# VGEA: An RNA viral assembly toolkit

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 [2], 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

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, bringing about the need to implement advanced bioinformatics tools to assemble consensus genomes well-representative of the viral population circulating in individual patients. Many tools have been developed to preprocess sequencing reads, carry-out *de novo* or reference-assisted assembly of viral genomes and assess the quality of the genomes obtained. Most of these tools however exist as standalone workflows and usually require huge computational resources. Here we present **VGEA** (**V**iral **G**enomes **E**asily **A**nalyzed), a Snakemake workflow for analyzing RNA viral genomes. VGEA enables users to map sequencing reads to the human genome to remove human contaminants, 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, map reads to a reference tailored to the sample using corrected contigs supplemented by the user's choice of reference sequences and evaluate/compare genome assemblies. We designed a project with the aim of creating a flexible, easy-to-use and all-in-one pipeline from existing/stand-alone bioinformatics tools for viral genome analysis that can be deployed on a personal computer. VGEA was built on the Snakemake workflow management system and utilizes existing tools for each step: **fastp** (*Chen et al., 2018)* for read trimming and read-level quality control, **BWA** (*Li and Durbin, 2009*) for

mapping sequencing reads to the human reference genome, **SAMtools (***Li et al., 2009)* for extracting unmapped reads and also for splitting bam files into fastq files, **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, **SeqKit** (*Shen et al., 2016)* for cleaning shiver assembly for QUAST, **QUAST** (*Gurevich et al., 2013*) to evaluate/assess the quality of genome assemblies and **MultiQC** (*Ewels et al., 2016)* for aggregation of the results from fastp, BWA and QUAST. Our pipeline was successfully tested and validated with SARS-CoV-2 (n = 20), HIV-1 (n = 20) and Lassa Virus (n = 20) datasets all of which have been made publicly available. VGEA is freely available on GitHub at: https://github.com/pauloluniyi/VGEA under the GNU General Public License.

1    **VGEA: An RNA viral assembly toolkit**

3    Paul E. Oluniyi[1,2], Fehintola V. Ajogbasile[1,2], Judith U. Oguzie[1,2], Jessica N. Uwanibe[1,2],

4    Adeyemi T. Kayode[1,2], Anise N. Happi[2], Chinedu A. Ugwu[1,2], Testimony J. Olumade[1,2], Olusola

5    Ogunsanya[3], Philomena E. Eromon[2], Onikepe A. Folarin[1,2], Simon D.W. Frost[4,5], Jonathan L.

6    Heeney[6], Christian T. Happi[1,2]*

8    1. Department of Biological Sciences, Faculty of Natural Sciences, Redeemer's University, Ede,

9    Osun State, Nigeria.

10    2. African Centre of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's

11    University, Ede, Osun State, Nigeria.

12    3. Department of Veterinary Pathology, Faculty of Veterinary Medicine, University of Ibadan,

13    Ibadan, Nigeria.

14    4. Microsoft Research, Redmond, 98052, WA, USA.

15    5. London School of Hygiene & Tropical Medicine, London, UK.

16    6. Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom.

19    *Corresponding Author:

20    Christian Happi[1,2]

21    Email address: happic@run.edu.ng

24   **ABSTRACT**

25   Next generation sequencing (NGS)-based studies have vastly increased our understanding of

26   viral diversity. Viral sequence data obtained from NGS experiments are a rich source of

27   information, these data can be used to study their epidemiology, evolution, transmission patterns,

28   and can also inform drug and vaccine design. Viral genomes however represent a great challenge

29   to bioinformatics due to their high mutation rate and forming quasispecies in the same infected

30   host, bringing about the need to implement advanced bioinformatics tools to assemble consensus

31   genomes well-representative of the viral population circulating in individual patients. Many tools

32   have been developed to preprocess sequencing reads, carry-out *de novo* or reference-assisted

33   assembly of viral genomes and assess the quality of the genomes obtained. Most of these tools

34   however exist as standalone workflows and usually require huge computational resources.  Here

35   we present **VGEA** (**V**iral **G**enomes **E**asily **A**nalyzed), a Snakemake workflow for analyzing

36   RNA viral genomes. VGEA enables users to map sequencing reads to the human genome to

37   remove human contaminants, split bam files into forward and reverse reads, carry out *de novo*

38   assembly of forward and reverse reads to generate contigs, pre-process reads for quality and

39   contamination, map reads to a reference tailored to the sample using corrected contigs

40   supplemented by the user's choice of reference sequences and evaluate/compare genome

41   assemblies. We designed a project with the aim of creating a flexible, easy-to-use and all-in-one

42   pipeline from existing/stand-alone bioinformatics tools for viral genome analysis that can be

43   deployed on a personal computer. VGEA was built on the Snakemake workflow management

44   system and utilizes existing tools for each step: **fastp** (*Chen et al., 2018*) for read trimming and

45   read-level quality control,  **BWA** (*Li and Durbin, 2009*) for mapping sequencing reads to the

46   human reference genome, **SAMtools (***Li et al., 2009)* for extracting unmapped reads and also for

47    splitting bam files into fastq files, **IVA (***Hunt et al., 2015)* for *de novo* assembly to generate

48    contigs, **shiver** (*Wymant et al., 2018*) to pre-process reads for quality and contamination, then

49    map to a reference tailored to the sample using corrected contigs supplemented with the user's

50    choice of existing reference sequences, **SeqKit** (*Shen et al., 2016)* for cleaning shiver assembly

51    for QUAST,  **QUAST** (*Gurevich et al., 2013)* to evaluate/assess the quality of genome

52    assemblies and **MultiQC** (*Ewels et al., 2016)* for aggregation of the results from fastp, BWA and

53    QUAST. Our pipeline was successfully tested and validated with SARS-CoV-2 (n = 20), HIV-1

54    (n = 20) and Lassa Virus (n = 20) datasets all of which have been made publicly available.

55    VGEA is freely available on GitHub at: https://github.com/pauloluniyi/VGEA under the GNU

56    General Public License.

57

58    **INTRODUCTION**

59    The most abundant biological entities on Earth are viruses as they can be found among all

60    cellular forms of life. So far, over four thousand five hundred viral species have been discovered,

61    from which a huge amount of sequence information has been collected by researchers and

62    scientists all over the world (*Pickett et al., 2012; Sharma et al., 2015; Brister et al., 2015*). In

63    recent times (past two decades), a number of these viruses have emerged in the human

64    population causing disease outbreaks and sometimes pandemics. These viruses include mainly:

65    Influenza virus, Severe Acute Respiratory Syndrome (SARS) coronavirus, Middle East

66    Respiratory Syndrome (MERS) coronavirus, Ebola virus, Yellow fever virus, Lassa virus

67    (LASV), Zika virus (*Chan, 2002; Bean et al., 2013; Folarin et al., 2016; Grubaugh et al., 2017;*

68    *Metsky et al., 2017; Siddle et al., 2018; Ajogbasile et al., 2020*) and  SARS-CoV-2 (*Chen et al.,*

69    *2020; Holshue et al., 2020; Sohrabi et al., 2020*). During these outbreaks and pandemics,

70    genomic sequencing for identification and characterization of the transmission and evolution of

71    the causative agents have proved to be critical in helping inform disease surveillance and

72    epidemiology.

