REALISTICALLY SIMULATING SARS-COV-2 WASTEWATER METAGENOME SEQUENCING DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

FATMA RABİA FİDAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOLOGY

SEPTEMBER 2022

Approval of the thesis:

REALISTICALLY SIMULATING SARS-COV-2 WASTEWATER METAGENOME SEQUENCING DATA

submitted by **FATMA RABİA FİDAN** in partial fulfillment of the requirements for the degree of **Master of Science in Biology Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ayşe Gül Gözen
Head of Department, **Biology** _____

Prof. Dr. Mehmet Somel
Supervisor, **Biological Sciences, METU** _____

Dr. Nick Goldman
Co-supervisor, **European Bioinformatics Institute, EMBL** _____

Examining Committee Members:

Assist. Prof. Dr. Aybar Can Acar
Health Informatics, METU _____

Prof. Dr. Mehmet Somel
Biological Sciences, METU _____

Assoc. Prof. Dr. Emre Keskin
Aquacultural Engineering, Ankara University _____

_____

_____

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: FATMA RABİA FİDAN

Signature :

ABSTRACT

REALISTICALLY SIMULATING SARS-COV-2 WASTEWATER METAGENOME SEQUENCING DATA

FİDAN, FATMA RABİA
M.S., Department of Biology
Supervisor: Prof. Dr. Mehmet Somel
Co-Supervisor: Dr. Nick Goldman

September 2022, 46 pages

Wastewater surveillance for SARS-CoV-2 is seeing increasingly widespread use as it proved useful in tracking variants and their prevalence in an unbiased manner. It has been shown that it is possible to detect an emerging variant from wastewater samples up to two weeks earlier than its detection at hospital clinics (Karthikeyan et al., 2021). Such data are critical for policies regarding the measures taken against variants of concern. Since such surveillance has important consequences, it is also vital to test and validate the surveillance methodologies and software packages, which in turn creates a need for a realistic SARS-CoV-2 wastewater metagenome sequencing data simulator. We stepped up to develop a prototype simulator, modelling many unusual features of the data, such as differential SARS-CoV-2 variant abundance, amplicon architecture, differential amplicon abundance of a primer set and major error components. By investigating wastewater metagenomic SARS-CoV-2 datasets, we identified high-frequency errors where many reads from the same sample wrongly supported the same artifactual mutation. This kind of error likely stemmed from RNA-degradation and PCR amplification processes, as the most significant source of noise

in wastewater metagenomic SARS-CoV-2 data analysis. This makes it crucial to realistically model high-frequency errors within inference and simulation frameworks for this type of data. To achieve this, we study the error characteristics of SARS-CoV-2 wastewater sequencing data, model the major high-frequency error components, and realistically implement these models into our simulator. We also aim to display some use cases of the simulated data in downstream applications such as the benchmarking of software for individual variant resolution. Moreover, comparisons involving results from wastewater and clinical data will allow us to see the differences in error characteristics of the clinical and wastewater data.

Keywords: SARS-CoV-2, pandemic, public healthcare, wastewater, amplicon sequencing, simulation

ÖZ

SARS-COV-2 ATIK SU METAGENOM VERİSİNİN GERÇEKÇİ SİMÜLASYONU

FİDAN, FATMA RABİA
Yüksek Lisans, Biyoloji Bölümü
Tez Yöneticisi: Prof. Dr. Mehmet Somel
Ortak Tez Yöneticisi: Dr. Nick Goldman

Eylül 2022 , 46 sayfa

Yapılan çalışmalarla yeni bir SARS-CoV-2 varyantının atık sularda hastane kliniklerinden 2 hafta kadar daha erken gözlemlenebildiğinin gösterilmesiyle (Karthikeyan et al., 2021) varyant takibi için atık su izlemesi tarafsız ve işe yarar bir metod olarak yaygınlaşmaya başladı. Bu şekilde bir varyant takibinin tehlikeli varyantlara karşı alınan önlem politikalarında oynayacağı rol ve bunların yol açabileceği büyük düzenlemeler yüzünden bu çalışmalarda kullanılan metodların ve yazılımların test edilmesi ve doğrulanması büyük önem taşımaktadır. Bu durum, gerçekçi bir atık su SARS-CoV-2 metagenom simulatörü ihtiyacını doğurmaktadır. Biz de gerçek atık sudan gelen verinin farklı SARS-CoV-2 varyant yoğunluğu, pirmer setine özel farklı amplikon yoğunlukları ve temel hata bileşenleri gibi en önemli özelliklerini yansıtan prototip bir simülatör yapmak için adım attık. Gerçek veriye baktığımızda bazı yapay mutasyonların verisetinde bir çok okuma tarafından desteklendiğini gördük. Bu tip hataların başlıca sebepleri arasında RNA'nın örneklemesine kadar geçen sürede su içinde beklemesinden kaynaklı RNA bozulmaları ve PCR hataları yer almaktadır. Bu

durum bu tarz yüksek sıklıklı hataların da simülasyonun bir parçası olmasını gerekli kılmaktadır. Biz de bu çalışmada bunu başarmak için gerçek verinin yüksek sıklıklı hata karakteristiklerini çalışıp başlıca yüksek sıklıklı hata bileşenlerini gerçekçi bir şekilde simülatörümüze uyguluyoruz. Ayrıca simülatör çıktısı verinin olası kullanım alanlarını göstermenin yanı sıra bireysel korona varyantlarını tespit etme gibi uygulama programlarında nasıl davrandığını gösteriyoruz. Atık su verisi ile kilinik veriyi karşılaştırarak iki verinin hata karakteristiklerinin farklı olduğunu gösteriyoruz.

Anahtar Kelimeler: SARS-CoV-2, pandemi, toplum sağlığı, atık su, amplikon sekanslama, simülayon

To my family

ACKNOWLEDGMENTS

Throughout my academic journey, I received much kindness, help, sympathy and support from wonderful people around me, without which this would have been far too difficult, if possible at all.

I want to start by thanking my mother Aynur Fidan, my father Ahmet Fidan and my sisters Sevde Fidan and Zehra Fidan for believing in my passion for science and supporting me through the years. I want to thank my furry children for being a source of happiness in my life.

I want to thank my "bacımlar", my dearest friends, Ebrar Sinmez, Seyda Balkan, Peri Beşarat, Nursu Çakır, Serena Mahnoor, Etkin Tarlan and Ahmet Körpınar for providing me with the best friend group, suffering lab reports, midterms, finals and registrations together, filling every possible moment with laughter, and being there for me.

I was lucky enough to work with two of the most amazing supervisors: Mehmet Somel and Nick Goldman. I thank Mehmet Somel with whom I worked for 5 happy years. He is a role model with his hard work, positivity and admirable personality. He took a rookie second grader and helped me become a competent master's student. I thank Nick Goldman for his immense kindness, mentorship and for the great oppurtunity he gave me, which is a keystone in my career. My remote and in-person experience in the group was nothing but wonderful. I hope to work with both Mehmet and Nick in the future again.

I want to thank all the members of Compevo lab, especially Sevim Seda Çokoğlu, who was my first research project partner, Dilek Koptekin, who was my first supervisor in Compevo lab, Ekin Sağlıcan who supervised me through another project and was very understanding when I needed time for the assistantship exam preparations and Melike Dönertaş, who kindly and eagerly provided me with mentorship and guidence that I needed as a firt-gen scientist.

I want to thank all Goldman Group members, especially Nicola De Maio, who co-supervised this project, Will Boulton, who created and handed the precious SWAMPy over to me. Both Will and Nicola were incredibly helpful during my adaptation period to the project. I thank Conor Walker, who was my first supervisor in the Goldman group and a great mentor.

I thank Josh Quick, members of the JBC-led Wastewater Genomics collaboration, in particular Terry Burke, Hubert Denise, Steve Paterson, Christopher Quince and Sébastien Raguideau for their contributions to this project, as well as Charlotte West, Lukas Weilguny, Sevim Seda Çokoğlu, Kıvılcım Başak Vural, Melih Yıldız and Gözde Atağ for helping me with SWAMPy program testing.

Finally, I thank my thesis examining committee members Aybar Can Acar and Emre Keskin for their contributions and enthusiasm in joining my thesis defense.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

VOC	Variant of concern
VUI	Variant under investigation
SWAMPy	Simulating SARS-CoV-2 Wastewater Amplicon Metagenomes in Python
RT-qPCR	Quantitative Reverse Transcription polymerase chain reaction
RC-PCR	Reverse complement polymerase chain reaction
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
WBE	Wastewater-based epidemiology
SNP	Single nucleotide polymorphism
VAF	Variant allele frequency
RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid

CHAPTER 1

INTRODUCTION

1.1 Summary

Wastewater surveillance has been an important tool for dealing with the SARS-CoV-2 pandemic. With the acceleration of wastewater sequencing studies in the pandemic, researchers create new softwares designed to be used with wastewater sequencing data. Benchmarking these softwares requires a simulator which can create realistic data accounting for the specialised set of characteristics we observe in wastewater sequencing data, among which a high rate of RNA degradation and PCR errors are prominent as a result of their environmental exposure in the sewage. In this study, we analysed the error characteristics of wastewater sequencing data and implemented our findings in our simulator tool SWAMPy. As a comparison, we also analysed clinical sequencing data which validated that a high error rate is a characteristic feature of wastewater sequencing data. We ran SWAMPy-simulated data through a downstream application software to show that it captures information as expected. SWAMPy thus accounts for the important characteristics of the data that were not covered by the existing simulators. It, therefore, provides a tool for creating control case data for researchers who work on SARS-CoV-2 wastewater studies and consequently contributes towards dealing with the SARS-CoV-2 pandemic.

1.2 Wastewater-Based Epidemiology Before the SARS-CoV-2 Pandemic

Before the SARS-CoV-2 pandemic, wastewater-based epidemiology had been used for tracking chemicals and pathogens by analysing samples taken from the sewage

systems. Tracking caffeine, alcohol and tobacco consumption and use of medications and illicit drugs have been the target of wastewater-based epidemiology studies [1], as have the tracking of viruses like hepatitis [2, 3] and poliovirus [4, 5, 6], and worldwide antimicrobial resistance [7]. Even prior to the SARS-CoV-2 pandemic, the presence of coronaviruses in faeces and their survival in the water had been shown [8].

For substance surveillance, wastewater is a valuable and practical source of information in terms of ethics because it does not require consent to collect the data [7] and it bypasses the biases coming from data collected via questionnaires about topics like drug use [1]. When it comes to public health concerns and tracking pathogens, it has been reported that wastewater surveillance is sometimes advantageous over clinical surveillance because the wastewater data coming from a population provides a wider picture of pathogens circulating in the population and is less biased and faster compared to clinical surveillance where only symptomatic people register to clinics for treatment and only after they start to show symptoms. This causes a delay in tracking symptomatic patients and misses asymptomatic people who may be important contributors to the pathogen circulation. [7].

There are different approaches for studying pathogens from sewer systems depending on the question and the organism(s) of interest. First of all, microbiological methods which require culturing the organism are used for routine controls of the wastewater treatment facilities where the composition of the bacterial community directly affects the treatment products [9, 10]. For detecting and quantifying a specific pathogen, DNA/RNA isolation followed by targeting a specific region with qPCR is a common approach [3, 6]. For discovering novel microorganisms and other metagenomic purposes, whole metagenome sequencing is used [9]. Also, when there is a database of known organisms, 16s rDNA sequencing can be used for characterising microbial communities [9, 11, 12].

