Xinhui Cai Thesis

Computational identification of within-host diversity of SARS-CoV-2 and its benefits for mRNA vaccine design

Xinhui Cai

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney

Dec 28th, 2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Xinhui Cai, declare that this thesis is submitted in fulfillment of the requirements for the award of Master of Analytics, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 28/12/2022

Acknowledgment

Through out the time of this research I have received a great amount of help and support from many people, without them this research would be impossible. The first person I'd like to thank is my principle supervisor, prof.Jinyan Li. During my master studies, prof.Li helps me form the idea of identifying co-existing strains from the early SARS-CoV-2 sample and provides me valuable feedback and guidance at each stage of my research and helping me with the scholarship which reduces my financial burden considerably. I would also like to thank the other members in the team for their support during my research, especially Tian Lan and Pengyao Ping. Tian made a great contribution in obtaining the protein docking information and Pengyao develops an important part of the double-model error correction method. Dr.Xuan Zhang developed the InsEC which is a source of inspiration for the strain identification algorithm I designed, she also has given me some useful advice from time to time. My parents support me firmly during my research which I appreciate very much. The sequencing read information of SARS-CoV-2 reference genome sequence provided by Dr.Si Haorui from Prof.Zhengli Shi's lab is very important to my research and I appreciate it very much.

I would like to express my gratitude to my co-supervisor Prof.Brian Oliver for providing very helpful insights for both of my papers and playing an important role in forming some of the ideas presented in the papers. His knowledge in virology and immune system has been very helpful.

The Australian Research Council has provided financial support to this research through the scholarship I received. I would like to thank the organization for its recognition and generosity.

I also want to thank the panel of my candidate assessments, Prof.Wei Liu, Prof.Longbing Cao, Prof.Jinyan Li and Prof.Brian Oliver, who have shared their valuable opinions on each stage of my research.

Finally, for people special to me and support me throughout my study and life, I am very grateful and happy to have them.

Xinhui Cai

December 2022 in Sydney

Abstract

Since the emergence of COVID-19 in Wuhan in late 2019, the world has been largely affected. Despite extensive research about the virus itself (SARS-CoV-2) causing the pandemic, some questions regarding its emergence are still unaccounted for. This paper contains an investigation and analysis of the sequencing reads and the SARS-CoV-2 reference genome sequence assembled from those reads. By retracing the steps of assembling the reference sequence, a probability that multiple strains of SARS-CoV-2 co-existed inside the patient's body is found. The assembly tool, MEGAHIT, applies an assembly process that tends to ignore multiple routes to form contigs. Therefore, we design a workflow that could rectify this potential issue and identify multiple strains from the COVID-19 patient sample. It involves error correction, extracting relevant reads, strain identification, phylogenetic study, and protein structure analysis. The results indicate that more than one strain of SARS-CoV-2 could be produced from the sample that was used to produce the reference sequence. Their binding affinity and phylogenetic relationships with the published SARS-CoV-2 reference genome, SARS-CoV, and some other variants of SARS-CoV-2 are also revealed. The discovered strains show some possible structural differences that affect the protein binding affinity between the spike protein and human ACE2. Consequently, this workflow for identifying within-host diversity highlights the existence of co-existing strains with distinct nucleotide sequences, emphasizing the importance of considering these variations when designing mRNA vaccines.

Keywords: SARS-CoV-2 strains; error correction; Illumina short reads; phylogenetic study;mRNA vaccine design

Contents

Chapter 1

Introduction

1.1 Background

The ongoing COVID-19 pandemic is one of the most influential public health events in human history. Extensive research have been carried out on the genome sequences of SARS-CoV-2, the virus that caused the pandemic. The goal is to understand the mechanism of its infection, various therapies against the infection, and vaccine design. The majority of these research works are based on the reference genome sequence obtained by Zhou et al. [71] (GISAID accession ID EPI_ISL_402124) and on the one acquired by Wu et al. [62] (NCBI accession ID NC045512.2). For example, the reference genome sequence has been used for the mRNA vaccine design of BNT162b2 [59] and mRNA-1273 [11]. The reference genome sequence was assembled by MEGAHIT ([34]) from the sequencing reads (SRR11092062) produced from a BALF sample of an early patient of the pandemic. However, as the virus' spike protein mutates further from the original SARS-CoV-2 strain spike protein, the efficacy of mRNA vaccine declines[55]. This would also be the case if the selected sequence is different from the major circulating strain because the sequences will also be different.

1.2 Research Objectives

In this paper, I present a workflow that identifies the within-host diversity of SARS-CoV-2 from the BALF sample that produced the reference genome sequence. The whole process started with a two-model error correction with miREC and Karect([1]). The objective here is to rule out the errors in next-generation sequencing (NGS) that could confuse the strain identification process. After that, the algorithm was applied to the reference genome sequence and the reads to produce different strains of SARS-CoV-2 that coexisted in the BALF samples. The produced strains, along with the reference genome sequence, related coronaviruses to SARS-CoV-2 and other major variants were included in a phylogenetic tree to reveal the strains' phylogenetic relationships. The coexisting strains' structural properties like binding affinity will also be presented with computational tools.

1.3 Research Significance

The revealing of coexisting strains in the BALF sample taken in the early days of COVID-19 might present a clearer path of how SARS-CoV-2 emerges. By constructing a phylogenetic tree with the whole genome sequences of related coronaviruses and their spike gene sequences, it is easier to identify the possible ancestry relationship. When more closely related coronavirus sequences are discovered later, the discovered strains could provide more support for finding the correct ancestors for SARS-CoV-2 and inferring the evolution process.

This research will also demonstrate an algorithm that could assist in finding coexisting sequences that are missed by normal de Bruijn graphs based on de-novo assembly tools. When assemblers based on this algorithm are forming contigs by traversing through k-mer edges, the less confident edges are usually discarded as errors. Some de-Bruijn-graph-based tools try to solve this error by forming alternative paths, like Haploflow and PEHaplo, but did not work well on all datasets. The variant calling tools such as gatk instead focus on local sites instead of the whole sequence and generally present a set of isolated base changes. My method aims to identify local variant sites and also retain the global focus. The resulting contigs are produced from combining compatible local variant sites, which is determined by supporting reads. Therefore, by applying the workflow proposed in this research, within-host diversity will be reliably identified for other virus sequences and their corresponding samples. This is important for widespread RNA viruses such as influenza because new strains generally come from mutations that happen in vivo and

are transmitted to other hosts. Identifying co-existing strains from the sample could give researchers a head start on potential circulating strains in the future.

An important issue for within-host diversity identification is distinguishing between low-frequency variants and sequencing errors. Some tools like PEHaplo try to utilize error correction to preprocess the reads in order to remove errors in the variant calling process. In this research, a novel two-step error correction process is implemented to increase the reliability of reads as much as possible. The two error correction methods are proportional correction and Karect. However, each of the two has limits. The proportional correction could not correct reads with a frequency equal to one, which means the reads that only appear once; while Karect tends to introduce non-existent reads after correction. By combining the proportional correction method and Karect together, this process ensures no non-existing reads will be brought into the read set by using Karect alone and the reads that appear only once are also corrected.

Another important aspect of this research is applying protein binding affinity study to the identified co-existing strains' spike protein. The 3D structures of the spike proteins are obtained by AlphaFold2, an accurate machine-learning-based protein folding prediction tool. With the obtained protein structures, the docking process between spike protein and human ACE2 (hACE2) can be studied. Apart from identifying co-existing strains, this research also provides information about their binding affinities with hACE2 to show how transmissible they are compared to the original SARS-Cov-2 strain.

The identified strains could also influence vaccine designs, especially for the ones that use genome sequences like mRNA vaccines. If a more transmissible strain is identified inside the host sample, it would allow researchers to prepare earlier for possible future circulating strains by designing with the more transmissible strains' sequences. From the history of SARS-CoV-2 evolution, the more transmissible strain always replaced the less transmissible one, like Delta replaced the original strain, Omicron (BA.1) replaced Delta and later Omicron variants such as BA.2.75 and BA.2 replaced the original Omicron BA.1.

Other than SARS-CoV-2, I also applied the workflow to a Foot-to-Mouth Disease Virus (FMDV) sample, and identified multiple co-existing strains in it. This also demonstrates the workflow could be applied to other RNA viruses as well.

1.4 Motivation

COVID-19 has caused millions of deaths and massive socioeconomic damage. Even with the increasing vaccination rate and the decreasing severity of new dominating Omicron

sub-variants, it still poses pressure on the health system. Moreover, some questions about the virus behind this global pandemic, SARS-CoV-2, remain unanswered. One of the most controversial questions is the origin of this virus. Multiple assumptions have been made including lab manufacturing, genetic recombination in Huanan Seafood Market, and natural mutation happening in wild bats carrying other coronaviruses. I would like to study more about SARS-CoV-2's emergence and provide more information for solving the emergence of SARS-CoV-2. It could also help to prevent the next pandemic from resulting in such a great loss we experienced during COVID-19 by taking measures that could lower the chances of future potential pathogens. Few infectious diseases have caused global transmission like COVID-19 in human history, and the information about those pathogens has been lost as time elapsed or could not be obtained due to technological limitations. While for SARS-CoV-2, the sequencing technology and online research platforms like NCBI have made information unprecedentedly available and lasting, offering a great opportunity for studying this virus.

Vaccination is key to reducing the damage of SARS-CoV-2. Many debates revolved around different kinds of vaccines and vaccine platforms. I am interested in how different vaccine technologies work, and which one has the most potential to control COVID-19. The most crucial factor is efficiency, but other aspects including safety, versatility, manufacturing, delivery, and cost are also important. As the development of the SARS-CoV-2 vaccine becomes a global focus, almost all vaccine platforms are involved to design the vaccine, which offers a great opportunity for comparing them as well. Some of the latest technologies like mRNA vaccine and viral-vector vaccine will also have a testing ground. Also by studying different kinds of vaccines, I hope I could utilize more computational methods in improving vaccine designs.

The most commonly used vaccine around the world is the mRNA vaccine BNT162b2 produced by Pfizer. The design of BNT162b2 is based on the spike gene sequence of SARS-CoV-2 and comes with a high efficiency of preventing the infection of the original SARS-CoV-2 strain and preventing severe illness and death for different variants. The genome sequence is essential for mRNA vaccine design, as it is used as an antigen. This is why the correctness of pathogen sequences will determine the efficacy of mRNA vaccines. When new variants emerge from mutations, designing the corresponding sequence for mRNA vaccine quickly could largely accelerate achieving herd immunity and prevention of infection. Most mutations happen within the host, so identifying within-host diversity allows a quick start of vaccine design against potentially threatening variants.

1.5 Thesis Structure

This thesis contains 5 chapters. Chapter 1 is the background and introduction to my research, with the significance and motivations. Chapter 2 contains related works for SARS-CoV-2 vaccine design and within-host diversity. Chapter 3 describes the method I developed to identify within-host diversity from sequencing reads. Chapter 4 contains the result of the identification of SARS-CoV-2 within-host diversity from the BALF sample that produced the SARS-CoV-2 reference genome sequence and the downstream phylogenetic and protein binding studies of the discovered co-existing strains. Chapter 5 is the summary of the thesis and my research throughout the two years of my studying at UTS.