73    Next Generation Sequencing (NGS) platforms have been widely accepted as high-throughput,

74    open view technologies that have many attractive features for virus detection and assembly

75    (*Tang & Chiu, 2010; Mokili et al., 2012*). NGS-based studies have vastly increased our

76    understanding of viral diversity (*Reyes et al., 2010; Cantalupo et al., 2011*). Pathogen sequence

77    data obtained from NGS experiments are a rich source of information, these data can be used to

78    study their epidemiology, evolution, transmission patterns, and can also inform drug and vaccine

79    design. The field of genomics, especially pathogen genomics has been transformed by NGS, with

80    costs constantly decreasing, equipment becoming more portable/field deployable during

81    outbreaks and remarkable increase in data availability.

82    The huge amount of data being generated requires various processing steps such as removal of

83    primers and adapters, quality filtering and control which is usually crucial for various

84    downstream analysis. Several tools have been developed for these purposes, such as fastp (*Chen*

85    *et al., 2018*) and Trimmomatic (*Bolger et al., 2014)*.

86    Reconstructing viral genomes from NGS data is usually achieved through *de novo* assembly

87    (which is the process of assembling genomes using overlapping sequencing reads), or through a

88    reference-guided approach (which involves mapping sequence reads to a reference genome).

89    Numerous tools have been developed for these purposes; SPAdes (*Bankevich et al., 2012*),

90    Burrows-Wheeler Alignment tool (BWA), V-GAP (*Nakamura, 2016)*, VirusTAP (*Yamashita,*

91    *2016*), V-Pipe (Posada-Céspedes, 2021) and viral-ngs (https://github.com/broadinstitute/viral-

92   ngs), amongst others. Contigs generated by *de novo* assembly however do not provide a

93   complete summary of reads, misassembly can result in the contigs having an incorrect structure,

94   and for parts of the genome where contigs could not be assembled, no information is available.

95   In addition, reference-guided assembly of viral genomes can lead to biased loss of information

96   which can then skew epidemiological and evolutionary conclusions (*Wymant et al., 2018)*.

97   Variant analysis and genome quality assessment to detect variants and changes occurring across

98   the genome of a virus is also a key step in viral genome analysis as viruses (especially RNA

99   viruses) are known to have high mutation rates (*Duffy, 2018)*. Variant analysis is important for

100  detecting outbreak origins and for phylogenetic/phylogeographic studies and best practices for

101  variant identification in microbial genomes have been proposed in literature and adopted to a

102  large extent (*Van der Auwera et al., 2013)*.

103  A number of pipelines that have been developed for downstream analysis of viral genomes

104  require high performance computing (HPC) clusters and/or cloud-based systems e.g. the V-pipe

105  authors recommend running V-pipe on clusters because for most applications, running V-pipe on

106  a local machine may not be efficient (https://github.com/cbg-ethz/V-pipe/wiki/advanced) and

107  some of these pipelines are only web-based such as VirAmp (*Wan et al., 2015*) and VirusTAP

108  (*Yamashita, 2016)*. Also, some pipelines have many dependencies to be installed especially if the

109  analysis requires multiple tasks to be performed. In low-and-middle income countries (LMICs)

110  where most scientists do not have access to HPC clusters or cloud-based systems and where

111  internet connection is too unstable to regularly make use of web-based platforms for analysis,

112  this can be a daunting task.

113  The challenges listed above motivated the development of VGEA (Viral Genomes Easily

114  Analyzed, available online at https://github.com/pauloluniyi/VGEA). VGEA makes use of

115  existing bioinformatics pipeline/tools to carry out various viral genome analysis tasks and is built

116  on an advanced workflow management system, **Snakemake** (*Köster & Rahmann, 2012)*.

117  **MATERIALS AND METHODS**

118  **Datasets**

119  We successfully tested and validated VGEA with SARS-CoV-2 (n = 20) and Lassa Virus (n =

120  20) datasets sequenced on the illumina MiSeq and illumina FGx sequencing machines in our

121  laboratory at the African Centre of Excellence for Genomics of Infectious Diseases (ACEGID),

122  Redeemer's University, Ede, Nigeria. Briefly, samples were inactivated in buffer AVL and viral

123  RNA was extracted according to the QiAmp viral RNA mini kit (Qiagen) manufacturer's

124  instructions. Extracted RNA was treated with Turbo DNase to remove contaminating DNA,

125  followed by cDNA synthesis with random hexamers. Sequencing libraries were prepared using

126  the Nextera XT kit (Illumina) as previously described (*Matranga et al., 2016*) and sequenced on

127  the Illumina Miseq platform with 101 base pair paired-end reads. We also tested and validated

128  VGEA with HIV-1 datasets sequenced on the illumina HiSeq 2500 obtained from NCBI

129  Sequence Read Archive (SRA). We made use of 60 test datasets [Lassa Virus (20), SARS-CoV-

130  2 (20) and HIV-1 (20)] for the validation of the VGEA pipeline. All our test datasets are

131  available on figshare (https://doi.org/10.6084/m9.figshare.13009997).

132

133

134   **Implementation**

135   The installation of VGEA requires the pipeline to be downloaded onto a personal computer and

136   creation of a conda environment to set up all dependencies. Complete installation steps are in the

137   github README file:

138   https://github.com/pauloluniyi/VGEA/blob/master/README.md

139   The analysis of VGEA is broken down into a set of 'rules' that links the output file of an analysis

140   into the input of the next task in the general workflow (Figure 1). The dependencies are **fastp** for

141   read trimming and read-level quality control, **BWA** for mapping sequencing reads to the human

142   reference genome, **SAMtools**  for extracting unmapped reads and also for splitting bam files into

143   fastq files, **IVA**  for *de novo* assembly to generate contigs, **shiver** to pre-process reads for quality

144   and contamination, then map to a reference tailored to the sample using corrected contigs

145   supplemented with the user's choice of existing reference sequences, **SeqKit** for cleaning shiver

146   assembly for QUAST, **QUAST** to evaluate/assess the quality of genome assemblies and

147   **MultiQC** for aggregation of the results from fastp, BWA and QUAST

148
149   All of these tools can be installed using a bioconda channel (*Grüning et al., 2018)*. The input

150   files for VGEA are paired-end fastq files**.** VGEA allows full customization of the pipeline, so

151   users can modify the parameters used in running their samples. It is possible to modify every

152   step of the workflow to suit the samples being processed. Users can also add more steps to the

153   pipeline as they see fit. The pipeline runs on Linux/Unix and Mac. No prior programming is

154   required however to run the pipeline and once the user supplies the input, the whole workflow

155   can run automatically from beginning to end.