On the other hand, some wetlab methods that enable easy sequencing of low abundance viruses directly from biological samples without first culturing, namely whole genome amplicon sequencing, have been become available. This method is ideal especially for viruses that are not suitable for metagenomics methods [13] and will be mentioned in detail in the following sections.

1.3 SARS-CoV-2 Surveillance in the Pandemic

After the first case of SARS-CoV-2 was reported in December 2019 and the disease started to spread, first studies about SARS-CoV-2 were published in a short time [14, 15]. These were the first sequencing studies of SARS-CoV-2 to our knowledge mainly to identify the culprit as a Betacoronavirus similar to previously identified ones in bats with high sequence similarity. In these studies, the authors performed total RNA extraction from samples taken from patients, high-throughput sequencing and *de novo* genome assembly. They described the genome structure of SARS-CoV-2.

After learning about the identity of the culprit, surveillance efforts were put into action employing many different methods. The main targets of surveillance are monitoring SARS-CoV-2 incidence and assessing the severity of Covid-19 in different social groups, tracking changes in the incubation period, fatality rate, recovery rate, hospitalisation and other epidemiological features, monitoring the circulating variants and detecting newly emerging variants. This information in turn guides public health actions like contact tracing, individual isolation and imposing quarantine on a population [16].

For detecting the existence of the virus, RT-qPCR is used with various kits targeting various SARS-CoV-2 genes like RdRP, E, N, S and ORF1ab [17]. Also, the S gene dropout method which takes advantage of the fact that some strains fail to respond to S gene targeting but are detected with other genes helped identify specific strains with certain mutations on the S gene [18]. In addition to these nucleic acid amplification tests, different kinds of serological tests including antigen detection and antibody detection tests are used [19], which can be faster and cheaper, but their specificity might not be ideal [20]. Finally, viral sequencing can provide more information than just detection and quantification of the virus. Partial sequencing of the genome can provide phylogeny and strain information [21] while whole genome sequencing enables tracking the mutations and newly emerging variants [22].

Among the different available sequencing protocols, amplicon sequencing stands out in SARS-CoV-2 studies. In whole genome amplicon sequencing, the genome of the

target organism is amplified in segments called amplicons (figure 1.1). Primers in a given primer set are designed so that the amplicons overlap, which results in whole genome coverage without gaps in between the amplicons, although genome ends are blind spots. Multiplex PCR or similar methods like RC-PCR [23] are used to both enrich and amplify the target segments in a single PCR step. This circumvents the necessity to deplete non-target RNAs or separately enrich for target RNAs prior to amplification [13]. Its targeted nature is suitable for heavily contaminated biological samples like the nasopharyngeal swabs contaminated with host DNA and wastewater samples containing host DNA along with other microbial and viral genomic material, and the test results are obtained in a relatively short time [24]. Apart from the practical advantages in wet lab procedures, in terms of the resulting data, it has been shown that amplicon sequencing can provide higher read-depth (although not uniform) even with low concentration input sample and with a higher percent of SARS-CoV-2-mapped reads and lower non-target organism-mapped reads [25]. The development of many commercially available, easily deployable kits for SARS-CoV-2 amplicon sequencing was followed by its widespread use in SARS-CoV-2 tracking.

Figure 1.1: Tiling amplicons in whole genome amplicon sequencing.

The biological samples for testing are generally collected from individuals who show symptoms and register at clinics. But there are also random testing and group testing approaches where samples from a group of people are pooled and tested together, which makes it possible to track more people with the available resources and guides the individual testing especially in communal living areas like long-term care facilities [26]. Wastewater became an important data source as will be elaborated on in the next section.

1.4 SARS-CoV-2 Wastewater Surveillance and Sequencing

Early in the pandemic, researchers tried to understand the mode and period of infection and other virological and epidemiological aspects of the virus. Studies like [27] and [28] showed that although being a respiratory tract virus, SARS-CoV-2 was shed in faeces and this was irrespective of being symptomatic or the severity of the illness. Moreover, it was shown that stool samples kept testing positive even after about a week later than pharyngeal swab samples turned negative, although stool viruses were not viable. The existence of the virus in the faeces paved the way to monitor the virus from sewage systems, which was an exciting idea. Because wastewater data represents a population of any size from a university campus to a whole city. This made it an ideal tracking method since it is impossible to individually sequence everyone in a population periodically due to financial and practical constraints.

Wastewater surveillance can help with some of the aims of SARS-CoV-2 surveillance mentioned in the previous section. Although we cannot exactly know many epidemiological features solely from the wastewater surveillance data, we can estimate disease prevalence and variant abundance change trends on a population level. By choosing strategic sampling points, we can also infer about different social groups. Although it provides real-time monitoring of the population in a feasible way, it has limitations because of the uncertainties in the data stemming from fluctuations in population size and wastewater flow over time as well as a low target/contaminant ratio [29]. In this sense, it is complementary to clinical surveillance. When combined with information coming from other surveillance methods wastewater surveillance can be very informative. For example in Hong Kong, they could detect a single Delta variant carrier individual in a population of 33,000 people thanks to routine wastewater surveillance, which then led to strategic upstream wastewater sampling upon Delta variant detection, which finally led to compulsory individual testing in a small area [30].

SARS-CoV-2 wastewater surveillance started mainly with detecting the existence of SARS-CoV-2 with RT-qPCR, quantifying the viral load and estimating the disease prevalence, while a small number of partial or whole genome sequencing were performed to validate what RT-qPCR detects was actually SARS-CoV-2 [31, 32, 33]. They showed that viral load estimations obtained from wastewater correlate with the

clinically reported case numbers and even preceded them, which makes wastewater a potential early warning system [34]. These promising results showed wastewater to be a valid data source in SARS-CoV-2 tracking and over 55 countries around the world (including Turkey) started watching their wastewaters for SARS-CoV-2 [35]. Then it became possible to track some specific variants from wastewater with variant-specific tests such as allele-specific RT-qPCR, and there have been case studies where wastewater surveillance had a significant impact on public healthcare decisions as mentioned before [30].

Later on, it was shown that whole-genome high-throughput sequencing of wastewater holds valuable information. In early sequencing studies, it was shown that it is possible to detect individual mutations in SARS-CoV-2 genomes and infer about VOCs and VUIs bearing those mutations, although wastewater being a mixture complicated the interpretation [36]. Then software programs like SAM Refiner [37], and Freyja [38] were developed which were designed to be used with wastewater sequencing data and can ascertain individual variants within the mixture and even estimate their relative proportions. Similar to the disease prevalence, variant abundances estimated from wastewater were shown to be in line with the clinically reported ones, and newly emerging variants were detected in wastewater significantly earlier than their detection in clinics [38], which is again a piece of critical information for public healthcare decisions.

1.5 Characteristics of Wastewater-derived Data

There is a specialised set of characteristics that we observe in the majority of SARS-CoV-2 wastewater sequencing data and ideally should also be represented in simulated data. First of all, a sample taken from wastewater contains biological matter from multiple people. Consequently, as opposed to a clinical sample, sequencing data coming from a single run will frequently contain information from multiple SARS-CoV-2 variants. Furthermore, these variants may be present in different proportions. Secondly, as mentioned previously, amplicon sequencing is the dominant method for SARS-CoV-2 sequencing studies in general and also specifically in wastewater. As of July 2022, out of 4,840,834 raw SARS-CoV-2 sequences listed in COVID-19 data

portal by ENA (https://www.covid19dataportal.org), 4,692,588 of them (96.9%) are derived from amplicon sequencing, which contributes with its characteristic features to the data. Amplicon sequencing data typically have patchy read depth across the genome as a result of differential binding efficiency of primers within a given primer set [25] and possible differential amplification success of the amplicons. Moreover, the read depth pattern mostly depends on the specific primer set in use, but a poor sample quality often inflates the variation across amplicons [25]. The third characteristic of the data is RNA degradation, PCR and other library preparation errors, collectively referred to as high-frequency errors from now on (as opposed to sequencing errors which are typically at low frequency). This is because viral RNA in wastewater is exposed to all sorts of environmental factors such as heat, chemicals and physical strains from the point it leaves the human body to sample collection, contrary to clinical samples. Finally, sequencing errors are a major component of the data. The most frequently used sequencing platform is Illumina with a paired-end library layout, which comes with its unique sequencing error patterns.

1.6 Aim

Testing and benchmarking are crucial for any software and method. Although there is truth set data for some specific bioinformatics applications, like Genome in a Bottle for variant calling (https://www.nist.gov/programs-projects/genome-bottle), it is not the case for many others. This may be because sometimes the truth data is impossible to obtain (like human evolution and population genomics studies) or is not feasible because of the time and resources it requires (like creating many non-existent potential variants of a virus). Researchers often rely on simulation for creating benchmarking data where they can control many parameters and produce plenty of data in a short time. In the context of SARS-CoV-2 wastewater sequencing, the software programs that are specifically designed to be used with this type of data need benchmarking. But as previously mentioned, the data has specific characteristics. Some existing sequencing simulators can account for some elements (like sequencing errors in ART [39] and many other sequencing simulators) while some data characteristics are overlooked altogether or not in a usable form for our data including read depth variation

across amplicons and high-frequency errors. We aim to create a realistic and easily usable SARS-CoV-2 wastewater sequencing simulator by exploring the overlooked characteristics and creating a program that takes advantage of the novel knowledge to account for the overlooked characteristics and making use of pieces of existing software for the elements that are already thoroughly studied.

1.7 Contributions

This dissertation reports my contributions towards a simulator, namely SWAMPy (Simulating SARS-CoV-2 Wastewater Amplicon Metagenomes in Python), that can provide realistic simulated data of SARS-CoV-2 paired-end amplicon sequencing data coming from wastewater and sequenced on an Illumina machine.

Analyses regarding the differential amplicon distribution were done by Will Boulton, who also created the first prototype version of SWAMPy, which generated a realistic read-count simulation per source genome and per amplicon among other features.

I, F. Rabia Fidan, did analyses regarding the high-frequency errors, statistically modelled the findings and implemented an error introduction functionality to SWAMPy. I also contributed with performance enhancements and release preparations, and improved the documentation. My work moved SWAMPy from prototype to publicly released software, with an accompanying paper being prepared for submission to an international peer-reviewed journal. My contributions will be detailed in the following chapters.

Nick Goldman and Nicola De Maio (EMBL-European Bioinformatics Institute) supervised, guided and actively participated in each step of the study.

Mehmet Somel reviewed the work and contributed with valuable discussions, thanks to which I added the clinical analyses.

Chapter 2 covers the analyses on high-frequency errors which constitutes error characterisation on real sequencing experiments and statistical modelling of different aspects of the errors coming from wastewater and clinical samples. Chapter 3 covers the implementation of the findings into SWAMPy together with the SWAMPy workflow

and working principles.