The structure of the thesis is outlined below:

1. Chapter 1: Introduction
    (a) Background
    (b) Research Objectives
    (c) Research Significance
    (d) Motivation
    (e) Thesis Structure
2. Chapter 2: Literature Review
    (a) Error Correction Algorithm
    (b) mRNA vaccines design and properties
    (c) Existed Within-host Diversity Identification Tools
3. Chapter 3: Computational discovery of intra-host coexisting strains of the SARS-CoV-2 reference strain: Motivation and methods
    (a) Data Accessibility
    (b) Workflow
4. Chapter 4: Computational discovery of intra-host coexisting strains of the SARS-CoV-2 reference strain: Results
    (a) MEGAHIT Analysis
    (b) Discovered Strain Differences
    (c) Comparison with Other Tools

(d) Phylogenetic Relationship Analysis

(e) Sequences Comparison

(f) Protein Binding Affinity Analysis

(g) Case Study: Wastewater Sample Co-existing Strain Identification

5. Chapter 5: Conclusions and Future Research Perspectives

1.6 Research Contribution

This research provides an optimization to the de-Bruijn-based genome assembler by identifying possible co-existing strains within the same sample with high sensitivity and reliability. The workflow also aims to call low frequency variants after double model error correction, and provide highly confident sequence that have full coverage with no gaps. The downstream analysis involves phylogenetic study of the identified strains and other variants of concerns and some batcoronaviruses, and the prediction of protein binding affinity between the identified strains' spike protein and human ACE2. With all the information obtained, an inference of the relationship between the identified strains and original SARS-CoV-2 strain becomes more based. This workflow is also valuable for other RNA virus within-host diversity identification, and could provide aid to design more effective mRNA vaccine through optimising antigen selection and sequence design.

Chapter 2

Literature review

Before starting to develop the workflow, I researched how the SARS-CoV-2 reference sequence was acquired and the design of different SARS-CoV-2 vaccines.

2.1 Error Correction Algorithm

There have been three generations of sequencing technologies developed to acquire genetic information by producing reads from samples. However, none of the existing methods are error-proof. Efforts have been made continuously to correct errors in the reads, as they are randomly distributed and will likely affect the assembled genomes.

2.1.1 Instance-based Error Correction

An instance-based error correction algorithm (InsEC) gave me the initial idea of how different strains' reads will be in the sam file produced by alignment against a reference sequence. InsEC forms a matrix from the sam file. Such a method is also used in the verification process in my workflow.

[69] proposed that some of the current correction methods take a global approach with genome-wide patterns and statistics that have suboptimal efficiency and the likelihood to introduce new errors into reads. This is likely due to non-uniform distribution in sequencing depth, errors, and repetitive regions in the genome sequences. Because global approaches are either k-mer-based or multiple alignment methods. The former determines strong and weak k-mers by their frequencies, and the latter forms a consensus string from grouped reads with shared k-mers. The authors suggested using accurate instance-based error correction methods instead. It corrects short reads of specific genes instead of performing a global correction, so the computational load is much lighter and the accuracy

improves.

[69] designed an accurate instance-based algorithm that corrects reads for any nucleotide sequence of interest. As an instance-based method, it could notice local features of the instance reads which the global approaches usually neglect. Another merit is that InsEC could reduce the error it introduces by conservatively combining patterns and statistics. It does not require a strong k-mer definition, so it is not affected by various sequencing depths. The pseudocode of InsEC is in 2.1. The essential idea is to construct a matrix of aligned reads to a reference sequence, then find out the dominant bases at each position, then update the reads with errors found. It could also remove irrelevant reads that show more errors than the threshold value.

Algorithm 1: Error Correction

Input: An extracted read set $\mathcal{R} = \{R^1, R^2, \cdots, R^m\}$, their corresponding alignment position $\{u_i\}_{i=1}^m$ and threshold λ

Output: A corrected read set $\mathcal{S}'$ and an updated nucleotide sequence H

Function ERROR_CORRECTION $(\mathcal{R}, \{u_i\}_{i=1}^m, \lambda)$

begin

- $q \leftarrow \max(u_i)$
- $H \leftarrow$ empty array ; ▷ *Store the updated nucleotide sequence*
- **for** $j = 1$ **to** $(q + n - 1)$ **do**
  - $(c, d) \leftarrow (0, 0)$
  - **for** $i = 1$ **to** m **do**
    - **if** $u_i \leq j \leq (u_i + n - 1)$ **then**
      - $f(j, r^i_{(j-u_i+1)}) \leftarrow f(j, r^i_{(j-u_i+1)}) + 1$
      - $c \leftarrow c + 1$
  - $H(j) \leftarrow \underset{x \in \{A,G,T,C\}}{\text{argmax}} (f(j, x))$
  - **foreach** $x \in \{A, T, C, G\}$ **do**
    - **if** $f(j, x) > 0$ **then**
      - $d \leftarrow d + 1$
  - **foreach** $x \in \{A, T, C, G\}$ **do**
    - $f(j, x) \leftarrow f(j, x)/c + d * 0.1$
- $\mathcal{S}' \leftarrow \emptyset$; ▷ *Store the corrected reads*
- **for** $i = 1$ **to** m **do**
  - **for** $j = 1$ **to** n **do**
    - $t(j) \leftarrow f(u_i + j - 1, r^i_j)$
  - $[B, I] \leftarrow$ sort$(t(1...n))$; ▷ *ascending order; B is sorted array and I is an index array; the k-th smallest is $t(I(k))$*
  - **if** $B(3) > \lambda$ **then**
    - $r^i_{I(1)} \leftarrow H(I(1) + u_i - 1)$
    - $r^i_{I(2)} \leftarrow H(I(2) + u_i - 1)$
    - Append R^i to $\mathcal{S}'$
- **return** $\mathcal{S}'$, H

Figure 2.1: Psuedocode for InsEC Algorithm

The authors evaluated the performance of InsEC by both error correction and the quality of the assembled result. The test dataset was downloaded from SRA, a set of Illumina reads of lung cancer-associated genes. A simulated dataset was also generated for testing using ART[24]. Three metrics were used, precision which shows the fraction of truly corrected bases in all changed bases, recall which shows the fraction of truly corrected

bases in all bases that should be corrected, and gain which shows the fraction of errors removed without inducing additional errors. By measuring the precision, recall, and gain of InsEC and existing models including Bcool, BFC, and Coral, InsEC exhibited a lead in all models in terms of all the metrics, especially in gain, for both single-end and paired-end reads. However, InsEC was ranked third in terms of running time, about 1.5 - 2 seconds longer than Coral and BFC, Bcool took 18.92 seconds. After selecting a subset of reads related to disease-causing genes and used as inputs for Coral, BFC, and Bcool, the gain for the three models improved moderately from 2.56 to 7.61%. This proves the working idea of instance-based error correction. As for assembly results, InsEC showed the closest assembled contigs for 5 out of 6 cases from the error-free datasets, and in three of these error-free datasets, the identical contigs were assembled.

2.1.2 karect

Karect is an error correction method based on multiple alignment developed by Allam et al. [1]. It is one of the most used error correction models nowadays. The first step of Karect correction is selecting a set of candidate reads with significant overlap with every read as a reference read in the read set using an improved heuristic method. This method involves breaking down the reference read and all candidate reads into k-mers, then keeping candidate reads that share at least a threshold number of k-mers with the reference read. If too few candidate reads are included, then the criteria of inclusion will lower to allow more mismatches or break down k-mer to smaller k-mers. The next step is aligning candidate reads to the reference read, and performing a normalization process with candidate reads that contain insertion/deletion. Normalization means generating a set of operations to transform a candidate read into the reference read and shift all operations as right as possible. The normalised alignment is then stored in a partial order graph (POG), along with the reference read. The weight of the edge is the number of alignments passing through the two nodes of the edge. Following the generation of POG, the corrected reference read is formed by having a path within the POG that has the best overall weight found by dynamic programming.

The authors performed tests on sequencing reads of 454, Ion Torrent, and Illumina and compared them with other tools including DAGCon, Fiona, Blue, Coral, MuffinKmeans, and HSHREC in 6 metrics. Those metrics include recall, precision, FScore, gain, time, and memory. FScore is 2 * precision * recall/(precision + recall). The authors believed that measuring correct individual bases correction is more suitable. Karect performed better

than other tools in terms of precision, Fscore and gain, while slightly worse than DAG-CON in recall in three datasets containing insertion/deletion errors, and outperformed all other tools in all metrics in 2 of the remaining 3 Illumina datasets with only substitution errors and with lower precision than SGA in one dataset. The authors also measured the corrected reads' effect on de-novo assembly with assemblers including Celera and Newbler for 454 and Ion Torrent datasets and Velvet and Celera for Illumina datasets. Karect is shown to improve the NGA50 of the assembled sequence.

2.2 mRNA vaccines design and properties

2.2.1 mRNA vaccines properties

Two major types of mRNA have been studied as vaccines: non-replicating mRNA and virally-derived, self-amplifying mRNA [58]. A key advantage of mRNA as a source of antigen is its efficiency in evoking MHC-I presentation and eliciting cytotoxic T-lymphocyte responses. This allows a high degree of versatility in the type and number of antigenic determinants, allowing fast vaccine development. The basic structure of *in vitro* transcribed (IVT) mRNA consists of a protein-encoding open reading frame (ORF) flanked by 5' and 3' untranslated regions (UTRs), ending with a 7-methyl guanosine 5' cap structure and a 3' poly(A) tail. Upon entering the host cell, the mRNA is translated by the cell to produce proteins. This makes mRNA vaccines suitable for delivering cytosolic or transmembrane proteins to the correct cellular compartments [47]. Figure 2.2 is an illustration of the mechanism by which mRNA vaccines mediate immune responses. Other advantages of mRNA vaccines include their generic production procedure, ease of design, and relative safety. These are discussed further in the following section.

The common administration routes for mRNA vaccines are needle-syringe and intramuscular injection [67]. Other methods such as needle-free and intra-dermal injection are also possible and have been shown to have a similar level of effectiveness as traditional administration routes in experimental studies [67]. Lipid nanoparticles (LNPs), as advanced delivery systems, are a viable approach to increase the efficacy and stability of mRNA vaccines [58, 67]. Compared to naked mRNA, encapsulation within nanoparticles offers much better protection against degradation, and also allows flexibility in controlling biodistribution, cellular targeting, and cellular uptake mechanisms. Charge-altering releasable transporters (CART) is another type of delivery system that has proven effective in some cases [22], but is not commonly used in prophylactic vaccines.