156

157 **RESULTS**

158

159 VGEA carries out read trimming and quality control tasks on input FASTQ data using fastp

160 (Figure 2). This increases the quality of data used for subsequent steps of the pipeline.

161 VGEA then maps reads to the human reference genome in order to remove human contaminants,

162 the pipeline carries out this step using BWA. Genome assembly and consensus sequence

163 generation is carried out, together with the generation of summary minority-variant information

164 (base frequencies at each position) and detailed minority-variant information (all reads aligned to

165 their correct position in the genome). VGEA carries out assembly using **IVA** and generates

166 consensus sequences using **shiver**. Previous study by the shiver developers has shown the

167 systematic superiority of mapping to shiver's constructed reference compared with mapping the

168 same reads to the closest of 3,249 references: median values of 13 bases called differently and

169 more accurately, zero bases called differently and less accurately, and 205 bases of missing

170 sequence recovered (*Wymant et al.,* 2018).

171

172 VGEA also assesses the quality of genome assemblies using QUAST. QUAST evaluates metrics

173 such as contig sizes, misassemblies and structural variations, genome representation and its

174 functional elements, variations of N50 based on aligned blocks and then presents these statistics

175 in graphical form. QUAST also makes a histogram of several metrics including the number of

176 complete genes, operons and the genome fraction (%). Finally, VGEA compiles the results of

177 BWA, fastp and QUAST into a single MultiQC report (Figure 3).

178

179

180 **Performance Evaluation**

181

182 VGEA makes use of Snakemake's benchmarking feature which allows the measurement of the

183 CPU usage and wall clock time of each rule in the pipeline. This allows the user to know which

184 step of the pipeline requires the least and highest amount of computational resources. Knowledge

185 of this can help the user decide on the number of threads to dedicate to each rule as VGEA also

186 makes use of Snakemake's multi-threading feature. Table 1 shows the benchmarking values for a

187 sample SARS-CoV-2 dataset analyzed using VGEA.

188

189 We compared the contigs generated by VGEA's assembly step with contigs generated using two

190 other standalone and commonly used assembly pipelines, SPAdes (*Bankevich et al., 2012*) and

191 Velvet (*Zerbino and Birney, 2008*). We compared against these two pipelines because most

192 commonly used assembly workflows like viral-ngs and VirAmp are built on them. We carried

193 out this comparison by making use of five different SARS-CoV-2 test datasets (namely CV18,

194 CV29, CV45, CV115 and CV145 datasets available on FigShare and NCBI). We compared the

195 assemblies to the SARS-CoV-2 reference genome, and N50/NG50, mis-assembly, mismatches

196 and indel scores were used to evaluate the performance of each assembly method as

197 recommended by Assemblathon 2 (*Bradnam et al., 2013*) (Table 2). Basic statistics were

198 calculated using QUAST. All results of our performance evaluation and comparison are provided

199 as Supplementary File 2. All analyses were run on a 64-bit personal computer with 16GB RAM

200 using four threads. SPAdes version 3.15.2 and Velvet version 1.2.10 were used for the

201 comparison purposes using the default parameters.

202

203 Evaluation statistics showed that contigs generated by VGEA had the highest NG50 score for

204 four of the five datasets and the highest N50 scores across all five datasets. In all five datasets,

205 VGEA's contigs had the highest genome fraction covering greater than 95% in four.

206 Comparison of maximum RAM used by VGEA, SPAdes and Velvet showed that VGEA used

207 the least amount of RAM for the analyses of all five datasets used for comparison. SPAdes and

208 Velvet however ran faster than VGEA for all analyses.

209

210 **DISCUSSION**

211

212 VGEA is built on the snakemake workflow management system (*Köster & Rahmann, 2012*), a

213 workflow management system that allows the effortless deployment and execution of complex

214 distributed computational workflows in any UNIX-based system, from local machines to high-

215 performance computing clusters. It is a user-friendly, customizable and reproducible pipeline

216 which can be deployed on a personal computer and which can run from start to finish with a

217 single command.

218

219 VGEA was designed with ease-of-use in mind and so all its dependencies can be installed in a

220 conda environment under the bioconda channel (*Grüning et al., 2018)* making it particularly

221 useful for scientists with little or no computational background and for scientists in LMICs who

222 don't have much access to high-performance computing clusters or cloud-computing resources.

223 VGEA capitalizes on Snakemake's multi-threading feature so that makes it possible for it to be

224 deployed on laptops with greater computing performance or a computing server to improve its

225 speed. The pipeline was tested with paired-end short-read sequencing data produced by the

226 illumina platform (MiSeq, MiSeq FGx and HiSeq 2500).

227

228 The results generated by the major steps of the VGEA pipeline are summed up together into a

229 MultiQC report which can be easily interpreted and understood by anyone with little or no

230 knowledge of bioinformatics.

231

232

233 **CONCLUSION**

234 VGEA was built primarily by biologists and in a manner that is easy to be employed by users

235 without significant computational background. As new and innovative tools for viral genome

236 analysis and assembly are increasingly being developed, these can easily be incorporated into the

237 VGEA pipeline. We hope that other scientists can build upon and improve VGEA as a tool to

238 extract more qualitative and quantitative information from viral genomes.

239

240

241 **ACKNOWLEDGEMENTS**

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

243 University. We especially appreciate Dr. Finlay Maguire and Dr. Gerry Tonkin-Hill for helpful

244 discussions and for making necessary changes to the pipeline. Also, thanks to Dr. Andreas Wilm

245 and Christopher Tomkins-Tinch for helpful comments and suggestions.