CHAPTER 2

ERROR CHARACTERISATION

2.1 Introduction

As described in Chapter 1, errors that are high in frequency in the data that likely stem from RNA degradation and library preparation errors we collectively refer to as high-frequency errors. In this part of the study, we classified different kinds of high-frequency errors, modelled error rates, error lengths and variant allele frequencies and then estimated model parameters from the data. We repeated the same analysis with clinical data to see if our assumption that wastewater data has more high-frequency errors is supported. We compared clinical, and wastewater error characteristics which support that high-frequency errors are indeed an important feature of wastewater data.

2.2 Methods

2.2.1 Classification

For error classification and characterisation we used real wastewater sequencing data from 121 experiments (see Section 2.5). We mapped raw reads to the Wuhan-Hu-1 [15] reference genome using bowtie2 [40] (version 2.4.4). We then used bcftools mpileup [41] (version 1.13) to obtain vcf files. We did not perform a separate variant calling as we are interested in errors and needed every discrepancy between the reference genome and our sample, which henceforth will be referred as variants. We filtered out positions with a read depth (DP) <10 and with <5 reads supporting the alternative allele (AD). The remaining variants are classified into different categories

as summarised in fig. 2.1. First of all, if the variant allele frequency (VAF) of a variant is smaller than 0.02, we classified them as sequencing errors since sequencing error rates are typically $< 2\%$ for Illumina devices [42]. Since the sequencing errors are modelled and simulated very well with the existing simulators, we did not model them but used an external software ART[39] for this purpose (see Chapter 3). The remaining variants included real polymorphisms between different SARS-CoV-2 variants and high-frequency errors. Distinguishing these two groups for the whole variant set proved challenging since there is no certain way to tell if a variant is a SNP or a high-frequency error. For this reason, we looked at a subset of the variants, namely, nonviable mutations: reasoning that if a variant is nonviable, it cannot be coming from a real organism; hence, it must be a (high-frequency) error. For this analysis, we only considered ORF1ab and S open reading frames of SARS-CoV-2 since these genes cover more than two-thirds of the genome, their function is better understood and they do not contain a reported stop codon in them to our knowledge, while this is not the case for other smaller open reading frames [43, 44]. We also excluded a portion from the 3' ends of these open reading frames due to the fact that some supposedly nonviable variants could be tolerable there since a few less codons from the 3' end might not disrupt its function completely. Nonviable meant for substitutions it is a nonsense mutation and for indels, its length is not a multiple of three. We further divided the high-frequency errors as either recurrent or unique based on if they appear in more than one wastewater sample and only one wastewater sample respectively. Then we divided both recurrent and unique errors as insertions, deletions and substitutions.

For clinical data analyses the same workflow is applied on 300 COG-UK clinical sequencing data generated with matching technical properties, i.e. amplicon sequencing, Illumina sequencing platform and paired-end library layout.

2.2.2 Modelling and Parameter Estimation

We modelled error rate as genome-wide error rate per locus for different type of errors. Error rates were corrected for the missing data according to error length models.

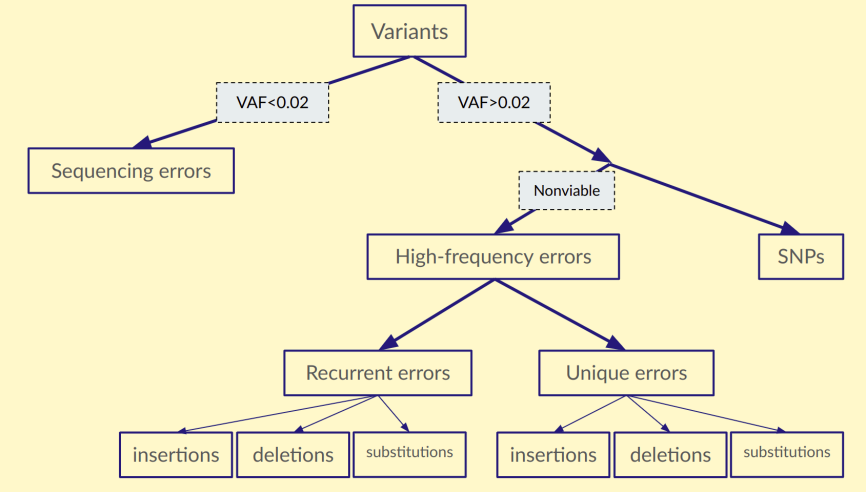

Figure 2.1: Variant classification criteria

2.2.2.1 Error length

Error lengths are modelled as follows based on the observation from the real data (see figure 2.2):

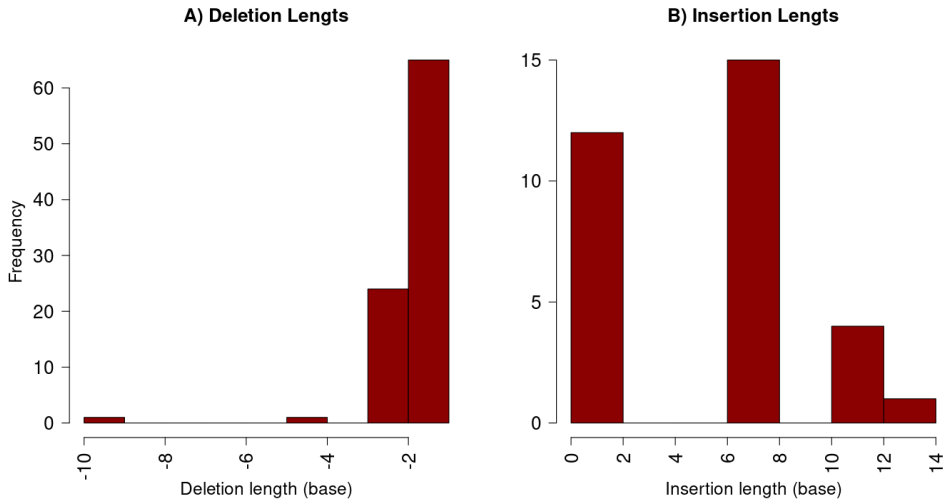

Figure 2.2: Histograms of indel length distributions. A) Deletions show a geometric-like length distribution. B) Insertion lengths display a uniform distribution

- substitution: 1
- insertion: $Uniform(max = L)$

- deletion: $Geometric(p)$

For insertions, the maximum value of the uniform distribution is the maximum observed insertion length in our dataset, 14. For deletions, we use the geometric distribution.

Given a deletion is initiated we model its length via:

$P(termination) = p$
$P(extension) = 1 - p$

For estimating the geometric distribution parameter p we could not use a standard distribution fitting technique because we cannot observe deletions with $len = 3j$ for $j = 1, 2, 3...$ Also, longer deletions are not present in our dataset. We wanted to avoid the bias coming from the missing observations by using only the deletions of length one and two thanks to the equation 2.1 and equate it with the observed deletions $\frac{\#length2}{\#length1}$.

$$\frac{P(len = 2)}{P(len = 1)} = \frac{(1-p)p}{p} = 1 - p \tag{2.1}$$

So, our estimator becomes:

$$\hat{p} = 1 - \frac{\#length2}{\#length1} \tag{2.2}$$

2.2.2.2 Error rate

We calculated the error rates separately for all six classes of high-frequency errors in both clinical and wastewater data. We applied corrections for the fact that we are using a subset of the genome. Error rates are calculated as follows where NS is the observed nonsense mutation count, ID is the observed indel count and L is the total loci count across all experiments:

- Substitution: $C\frac{NS}{L}$

- Indel rate: $C\frac{ID}{L}$

C values are correction factors. For substitutions, it corrects for the fact that only the loci along the genome which are capable of creating a stop codon by a single nucleotide change are visible to us in this analysis. I calculated the correction factor as:

correction factor = $\frac{\text{\#all possible mutations}}{\text{\#all possible nonsense mutations}}$

from the reference SARS-CoV-2 genome (Wuhan-Hu-1 [15]) by changing every reference nucleotide with the other three one by one and checking if that mutation is a nonsense mutation. I did this locus by locus, returning to the reference allele each time before moving on to the next locus. Here we did not take into account the potential differences in the rates of different mutations (i.e. whether it is a A->T or C->T). The correction factor value is approximately 21. For insertions, the correction factor is simply $3/2$ because of the uniform distribution and lack the deletions of length multiples of three. Finally, for deletions, we made use of the geometric distribution definition, and its calculation details can be seen in Appendix A

2.2.2.3 VAF

For modelling VAF of all types of errors, we used Beta distribution, and the two parameters were estimated with the method of moments.

2.3 Results

Numerical values of the different kinds of error rates can be seen in Table 2.1a and an extended version of this table including the error counts and correction factors can be seen in Table A.1 in Appendix A.

In all except one error category, the wastewater error rate was higher than the clinical error rate. In particular, unique and recurrent substitutions and unique deletions error rates were 74, 67 and 43 times the clinical ones.

Table 2.1: Table1: Error error rates of different categories of errors and error rate ratio of wastewater to clinical

	Ww rate (rWw)	**Clinical rate (rC)**	**rWw/ rC**
rec ins	0.000036	0.000029	1.251549
rec del	0	0.000008	0
rec subs	0.003357	0.00005	66.924926
u ins	0.000025	0.000003	7.904519
u del	0.000125	0.000003	42.704206
u subs	0.002485	0.000033	74.302477

(a) Ww: wastewater, u: unique, rec: recurrent, ins: insertion, del: deletion, subs: substitution

Deletion length model's geometric distribution parameter estimation on our dataset yielded the estimations of 0.63 for wastewater and 0.86 for clinical data.

2.4 Discussion

In each category of errors, we saw some recurrent errors, that is errors that are observed in multiple independent sequencing runs. Although the biological basis of these unique and recurrent errors is currently unknown to our knowledge, recurrent errors might originate from the positions along the genome that are more susceptible to degradation or to context-dependent PCR errors. In this sense, we associate recurrent errors more with RNA degradation and unique errors with random PCR errors. But this question needs to be further explored in the future.

Our error rate ratios support the hypothesis that wastewater data has higher high-frequency error rates and an ideal wastewater sequencing simulator should account for them in addition to sequencing errors.

There are a few other error elements that can be taken into account in the future to improve our error model such as PCR-mediated recombination products, also known as PCR chimeras [45, 46]. Similarly, the context of the errors (for example a homopolymer region or a structured part of the DNA) might affect PCR errors [46], which is

not currently accounted for.

2.5 Dataset availability

Wastewater sequencing experiments were conducted by members of JBC-led Wastewater Genomics collaboration and the data was internally shared with EBI and other collaborators. 12 of the 121 samples are mixtures of synthetically produced SARS-CoV-2 variant genomes, which mimic wastewater samples. 109 of them were sampled from wastewaters in the UK.

Clinical data is clinical sequencing experiments by COGUK. ENA accessions of all samples can be seen in Appendix A.

CHAPTER 3

SWAMPY

3.1 Introduction

SWAMPy (Simulating SARS-CoV-2 Wastewater Amplicon Metagenomes in Python) is a Python program designed to simulate wastewater amplicon sequencing data coming from an Illumina sequencing platform.

This chapter of the thesis explains the general workflow of our simulator tool SWAMPy and its methods with an emphasis on the specific functionalities I designed and implemented. It also shows how SWAMPy-simulated data behaves in a downstream application and mentions my other contributions to SWAMPy.

