeki TM, Pillai SK, Washington State 2019-nCoV Case Investigation Team. 2020. First Case of 2019 Novel Coronavirus in the United States. *The New England journal of medicine*. DOI: 10.1056/NEJMoa2001191.

Hunt M, Gall A, Ong SH, Brener J, Ferns B, Goulder P, Nastouli E, Keane JA, Kellam P, Otto TD. 2015. IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics* 31:2374–2376. DOI: 10.1093/bioinformatics/btv120.

Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple

sequence alignment based on fast Fourier transform. *Nucleic acids research* 30:3059–3066. DOI: 10.1093/nar/gkf436.

Kokot M, Dlugosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33:2759–2761. DOI: 10.1093/bioinformatics/btx304.

Köster J, Rahmann S. 2012. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* 28:2520–2522. DOI: 10.1093/bioinformatics/bts480.

Kurtzer GM, Sochat V, Bauer MW. 2017. Singularity: Scientific containers for mobility of compute. *PloS one* 12:e0177459. DOI: 10.1371/journal.pone.0177459.

Langmead B. 2010. Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis... [et al.]* Chapter 11:Unit 11.7. DOI: 10.1002/0471250953.bi1107s32.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760. DOI: 10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. DOI: 10.1093/bioinformatics/btp352.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology* 14:e1005944. DOI: 10.1371/journal.pcbi.1005944.

Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, West K, Qu J, Baniecki ML, Gladden-Young A, Lin AE, Tomkins-Tinch CH, Ye SH, Park DJ, Luo CY, Barnes KG, Shah RR, Chak B, Barbosa-Lima G, Delatorre E, Vieira YR, Paul LM, Tan AL, Barcellona CM, Porcelli MC, Vasquez C, Cannons AC, Cone MR, Hogan KN, Kopp EW, Anzinger JJ, Garcia KF, Parham LA, Ramírez RMG, Montoya MCM, Rojas DP, Brown CM, Hennigan S, Sabina B, Scotland S, Gangavarapu K, Grubaugh ND, Oliveira G, Robles-Sikisaka R, Rambaut A, Gehrke L, Smole S, Halloran ME, Villar L, Mattar S, Lorenzana I, Cerbino-Neto J, Valim C, Degrave W, Bozza PT, Gnirke A, Andersen KG, Isern S, Michael SF, Bozza FA, Souza TML, Bosch I, Yozwiak NL, MacInnis BL, Sabeti PC. 2017. Zika virus evolution and spread in the Americas. *Nature* 546:411–415. DOI: 10.1038/nature22402.

Mokili JL, Rohwer F, Dutilh BE. 2012. Metagenomics and future perspectives in virus discovery. *Current opinion in virology* 2:63–77. DOI: 10.1016/j.coviro.2011.12.004.

Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, 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 P-P, Lin AE, Schaffner SF, Bird B, Towner J, Mamoh M, Gbakie M, Kanneh L, Kargbo D, Massally JLB, 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. DOI: 10.1016/j.cell.2015.06.007.

Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, Liu M, Kumar S, Zaremba S, Gu Z, Zhou L, Larson CN, Dietrich J, Klem EB, Scheuermann RH. 2012. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic acids research* 40:D593–8. DOI: 10.1093/nar/gkr859.

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 466:334–338. DOI: 10.1038/nature09199.

Sharma D, Priyadarshini P, Vrati S. 2015. Unraveling the web of viroinformatics: computational tools and databases in virus research. *Journal of virology* 89:1489–1501. DOI: 10.1128/JVI.02027-14.

Siddle KJ, Eromon P, Barnes KG, Mehta S, Oguzie JU, Odia I, Schaffner SF, Winnicki SM, Shah RR, Qu J, Wohl S, Brehio P, Iruolagbe C, Aiyepada J, Uyigue E, Akhilomen P, Okonofua G, Ye S, Kayode T, Ajogbasile F, Uwanibe J, Gaye A, Momoh M, Chak B, Kotliar D, Carter A, Gladden-Young A, Freije CA, Omoregie O, Osiemi B, Muoebonam EB, Airende M, Enigbe R, Ebo B, Nosamiefan I, Oluniyi P, Nekoui M, Ogbaini-Emovon E, Garry RF, Andersen KG, Park DJ, Yozwiak NL, Akpede G, Ihekweazu C, Tomori O, Okogbenin S, Folarin OA, Okokhere PO, MacInnis BL, Sabeti PC, Happi CT. 2018. Genomic Analysis of Lassa Virus during an

Increase in Cases in Nigeria in 2018. *The New England journal of medicine* 379:1745–1753. DOI: 10.1056/NEJMoa1804498.

Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, Iosifidis C, Agha R. 2020. World Health Organization declares Global Emergency: A review of the 2019 Novel Coronavirus (COVID-19). *International journal of surgery* . DOI: 10.1016/j.ijsu.2020.02.034.