246

247

248 **Abbreviations**

249 **VGEA** Viral Genomes Easily Assembled

250 **NGS** Next generation sequencing

251 **RNA** Ribonucleic acid

252 **SARS** Severe Acute Respiratory Syndrome

253 **MERS** Middle East Respiratory Syndrome

254 **IVA** Iterative Virus Assembler

255 **SHIVER** Sequences from HIV Easily Reconstructed

256 **HPC** High Performance Computing

257 **Data Availability**

258 The following information was supplied regarding data availability:

259     VGEA is freely available on GitHub at: https://github.com/pauloluniyi/VGEA under the

260     GNU General Public License.

261

262 All primary test datasets used for the validation of the VGEA pipeline are available on figshare

263 (https://doi.org/10.6084/m9.figshare.13009997). All SARS-CoV-2 and Lassa virus test datasets

264 have been submitted to NCBI SRA (BioProject accession numbers, PRJNA666685 and

265 PRJNA666664). All HIV-1 test datasets are available on NCBI SRA (accession numbers:

266 ERR3953696, ERR3953853, ERR3953893, ERR3953891, ERR3953866, ERR3953846,

267 ERR3953756, ERR3953877, ERR3953876, ERR3953750, ERR3953741, ERR3953697,

268 ERR3953699, ERR3953706, ERR3953708, ERR3953710, ERR3953712, ERR3953716,

269 ERR3953295, ERR3953693).

270

271

272

273    **REFERENCES**

274

275    **Ajogbasile FV, Oguzie JU, Oluniyi PE, Eromon PE, Uwanibe JN, Mehta SB, Siddle KJ,**

276         **Odia I, Winnicki SM, Akpede N, Akpede G, Okogbenin S, Ogbaini-Emovon E,**

277         **MacInnis BL, Folarin OA, Modjarrad K, Schaffner SF, Tomori O, Ihekweazu C,**

278         **Sabeti PC, Happi CT. 2020.** Real-time Metagenomic Analysis of Undiagnosed Fever

279         Cases Unveils a Yellow Fever Outbreak in Edo State, Nigeria. *Scientific reports* 10:3180.

280         DOI: 10.1038/s41598-020-59880-w.

281

282    **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990.** Basic local alignment search

283         tool. *Journal of molecular biology* 215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.

284

285    **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,**

286         **Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler**

287         **G, Alekseyev MA, Pevzner PA. 2012.** SPAdes: a new genome assembly algorithm and

288         its applications to single-cell sequencing. *Journal of computational biology: a journal of*

289         *computational molecular cell biology* 19:455–477. DOI: 10.1089/cmb.2012.0021.

290

291    **Bean AGD, Baker ML, Stewart CR, Cowled C, Deffrasnes C, Wang L-F, Lowenthal JW.**

292    **2013.** Studying immunity to zoonotic diseases in the natural host - keeping it real. *Nature*

293    *reviews. Immunology* 13:851–861. DOI: 10.1038/nri3551.

294

295    **Bolger AM, Lohse M, Usadel B. 2014.** Trimmomatic: a flexible trimmer for Illumina sequence

296    data. *Bioinformatics* 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.

297

298    **Brister JR, Ako-Adjei D, Bao Y, Blinkova O. 2015.** NCBI viral genomes resource. *Nucleic*

299    *acids research* 43:D571–7. DOI: 10.1093/nar/gku1207.

300

301    **Cantalupo PG, Calgua B, Zhao G, Hundesa A, Wier AD, Katz JP, Grabe M, Hendrix RW,**

302    **Girones R, Wang D, Pipas JM. 2011.** Raw sewage harbors diverse viral populations.

303    *mBio* 2. DOI: 10.1128/mBio.00180-11.

304

305    **Chan PKS. 2002.** Outbreak of avian influenza A(H5N1) virus infection in Hong Kong in 1997.

306    *Clinical infectious diseases: an official publication of the Infectious Diseases Society of*

307    *America* 34 Suppl 2:S58–64. DOI: 10.1086/338820.

308

309    **Chen S, Zhou Y, Chen Y, Gu J. 2018.** fastp: an ultra-fast all-in-one FASTQ preprocessor.

310    *Bioinformatics* 34:i884–i890. DOI: 10.1093/bioinformatics/bty560.

311

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

313     **Yu T, Zhang X, Zhang L. 2020.** Epidemiological and clinical characteristics of 99 cases

314     of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*

315     395:507–513. DOI: 10.1016/S0140-6736(20)30211-7.

316

317   **Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T,**

318     **Kauff F, Wilczynski B, de Hoon MJL. 2009.** Biopython: freely available Python tools

319     for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.

320     DOI: 10.1093/bioinformatics/btp163.

321

322   **Duffy S**. **2018**. Why are RNA virus mutation rates so damn high?. *PLoS biology*, *16*(8),

323     e3000003. https://doi.org/10.1371/journal.pbio.3000003.

324

325   **Ewels P, Magnusson M, Lundin S, Käller M. 2016.** MultiQC: summarize analysis results for

326     multiple tools and samples in a single report. *Bioinformatics, 32*(19):3047-3048.

327     DOI:10.1093/bioinformatics/btw354.

328

329   **Folarin OA, Ehichioya D, Schaffner SF, Winnicki SM, Wohl S, Eromon P, West KL,**

330     **Gladden-Young A, Oyejide NE, Matranga CB, Deme AB, James A, Tomkins-Tinch**

331     **C, Onyewurunwa K, Ladner JT, Palacios G, Nosamiefan I, Andersen KG, Omilabu**

332     **S, Park DJ, Yozwiak NL, Nasidi A, Garry RF, Tomori O, Sabeti PC, Happi CT.**

333     **2016.** Ebola Virus Epidemiology and Evolution in Nigeria. *The Journal of infectious*

334     *diseases* 214:S102–S109. DOI: 10.1093/infdis/jiw190.

335

336 **Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K, Wiley**

337       **MR, White S, Thézé J, Magnani DM, Prieto K, Reyes D, Bingham AM, Paul LM,**

338       **Robles-Sikisaka R, Oliveira G, Pronty D, Barcellona CM, Metsky HC, Baniecki**

339       **ML, Barnes KG, Chak B, Freije CA, Gladden-Young A, Gnirke A, Luo C,**

340       **MacInnis B, Matranga CB, Park DJ, Qu J, Schaffner SF, Tomkins-Tinch C, West**

341       **KL, Winnicki SM, Wohl S, Yozwiak NL, Quick J, Fauver JR, Khan K, Brent SE,**

342       **Reiner RC Jr, Lichtenberger PN, Ricciardi MJ, Bailey VK, Watkins DI, Cone MR,**

343       **Kopp EW 4th, Hogan KN, Cannons AC, Jean R, Monaghan AJ, Garry RF, Loman**

344       **NJ, Faria NR, Porcelli MC, Vasquez C, Nagle ER, Cummings DAT, Stanek D,**

345       **Rambaut A, Sanchez-Lockhart M, Sabeti PC, Gillis LD, Michael SF, Bedford T,**

346       **Pybus OG, Isern S, Palacios G, Andersen KG. 2017.** Genomic epidemiology reveals

347       multiple introductions of Zika virus into the United States. *Nature* 546:401–405. DOI:

348       10.1038/nature22400.