3.2 Contributions

Step 2 of the workflow described below was implemented by me. Steps 1,3 and 4 of the workflow were implemented by Will Boulton in the prototype simulator and some of them were then later improved by me along with bug fixes, as will be explained in the following sections. Program testing was done by me with the help of voluntary testers on multiple devices with Linux and macOS operating systems. I added a compatible Conda environment to the program repository, listed OS-related dependencies, prepared and added sample data to the repository, documented new functionalities and improved the existing documentation. I also performed analyses using a downstream application with SWAMPy-simulated data.

3.3 Methods

3.3.1 Step 1 - Creating Amplicons

The SWAMPy workflow starts with taking inputs which are the genomes of the SARS-CoV-2 variants (in a multi-FASTA format) that will be present in the simulated wastewater sample, their proportions in the mixture and the total target number of sequencing reads to be simulated. The user also chooses one of the supported primer sets in the program, which are ARTIC v1, ARTIC v4 [47] and Nimagen v2 [23].

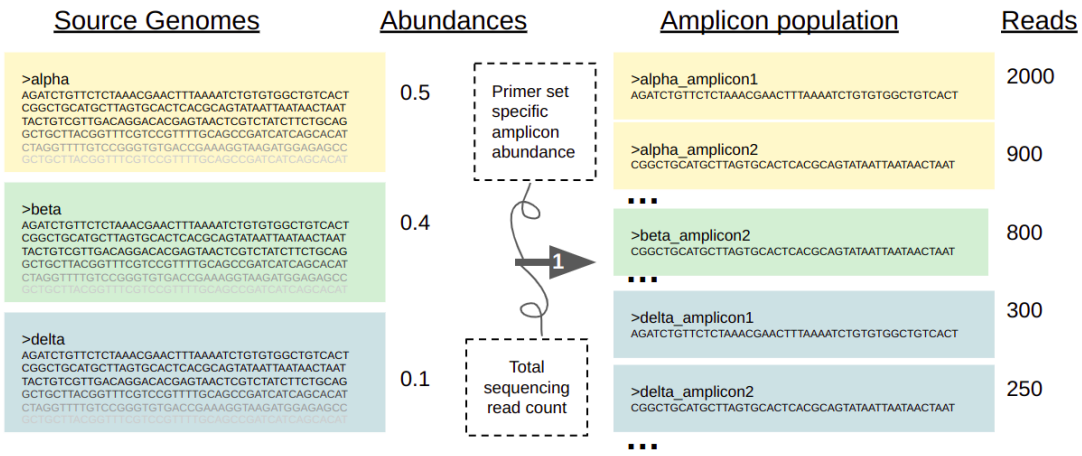

Figure 3.1: Step 1 of SWAMPy workflow.

SWAMPy starts by aligning the primers to the source genomes and slicing them from the primer binding positions to create amplicons of the source genomes (Figure 3.1). Then read counts per amplicon per genome are simulated using a Dirichlet distribution to take into account the read depth variation across the amplicons of a primer set and stochasticity across different experiments. Specific amplicon abundance profiles of individual primer sets obtained from real experiments were used to derive the default Dirichlet parameters. Relative variant abundance input by the user is also taken into account. For the details of the read count simulation, see Boulton & Fidan et al. (2022) (in preparation). At the end of step 1, we have amplicons of the source variant genomes, and the number of sequencing reads that we want to simulate from each.

3.3.2 Step 2 - High-frequency errors

After obtaining the initial amplicon population at step 1, SWAMPy diversifies it with mutant versions of the amplicons (Figure 3.2). This includes simulating individual errors, creating combinations of errors within individual amplicons to create different versions, and then distributing the read count of the wild-type amplicon between all versions (wild-type and mutant).

The natural way to simulate PCR errors is to simulate the inheritance of the errors from the previous PCR cycle and the addition of the new errors at each replication step. But simulating an error tree for each error would be computationally too expensive and would be infeasible in terms of run times. Also, we would lose some degree of control over the VAF of the individual errors. Instead, we opted for a more feasible and good enough approach where we simulated the errors individually and then grouped them randomly (keeping the control over their individual variant allele frequencies).

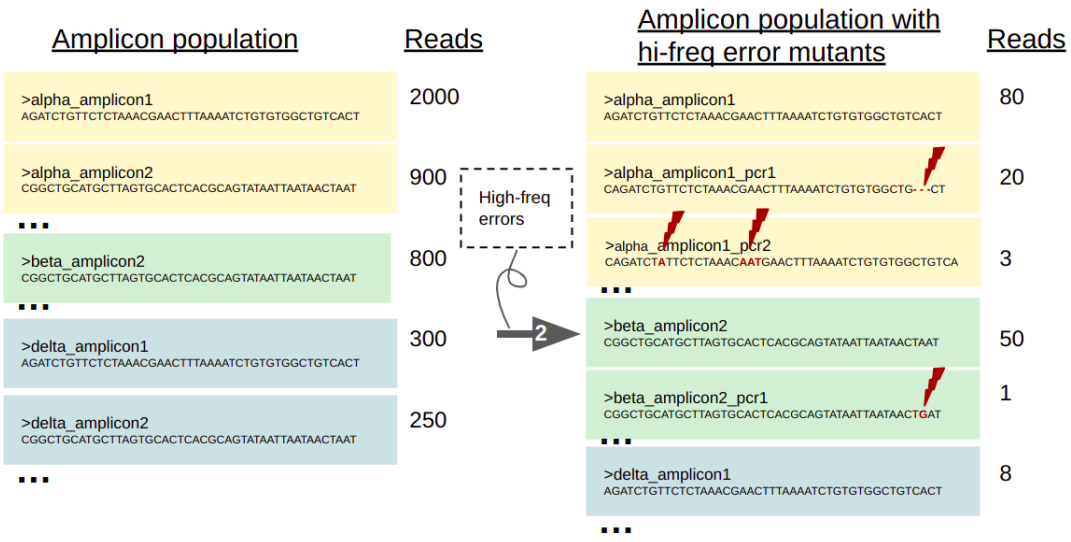

Figure 3.2: Step 2 of SWAMPy workflow.

3.3.2.1 Simulating individual errors

We use the error models and estimated parameters from Chapter 2 at this stage. We create a table like that shown in Table 3.1 that contains all targeted errors to introduce:

1. The number of each type of error to be introduced is sampled from Poisson($L \times R$) where L is the length of the Wuhan reference genome Wuhan-Hu-1 [15] and R is the error rate. Error rates are user-definable for each of the six types of error, with default values estimated from real wastewater experiments as described in Chapter 2 (See table 2.1a).

2. A genome position for each error is sampled randomly without replacement from Wuhan-Hu-1. For unique errors, one of the source genomes is randomly assigned with sampling weights equal to the genome abundances in the mixture. Moreover, if more than one amplicon spans the previously determined error position, a unique error is assigned to only one of them.

3. An error length is assigned to each error. The error length is always 1 for substitutions while it is sampled from a geometric distribution, Geometric(p), for deletions. p is the probability parameter of the geometric distribution and higher p will result in shorter deletions. For insertions, it is sampled from Uniform(*max*) where *max* is the maximum insertion length. Error length parameters p and *max* can be defined by the user, with their default values obtained from real data (see Chapter 2).

4. An alternative allele is created for each error. For substitutions, this is a random single nucleotide that is different from the reference genome, and for insertions, it is a sequence of randomly sampled nucleotides of the previously determined error length.

5. A variant allele frequency (VAF) is sampled for each error from a Beta distribution as (*VAF*, $1-$*VAF*) $\sim$ Beta(α, β) (see Chapter 2). Similarly, Beta parameters are user-definable separately for unique and recurrent substitutions, insertions and deletions. Assigned VAF values are the expected VAF of the recurrent errors in the final mixture, while for unique errors, the expected value of the VAF in the final mixture will be *(assigned VAF)* $\times$ *(amplicon abundance)*

Table 3.1: Examples of High-frequency Errors

type	rec/u	genome	len	pos	ref	alt	VAF	amp
subs	rec	g1,g2,g3	1	20,000	A	T	0.1	70,71
subs	u	g2	1	530	T	G	0.2	3
ins	rec	g1,g2,g3	7	245	A	AGCG	0.9	2
del	u	g3	3	230	AGCT	A	0.6	2
...								

(a) Abbreviations: rec: recurrent, u: unique, subs: substitution, del: deletion, ins: insertion, amp: amplicon number, len: length, alt: alternative allele, pos: genomic position, gX: SARS-CoV-2 variant genome identifier

3.3.2.2 Creating mutant versions

After we compile the table that contains all simulated errors, one by one, we process all amplicons in the amplicon population that we previously created. For each amplicon:

1. Errors that correspond to this amplicon are selected from the error table.

2. Because simulated error positions are based on the Wuhan-Hu-1 reference and a variant amplicon in a wastewater sample may contain indels, the amplicon sequences are aligned to Wuhan-Hu-1 using Bowtie 2 [40].

3. Positions of each error in the amplicon relative to the amplicon start are determined taking into account indels that the variant may have and genomic position of the errors.

4. The number of reads in which each individual error will be present in the mixture is determined by sampling a read count n_e for each error from Binomial(N, *VAF*) where N is the read count of the amplicon and *VAF* is the variant allele frequency of the error. At this stage, the sum of error read counts can be larger

than the total read count of the amplicon because some reads will contain multiple errors.

5. Different mutant versions i of this amplicon bearing different error combinations, and their read count n_i is determined. To determine the n_i, all errors are allowed to separately and randomly sample n_e reads from the reads of the amplicon. The result of this sampling is grouped first by the reads (because the same read might be sampled by multiple errors), and then different combinations of errors to finally obtain a read count for each combination. Then the read count of the wild-type amplicon is $N - \sum_{i=1}^{p} n_i$ where N is the total read count of that amplicon and n_i are the read counts of the mutant versions where there are p mutant versions of the amplicon.

6. Finally, the amplicon sequence is modified with the corresponding sets of errors to create mutant versions.

3.3.3 Step 3 - Creating sequencing reads

Now that the diversified amplicon population is known, the next step is to create sequencing reads (Figure 3.3). At this stage, we need to add sequencing errors and base quality simulation to the data. This is sequencing platform-dependent and is thoroughly studied by other researchers. Existing tools are good at performing this task, and hence we used an external tool within the SWAMPy workflow at this step. We use ART [39] in amplicon mode to create paired-end sequencing reads from each amplicon. We also suppressed creating alignment files (because ART produces alignment files as default, but SWAMPy only needs FASTQ files) with "noALN" flag and also used "maskN" flag to faithfully transcribe the N characters appearing in the source genome files. Other ART options such as read length (default:150), sequencing machine type (default:MiSeq V3), total read count (default:100K) and random seed can all be taken as inputs by SWAMPy and passed to ART. At the end of step 3, we have a pool of forward and reverse FASTQ files created by ART.