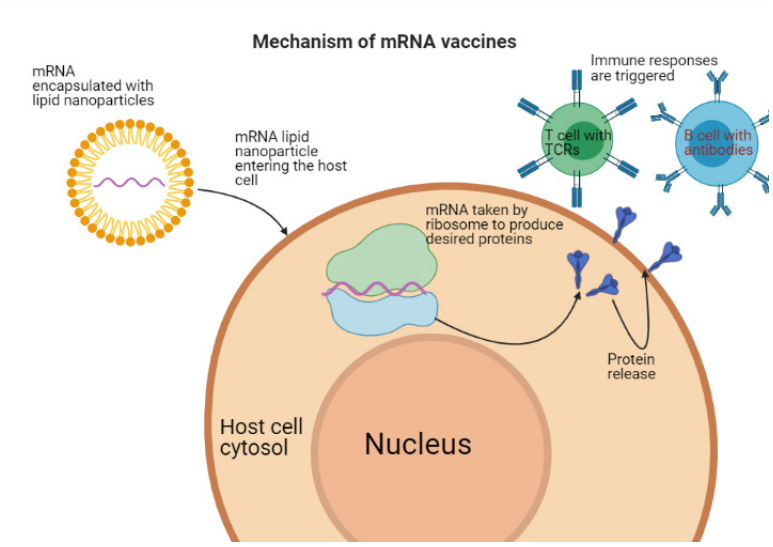

Figure 2.2: Mechanism of mRNA vaccines

Despite promising results in animal experiments and some human trials, there are still concerns about the mRNA vaccines' efficiency due to disadvantages such as short exposure time to the immune system, low stability of mRNA, the need for adjuvants, and immunodominant effects. Thess et al. [56] engineered mRNA sequences to enhance protein expression and suppress cytokine secretion, and the unmodified engineered mRNA was found to outperform its pseudouridine-modified counterpart. In contrast, Liang et al. [37] proposed that modified mRNA vaccines can offer precision in antigen design, as well as good tolerability and broad immune responses. The common belief in the research community is that unmodified sequences do not provide attractive features, while modification offers much greater potential in improving immunogenicity and pharmacokinetics [27]. Another common measure to increase the immune response for mRNA vaccines is to use adjuvants, but this topic is beyond the scope of the current review.

2.2.2 Comparison of mRNA vaccines with other types of vaccines

To better understand the advantages of mRNA vaccines, this section compares their characteristics with other existing vaccine types such as live attenuated vaccines and subunit vaccines.

Safety Profile

The safety profile is arguably the most crucial property of a vaccine. mRNA vaccines typically possess a good safety profile as they do not contain live viruses. As recent research [29, 41, 45, 58] pointed out, mRNA vaccines also do not pose any risks of genetic inte-

gration. Through experiments with SARS-CoV-2 vaccines, Krammer [28] noted that the adverse effects of mRNA vaccines were dose-dependent. The two tested mRNA vaccines in this study showed a superior safety profile, especially when compared to viral vector ChAdOxnCoV-19 or Ad5-based vaccines. However, when Liu [38] compared mRNA vaccines with plasmid DNA vaccines, the author noted potential toxicity issues with mRNA vaccines that were not predicted by preclinical safety testing due to species differences between humans and the animal models used.

Efficacy

Efficacy is the primary measure for the success of a vaccine. Since their conception and development, mRNA vaccines have been criticized for their instability and low translation efficiency [58]. However, many technologies such as LNPs and adjuvants have largely resolved these concerns. For SARS-CoV-2 vaccine candidates, Krammer [28] stated that adjuvanted protein-based vaccines have the best performance in terms of immunogenicity, followed by mRNA vaccines, the viral vector ChAdOx1, and finally the AdV5-based vaccines. Pandey et al. [45] suggested that recombinant vector vaccines could better trigger immune responses than mRNA vaccines due to their longer exposure time *in vivo*. Liu [38] noted that DNA vaccines could persist for a longer period of time than mRNA, but their translation efficiency might be reduced due to the need to enter the nucleus. Nevertheless, in some animal experiments, mRNA vaccines were shown to be more effective than DNA vaccines [29, 46]. Immunization strategies such as escalated antigen delivery and cross-reactive antigen design proposed by Burton and Walker [5] could help to increase the performance of mRNA vaccines.

Design, Production, and Distribution

The flexible and straightforward design process of mRNA vaccines has enabled their timely development against rapidly evolving infections such as COVID-19. Traditional vaccines such as live attenuated or inactivated vaccines require thorough research to develop adequate processes for culturing the target virus and killing or attenuating the virus before production [21]. In comparison, mRNA vaccines only require the viral sequence and structural information for vaccine design [10], using a universal platform with the only difference being in the ORF for different viruses or strains. As an example, the two leading mRNA vaccines for SARS-CoV-2 use LNP encapsulation, together with a generic mRNA vaccine IVT process [11, 59]. This greatly increases the speed of mRNA vaccine design and de-

velopment, as seen in Table 2.1 where mRNA-1273 was the first vaccine for SARS-CoV-2 to enter clinical trials.

The requirements for production and distribution are often overlooked in designing vaccines but are important considerations for their widespread applications. mRNA vaccines require frozen delivery chains and storage, creating a significant challenge for large-scale immunization, particularly in less developed regions. These practical restrictions are a serious shortcoming of mRNA vaccines compared to other vaccine types [28]. Nevertheless, challenges in the distribution of mRNA vaccines are somewhat offset by their simple and straightforward production [41, 45]. Faced with the current massive demand for SARS-CoV-2 vaccines, the ability of mRNA vaccines to offer efficient production and up-scaling is a distinct advantage. However, most mRNA vaccines have strict storage and transportation requirements, but there have been some successful attempts to lower such requirements and make mRNA vaccines more viable and accessible Zhang et al. [68].

The ability of mRNA vaccines to provide long-term global-scale immunization is difficult to determine since most are yet to be tested in large-scale human trials. Although mRNA vaccines may not be suitable for use against all diseases, they do present clear advantages in combining a good safety profile, acceptable immunogenicity, and quick design and development. In light of the urgent need to develop a vaccine for a fast-evolving and unstable RNA virus that has caused a global pandemic, mRNA vaccines could provide a viable solution to be used in the fight against SARS-CoV-2.

2.3 mRNA vaccines designed for COVID-19

Several mRNA vaccines have been designed for immunization against COVID-19, some of which have been approved for public use after demonstrating superior safety and efficacy in clinical trials. Of the vaccine candidates reported in April 2020 (Table 2.1), those with the most advanced progress are the BNT162b1 from Mulligan et al. [42], BNT162b2 from Walsh et al. [59], and mRNA-1273 from Corbett et al. [11], all of which had quickly and safely entered phase 3 clinical trials. Among these, the Moderna vaccine Corbett et al. [11] was the first to commence clinical trials and be implemented for large-scale immunization, including in the US, Canada, European Union, Israel, and Switzerland (list expanding). BNT162b2 has been approved for use in the UK, US, Canada, Singapore, Saudi Arabia, Argentina, Switzerland, the European Union, and some other Latin American countries (list expanding).

On November 9, 2020, Pfizer and BioNTech announced that BNT162b2 was found to

be more than 90% effective during phase 3 clinical trials. However, some researchers have raised doubts about the actual efficacy of BNT162b2 since a number of suspected cases of COVID-19 in the study population were not included in the efficacy analysis, and Pfizer had been prompted to provide further clarification on this issue. Nevertheless, in Israel, mass vaccination on a nationwide level with over 1.5 million participants using two doses of BNT162b2 showed 92% effectiveness for documented infection [14]. On November 16, 2020, Moderna announced a 94.5% effective result in the phase 3 trial of mRNA-1273. Both BNT162b2 and mRNA-1273 mRNA vaccines adopted the same design principle [60], using the LNP approach with similar encoding regions, which comprise the spike protein with two proline mutations to lock in the prefusion conformation.

Institutions	**Approaches**	**April 2020 Progress**	**May 2021 progress**
Moderna (mRNA-1273)	mRNA	phase 1 clinical trial	completed trials and immunised massively
Curevac	mRNA	pre-clinical phase	phase 2b/3 trial
ExpresS2ion	recombinant-protein	pre-clinical phase	clinical trial
iBio	recombinant-protein	pre-clinical phase	completed toxicity studies
Novavax	recombinant-protein	pre-clinical phase	completed clinical trial and in production
Baylor College of Medicine	recombinant-protein	pre-clinical phase	clinical trial
University of Queensland	recombinant-protein	pre-clinical phase	stopped at phase 1 trial
Sichuan Clover Biopharmaceuticals	recombinant-protein	pre-clinical phase	phase 2/3 trial

continues on next page

Vaxart	viral-vector	pre-clinical phase	completed phase 1 trial
Geovax	viral-vector	pre-clinical phase	animal testing
University of Oxford	viral-vector	pre-clinical phase	completed trials and immunised massively, stopped in some countries
Cansino Biologics	viral-vector	pre-clinical phase	phase 3 trial
Inovio	DNA	pre-clinical phase	phase 2/3 trial
Applied DNA Sciences	DNA	pre-clinical phase	completed phase 1 trial
Codagenix	live-attenuated vaccine	pre-clinical phase	phase 1 trial and approved in 40 countries
Pfizer	mRNA	phase 1 clinical trial	completed trials and immunised massively
Anhui Zhifei Longcom	viral vector	pre-clinical phase	completed clinical trial and approved in 2 countries
Bharat Biotech	inactivated	pre-clinical phase	completed clinical trial and approved in 9 countries
Chumakov Center	inactivated	NA	approved in one country
FBRI	protein subunit	pre-clinical phase	approved in one country

continues on next page

Gamaleya	viral-vector	NA	phase 3 trial
Janssen	viral-vector	pre-clinical trial	completed clinical trial and immunised massively, stopped in some countries
RIBSP Kazakhstan	inactivated	NA	phase 3 clinical trial approved in 1 country
Beijing/Wuhan Institute of Biological Products	inactivated	phase 1 trial	completed trials and immunised massively
Sinovac	inactivated	phase 1 trial	completed trials and immunised massively

Table 2.1: Progress of SARS-CoV-2 vaccines in April 2020 and May 2021 (NA means no data)

According to Feikin et al. [17], two mRNA vaccines for SARS-CoV-2 have better duration than the other two viral vector vaccines (Ad26.COV2.S (Janssen) and Vaxzevria (University of Oxford)). The efficacy of the two mRNA vaccines retain better after more than four to six months than their two counterparts. The authors of this paper investigate 18 studies and such finding is very much consistent among these studies.

Mutations of SARS-CoV-2 is rapid as it is a highly-contagious RNA virus. According to CDC, there are 11 variants that are being monitored and the Omicron variant being the variant of concern(VOC)[7]. The efficacy of vaccines is affected by the mutations of each variant,but mRNA-1273 retains good effectiveness across different variants tested [4]. This is because the changes of these variants in spike protein does not largely affect the binding of vaccine-generated antibodies, but that is not the case for Omicron variant [55]. The protection rate against SARS-CoV-2 Omicron symptomatic infections is much lower than it is for other variants albeit the the mRNA vaccines still provide solid protection against hospitalization and severe cases.