Tang P, Chiu C. 2010. Metagenomics for the discovery of novel human viruses. *Future microbiology* 5:177–189. DOI: 10.2217/fmb.09.120.

Wymant C, Blanquart F, Golubchik T, Gall A, Bakker M, Bezemer D, Croucher NJ, Hall M, Hillebregt M, Ong SH, Ratmann O, Albert J, Bannert N, Fellay J, Fransen K, Gourlay A, Grabowski MK, Gunsenheimer-Bartmeyer B, Günthard HF, Kivelä P, Kouyos R, Laeyendecker O, Liitsola K, Meyer L, Porter K, Ristola M, van Sighem A, Berkhout B, Cornelissen M, Kellam P, Reiss P, Fraser C, BEEHIVE Collaboration. 2018. Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus evolution* 4:vey007. DOI: 10.1093/ve/vey007.

Table 1(on next page)

Comparison between Lassa virus whole genomes (S segment) obtained with VGEA and viral-ngs

Table 1: Comparison between Lassa virus whole genomes (S segment) obtained with **viral-ngs** and **VGEA**

S/N	Sample ID	Lassa Virus Genome Segment	Genome length with viral-ngs	Genome length with VGEA
1.	758	S	3392	3413
2.	852	S	3391	3413
3.	934	S	3394	3413
4.	1004	S	3392	3413
5.	1078	S	3382	3414
6.	1126	S	3394	3413
7.	A4	S	3360	3413
8.	A7	S	3380	3413
9.	J1	S	3374	3413
10.	K7	S	3389	3412
11.	0202C	S	3148	3413
12.	1177	S	3389	3413
13.	1801	S	3379	3412
14.	1880	S	3386	3413
15.	540	S	3294	3406
16.	D2	S	3394	3413
17.	F8	S	3391	3413
18.	J4	S	3386	3413
19.	O1	S	3393	3413

20.	O2	S	3381	3413

Table 2(on next page)

Comparison between Lassa virus whole genomes (L segment) obtained with VGEA and viral-ngs

1 **Table 2:** Comparison between Lassa virus whole genomes (L segment) obtained with **viral-ngs**
2 and **VGEA**
3

S/N	Sample ID	Lassa Virus Genome Segment	Genome length with viral-ngs	Genome length with VGEA
1.	758	L	7225	7271
2.	852	L	7239	7268
3.	934	L	7245	7272
4.	1004	L	7235	7269
5.	1078	L	7225	7272
6.	1126	L	7234	7270
7.	A4	L	7150	7271
8.	A7	L	7214	7273
9.	J1	L	7154	7271
10.	K7	L	7187	7272
11.	0202C	L	6418	7272
12.	1177	L	7121	7271
13.	1801	L	7221	7269
14.	1880	L	7195	7272
15.	540	L	7020	7225
16.	D2	L	7237	7271
17.	F8	L	7230	7272
18.	J4	L	7224	7273

19.	O1	L	7246	7273
20.	O2	L	7224	7275

4

Table 3(on next page)

Comparison between SARS-CoV-2 whole genomes obtained with VGEA and viral-ngs

Table 3: Comparison between SARS-CoV-2 whole genomes obtained with **viral-ngs** and **VGEA**

S/N	Sample ID	Genome length with viral-ngs	Genome length with VGEA
1.	CV18	29858	29903
2.	CV29	29865	29986
3.	CV43	29875	29903
4.	CV45	29897	30002
5.	CV47	29900	29903
6.	CV48	29859	30145
7.	CV50	29897	29903
8.	CV55	29898	30636
9.	CV57	29895	30129
10.	CV115	29859	30180
11.	CV145	28898	30250
12.	CV153	29848	29903
13.	CV155	29870	29903
14.	CV156	29856	29903
15.	CV163	29864	30356
16.	CV165	29872	29958
17.	CV167	29894	29900
18.	CV170	29897	29903
19.	CV185	29870	29903
20.	CV192	29897	29903

Table 4(on next page)

Comparison between HIV-1 whole genomes obtained with VGEA and viral-ngs

Table 4: Comparison between HIV-1 whole genomes obtained with **viral-ngs** and **VGEA**

S/N	SRA Accession Number	Genome length with viral-ngs	Genome length with VGEA
1.	ERR3953696	8927	9877
2.	ERR3953853	9054	9872
3.	ERR3953893	9083	9880
4.	ERR3953891	7827	9818
5.	ERR3953866	9088	9872
6.	ERR3953846	9148	9917
7.	ERR3953756	8963	9913
8.	ERR3953877	9035	9802
9.	ERR3953876	7317	9824
10.	ERR3953750	5532	9860
11.	ERR3953741	7376	9822
12.	ERR3953697	6728	9798
13.	ERR3953699	9054	9837
14.	ERR3953706	7317	9826
15.	ERR3953708	7393	9839
16.	ERR3953710	6705	9808
17.	ERR3953712	9108	9846
18.	ERR3953716	9134	9835
19.	ERR3953295	8676	9882
20.	ERR3953693	8885	9857