349

350 **Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H.,**

351       **Valieris, R., Köster, J., & Bioconda Team. 2018**. Bioconda: sustainable and

352       comprehensive software distribution for the life sciences. *Nature methods*, *15*(7),

353       475–476. https://doi.org/10.1038/s41592-018-0046-7.

354

355 **Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. 2013**. QUAST: quality assessment

356       tool for genome assemblies. *Bioinformatics (Oxford, England)*, *29*(8), 1072–1075.

357       https://doi.org/10.1093/bioinformatics/btt086.

358

359 **Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Spitters C, Ericson**

360 **K, Wilkerson S, Tural A, Diaz G, Cohn A, Fox L, Patel A, Gerber SI, Kim L, Tong**

361 **S, Lu X, Lindstrom S, Pallansch MA, Weldon WC, Biggs HM, Uyeki TM, Pillai SK,**

362 **Washington State 2019-nCoV Case Investigation Team. 2020.** First Case of 2019

363 Novel Coronavirus in the United States. *The New England journal of medicine*. DOI:

364 10.1056/NEJMoa2001191.

365

366 **Hunt M, Gall A, Ong SH, Brener J, Ferns B, Goulder P, Nastouli E, Keane JA, Kellam P,**

367 **Otto TD. 2015.** IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics*

368 31:2374–2376. DOI: 10.1093/bioinformatics/btv120.

369

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

371 sequence alignment based on fast Fourier transform. *Nucleic acids research* 30:3059–

372 3066. DOI: 10.1093/nar/gkf436.

373

374 **Kokot M, Dlugosz M, Deorowicz S. 2017.** KMC 3: counting and manipulating k-mer

375 statistics. *Bioinformatics* 33:2759–2761. DOI: 10.1093/bioinformatics/btx304.

376

377 **Köster J, Rahmann S. 2012.** Snakemake--a scalable bioinformatics workflow engine.

378 *Bioinformatics* 28:2520–2522. DOI: 10.1093/bioinformatics/bts480.

379

380    **Langmead B. 2010.** Aligning short sequencing reads with Bowtie. *Current protocols in*

381        *bioinformatics / editoral board, Andreas D. Baxevanis... [et al.]* Chapter 11:Unit 11.7.

382        DOI: 10.1002/0471250953.bi1107s32.

383

384    **Li H, Durbin R. 2009.** Fast and accurate short read alignment with Burrows-Wheeler

385        transform. *Bioinformatics* 25:1754–1760. DOI: 10.1093/bioinformatics/btp324.

386

387    **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,**

388        **Durbin R, 1000 Genome Project Data Processing Subgroup. 2009.** The Sequence

389        Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. DOI:

390        10.1093/bioinformatics/btp352.

391

392    **Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018.** MUMmer4:

393        A fast and versatile genome alignment system. *PLoS computational biology*

394        14:e1005944. DOI: 10.1371/journal.pcbi.1005944.

395

396    **Matranga, C. B., Gladden-Young, A., Qu, J., Winnicki, S., Nosamiefan, D., Levin, J.**

397        **Z., & Sabeti, P. C. 2016**. Unbiased Deep Sequencing of RNA Viruses from

398        Clinical Samples. *Journal of visualized experiments : JoVE*, (113), 54117.

399        https://doi.org/10.3791/54117.

400

401    **Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, West K, Qu J,**

402       **Baniecki ML, Gladden-Young A, Lin AE, Tomkins-Tinch CH, Ye SH, Park DJ, Luo**

403       **CY, Barnes KG, Shah RR, Chak B, Barbosa-Lima G, Delatorre E, Vieira YR, Paul**

404       **LM, Tan AL, Barcellona CM, Porcelli MC, Vasquez C, Cannons AC, Cone MR,**

405       **Hogan KN, Kopp EW, Anzinger JJ, Garcia KF, Parham LA, Ramírez RMG,**

406       **Montoya MCM, Rojas DP, Brown CM, Hennigan S, Sabina B, Scotland S,**

407       **Gangavarapu K, Grubaugh ND, Oliveira G, Robles-Sikisaka R, Rambaut A,**

408       **Gehrke L, Smole S, Halloran ME, Villar L, Mattar S, Lorenzana I, Cerbino-Neto J,**

409       **Valim C, Degrave W, Bozza PT, Gnirke A, Andersen KG, Isern S, Michael SF,**

410       **Bozza FA, Souza TML, Bosch I, Yozwiak NL, MacInnis BL, Sabeti PC. 2017.** Zika

411       virus evolution and spread in the Americas. *Nature* 546:411–415. DOI:

412       10.1038/nature22402.

413

414     **Mokili JL, Rohwer F, Dutilh BE. 2012.** Metagenomics and future perspectives in virus

415       discovery. *Current opinion in virology* 2:63–77. DOI: 10.1016/j.coviro.2011.12.004.

416

417     **Nakamura, Y., Yasuike, M., Nishiki, I., Iwasaki, Y., Fujiwara, A., Kawato, Y., Nakai,**

418       **T., Nagai, S., Kobayashi, T., Gojobori, T., & Ototake, M. 2016**. V-GAP: Viral

419       genome assembly pipeline. *Gene*, *576*(2 Pt 1), 676–680.

420       https://doi.org/10.1016/j.gene.2015.10.029.

421

422     **Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG, Sealfon RS, Ladner**

423    **JT, Kugelman JR, Matranga CB, Winnicki SM, Qu J, Gire SK, Gladden-Young A,**

424    **Jalloh S, Nosamiefan D, Yozwiak NL, Moses LM, Jiang P-P, Lin AE, Schaffner SF,**

425    **Bird B, Towner J, Mamoh M, Gbakie M, Kanneh L, Kargbo D, Massally JLB,**

426    **Kamara FK, Konuwa E, Sellu J, Jalloh AA, Mustapha I, Foday M, Yillah M,**

427    **Erickson BR, Sealy T, Blau D, Paddock C, Brault A, Amman B, Basile J, Bearden S,**

428    **Belser J, Bergeron E, Campbell S, Chakrabarti A, Dodd K, Flint M, Gibbons A,**

429    **Goodman C, Klena J, McMullan L, Morgan L, Russell B, Salzer J, Sanchez A,**

430    **Wang D, Jungreis I, Tomkins-Tinch C, Kislyuk A, Lin MF, Chapman S, MacInnis**

431    **B, Matthews A, Bochicchio J, Hensley LE, Kuhn JH, Nusbaum C, Schieffelin JS,**

432    **Birren BW, Forget M, Nichol ST, Palacios GF, Ndiaye D, Happi C, Gevao SM,**

433    **Vandi MA, Kargbo B, Holmes EC, Bedford T, Gnirke A, Ströher U, Rambaut A,**

434    **Garry RF, Sabeti PC. 2015.** Ebola Virus Epidemiology, Transmission, and Evolution

435    during Seven Months in Sierra Leone. *Cell* 161:1516–1526. DOI:

436    10.1016/j.cell.2015.06.007.