2.3.1 mRNA vaccines designed for preventing other diseases

Though not yet common, mRNA vaccines have already been developed and tested to prevent several viral infections such as Zika virus (ZIKV), dengue virus (DENV), respiratory syncytial virus (RSV), influenza virus (specifically H10N8 and H7N9), and flavivirus. Other targets of mRNA vaccines include non-infectious diseases such as diabetes and cancer. It is worth noting that several mRNA vaccines developed to treat cancer used an *ex vivo* dendritic cell (DC) loading approach instead of nanoparticles or LNPs, a technique that is rarely used in other types of vaccine designs for infectious diseases. A list of the mRNA vaccines discussed in this section is presented in Table 2.2.

Epitope studies proven to be helpful in the understanding of SARS-CoV-2 [53] could similarly inform vaccine design for other diseases. In this section, we discuss examples of the tools and delivery systems, as well as testing methods and outputs used in the development of mRNA vaccines against infectious and non-infectious diseases (other than COVID-19).

Author	**Target**	**Design of Encodings**	**Experiment Subjects**
Richner et al. [50]	ZIKV	full length prM and E gene	mouse
Roth et al. [51]	DENV	DENV1-NS poly-epitope	mouse
Espeseth et al. [15]	RSV	mF RNA	rat and mouse
Feldman et al. [18]	H10N8 & H7N9	HA glycoproein	human
Firdessa-Fite and Creusot [19]	diabetes	epitope sequence	mouse
VanBlargan et al. [57]	flavivirus	prM and E genes	mouse
Cafri et al. [6]	cancer	neoantigens	human
Chahal et al. [8]	ZIKV	prM and E proteins	mouse

Table 2.2: mRNA vaccines for infectious and non-infectious diseases (other than COVID-19)

Zika virus

ZIKV is an mRNA virus that can be spread by mosquito bites, for which there is no vaccine or specific treatment other than supportive therapy. Richner et al. [50] designed an mRNA vaccine that targets ZIKV infection with optimised LNPs encapsulating modified mRNA to induce high levels of protein expression *in vivo*. It included full-length prM and E genes (encoding prM and E proteins; both are crucial to ZIKV) from an Asian ZIKV strain. Due to the similarities between ZIKV and DENV, there is a theoretical concern that vaccine-induced cross-reactive antibodies could cause more severe DENV infection through antibody-dependent enhancement. As a result, the researchers hence modified the ZIKV mRNA vaccine on four mutations in or near the E-DII-FL epitope (an epitope on the DENV E protein). This modification was shown to be protective against ZIKV while diminishing the production of antibodies enhancing DENV infection in cells or mice. Richner et al. [50] assessed the immunogenicity and protective activity of their mRNA vaccine by challenging AG129 mice with ZIKV virus 42 days after vaccination. With the exception of one animal, all the mice that received mRNA vaccines, with or without a boost, survived.

Chahal et al. [8] encoded prM and E proteins of ZIKV into an RNA replicon vector as ORF, and used IVT to make the vaccine. The researchers generated an overlapping 15-mer peptide library spanning amino acids 105 to 713 of the ZIKV polyprotein, along with 7 peptides to induce T-cell response. They evaluated H-2Db- and H-2Kb-binding epitopes from all 7 individual 'hits' with ANN-based models from IEDB. The two best-ranked peptides were selected for solid-phase peptide synthesis (SPPS). Chahal et al. [8] used mice to test the vaccine efficacy, where C57BL/6 mice were immunised by intra-muscular injection of the test vaccine, and the control group was immunised with a similar RNA replicon vaccine encoding the Ebola virus glycoprotein. Only two mice in the control group exhibited seropositivity above the detection limit of the assay, compared to all animals in the vaccine group.

Although both mRNA vaccines chose to encode the prM-E proteins as the immunogen, they are structurally different. While the design from Richner et al. [50] used LNP as encapsulation and required two doses, and Chahal et al. [8] designed a modified dendrimer nanoparticle encapsulated mRNA vaccine that only required one dose. The LNP encapsulation has been the preferred choice due to its maturity and effectiveness, but the modified dendrimer nanoparticle encapsulation is also worth exploiting. Additionally, the work of Richner et al. [50] measured the efficacy of serum neutralization titers, while Cha-

hal et al. [8] targeted T-cell responses. In addition to immunity, Richner et al. [50] took reducing the cross-reactivity issue with DENV infection into consideration and tackled it by modifying the epitopes.

Dengue virus

DENV is another RNA virus that has become a global public health threat in recent decades. Roth et al. [51] proposed that $CD4^+$ and $CD8^+$ T-cells contribute to protection against DENV, with each targeting different structural/non-structural proteins. A prototype consensus sequence based on epidemic strains of DENV1 was selected, which was the DENV1-NS T-cell poly-epitope based on the analysis of $CD8^+$ T-cell epitopes from human donors. Using this, an LNP-encapsulated modified mRNA vaccine encoding DENV1-NS was prepared. The vaccine and boost were given at a 4-week interval, with the mice challenged four weeks after the boost to test the vaccine efficacy. The assessment included quantification of viral loads, interferon (IFN) splenocytes, and neutralisation assay. Two injections of the mRNA vaccine were found to provide significant protection against the virus.

Respiratory syncytial virus (RSV)

RSV is a common cause of respiratory infection in children. Espeseth et al. [15] designed a modified mRNA LNP-based vaccine against RSV infection. Due to its conserved nature among serotypes, neutralising antibodies induced by natural RSV infection predominantly target the RSV F protein, providing an attractive vaccine target. Since RSV F protein is difficult to express and purify, mRNA vaccines offer the ability to efficiently explore multiple antigen designs. Espeseth et al. [15] injected cotton rats challenged with the RSV-A2 virus with their designed mRNA vaccines, expressing either prefusion stabilised or native forms of RSV F protein. The mRNA vaccines showed better immunogenicity compared to a protein subunit vaccine. Since the F protein is conserved across strains, the mRNA vaccines designed for the RSV-A strain also provided cross-protection for RSV-B. The rats injected with mRNA vaccines showed no signs of vaccine-elicited respiratory disease (VERD), which was a potential safety concern.

Influenza H10N8 and H7N9

Influenza is highly contagious and the most common infectious disease worldwide. Feldman et al. [18] evaluated the immunogenicity and safety of the first mRNA vaccines against

H10N8 and H7N9 influenza viruses. These vaccines consisted of chemically modified mRNAs encoding the full-length, membrane-bound form of the hemagglutinin (HA) glycoprotein from the H10N8 or H7N9 influenza strains. Both used an LNP delivery system. In phase 1 randomised, placebo-controlled, double-blinded clinical trials, participants were tested for immunogenicity by hemagglutination inhibition (HAI) and micro-neutralisation assays, as well as peripheral blood mononuclear cell (PBMC) persistence. After 6 months of immunisation, 22 of 23 participants receiving the H10N8 vaccine remained seropositive, while for the H7N9 vaccine the protection rate was 52%.

Flavivirus

Powassan virus (POWV) is a flavivirus transmitted by ticks, which may cause life-threatening encephalitis. VanBlargan et al. [57] designed an LNP-encapsulated modified mRNA vaccine, consisting of base-modified mRNA encoding the prM and E genes of POWV Spooner, a POWV strain. This mRNA was preceded by the prM signal sequence (also known as leader peptide or transit peptide) of POWV, or signal sequence of the Japanese encephalitis virus (JEV). C57BL/6 mice were immunised with POWV, JEV, or placebo LNPs followed by a second boost. When challenged with POWV Spooner, 100% of the mice in the placebo-controlled group died, while all POWV and JEV immunised mice were survived.

Diabetes

Firdessa-Fite and Creusot [19] compared LNPs with DCs as two major types of delivery vehicles for mRNA vaccines. The mRNA construct encoded multiple epitopes from different antigens. A potential application for this platform is antigen-specific immunotherapy for type I diabetes. The two delivery routes could be used to target different lymphoid tissues. After intravenous injection, the mRNA delivered by DCs and LNPs was localised mainly in the lungs and spleen, respectively. After local intradermal administration, both delivery routes resulted in mRNA expression at the injection site, as well as robust T-cell responses in draining lymph nodes.

Cancer

Based on the knowledge that T-cells recognising neoantigens are present in most cancers, Cafri et al. [6] developed an mRNA vaccine for cancer treatment. The vaccine named mRNA-4650 was composed of an mRNA backbone that encoded up to 20 different

neoantigens expressed by the autologous cancer. Mutations generated from *in silico* prediction were included in the mRNA construct. The vaccine sequences were electronically submitted to Moderna Therapeutics for manufacturing, and used to treat 4 patients with metastatic gastrointestinal cancer. The mRNA vaccine for each patient encoded different predicted neoantigens based on sequencing data. The T-cell analysis from patients showed increased mutation-specific responses against predicted neoantigens that were not detected before the vaccination. The vaccine was shown to be safe, demonstrating potential for future personalised treatment of patients with common epithelial cancers.

2.4 Existed Within-host Diversity Identification Tools

There are several tools with the functionality of identifying multiple strains within a sample. Some of them are also based on the de-Bruijn graph algorithm like megahit.

Haploflow [20] is a de novo genome assembler with a novel flow algorithm. It constructs a de-Bruijn graph from reads with k-mer hashing. Then it takes the weakly connected components and transformed each one individually into condensed versions of de-Bruijn graphs called unitig graphs. From the constructed unitig graph, the algorithm will infer a set of contigs by putting paths with similar coverages throughout the graph together. After that, it executes the flow algorithm repetitively with 3 steps: finding paths through the graph, assigning flow values to these paths, and determining the path sequences. The process stops when the who graph has been resolved into contigs. The authors used wastewater metagenome data to reconstruct viral haplotypes from 8 SARS-CoV-2 samples and managed to identify 2 strains from GISAID in 6 of these samples. The authors also provide a HIV read set as sample data.

PEHaplo [9] is a more complex tool comprising five major components to form a pipeline. The input reads are processed with Karect for error correction first to remove low-quality reads. PEHaplo then uses corrected reads to construct the overlap graph and applied node collapsing to simplify the graph. A path will be constructed from the pruned graph by extending the current path with the successor node that has the most paired-end connections. After PEHpalo finishes generating contigs, PECC is used to correct the contigs. The authors evaluated the program with an HIV sample, an Influenza sample, and several simulated datasets. The evaluation process is implemented with MetaQuast, a tool that reports various statistics like N50. In terms of genome covered percentage, unaligned length, mismatch rate, and indels (insertion/deletion percentage), PEHaplo is on par with another de novo assembly tool SAVAGE in the real Influenza and HIV datasets, with a

much shorter computation time (SAVAGE used 127h while PEHaplo used 23 minutes). In the simulated datasets, PEHaplo performed a little better than SAVAGE.

The gatk (Genome Analysis Toolkit) [48] is an analysis toolkit that consists of multiple tools useful for sequencing data analysis. HaplotypeCaller is the recommended variant calling tool gatk contains. It identifies variants from a sample through a four-step process. Firstly, HaplotypeCaller (HC) determines a region with significant evidence for a variation called the active region. After that, HC reassembles possible haplotypes in the active region by building a de-Bruijn graph. Each possible haplotype will be assigned a likelihood value by aligning reads against it and constructing a matrix with PairHMM algorithm. Finally, the most likely genotype determined by Bayes' rule is assigned to the sample.