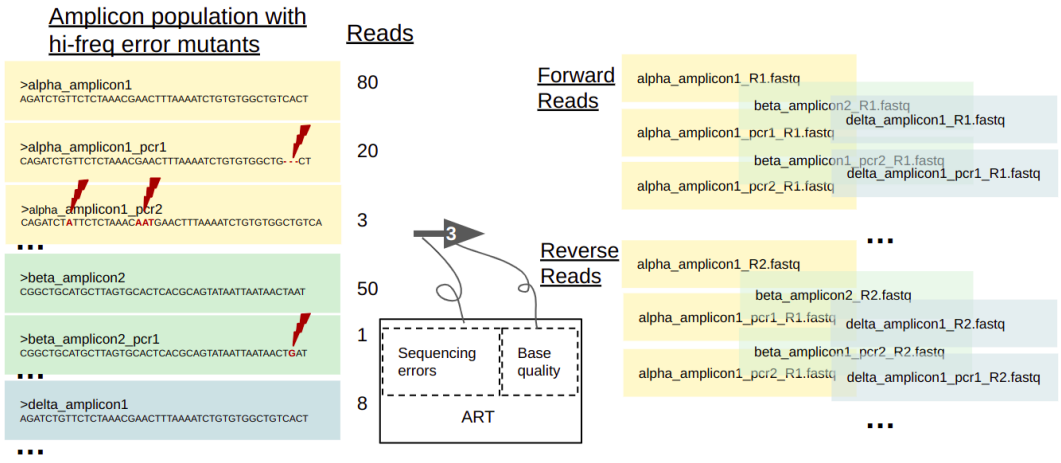

Figure 3.3: Step 3 of SWAMPy workflow.

3.3.4 Step 4 - Merging and shuffling

At this last step (Figure 3.4), the program merges and shuffles all the forward and the reverse reads separately to create a single forward and a single reverse FASTQ file to avoid potential biases in case a downstream application software uses reads with the order in the file.

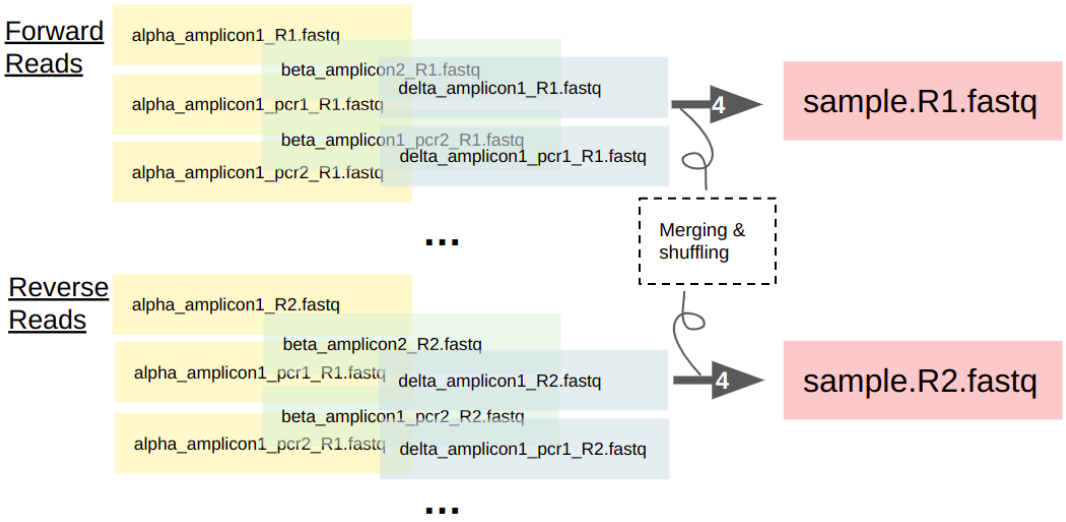

Figure 3.4: Step 4 of SWAMPy workflow.

3.3.5 Downstream application

We wanted to see if SWAMPy can be potentially useful in scenarios where we are testing a demixing method over a realistic series of samples. For that, I wanted to test the program called Freyja [38] which can demix a mixed wastewater sample into individual variants and estimate their relative abundances. I used SWAMPy to simulate 73 time points throughout the course of a hypothetical SARS-CoV-2 pandemic where the Alpha variant (B.1.1.7) starts out dominant before Delta (AY.4) rises in frequency and then Omicron (BA.1.1) emerges and takes over. The exact abundances at each time point can be seen in Appendix B. Then I used Frejya to demix the simulated samples and estimate variant abundances. I then compared simulation abundaces with Frejya estimations.

3.4 Results

3.4.1 SWAMPy

The source code of our python implementation of SWAMPy, together with the program documentation and example files is available under the GPL-v3 license at: https://github.com/goldman-gp-ebi/sars-cov-2-metagenomic-simulator.

SWAMPy takes as input a multi-FASTA file (figure 3.5) containing the SARS-CoV-2 variant genomes that will be present in the simulated wastewater sample, as well as a file that contains the relative abundances of these variants in the mixture as input (figure 3.6). For ease of use, other input files (primer set BED, FASTQ, and primer set specific amplicon distribution files) were wrapped with a single "–primer-set" parameter which loads the corresponding input files of the chosen primer set. As of August 2022, there are three supported primer sets: ARTIC V1, ARTIC V4 [47] and Nimagen V2 [23]. There are many command line parameters that allow fine control of the program such as the parameter c that reflects the quality of the wastewater sample, the total number of simulated reads, and error rates, VAF and lengths of high-frequency errors. The full list of command line interface arguments and their explanations are available on the GitHub wiki page "CLI-arguments".

Figure 3.5: Example genomes multi-fasta file showing the variants from the section 3.3.5. (The first 2 variants are randomly shortened, and the last variant is truncated)

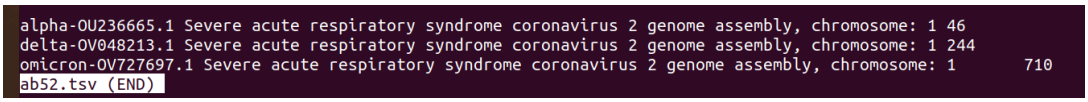

Figure 3.6: Example abundances file showing the 53rd time point from section 3.3.5

SWAMPy produces five output files as default:

- A forward and a reverse FASTQ file of the simulated reads, matching Illumina standards
- A table that shows the abundance of each wild-type amplicon after the randomness in amplicon copy number sampling (as described in 3.2) took effect
- A VCF file that contains all the targeted high-frequency errors from the error table described in Chapter 2.
- A log file

As an example, it is visible in the alignment images of 53rd time point of pandemic simulation output that the major characteristics of wastewater data are present in the simulated data. Overlapping amplicon structure and the variation in coverage across

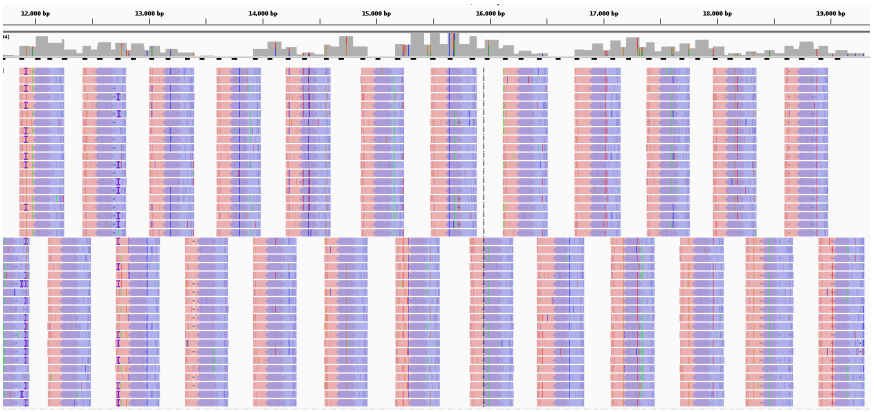

Figure 3.7: IGV images of SWAMPy simulated reads. The gray track at the top shows the coverage. Pink and blue chunks are forward and reverse reads respectively. Paired-end overlapping forward and reverse reads make up one amplicon. Amplicons overlap not to leave gaps.

the genome is visible in figure 3.7. SNPs between the variants in the mixture is also visible in the data. Figure 3.8A shows a C->T SNP at 7124th position on the Delta variant but not present in the Alpha and the Omicron variants. In figure 3.8B Sequencing errors added by ART [39] are visible in high density towards the read ends as a trademark of the Illumina machines. Figure 3.8C shows the unique high-frequency errors which are characterised by their presence in only one of the source genomes in the mixture and not crossing the amplicon boundary if it falls on an amplicon overlap region. Figure 3.8D shows the recurrent high-frequency errors that are characterised by their presence in all source genomes and both amplicons of an overlap region. Finally, the source genome of a read can be seen in IGV (figure 3.9) as well as in FASTQ files, which of course is not possible in real wastewater sequencing experiments.

3.4.2 Downstream application

Results suggest that SWAMPy can be useful in testing the methods of downstream applications. We saw that Freyja is quite successful in demixing the simulated data overall and finds all major features, though it sometimes finds in relatively high frequencies variants that are not present in the simulated mixture (Figure 3.10). This stems from misclassification of some variants as others probably because of some

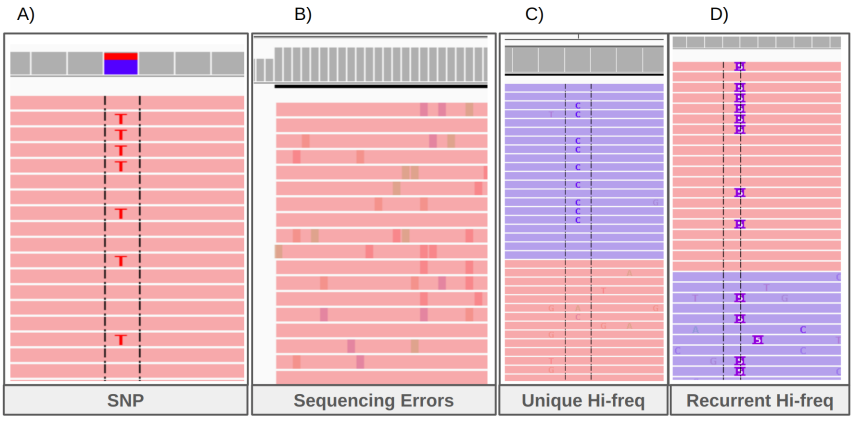

Figure 3.8: IGV images of SWAMPy simulated reads. A) A real SNP between different SARS-CoV-2 variants. B) Sequencing errors added by ART. This image is from a read end where the sequencing error density is higher. C) A unique high-frequency error on an amplicon overlap region. It is seen that only one of the amplicons carries the error. D) Recurrent high-frequency error on an amplicon overlap region. Both amplicons carry the error.

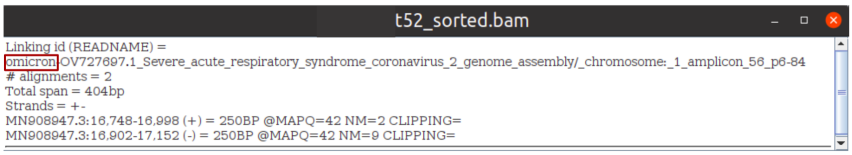

Figure 3.9: Information regarding the source genome of a particular read can be retrieved from the SWAMPy output.

high-frequency errors in the simulated data. This feature of Freyja can be further investigated by comparing Freja identifications with actual individual mutations in the simulated data.