437

438    **Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, Liu M, Kumar S,**

439    **Zaremba S, Gu Z, Zhou L, Larson CN, Dietrich J, Klem EB, Scheuermann RH.**

440    **2012.** ViPR: an open bioinformatics database and analysis resource for virology research.

441    *Nucleic acids research* 40:D593–8. DOI: 10.1093/nar/gkr859.

442

443    **Posada-Céspedes, S., Seifert, D., Topolsky, I., Jablonski, K. P., Metzner, K. J., &**

444    **Beerenwinkel, N. 2021**. V-pipe: a computational pipeline for assessing viral

445    genetic diversity from high-throughput data. *Bioinformatics (Oxford, England),*

446    btab015. Advance online publication.

447    https://doi.org/10.1093/bioinformatics/btab015

448

449

450    **Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. 2010.** Viruses

451    in the faecal microbiota of monozygotic twins and their mothers. *Nature* 466:334–338.

452    DOI: 10.1038/nature09199.

453

454    **Sharma D, Priyadarshini P, Vrati S. 2015.** Unraveling the web of viroinformatics:

455    computational tools and databases in virus research. *Journal of virology* 89:1489–1501.

456    DOI: 10.1128/JVI.02027-14.

457

458    **Shen W, Le S, Li Y, Hu F. 2016.** SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q

459    File Manipulation. *PLoS One* 11(10):e0163962. DOI:10.1371/journal.pone.0163962.

460

461    **Siddle KJ, Eromon P, Barnes KG, Mehta S, Oguzie JU, Odia I, Schaffner SF, Winnicki**

462    **SM, Shah RR, Qu J, Wohl S, Brehio P, Iruolagbe C, Aiyepada J, Uyigue E,**

463    **Akhilomen P, Okonofua G, Ye S, Kayode T, Ajogbasile F, Uwanibe J, Gaye A,**

464    **Momoh M, Chak B, Kotliar D, Carter A, Gladden-Young A, Freije CA, Omoregie**

465    **O, Osiemi B, Muoebonam EB, Airende M, Enigbe R, Ebo B, Nosamiefan I, Oluniyi**

466    **P, Nekoui M, Ogbaini-Emovon E, Garry RF, Andersen KG, Park DJ, Yozwiak NL,**

467    **Akpede G, Ihekweazu C, Tomori O, Okogbenin S, Folarin OA, Okokhere PO,**

468        **MacInnis BL, Sabeti PC, Happi CT. 2018.** Genomic Analysis of Lassa Virus during an

469        Increase in Cases in Nigeria in 2018. *The New England journal of medicine* 379:1745–

470        1753. DOI: 10.1056/NEJMoa1804498.

471

472  **Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, Iosifidis C, Agha R. 2020.**

473        World Health Organization declares Global Emergency: A review of the 2019 Novel

474        Coronavirus (COVID-19). *International journal of surgery* . DOI:

475        10.1016/j.ijsu.2020.02.034.

476

477  **Tang P, Chiu C. 2010.** Metagenomics for the discovery of novel human viruses. *Future*

478        *microbiology* 5:177–189. DOI: 10.2217/fmb.09.120.

479

480  **Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine**

481        **A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler**

482        **D, Gabriel S, DePristo MA**. **2013**. From FastQ data to high confidence variant

483        calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc*

484        *Bioinformatics*. 43(1110):11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43.

485

486  **Wan Y, Renner DW, Albert I, Szpara ML**. **2015.** VirAmp: a galaxy-based viral genome

487        assembly pipeline. *Gigascience*. 4:19. doi:10.1186/s13742-015-0060-y.

488

489  **Wymant C, Blanquart F, Golubchik T, Gall A, Bakker M, Bezemer D, Croucher NJ, Hall**

490       **M, Hillebregt M, Ong SH, Ratmann O, Albert J, Bannert N, Fellay J, Fransen K,**

491       **Gourlay A, Grabowski MK, Gunsenheimer-Bartmeyer B, Günthard HF, Kivelä P,**

492       **Kouyos R, Laeyendecker O, Liitsola K, Meyer L, Porter K, Ristola M, van Sighem**

493       **A, Berkhout B, Cornelissen M, Kellam P, Reiss P, Fraser C, BEEHIVE**

494       **Collaboration. 2018.** Easy and accurate reconstruction of whole HIV genomes from

495       short-read sequence data with shiver. *Virus evolution* 4:vey007. DOI: 10.1093/ve/vey007.

496

497    **Yamashita, A., Sekizuka, T., & Kuroda, M. 2016**. VirusTAP: Viral Genome-Targeted