2.5 RNA viruses' quasi-spices and within-host diversity

Lauring [32] provided an overview of within-host diversity study emphasizing the progress in the last decade. An important trend is the choice of sequencing technology changing from Sanger to next-generation sequencing (NGS). With Sanger sequencing, even though it is able to identify single-nucleotide variants (SNVs) and recombinant haplotypes, it does not provide accurate results for singletons and the frequency of variants is not represented properly. NGS can identify mutations and their respective frequencies within a population through repetitive sequencing. However, bioinformatics tools are often required to detect sequence haplotypes with NGS short reads. The authors recommended that a solid variant should have a 1/10 frequency of coverage. Identifying within-host diversity usually serves the purpose of investigating within-host mutation. The authors summarized some criteria for finding positive selection and provided case studies on the influenza virus, dengue virus, and cytomegalovirus. In quasispecies theory, when inside a host, RNA viruses are proposed to be best represented by a collection of mutants. Nevertheless, the authors believe this theory could not reflect the reality of infection in vivo. The virulence and pathogenesis were shown to correlate with viral diversity in various studies earlier, but the authors' team conducted experiments that show the determinant of virulence is replication speed instead of diversity. Another issue of linking viral diversity with virulence is lacking a systematic explanation of how the mutants with often less than 1% frequency influence virulence. Experiments in animals also demonstrate the with-in-host swarm is not a common phenomenon.

Within-host diversity was also observed in SARS-CoV-2 infection [39]. From a dataset consisting of 1390 SARS-CoV-2 nasopharyngeal samples collected from the first wave of

infection in the UK, Lythgoe et al. [39] observed intrahost diversity in multiple cases. The variants were detected with veSEQ protocol and for identifying variant cites has a minimum depth limit of 100 reads. Part of the called variants was reproduced by resequencing the sample and found that single nucleotide variants (SNV) with less than 2% of minor allele frequency (MAF) are generally indistinguishable from noise. The authors also estimated the transmission bottleneck size of SARS-CoV-2 to be between one and eight with the exact beta-binomial method. Within-host diversity was also observed in 1313 samples with high-viral load, a depth of more than 100 reads, a MAF of at least 3%, and not varying in synthetic RNA controls or appearing at low frequency in a large number of samples. In samples with >50,000 unique reads, SNV is observed in 66% of them. Only two samples showed a high number of SNVs (15 and 18 respectively). The authors also found that within-host diversity of SARS-CoV-2 is unlikely a result of coinfection with seasonal coronaviruses as only one sample showed positive for another betacoronavirus but the SARS-CoV-2 genome from this sample has only one SNV with no evidence of mixed-base calls at other positions. Open-read-frames (ORFs) 3a, 7a, and 8 and N protein have the highest densities of SNV sites. The ratio of non-synonymous substitutions and synonymous substitutions is less than 1, consistent with other studies.

The study of within-host diversity in influenza viruses has a longer history than SARS-CoV-2. Xue et al. [65] aimed to provide the substrate of the influenza virus' evolution within-host for global evolution with a review of recent progress in influenza virus' mutation over the course of infection. Deep-sequencing approaches are able to detect within-host diversity at above 1% frequencies of the total viral population but are still not ideal because the viral genetic material could be dwarfed by host and co-occurring microbes, and they cannot firmly determine linkages among the intrahost strains. For acute infections, the number of different strains identified within the same host is generally low as confirmed in multiple studies. While for chronic infections, extensive evolution can be displayed, with the possibility of emerging putative antigenic variants that are drug-resistant. The trend of evolution of the influenza virus inside the human body could be affected by antigenic selection, antiviral resistance, tissue specificity, and spatial structure, and state of coinfection. Viral coinfection could allow deleterious mutation to persist and cause genetic complementation. In cell culture observations, defective viruses with large gene deletions could quickly arise and spread through the population.

Chapter 3

Computational discovery of intra-host coexisting strains of the SARS-CoV-2 reference strain: Motivation and methods

3.1 Data Accessibility

The paired-end reads of SARS-CoV-2 reference sequence were acquired by asking the authors of the reference sequence. The read file is available on sequence read archive (SRA) [33] with accession number SRR11092062. The reference genome sequence is accessible in both GISAID (EPI_ISL_402124) and NCBI(MN996528.1).

3.2 Workflow

Our method has five main steps for the unveiling of the two coexisting strains of the SARS-CoV-2 reference genome. The five steps are extraction of non-human reads, error correction of these reads, identification of those reads containing mutations, phylogenetic study on the genome sequences and spike proteins, and binding affinity studies of the mutated spike proteins with the host receptors.

The nucleotide sequences and amino acid sequences used for phylogenetic analysis are downloaded from NCBI Genebank. The accession numbers are: MN996532.2 (RaTG13), EF203064.1 (HKU2), AT278488.2 (SARS-CoV), KC776174.1 (MERS-CoV), OM366054.1 (SARS-CoV-2 B.1.617.1), MZ724516.1 (SARS-CoV-2 B.1.617.2), MZ275302.1 (SARS-CoV-

2 C.37), MW642248.1 (SARS-CoV-2 P.1), MZ068161.1 (SARS-CoV-2 B.1.351), OM616632.1 (SARS-CoV-2 B.1.1.7), OM766152.1 (SARS-CoV-2 BA.1), MN996528.1 (SARS-CoV-2 WIV04), MZ937000.1 (BANAL-20-52), MZ937003.2 (BANAL-20-236).

The sequencing reads are downloaded from SRA[33] with accession numbers from SRR11092057 to SRR11092064. The data set SRR11092062 has the best quality and contains the most reads of SARS-CoV-2. It has been chosen as the primary dataset in this work.

A graphical representation of our method for coexisting strain identification is presented in Figure 3.1.

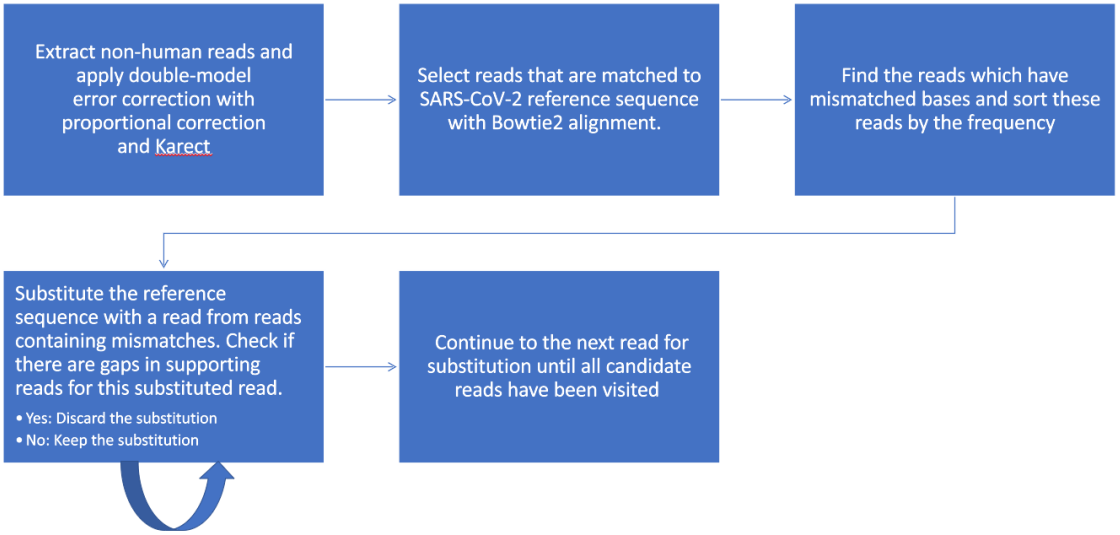

Figure 3.1: Strain identification process

3.2.1 Non-human reads extraction

The workflow begins by filtering out human-related reads from the raw sequencing data. SRR11092062 was sequenced from a BALF viral sample taken from a COVID-19 patient, which means it inevitably contains reads of the human genes. The authors of the reference genome sequence filtered out mammalian genes with a local database and used the remaining reads for the input of MEGAHIT[71]. Since the database was unpublished, reads of human genes were filtered out by Bowtie2 [30] in this work using the GRCh38.p13 human DNA sequence[44] as the reference. There are 122608060 reads in total in the raw paired-end read set. After the filtering, there are 11752058 reads left, trimming the total size of the files from 48.2GB to 4.2GB. This process is also applied to other readsets that are used in this study.

3.2.2 Error correction for the non-human reads

A double-model error correction mechanism is implemented to increase the reliability of reads used for new strain identification.SRR11092062 is paired-end NGS RNA-sequencing reads dataset with a possible error rate of 0.5%-2% [63]. Errors in reads could interfere with identifying new strains. We combine two error correction tools, each tackling a part of the possible erroneous reads.

The first tool is called proportional error correction. It is a model developed based on miREC [70] called smallRNAproper (Small RNA sequencing reads restoration)[64] which works with the whole reads instead of with k-mers. The idea is to judge potentially erroneous bases by frequency comparison. The reads which appear only once are likely to have errors due to the nature of PCR amplification. Reads that have multiple occurrences are more likely to be error-free, so the correction involves correcting erroneous reads based on correct reads. SmallRNApropor substitutes the erroneous reads with their correct counterparts (correcting reads with an edit distance of 1 with the given erroneous reads). The correct candidate of each read is achieved by creating a hash table with the key being the string of reads and the value being the frequency of the read. Each read with the frequency of one will be assigned the candidate read based on the highest frequency read within an editdistance of one. The smallRNAproper corrects each paired-end read file individually. The reads are 150-base-long, and the number of possible erroneous nucleotides is most likely less than 2 (the error rate of Miseq sequencing platform could be above 1.5% but less than 2% [52]). The original code of smallRNAproper works on FASTA files only, so I modified to code to work on FASTQ files by modifying data structures to handle

quality scores. This step corrects a large part of possible erroneous reads but may still leave some of the errors untouched because not all reads with one occurrence will have a correct counterpart. These reads were handled by the second correction tool, Karect [1].

All reads with one occurrence are separated from the read files and corrected with Karect as it deals with untouched, erroneous reads. After that, the frequencies of each unique corrected read are counted. Those corrected reads with a frequency larger than 1 and an edit-distance of 1 to its prior-correction state are kept, and others are reverted to the original form. Finally, these corrected reads are combined with other reads with multiple occurrences. The reason to use Karect after proportional correction is that Karect tends to create non-existing reads when non-singleton reads are involved, and the specific trait of Karect, multiple sequence alignment, is best suited to work with singleton reads.

3.2.3 New strain identification

The corrected reads were aligned to the reference strain sequence EPI_ISL_402124. The purpose is to further extract those reads that are only related to SARS-CoV-2. We chose the alignment tool Bowtie2[30] (with mode set to "local") to extract these relevant reads. This tool was chosen after an evaluation process that included running alignment with Bowtie2, minimap2[35], and bwa-mem2[36] on the SARS-CoV-2 reference genome sequence and the raw reads that produced the sequence. The documentation of bwa-mem2 (the tool Zhou et al. [71] also used for extracting non-mammal reads) is not very detailed, and adjusting arguments are causing irrelevant results to the reference genome sequence. We contacted the author of bwa-mem2 but but the author suggested using default settings because specifying other parameters did not work well. As for minimap2, its document states that the great part of it is the speed for long reads, but short spliced reads are not a good use case[35]. According to another benchmark from Musich et al. [43], Bowtie2 is one of the best aligners for NGS reads. It also has complete functionality, including adjusting mismatch punishment and changing the syntax of the CIGAR string. Since speed is not the concern in our scenario, Bowtie2 is the most suitable tool of the three.