3.5 Discussion

SWAMPy is an easy-to-use and reasonably fast wastewater SARS-CoV-2 simulator that takes into account the major characteristics of this type of data like a mixture of different variants in different proportions, differential amplicon abundance pattern of different primer sets and PCR and RNA degradation errors at high frequencies.

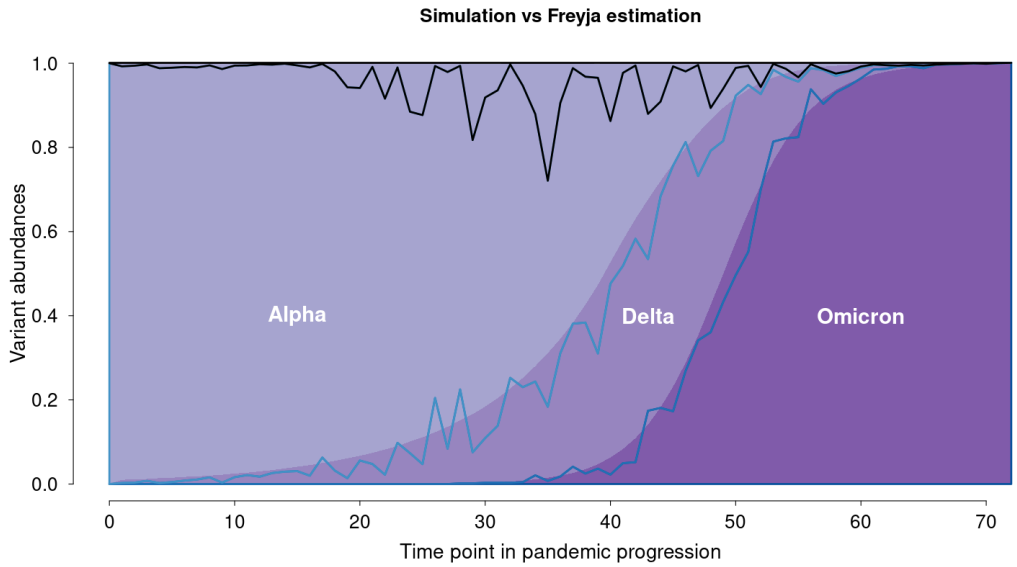

Figure 3.10: Progression of a simulated pandemic with 73 time points and Freyja estimations. Background colours represent the values given as input to SWAMPy simulations and lines represent the Freyja estimations.

SWAMPy fills a niche that other simulators do not and enables researchers working on wastewater SARS-CoV-2 studies to test and benchmark their methods.

In the future, SWAMPy can be improved by taking into account the following considerations:

In real data, it is expected that different mutant versions of the amplicons can be placed on a tree and error combinations are not random. However, since constructing a tree for each amplicon would be computationally too expensive, we took a more simplistic approach where we determined the error combinations randomly respecting the summary statistics such as VAF.

We assign an alternative allele to an error with respect to the Wuhan-Hu-1 reference instead of the variant genome present because otherwise, it would not be possible to create recurrent errors. But this means a small proportion of the errors will not show as a variant in the simulated data if the SARS-CoV-2 variant happens to share the same variation naturally. But this will occur only rarely on a proportion of all simulated errors, so it should not be a serious problem. Similarly, when an error is

unique, it is constrained in only one of the variant genomes in the mixture, which systematically understates the VAF of unique errors in the overall mixture. Research on the biological origins and mechanisms of the recurrent and unique errors can help us create better error models, which in turn can solve these problems.

Moreover, during amplicon creation in Step 1, we use default Bowtie 2 settings which can be too stringent and cause an amplicon to drop out when there is a variant in the primer binding site while in reality there might be some flexibility. In the future, primer binding conditions can be relaxed to allow some errors.

Finally, we want to add more primer set support to the program in the future.

3.6 Other improvements

I made many modifications to the program as well as bug fixes. Some important ones are as follows:

- I improved the performance in Step 3 by merging the runs of art_illumina of ART program. This decreased the number of times we had to invoke the program and saved a good amount of time on initializing and exiting the program. Table 3.2 shows the run times before and after the enhancement.
- I modified the way that internal temporary files are handled. This helped fix a bug that is caused by the special characters in variant genome names by creating an internal escaped version of the file. More importantly, tidying up temporary file handling enabled simultaneous executions of SWAMPy without interference from one run to another as long as they have unique "–temp_folder" parameters. This enabled using SWAMPy in workflows such as a Snakemake rule, or when submitting simultaneous LSF or SLURM jobs.
- I modified the command line arguments for ease of use, including for example wrapping three dependent and confusing arguments with a single and easy "–primer_set" argument.
- I prepared example input files from publicly available data. This enabled users to have a ready-to-run example case.

- I tested the program on multiple MAC and Linux devices. This enabled me to work out some conflicts caused by non-compatible versions of the dependency programs as well as OS-related dependencies. This information enabled me to prepare a compatible Conda environment and guide the user about the dependencies, which increased reproduciblity greatly.
- I improved documentation, organized already existing content in wiki pages and added missing content.

Table 3.2: Run times. The first row is before the enhancement. Other rows are after the enhancement. Columns show the number of genomes in the simulation mixture, hi-frequency error rates and run times of step 2 and step 3 of the workflow as well as the total run time. The decrease of step 3 run times is prominent after the enhancement

#genomes	error rate	Step 2	Step 3	total
15	default	2.5m	8m	11m
15	default	2m	0.5m	3m
3	default	0.5m	0.5m	1m 40s
3	subs x10	50s	0.5m	1.5m
3	subs x 100	5m	10s	6.5m

3.7

REFERENCES

[1] M. Lorenzo and Y. Picó, "Wastewater-based epidemiology: current status and future prospects," 2019.

[2] G. La Rosa, S. Della Libera, M. Iaconelli, A. R. Ciccaglione, R. Bruni, S. Taffon, M. Equestre, V. Alfonsi, C. Rizzo, M. E. Tosti, M. Chironna, L. Romanò, A. R. Zanetti, and M. Muscillo, "Surveillance of hepatitis A virus in urban sewages and comparison with cases notified in the course of an outbreak, Italy 2013," *BMC Infectious Diseases*, vol. 14, pp. 1–11, jul 2014.

[3] M. Hellmér, N. Paxéus, L. Magnius, L. Enache, B. Arnholm, A. Johansson, T. Bergström, and H. Norder, "Detection of pathogenic viruses in sewage provided early warnings of hepatitis A virus and norovirus outbreaks," *Applied and Environmental Microbiology*, vol. 80, no. 21, pp. 6771–6781, 2014.

[4] "Role of environmental poliovirus surveillance in global polio eradication and beyond," *Epidemiology & Infection*, vol. 140, pp. 1–13, jan 2012.

[5] H. Asghar, O. M. Diop, G. Weldegebriel, F. Malik, S. Shetty, L. E. Bassioni, A. O. Akande, E. A. Maamoun, S. Zaidi, A. J. Adeniji, C. C. Burns, J. Deshpande, M. S. Oberste, and S. A. Lowther, "Environmental Surveillance for Polioviruses in the Global Polio Eradication Initiative," *The Journal of Infectious Diseases*, vol. 210, pp. S294–S303, nov 2014.

[6] P. M. Lago, H. E. Gary, L. S. Pérez, V. Cáceres, J. B. Olivera, R. P. Puentes, M. B. Corredor, P. Jímenez, M. A. Pallansch, and R. G. Cruz, "Poliovirus detection in wastewater and stools following an immunization campaign in Havana, Cuba," *International Journal of Epidemiology*, vol. 32, pp. 772–777, oct 2003.

[7] R. S. Hendriksen, P. Munk, P. Njage, B. van Bunnik, L. McNally, O. Lukjancenko, T. Röder, D. Nieuwenhuijse, S. K. Pedersen, J. Kjeldgaard, R. S. Kaas, P. T. L. C. Clausen, J. K. Vogt, P. Leekitcharoenphon, M. G. van de Schans,

T. Zuidema, A. M. de Roda Husman, S. Rasmussen, B. Petersen, A. Bego, C. Rees, S. Cassar, K. Coventry, P. Collignon, F. Allerberger, T. O. Rahube, G. Oliveira, I. Ivanov, Y. Vuthy, T. Sopheak, C. K. Yost, C. Ke, H. Zheng, L. Baisheng, X. Jiao, P. Donado-Godoy, K. J. Coulibaly, M. Jergović, J. Hrenovic, R. Karpíšková, J. E. Villacis, M. Legesse, T. Eguale, A. Heikinheimo, L. Malania, A. Nitsche, A. Brinkmann, C. K. S. Saba, B. Kocsis, N. Solymosi, T. R. Thorsteinsdottir, A. M. Hatha, M. Alebouyeh, D. Morris, M. Cormican, L. O'Connor, J. Moran-Gilad, P. Alba, A. Battisti, Z. Shakenova, C. Kiiyukia, E. Ng'eno, L. Raka, J. Avsejenko, A. Bērziņš, V. Bartkevics, C. Penny, H. Rajandas, S. Parimannan, M. V. Haber, P. Pal, G. J. Jeunen, N. Gemmell, K. Fashae, R. Holmstad, R. Hasan, S. Shakoor, M. L. Z. Rojas, D. Wasyl, G. Bosevska, M. Kochubovski, C. Radu, A. Gassama, V. Radosavljevic, S. Wuertz, R. Zuniga-Montanez, M. Y. Tay, D. Gavačová, K. Pastuchova, P. Truska, M. Trkov, K. Esterhuyse, K. Keddy, M. Cerdà-Cuéllar, S. Pathirage, L. Norrgren, S. Örn, D. G. Larsson, T. V. der Heijden, H. H. Kumburu, B. Sanneh, P. Bidjada, B. M. Njanpop-Lafourcade, S. C. Nikiema-Pessinaba, B. Levent, J. S. Meschke, N. K. Beck, C. D. Van, N. D. Phuc, D. M. N. Tran, G. Kwenda, D. adjim Tabo, A. L. Wester, S. Cuadros-Orellana, C. Amid, G. Cochrane, T. Sicheritz-Ponten, H. Schmitt, J. R. M. Alvarez, A. Aidara-Kane, S. J. Pamp, O. Lund, T. Hald, M. Woolhouse, M. P. Koopmans, H. Vigre, T. N. Petersen, and F. M. Aarestrup, "Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage," *Nature Communications 2019 10:1*, vol. 10, pp. 1–12, mar 2019.

[8] "Survival of surrogate coronaviruses in water," *Water Research*, vol. 43, pp. 1893–1898, apr 2009.

[9] A. W. Chan, J. Naphtali, and H. E. Schellhorn, "High-throughput DNA sequencing technologies for water and wastewater analysis," *Science Progress*, vol. 102, pp. 351–376, dec 2019.

[10] Ł. Jatowiecki, J. M. Chojniak, E. Dorgeloh, B. Hegedusova, H. Ejhed, J. Magnér, and G. A. Płaza, "Microbial Community Profiles in Wastewaters from Onsite Wastewater Treatment Systems Technology," *PLOS ONE*, vol. 11, p. e0147725, jan 2016.

[11] C. O. Osunmakinde, R. Selvarajan, B. B. Mamba, and T. A. Msagati, "Profiling Bacterial Diversity and Potential Pathogens in Wastewater Treatment Plants Using High-Throughput Sequencing Analysis," *Microorganisms*, vol. 7, nov 2019.