498       Assembly Pipeline. *Frontiers in microbiology*, *7*, 32

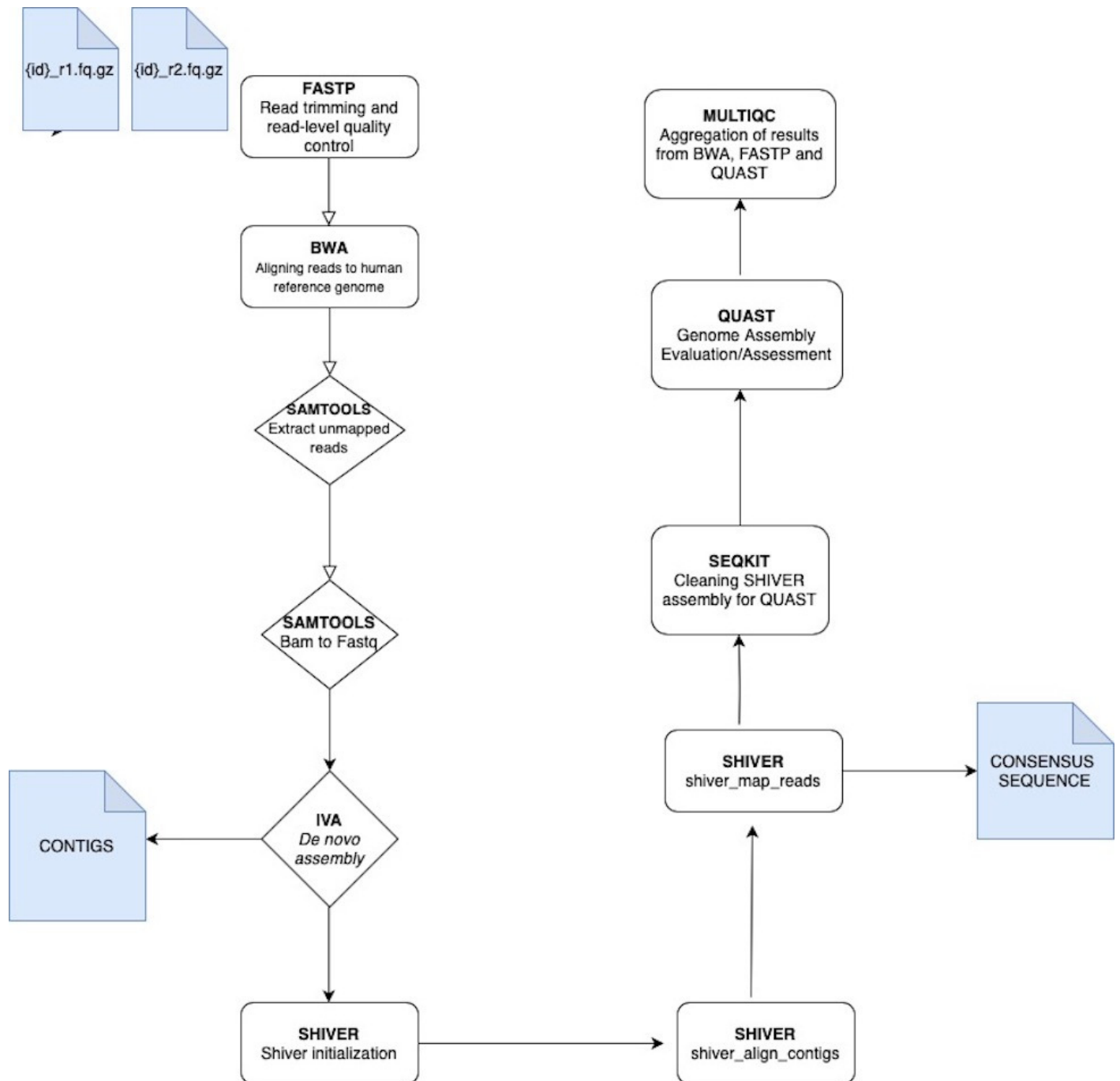
499       https://doi.org/10.3389/fmicb.2016.00032.