The resulting sam file of Bowtie2 contains all the reads, in my case, most of the reads are human genome or other microorganisms, it is essential to extract reads that are not relevant to SARS-CoV-2 for the following steps. A divide-and-conquer method is used to speed up the process of going through the entire sam file. The sam file is broken into N parts where N is the number of threads of the CPU. N threads will be created and each going through the corresponding part of the sam file to select and output reads that do not

show unmatched in the cigar string field into another file. After all the threads finish, the files they produced will be combined and then all the reads will be sorted by the starting indices.

After the SARS-CoV-2 related reads are extracted under an increased tolerance for the mismatched penalty, we focus on a subset of SARS-CoV-2 related reads that are slightly different from the reference sequence, and we consider them as candidate reads for substitution. These reads will be put into a file and sorted by their frequencies. Every difference in these reads is marked with their position and type. Each of these candidates reads containing different bases is then substituted into the reference genome according to its starting index position one by one in a non-overlapping manner. In this way, the previous changes cannot be overwritten. It is also because non-overlapping reads are more likely from the same strain, according to how enzymatic fragmentation works [23].

After each substitution, an alignment is run to ensure the substitution is "valid"; that is, to ensure that there exists a set of reads that support the updated sequence. The idea of supporting reads is that a paired-end read set should contain completely matched reads that could be connected to form the genome sequence it assembled. In this way, the assembled genome sequence can be seen as valid because it is actually assembled with all the reads that perfectly match it. In a paired-end readset, reads from both ends must both match completely to the genome sequence since this means the read pair is more reliable than just one perfect match read alone. The relevant reads of the original SARS-CoV-2 sequence will be aligned against the updated sequence and the sam file containing relevant reads of the updated sequence will go through the same multi-threading extraction process. A subset of reads that matched completely to the updated sequence and are paired (both of the paired-end reads are included). Then a matrix is formed by these reads, with the rows being the reads' bases and the columns being the position of the reference sequence. If there are no empty columns in this matrix, then the substitution is deemed valid and kept; otherwise, the substitution is reverted.

Besides the best-coverage sample SRR11092062, Zhou et al. [71] also uploaded four other samples from SARS-CoV-2 patients in Wuhan with their paper that published the reference sequence. These samples contain fewer reads of SARS-CoV-2, and the number of generated reads is also smaller than SRR11092062. The read-substitutions are verified on these 4 samples because these patients might also be infected with or developed the same within-host diversity with the discovered strains of SARS-CoV-2. The verification process also starts with separating SARS-CoV-2 reads with alignment against the original

strain reference sequence. Using the original strain as the reference sequence would not miss the reads from other strains of SARS-CoV-2 because the difference among strains is unlikely to be a long continuous streak of nucleotide bases. After separating SARS-CoV-2 reads, the next step is to inspect each substituted base to see whether there are reads that support this change (having the same base at the same position) or if they have appeared more than once. After finding a supporting read, the inspection continues forward and backward to find how many bases in this read match the discovered strain until a mismatch happens, which would create a segment spanning one or multiple reads. If the head or end of the read is reached without encountering a mismatch, then the program goes over to the read prior to or after the current read and starts from the same position relative to the discovered strain sequence. For each read the process traversed through, it should not have a conflict with the position of the current substituted base; otherwise, it would not be included for expanding the segment. If a substituted base has no support in other samples and the frequency of the read carrying it is only 1, and the segment of it is shorter than 50-base, then the substitution would be reverted from the resulting sequence, and another read would be placed with the same process of substitution.

The output of the process includes the FASTA files of two identified strains' complete nucleotide sequences, their spike gene sequences, and the lists of differences for each strain on the original strain sequence. Furthermore, the spike gene sequences of the discovered strains are translated into amino acid sequences to show which changes are non-synonymous (changing the translated amino acid sequences).

3.2.4 Phylogenetic Study

It is important to identify the phylogeny of the coexisting strains to the reference strain, major later variants of SARS-CoV-2, and other coronaviruses. A group of genome sequences from bat coronaviruses, SARS-CoV, MERS-CoV, and other major variants of SARS-CoV-2 is included in our phylogenetic analysis. We used MAFFT[26] to run a multi-sequence alignment (MSA) on these sequences and constructed two phylogenetic trees with FastTree[49] for the spike protein sequences and the whole genome sequences. The constructed phylogenetic trees are visualized by MEGA7 with an adjusted branch length ratio for clear presentation.

The choice of viruses used to build the tree is based on their potential relationship with the SARS-CoV-2 reference strain. Bat coronaviruses RaTG13, HKU2, BANAL-20-236, and BANAL-20-52 are included for the similarity of their spike genes to SARS-CoV or

SARS-CoV-2. The HKU2 bat coronavirus has a spike protein that is suspected to be from a common evolutionary origin of bat-CoV HKU2, bat-SARS-CoV, and SARS-CoV [31]. SARS-CoV and MERS-CoV are also included. The two BANAL bat coronaviruses are found to have a higher similarity in the spike protein, especially BANAL-20-52, which is 98.4% similar to the SARS-CoV-2 original strain spike protein. RaTG13 is a bat coronavirus suspected as the origin of SARS-CoV-2 [71], so it is also included. All the WHO variants of concern (B.1.1.7, B.1.617.2, P.1, B.1.351, and B.1.1.529/BA.1) and some other variants of interest (C.37 and B.1.617.1) are included as well. In this way, the phylogenetic tree we constructed could reflect how close our newly discovered coexisting strains are to SARS-CoV-2 major strains and other bat coronaviruses.

3.2.5 Protein binding affinity investigation

The entire nucleotide sequences of the coexisting strains are translated into amino acid sequences by exPasy[16]. Alphafold2 [25] is then used to predict the 3D folding structures of the spike protein amino acid sequences. The spike protein structure of the SARS-CoV-2 reference strain is acquired from Protein Data Bank (PDB)[3] under the accession number 7CWL, and the ACE2 structure under the accession number 6M1D for comparing the difference in the binding affinities among the coexisting strain 1, coexisting strain 2 and the reference strain. The average of the predicted local Distance Difference Test (PlDDT) on each site is used to find the most suitable model (see Figures 3.2, 3.3, 3.4) among the 5 protein structure models of the spike protein produced by Alphafold2, and model 3 is selected according to this metric.

The protein binding affinity is examined by assessing the docking energy scores. We use the tool HDOCK[66] to measure the scores between model 3 and human ACE2. The tool can provide protein docking scores in the format of predicted docking models.

• Original Strain plddt

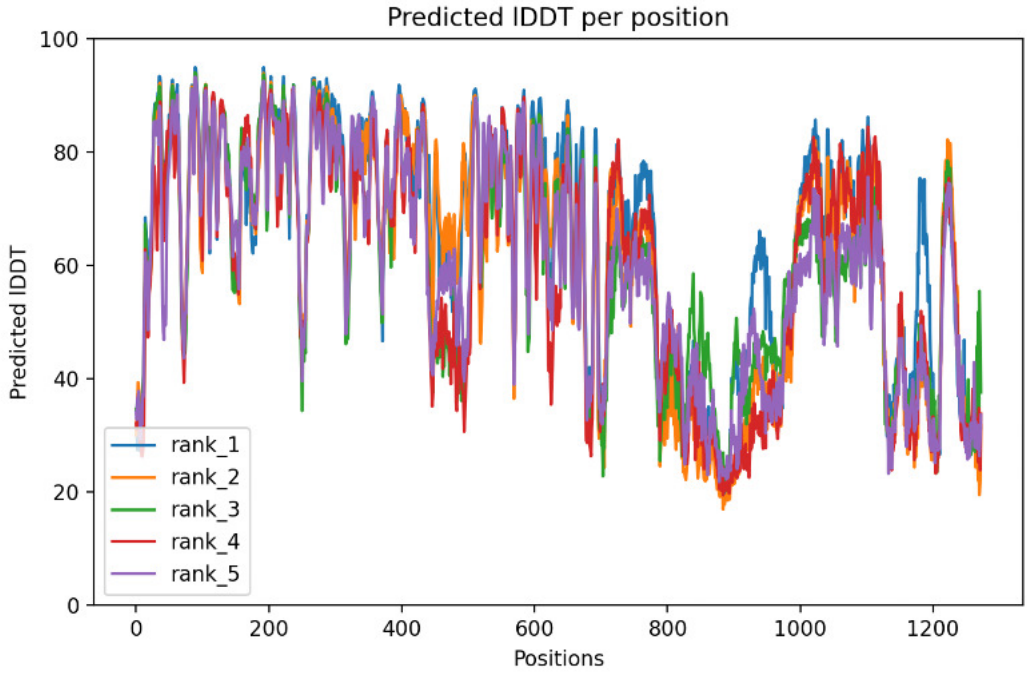

Figure 3.2: The highest PLDDT model 3 with an average score 63.4 is chosen among five predicted models by HDOCK for the original reference genome (WIV04) spike protein.

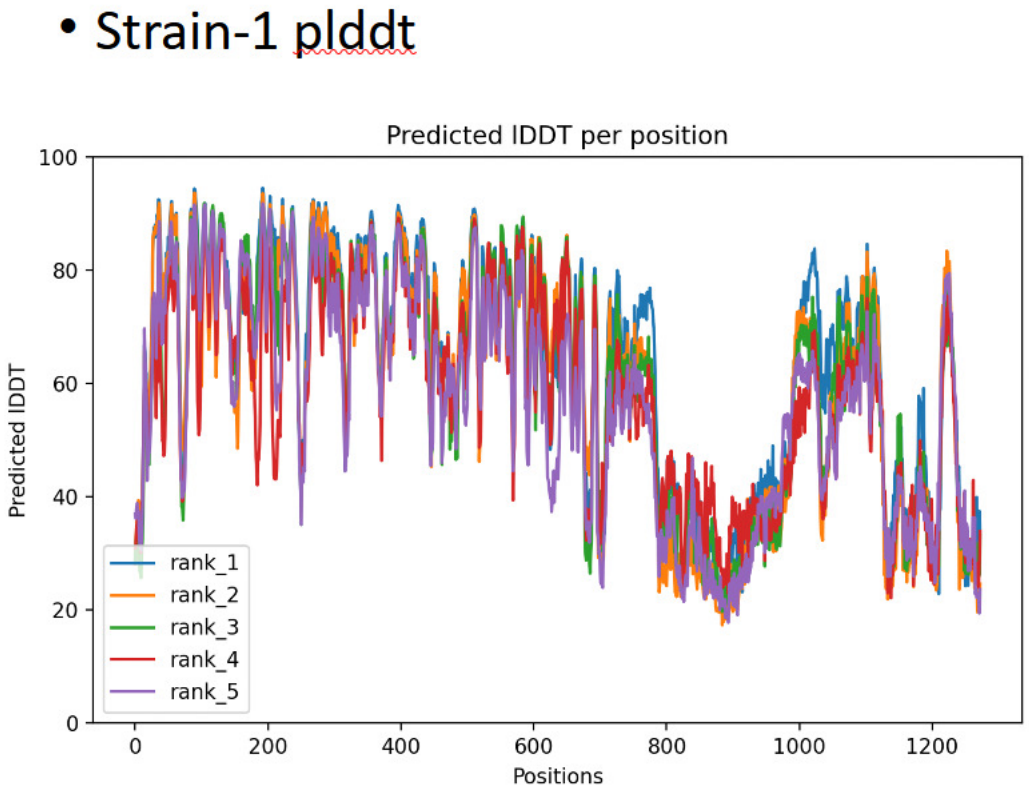

Figure 3.3: The highest PLDDT model 3 with an average score 63.4 is chosen among five predicted models by HDOCK for the discovered strain 1 spike protein.