[12] S. Begmatov, A. G. Dorofeev, V. V. Kadnikov, A. V. Beletsky, N. V. Pimenov, N. V. Ravin, and A. V. Mardanov, "The structure of microbial communities of activated sludge of large-scale wastewater treatment plants in the city of Moscow," *Scientific Reports 2022 12:1*, vol. 12, pp. 1–14, mar 2022.

[13] J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, D. R. Burton, L. L. Lewis-Ximenez, J. G. De Jesus, M. Giovanetti, S. C. Hill, A. Black, T. Bedford, M. W. Carroll, M. Nunes, L. C. Alcantara, E. C. Sabino, S. A. Baylis, N. R. Faria, M. Loose, J. T. Simpson, O. G. Pybus, K. G. Andersen, and N. J. Loman, "Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples," *Nature Protocols 2017 12:6*, vol. 12, pp. 1261–1276, may 2017.

[14] P. Zhou, X. L. Yang, X. G. Wang, B. Hu, L. Zhang, W. Zhang, H. R. Si, Y. Zhu, B. Li, C. L. Huang, H. D. Chen, J. Chen, Y. Luo, H. Guo, R. D. Jiang, M. Q. Liu, Y. Chen, X. R. Shen, X. Wang, X. S. Zheng, K. Zhao, Q. J. Chen, F. Deng, L. L. Liu, B. Yan, F. X. Zhan, Y. Y. Wang, G. F. Xiao, and Z. L. Shi, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature 2020 579:7798*, vol. 579, pp. 270–273, feb 2020.

[15] F. Wu, S. Zhao, B. Yu, Y. M. Chen, W. Wang, Z. G. Song, Y. Hu, Z. W. Tao, J. H. Tian, Y. Y. Pei, M. L. Yuan, Y. L. Zhang, F. H. Dai, Y. Liu, Q. M. Wang, J. J. Zheng, L. Xu, E. C. Holmes, and Y. Z. Zhang, "A new coronavirus associated with human respiratory disease in China," *Nature 2020 579:7798*, vol. 579, pp. 265–269, feb 2020.

[16] WHO, "Public Health Surveillance for COVID-19," *Interim guidance*, no. February, pp. 253–278, 2022.

[17] H. G. Barreto, F. A. de Pádua Milagres, G. C. de Araújo, M. M. Daúde, and V. A. Benedito, "Diagnosing the novel SARS-CoV-2 by quantitative RT-PCR: vari-

ations and opportunities," *Journal of Molecular Medicine (Berlin, Germany)*, vol. 98, p. 1727, dec 2020.

[18] A. Sheikh, J. McMenamin, B. Taylor, and C. Robertson, "SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness," *The Lancet*, vol. 397, pp. 2461–2462, jun 2021.

[19] World Health Organization, "Laboratory testing for coronavirus disease 2019 (COVID-19) in suspected human cases," no. March, pp. 1–7, 2020.

[20] R. S. Y. Wong, "COVID-19 testing and diagnosis: A comparison of current approaches," *Malays J Pathol*, vol. 43, no. 1, pp. 3–8, 2021.

[21] S. H. Lee, J. McGrath, S. P. Connolly, and J. Lambert, "Partial N Gene Sequencing for SARS-CoV-2 Verification and Pathway Tracing," *International Medical Case Reports Journal*, vol. 14, p. 1, 2021.

[22] J. Yavarian, A. Nejati, V. Salimi, N. Z. S. Jandaghi, K. Sadeghi, A. Abedi, A. S. Zarchi, M. M. Gouya, and T. Mokhtari-Azad, "Whole genome sequencing of SARS-CoV2 strains circulating in Iran during five waves of pandemic," *PLOS ONE*, vol. 17, p. e0267847, may 2022.

[23] J. P. Coolen, F. Wolters, A. Tostmann, L. F. van Groningen, C. P. Bleeker-Rovers, E. C. Tan, N. van der Geest-Blankert, J. L. Hautvast, J. Hopman, H. F. Wertheim, J. C. Rahamat-Langendoen, M. Storch, and W. J. Melchers, "SARS-CoV-2 whole-genome sequencing using reverse complement PCR: For easy, fast and accurate outbreak and variant analysis.," *Journal of Clinical Virology*, vol. 144, pp. 1386–6532, nov 2021.

[24] V. Hourdel, A. Kwasiborski, C. Balière, S. Matheus, C. F. Batéjat, J. C. Manuguerra, J. Vanhomwegen, and V. Caro, "Rapid Genomic Characterization of SARS-CoV-2 by Direct Amplicon-Based Sequencing Through Comparison of MinION and Illumina iSeq100TM System," *Frontiers in Microbiology*, vol. 11, p. 2354, sep 2020.

[25] T. Liu, Z. Chen, W. Chen, X. Chen, M. Hosseini, Z. Yang, J. Li, D. Ho, D. Turay, C. P. Gheorghe, W. Jones, and C. Wang, "A benchmarking study of SARS-

CoV-2 whole-genome sequencing protocols using COVID-19 patient samples," *iScience*, vol. 24, p. 102892, aug 2021.

[26] D. J. Toth and K. Khader, "Efficient SARS-CoV-2 surveillance strategies to prevent deadly outbreaks in vulnerable populations," *BMC Medicine*, vol. 19, dec 2021.

[27] R. Wölfel, V. M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M. A. Müller, D. Niemeyer, T. C. J. Kelly, P. Vollmar, C. Rothe, M. Hoelscher, T. Bleicker, S. Brünink, J. Schneider, R. Ehmann, K. Zwirglmaier, C. Drosten, and C. Wendtner, "Virological assessment of hospitalized cases of coronavirus disease 2019," *medRxiv*, p. 2020.03.05.20030502, mar 2020.

[28] Y. Chen, L. Chen, Q. Deng, G. Zhang, K. Wu, L. Ni, Y. Yang, B. Liu, W. Wang, C. Wei, J. Yang, G. Ye, and Z. Cheng, "The presence of SARS-CoV-2 RNA in the feces of COVID-19 patients," *Journal of medical virology*, vol. 92, pp. 833–840, jul 2020.

[29] N. Sims and B. Kasprzyk-Hordern, "Future perspectives of wastewater-based epidemiology: Monitoring infectious disease spread and resistance to the community level," *Environment International*, vol. 139, p. 105689, jun 2020.

[30] Y. Deng, X. Xu, X. Zheng, J. Ding, S. Li, H. kwong Chui, T. kin Wong, L. L. Poon, and T. Zhang, "Use of sewage surveillance for COVID-19 to guide public health response: A case study in Hong Kong," *The Science of the Total Environment*, vol. 821, p. 153250, may 2022.

[31] W. Ahmed, N. Angel, J. Edson, K. Bibby, A. Bivins, J. W. O'Brien, P. M. Choi, M. Kitajima, S. L. Simpson, J. Li, B. Tscharke, R. Verhagen, W. J. Smith, J. Zaugg, L. Dierens, P. Hugenholtz, K. V. Thomas, and J. F. Mueller, "First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community," *Science of The Total Environment*, vol. 728, p. 138764, aug 2020.

[32] D. S. Smyth, M. Trujillo, D. A. Gregory, K. Cheung, A. Gao, M. Graham, Y. Guan, C. Guldenpfennig, I. Hoxie, S. Kannoly, N. Kubota, T. D. Lyddon, M. Markman, C. Rushford, K. M. San, G. Sompanya, F. Spagnolo, R. Suarez,

E. Teixeiro, M. Daniels, M. C. Johnson, and J. J. Dennehy, "Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater," *Nature Communications 2022 13:1*, vol. 13, pp. 1–9, feb 2022.

[33] G. Medema, L. Heijnen, G. Elsinga, R. Italiaander, and A. Brouwer, "Presence of SARS-Coronavirus-2 in sewage," *medRxiv*, p. 2020.03.29.20045880, jul 2020.

[34] W. Randazzo, E. Cuevas-Ferrando, R. Sanjuán, P. Domingo-Calap, and G. Sánchez, "Metropolitan wastewater analysis for COVID-19 epidemiological surveillance," *International Journal of Hygiene and Environmental Health*, vol. 230, p. 113621, sep 2020.

[35] C. C. Naughton, F. A. Roman, A. F. Grace Alvarado, A. Q. Tariqi, M. A. Deeming, K. Bibby, A. Bivins, J. B. Rose, G. Medema, W. Ahmed, P. Katsivelis, V. Allan, R. Sinclair, Y. Zhang, M. N. Kinyua, and C. Author cnaughton, "Show us the Data: Global COVID-19 Wastewater Monitoring Efforts, Equity, and Gaps," *medRxiv*, p. 2021.03.14.21253564, nov 2021.

[36] M. R. Brown, M. J. Wade, S. McIntyre-Nolan, I. Bassano, H. Denise, D. Bass, J. Bentley, J. T. Bunce, J. Grimsley, A. Hart, T. Hoffmann, A. Jeffries, S. Paterson, M. Pollock, J. Porter, D. Smith, R. van Aerle, G. Watts, A. Engeli, and G. Henderson, "Wastewater Monitoring of SARS-CoV-2 Variants in England: Demonstration Case Study for Bristol (Dec 2020-March 2021)," no. March, pp. 1–9, 2021.

[37] D. A. Gregory, C. G. Wieberg, J. Wenzel, C. H. Lin, and M. C. Johnson, "Monitoring SARS-CoV-2 Populations in Wastewater by Amplicon Sequencing and Using the Novel Program SAM Refiner," *Viruses*, vol. 13, aug 2021.

[38] S. Karthikeyan, J. I. Levy, P. D. Hoff, G. Humphrey, A. Birmingham, K. Jepsen, S. Farmer, H. M. Tubb, T. Valles, C. E. Tribelhorn, R. Tsai, S. Aigner, S. Sathe, N. Moshiri, B. Henson, A. Hakim, N. A. Baer, T. Barber, P. Belda-Ferre, M. Chacón, W. Cheung, E. S. Cresini, E. R. Eisner, A. L. Lastrella, E. S. Lawrence, C. A. Marotz, T. T. Ngo, T. Ostrander, A. Plascencia, R. A. Salido, P. Seaver, E. W. Smoot, D. McDonald, R. M. Neuhard, A. L. Scioscia, A. M. Sat-

terlund, E. H. Simmons, C. M. Aceves, C. Anderson, K. Gangavarapu, E. Hufbauer, E. Kurzban, J. Lee, N. L. Matteson, E. Parker, S. A. Perkins, K. S. Ramesh, R. Robles-Sikisaka, M. A. Schwab, E. Spencer, S. Wohl, L. Nicholson, I. H. Mchardy, D. P. Dimmock, C. A. Hobbs, O. Bakhtar, A. Harding, A. Mendoza, A. Bolze, D. Becker, E. T. Cirulli, M. Isaksson, K. M. S. Barrett, N. L. Washington, J. D. Malone, A. M. Schafer, N. Gurfield, S. Stous, R. Fielding-Miller, R. Garfein, T. Gaines, C. Anderson, N. K. Martin, R. Schooley, B. Austin, S. F. Kingsmore, W. Lee, S. Shah, E. McDonald, M. Zeller, K. M. Fisch, L. Laurent, G. W. Yeo, K. G. Andersen, and R. Knight, "Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission," *medRxiv*, p. 2021.12.21.21268143, apr 2021.