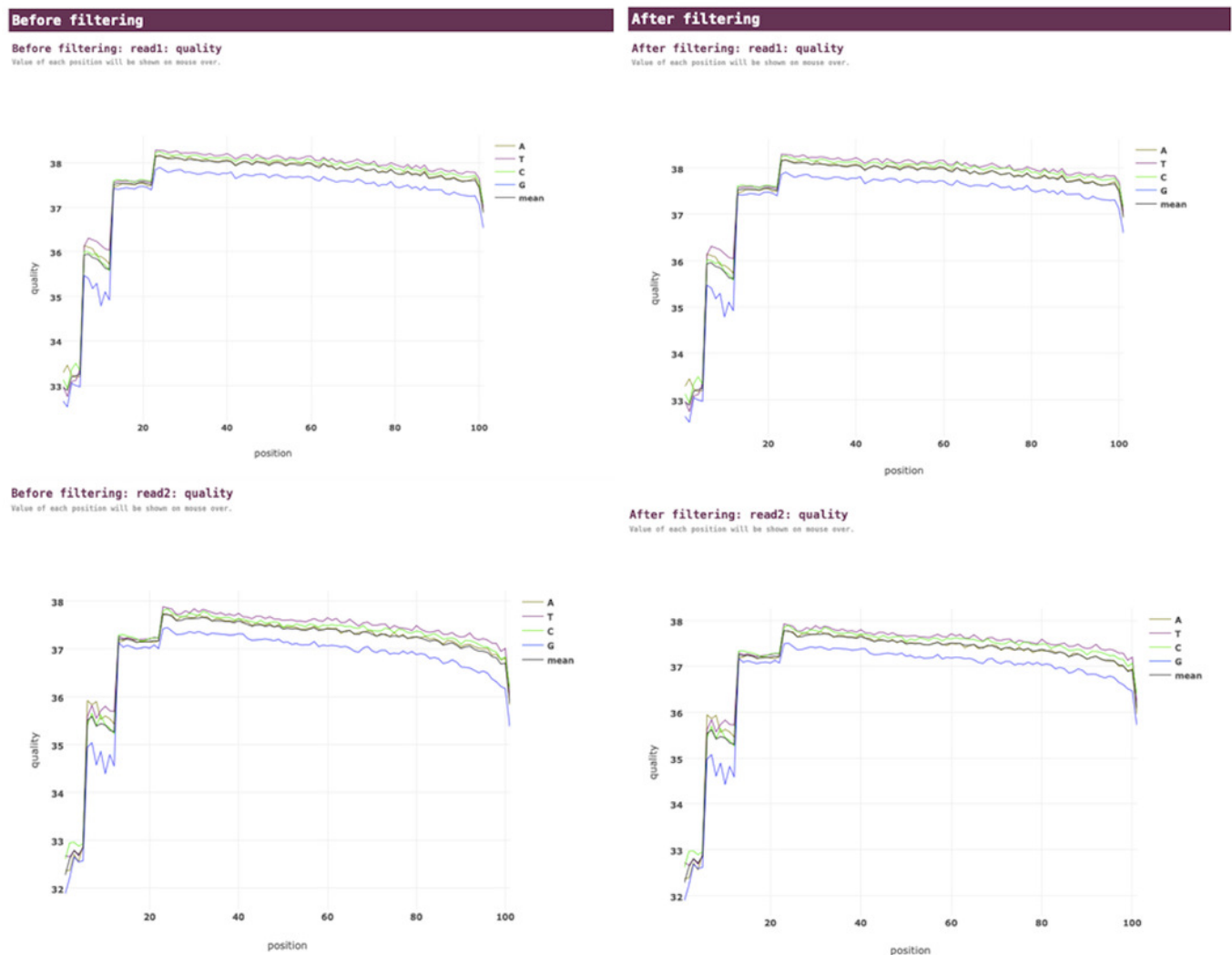
500

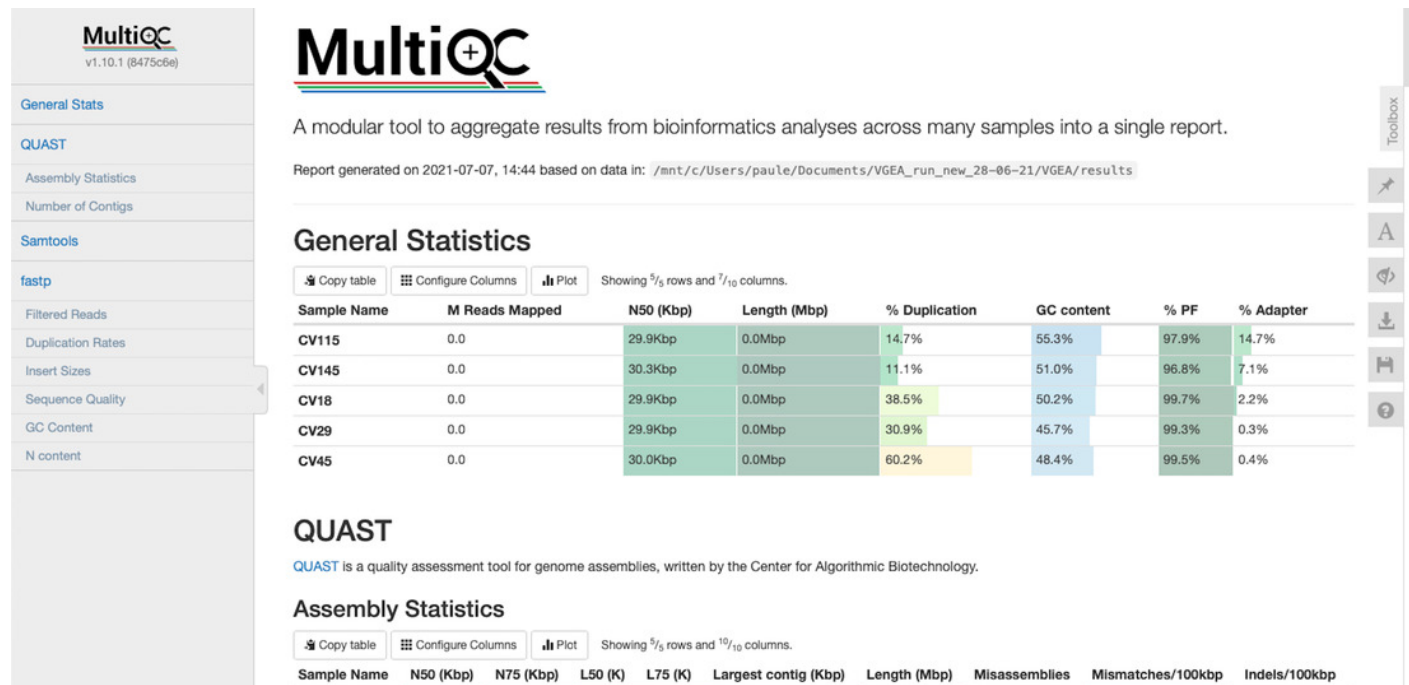
501
502
503

# Figure 1

A schematic workflow of VGEA

User-supplied paired-end fastq files are pre-processed and trimmed using **FASTP** followed by mapping to the human reference genome with **BWA**. Following mapping, a BAM file containing unaligned/unmapped reads is extracted using **SAMTOOLS**. This BAM file is then split into fastq files of forward and reverse reads also with **SAMTOOLS** after which *de novo* assembly is carried out using **IVA**. Following *de novo* assembly, **SHIVER** is used to map the reads and generate consensus sequences, and detailed minority variant information (full explanation of the shiver method is in supplementary file 1). **SEQKIT** is used to clean the SHIVER output for QUAST after which genome evaluation and assessment is carried out using **QUAST**. **MULTIQC** is then used for aggregation of results from BWA, FASTP and QUAST.

# Figure 2

fastp pre-processing report for a SARS-CoV-2 test dataset analyzed using VGEA

# Figure 3

MultiQC report of five SARS-CoV-2 datasets analyzed using VGEA

Benchmarking values (time and CPU usage) for a SARS-CoV-2 dataset analyzed using VGEA

| VGEA rule name | Time (h:m:s) | Maximum RAM used (MB) |
|---|---|---|
| human_reference_index | 1:01:53 | 4688.56 |
| fastp | 0:00:14 | 581.91 |
| bwa_human | 0:08:52 | 5960.95 |
| samtools_extract | 0:02:40 | 16.21 |
| bamtofastq | 0:01:39 | 6.61 |
| *iva | 8:19:11 | 238.57 |
| shiver_init | 0:00:53 | 64.97 |
| shiver_align_contigs | 0:04:37 | 2509.64 |
| shiver_map_reads | 0:31:51 | 567.27 |
| shiver_tidy | 0:00:00 | 1.06 |
| quast | 0:00:33 | 72.51 |

2 *IVA was run using one CPU core and two threads so if allowed more computational resources, the assembly time
3 will be even shorter.

Performance comparison using different assembly pipelines

2

| Sample ID | # reads (x 10$^6$) | Pipeline | # contigs | Largest contig (bp) | N50 | NG50 | Genome fraction (%) | Mis assemblies | Mismatches | Indels | Maximum RAM used (MB) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CV18 | 3.2 | VGEA | 42 | 29928 | 2294 | 29928 | 99.776 | 0 | 10 | 0 | 627 |
| | | SPAdes | 384 | 22141 | 1435 | 22141 | 99.652 | 1 | 18 | 1 | 2447 |
| | | Velvet | 68 | 1858 | 728 | 922 | 19.326 | 0 | 3 | 0 | 1544 |
| CV29 | 1.8 | VGEA | 31 | 7731 | 3065 | 7534 | 99.786 | 0 | 9 | 0 | 484 |
| | | SPAdes | 478 | 24904 | 1136 | 24904 | 99.632 | 0 | 7 | 0 | 2314 |
| | | Velvet | 66 | 2877 | 942 | 1380 | 1.729 | 0 | 0 | 0 | 807 |
| CV45 | 6.2 | VGEA | 30 | 16248 | 2603 | 16248 | 98.291 | 1 | 11 | 0 | 666 |
| | | SPAdes | 45 | 6779 | 1255 | 2447 | 94.957 | 0 | 35 | 12 | 2504 |
| | | Velvet | 535 | 5239 | 898 | 3030 | 14.256 | 0 | 0 | 0 | 1360 |
| CV115 | 2 | VGEA | 28 | 5225 | 2258 | 3060 | 96.957 | 0 | 12 | 0 | 177 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *SPAdes | 49 | 1942 | 1068 | 1828 | - | - | - | - | 1735 |
| | | Velvet | 41 | 2847 | 819 | 931 | 68.134 | 0 | 9 | 0 | 511 |
| CV145 | 4.4 | VGEA | 28 | 6807 | 2049 | 4214 | 73.093 | 0 | 14 | 0 | 635 |
| | | SPAdes | 188 | 3216 | 1190 | 2477 | 5.073 | 2 | 13 | 0 | 2547 |
| | | Velvet | 178 | 1798 | 682 | 1107 | 3.578 | 0 | 0 | 0 | 1459 |

4     *QUAST gave no genome fraction value for this sample