• Strain-2 plddt

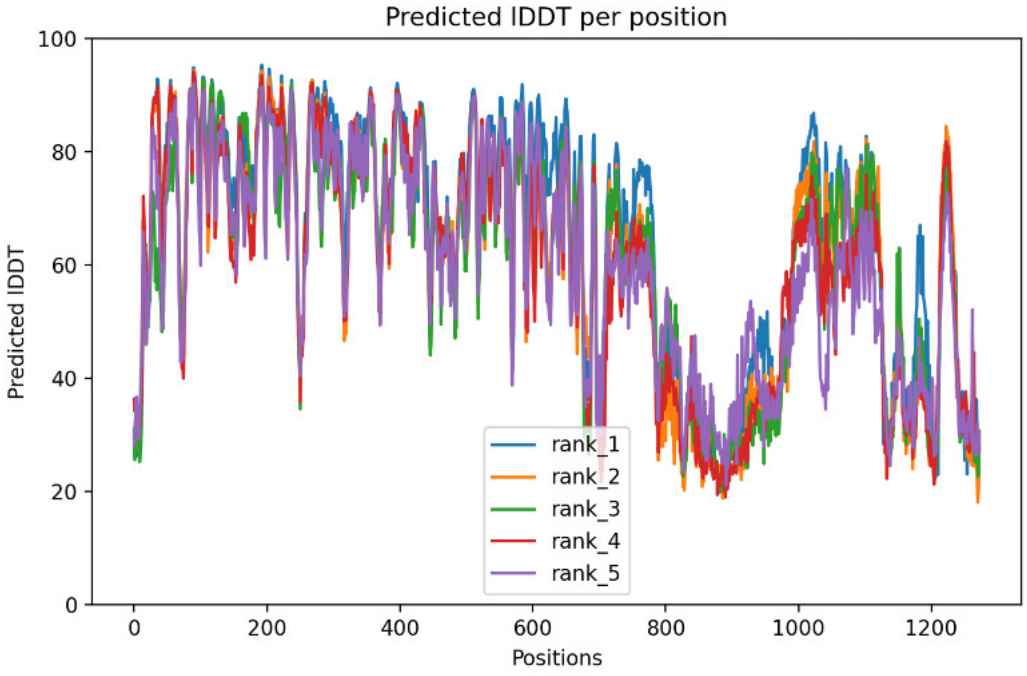

Figure 3.4: The highest PLDDT model 3 with an average score 63.4 is chosen among five predicted models by HDOCK for the discovered strain 2 spike protein .

Chapter 4

Computational discovery of intra-host coexisting strains of the SARS-CoV-2 reference strain: Results

Upon researching about MEGAHIT de-novo assembler, which was used to assemble the SARS-CoV-2 reference genome sequence, I discovered it could potentially eliminate two coexisting strains of the SARS-CoV-2 reference genome were derived using our designed workflow. The analysis presented in this section is based on the nucleotide sequences of the two discovered strains and their spike protein amino acid sequences.

4.1 MEGAHIT Analysis

With the help of my supervisor, I managed to discover that MEGAHIT might be unsuitable for discovering different strains. I tested this by selecting a subset of candidate reads that cover a part of the whole genome sequence and ran MEGAHIT on them. When the k-min value was small (<25), MEGAHIT produces the same sequence as the part in the reference genome sequence, despite the reads being different. However, when setting the k-min to a larger value like 38, MEGAHIT generated a different sequence with no supporting reads, which means it's invalid. MEGAHIT's lacking ability to find multiple strains is confirmed by other research as well[20].

Moreover, during the process of reproducing the SARS-CoV-2 reference genome sequence, I found that even the version of MEGAHIT could affect the resulting sequence.

In the earliest two researches that produced the SARS-CoV-2 reference genome sequence ([62, 71]), Wu et al. [62] used MEGAHIT version 1.1.3 and Zhou et al. [71] used MEGAHIT version 1.2.9. When reproducing the result, it is noted that different versions of MEGAHIT generated different contigs, despite the read set is the same. For read set SRR11092062 provided by Zhou et al. [71], MEGAHIT1.2.9 produced the longest contig with 30069 bases after removing human gene reads and MEGAHIT1.1.3 produced the longest contig with 30105 bases. The two contigs are close to each other with an edit distance of 39 and both are close to the reverse complement form of SARS-CoV-2 reference genome sequence with edit distances being 181 and 217 respectively. Except for the longest contigs, MEGAHIT also produced several long contigs with about 4000-10000 bases. Examining these shorter contigs showed close editdistances to the longest contigs. This means such contigs are possibly a part of other strains but could not be connected.

I also tried to produce the co-existing strain with MEGAHIT. The initial attempt was to find bubbles in the spike gene region from the de-Bruijn graph, which were merged by MEGAHIT. This is done by creating two sets of reads, one that contains only reads with only matched bases called real-narrowed-reads, and the other that contains either matching the reference genome sequence completely or only having mismatched bases called narrowed-reads. Then I calculated the sequencing depths (how many reads cover this position) at each position of the spike gene region for both read sets. The testing merged bubble is located by finding the position in which the sequencing depth that narrowed-reads minus the sequencing depths of real-narrowed-reads is the largest across the spike gene region. Then remove all the reads in real-narrowed-reads and their pairs in the whole read set and run MEGAHIT on the remaining reads. This attempt ends with a contig that does not differ from the original SARS-CoV-2 reference genome sequence except for dozens of bases at the beginning and end. This means MEGAHIT still merges the edges from the co-existing strains so the resulting contig does not change in the spike region. I also tried using a smaller k-min value to be more sensitive but the results remained the same. This means without rewriting MEGAHIT it would be almost impossible to identify co-existing strains with merged bubbles.

4.2 Discovered Strain Differences

Compared to the original strain of SARS-CoV-2, the coexisting strain 1 has different nucleotide bases in the spike gene located across the entire region. There are 22 non-synonymous differences (differences in nucleotide bases that affect the translated protein

amino acid sequence) and 4 synonymous differences in strain 1's spike gene region. 115 out of the 179 differences are located inside the ORF (opening reading frames) region, 5 in the nucleocapsid (N) protein, 12 in the M (membrane) protein, 1 in the E (envelope) protein, and the rest are in NS3 (non-structural 3) proteins.

For the coexisting strain 2, there are fewer different nucleotide bases across the whole sequence. In the spike region, there are 1 synonymous change and 9 non-synonymous changes. About 67 out of the 89 different bases are located in the ORF region. 2 in the N protein, and the rest are in NS8 proteins. The coexisting strain 2 has more differences in the N protein region than coexisting strain 1 but much fewer differences in the spike protein region. Much more details about the two coexisting strains' nucleotide sequences, spike protein sequences, and the differences in spike protein compared to the SARS-CoV-2 reference sequence are shown in the supplementary materials. The base differences were acquired using an alignment tool named EMBOSS Needle[40].

The amino acid mutations in the spike proteins of the discovered strains do not overlap with any mutations in later mutated strains of the virus (prior to Omicron), suggesting that our newly discovered strains are possible to be the ancestor or sibling of the SARS-CoV-2 reference strain instead of offsprings. Interestingly, in discovered strain 2, there is one change A27S overlap with some Omicron variants (BA.2, BA.4, BA.5, BA2.12.1, BA2.75). However, other differences in discovered strain 2 also do not match these Omicron variants.

A coverage chart of the original reference strain and our newly discovered two coexisting strains clearly shows that all three sequences can be fully supported by the reads with a similar pattern of coverage at each position (Figure 4.1).

4.3 Comparison with Other Tools

Three other quasi-species genome assembly tools, PEHaplow[9], Haploflow[20] and GATK HaplotypeCaller[48], have also been applied to SRR11092062.

Despite the fact that Haploflow previously detected variant strains of SARS-CoV-2 from wastewater samples, it did not produce a reliable result from SRR11092062. After more than 168h of supercomputer processing time, Haploflow failed to produce any contigs longer than 1500 bases; on our size-reduced and error-corrected version of SRR11092062 (non-human reads set), Haploflow produced similar results in a much shorter time. HaplotypeCaller reported three mutations, located at positions 306, 565, and 17825. Our approaches also detected all these sites. Running HaplotypeCaller with our error-corrected version of the read set revealed four nucleotide mutations: three are the same as the pre-

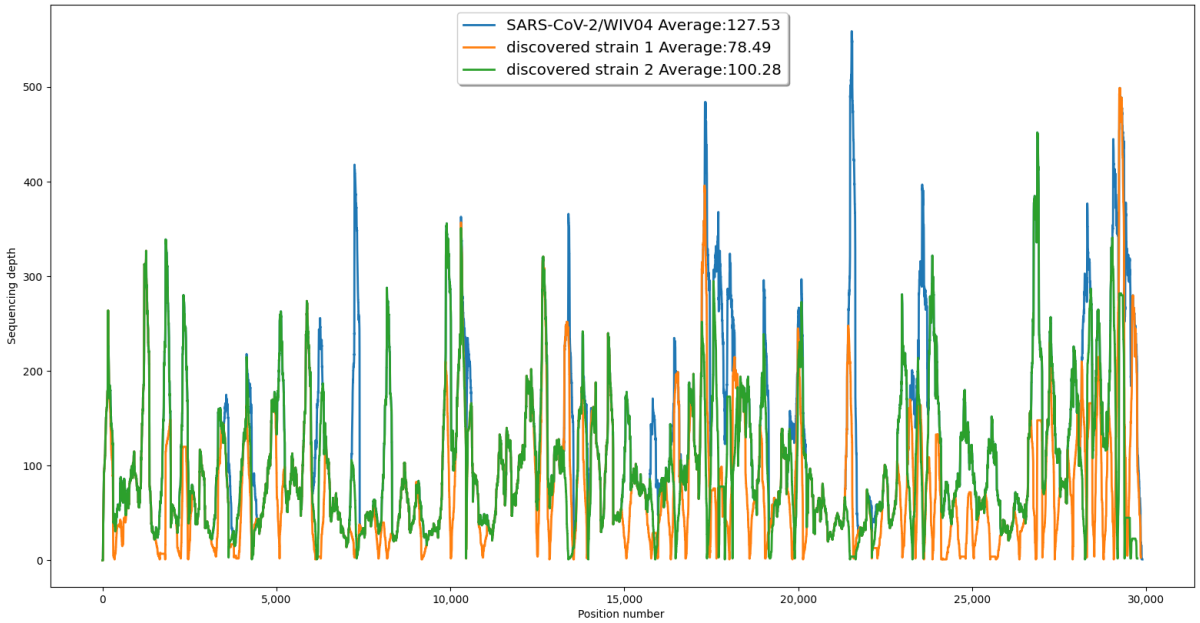

Figure 4.1: The coverage at each position of the sequences

corrected read set and one additional at position 3784. Although being able to identify several positions with possible mutation variations, HaplotypeCaller proposed a substitution that could not be supported by the reads. Position 306's substitution has led to a 0-coverage of its newly formed contig at positions 310-324. This means that there are no reads that can fully match the contig. As for PEHaplo, running it with default arguments and SRR11092062 reads return errors in handling the FASTQ file format. Other than that, preparing the environment for PEHaplo using conda[2] as the authors proposed took more than 48h.