[39] W. Huang, L. Li, J. R. Myers, and G. T. Marth, "ART: a next-generation sequencing read simulator," *BIOINFORMATICS APPLICATIONS NOTE*, vol. 28, no. 4, pp. 593–594, 2012.

[40] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, pp. 357–359, Apr 2012.

[41] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li, "Twelve years of SAMtools and BCFtools," *GigaScience*, vol. 10, 02 2021. giab008.

[42] N. Stoler and A. Nekrutenko, "Sequencing error profiles of Illumina sequencing instruments," *NAR Genomics and Bioinformatics*, vol. 3, no. 1, 2021.

[43] I. Jungreis, R. Sealfon, and M. Kellis, "SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes," *Research Square*, oct 2020.

[44] S. Delbue, S. D'Alessandro, L. Signorini, M. Dolci, E. Pariani, M. Bianchi, S. Fattori, A. Modenese, C. Galli, I. Eberini, and P. Ferrante, "Isolation of sars-cov-2 strains carrying a nucleotide mutation, leading to a stop codon in the orf 6 protein," *Emerging Microbes & Infections*, vol. 10, no. 1, pp. 252–255, 2021. PMID: 33525998.

[45] A. Meyerhans, J. P. Vartanian, and S. Wain-Hobson, "DNA recombination during PCR," *Nucleic Acids Research*, vol. 18, no. 7, pp. 1687–1691, 1990.

[46] V. Potapov and J. L. Ong, "Examining Sources of Error in PCR by Single-Molecule Sequencing," 2017.

[47] J. R. Tyson, P. James, D. Stoddart, N. Sparks, A. Wickenhagen, G. Hall, J. H. Choi, H. Lapointe, K. Kamelian, A. D. Smith, N. Prystajecky, I. Goodfellow, S. J. Wilson, R. Harrigan, T. P. Snutch, N. J. Loman, and J. Quick, "Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore," *bioRxiv*, 2020.

APPENDIX A

CHAPTER 2 APPENDICES

A.1 Deletion rate correction

Since we modelled deletions with geometric distribution, we correct for the lack of multiples of three as follows: First we show the sum of the geometric distribution probabilities equals to 1 in equation A.1. We find the missing part by summing the probabilities of multiples of three (equation A.2). We subtract this from one, which is the the proportion of the data visible to us. We take reciprocal of this value, which is the correction factor (equation A.3).

$$\sum_{i=1}^{\infty} p(1-p)^{i-1} = p(1 + (1-p) + (1-p)^2 + ...) = p\frac{1}{1-(1-p)} = 1 \qquad \text{(A.1)}$$

$$\begin{aligned} \sum_{i=1}^{\infty} p(1-p)^{3i-1} &= p(1-p)^2 \sum_{i=1}^{\infty} (1-p)^{3i-3} \\ &= p(1-p)^2 \sum_{i=1}^{\infty} \left((1-p)^3\right)^{i-1} \\ &= \frac{p(1-p)^2}{1-(1-p)^3} \end{aligned} \qquad \text{(A.2)}$$

$$\frac{1}{1-\frac{p(1-p)^2}{1-(1-p)^3}} = \frac{p^2-3p+3}{2-p} \qquad \text{(A.3)}$$

Table A.1: Extended Table1: Error counts, correction factors and error rates of different category of errors.

	Ww	Ww ratio	Clinical	Clinical ratio	Ww C	clinical C	Ww rate (rWw)	Clinical rate (rC)	rWw/ rC
rec ins	19	0,00002391688003	120	0,00001910982608	1,5	1,5	0,00003587532005	0,00002866473912	1,251548807
rec del	0	0	47	0,000007484681882	1,09	1,02	0	0,00000763437552	0
rec subs	127	0,0001598654613	15	0,00000238872826	21	21	0,003357174686	0,00005016329346	66,92492567
u ins	13	0,00001636418107	13	0,000002070231159	1,5	1,5	0,00002454627161	0,000003105346738	7,90451878
u del	91	0,0001145492675	18	0,000002866473912	1,09	1,02	0,0001248587016	0,00000292380339	42,70420576
u subs	94	0,000118325617	10	0,000001592485507	21	21	0,002484837957	0,00003344219564	74,30247653

A.2 Extended table: Error rates

A.3 Clinical data accessions

ERR8168754, ERR8170495, ERR8171952, ERR8172380, ERR8172664, ERR8173717, ERR8174217, ERR8174267, ERR8175020, ERR8177835, ERR8178034, ERR8178181, ERR8178260, ERR8178462, ERR8178707, ERR8179489, ERR8179838, ERR8179914, ERR8180400, ERR8180978, ERR8181059, ERR8181416, ERR8181818, ERR8182040, ERR8182190, ERR8182225, ERR8182466, ERR8182735, ERR8182804, ERR8183137, ERR8183172, ERR8183270, ERR8183459, ERR8183652, ERR8183684, ERR8183812, ERR8184057, ERR8184081, ERR8184103, ERR8184602, ERR8184624, ERR8184731, ERR8184801, ERR8184998, ERR8185096, ERR8185446, ERR8185448, ERR8185564, ERR8185572, ERR8185744, ERR8186224, ERR8186514, ERR8186625, ERR8186725, ERR8186996, ERR8187280, ERR8187434, ERR8187455, ERR8187774, ERR8187822, ERR8188724, ERR8188829, ERR8189531, ERR8189532, ERR8190510, ERR8190901, ERR8191040, ERR8193067, ERR8193281, ERR8193536, ERR8193795, ERR8193923, ERR8193959, ERR8193968, ERR8193977, ERR8193984, ERR8194147, ERR8194370, ERR8194382, ERR8194567, ERR8194577, ERR8194684, ERR8195947, ERR8196499, ERR8197775, ERR8198192, ERR8198441, ERR8198449, ERR8199245, ERR8199491, ERR8200597, ERR8201792, ERR8201869, ERR8201913, ERR8201954, ERR8202329, ERR8202440, ERR8202843, ERR8203330, ERR8203435, ERR8203451, ERR8203571, ERR8204099, ERR8204175, ERR8204507, ERR8204545, ERR8205435, ERR8205710, ERR8206053, ERR8206510, ERR8206707, ERR8207215, ERR8207662, ERR8207724, ERR8207873, ERR8208169, ERR8208303, ERR8208378, ERR8208649, ERR8208675, ERR8209565, ERR8209600, ERR8210045, ERR8210056, ERR8210151, ERR8210181,

ERR8210259, ERR8210422, ERR8210844, ERR8210936, ERR8211527, ERR8211720, ERR8211810, ERR8212822, ERR8213127, ERR8213366, ERR8213419, ERR8213936, ERR8214041, ERR8215136, ERR8215343, ERR8215408, ERR8215643, ERR8215683, ERR8215816, ERR8216059, ERR8216454, ERR8216840, ERR8216993, ERR8217589, ERR8218906, ERR8218933, ERR8219172, ERR8219433, ERR8219792, ERR8220069, ERR8220265, ERR8220616, ERR8220911, ERR8220984, ERR8221552, ERR8221765, ERR8222149, ERR8222237, ERR8222493, ERR8222665, ERR8222871, ERR8222975, ERR8223010, ERR8223065, ERR8223069, ERR8223174, ERR8223290, ERR8223468, ERR8223562, ERR8223942, ERR8224674, ERR8225068, ERR8225193, ERR8225379, ERR8225573, ERR8225709, ERR8226180, ERR8226467, ERR8226555, ERR8227407, ERR8227770, ERR8227844, ERR8228313, ERR8228534, ERR8229153, ERR8229287, ERR8229981, ERR8230047, ERR8230164, ERR8230240, ERR8230291, ERR8230393, ERR8230483, ERR8230752, ERR8230768, ERR8230998, ERR8231358, ERR8232003, ERR8232116, ERR8232197, ERR8232215, ERR8232376, ERR8233661, ERR8233872, ERR8233986, ERR8234411, ERR8234498, ERR8234577, ERR8234609, ERR8235269, ERR8235949, ERR8235960, ERR8236055, ERR8236339, ERR8236360, ERR8236491, ERR8236679, ERR8236838, ERR8236923, ERR8237113, ERR8237256, ERR8238442, ERR8238541, ERR8238570, ERR8238808, ERR8239173, ERR8239265, ERR8239309, ERR8239632, ERR8239679, ERR8239697, ERR8240076, ERR8240203, ERR8240204, ERR8240247, ERR8240629, ERR8240808, ERR8241013, ERR8241052, ERR8241592, ERR8241933, ERR8242242, ERR8242827, ERR8242973, ERR8243271, ERR8243426, ERR8243615, ERR8244023, ERR8244969, ERR8245425, ERR8245438, ERR8245447, ERR8245642, ERR8246607, ERR8246772, ERR8246838, ERR8246969, ERR8247014, ERR8247358, ERR8247908, ERR8248045, ERR8248194, ERR8248643, ERR8248758, ERR8248999, ERR8249202, ERR8249280, ERR8250040, ERR8250430, ERR8250520, ERR8250728, ERR8250768, ERR8250895, ERR8250908, ERR8251378, ERR8251689, ERR8251848, ERR8251993, ERR8253002, ERR8253265, ERR8253458, ERR8254626, ERR8254962, ERR8255106, ERR8256094, ERR8256624, ERR8256915, ERR8257351, ERR8257601, ERR8258828, ERR8258939, ERR8259013, ERR8259912, ERR8261840

APPENDIX B

CHAPTER 3 APPENDICES

B.1 Pandemic simulation table

Table B.1: Simulated genome proportions

time	f(Alpha)	f(Delta)	f(Omicron)	time	f(Alpha)	f(Delta)	f(Omicron)
0	1000	0	0	37	617	357	26
1	990	10	0	38	574	391	35
2	989	11	0	39	526	426	47
3	988	12	0	40	474	464	63
4	986	14	0	41	420	497	83
5	985	15	0	42	370	521	109
6	983	17	0	43	324	535	141
7	982	18	0	44	280	538	182
8	980	20	0	45	239	530	231
9	978	22	0	46	201	511	288
10	975	25	0	47	166	481	353
11	973	27	0	48	134	441	425
12	970	30	0	49	106	395	499
13	967	33	0	50	82	344	574
14	963	37	0	51	62	293	645
15	959	41	0	52	46	244	710
16	955	45	0	53	34	198	768
17	950	50	0	54	24	159	817
18	945	55	0	55	17	125	858
19	939	61	0	56	12	97	891
20	933	67	0	57	8	75	917
21	926	74	0	58	6	57	937
22	918	82	0	59	4	44	952
23	909	90	0	60	3	33	964
24	900	100	1	61	2	25	973
25	889	110	1	62	1	19	980
26	877	122	1	63	1	14	985
27	864	135	1	64	1	10	989
28	849	149	2	65	0	8	992
29	833	164	2	66	0	6	994
30	815	181	3	67	0	4	995
31	795	200	4	68	0	3	997
32	773	221	6	69	0	2	998
33	748	244	8	70	0	2	998
34	721	269	11	71	0	1	999
35	690	296	15	72	0	0	1000
36	655	325	20				