4.4 Phylogenetic Relationship Analysis

A phylogenetic tree constructed from the whole genome sequences is shown in Figure 4.2. Another phylogenetic tree is based on the spike proteins of the chosen coronaviruses, which have different levels of similarity among each other compared to the one constructed from the whole nucleotide sequences (see Figure 4.3). In both cases, the two discovered strains have a closer relationship to the original SARS-CoV-2 strain WIV04 than to other coronaviruses.

The spike protein similarity comparison results are shown in Table 4.1 and Table 4.2. The similarity rates between the two tables are low. The spike protein amino acid sequences are more suitable for similarity analysis than the whole nucleotide sequences because protein sequences are more stable across different sequences of the same virus.

The similarity comparison of spike protein sequences among the chosen coronaviruses suggests that the original SARS-CoV-2 strain (WIV04) shares more bases with all other coronaviruses than the discovered strain 1 and discovered strain 2. Discovered strain 1 has about 1% of bases that are different from the original SARS-CoV-2 strain, and discovered strain 2 has only about 0.3% of the bases being different. For the whole nucleotide sequences comparison, the percentage of differences is even smaller. The phylogenetic tree constructed from the spike protein sequences shows that our discovered coexisting strain 1 and strain 2 are closer to the SARS-CoV-2 viruses than the other bat coronaviruses. The similarity comparison table for spike protein sequences and the whole nucleotide sequences are summarized in Table 4.1 and Table 4.2.

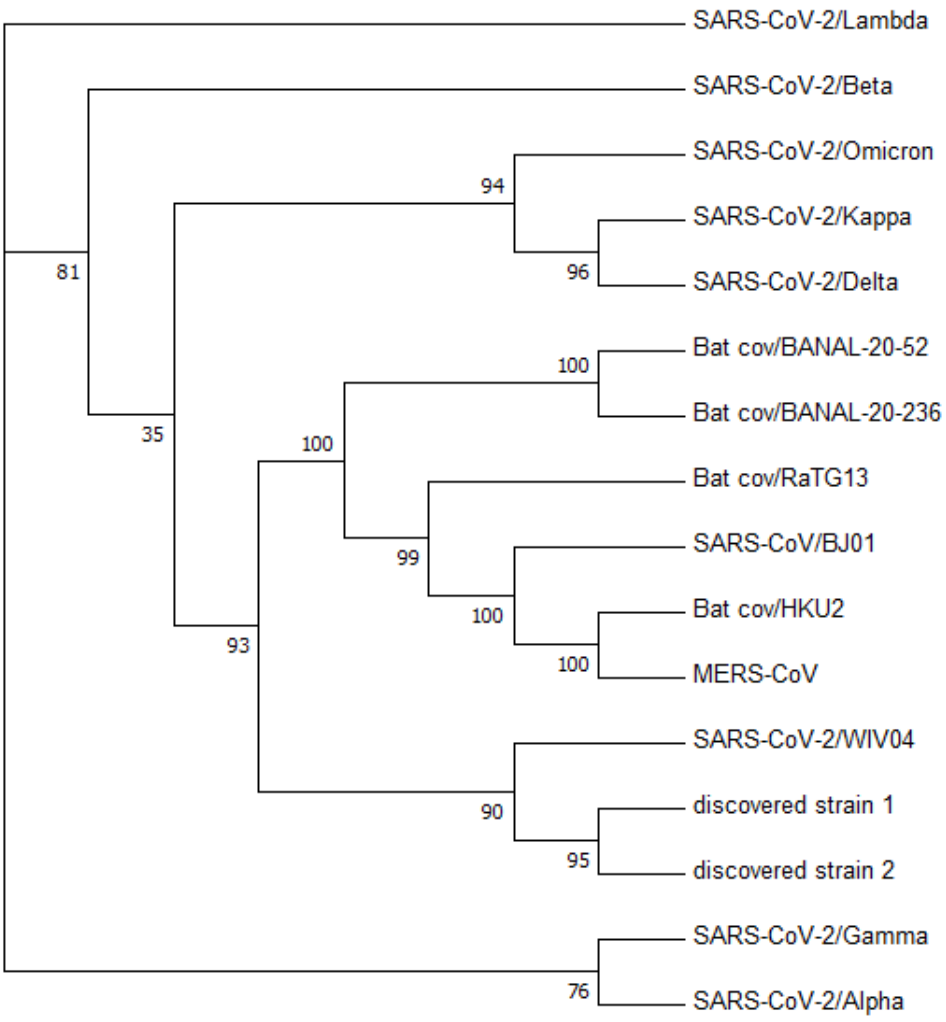

Figure 4.2: Phylogenetic tree constructed from nucleotide sequences

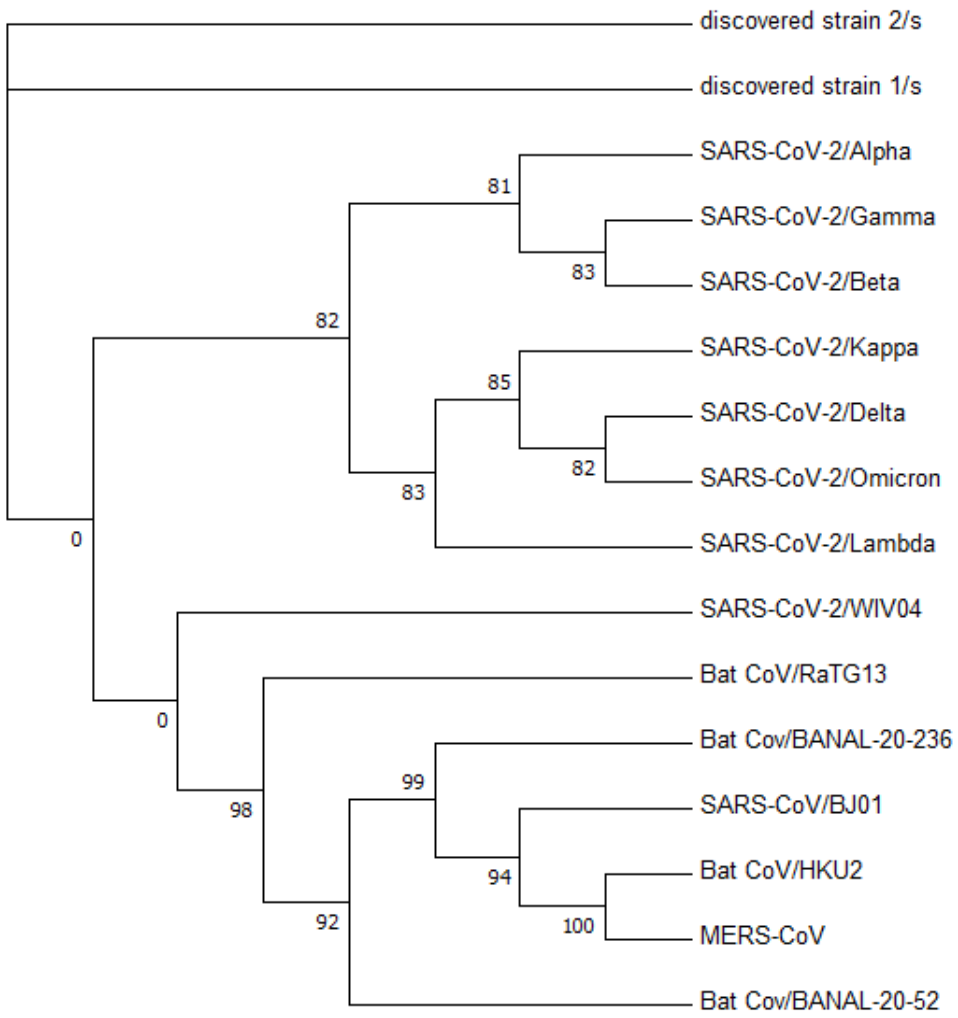

Figure 4.3: Phylogenetic tree constructed from the spike protein sequences

Virus name	SARS-CoV-2/WIV04	discovered_strain_1	discovered_strain_2
Bat_CoV/RaTG13	0.974	0.963	0.971
Bat_CoV/HKU2	0.188	0.184	0.188
SARS-CoV/BJ01	0.764	0.754	0.763
MERS-CoV	0.327	0.323	0.327
SARS-CoV-2/Kappa	0.993	0.982	0.99
SARS-CoV-2/Delta	0.993	0.982	0.99
SARS-CoV-2/Lambda	0.989	0.978	0.986
SARS-CoV-2/Gamma	0.991	0.98	0.988
SARS-CoV-2/Beta	0.991	0.98	0.99
SARS-CoV-2/Alpha	0.992	0.981	0.989
SARS-CoV-2/Omicron	0.995	0.984	0.992
Bat_Cov/BANAL-20-52	0.984	0.973	0.981
Bat_Cov/BANAL-20-236	0.907	0.897	0.907
SARS-CoV-2/WIV04	1.0	0.989	0.997
discovered_strain_1/s	0.989	1.0	0.986

continues on next page

discovered_strain_2/s	0.997	0.986	1.0

Table 4.1: Similarity comparison on the spike proteins amino acid sequences of SARS-CoV-2 original strain(WIV04), two discovered strains, and other related coronaviruses. The similarity is calculated as the length of the shared sub-sequence divided by the total length of the spike protein amino acid sequence.

Virus name	**SARS-CoV-2/WIV04**	**discovered_strain_1**	**discovered_strain_2**
Bat_cov/RaTG13	0.961	0.954	0.957
Bat_cov/HKU2	0.514	0.511	0.514
SARS-CoV/BJ01	0.798	0.793	0.796
MERS-CoV	0.57	0.568	0.57
SARS-CoV-2/Kappa	0.996	0.989	0.993
SARS-CoV-2/Delta	0.996	0.988	0.992
SARS-CoV-2/Lambda	0.997	0.989	0.993
SARS-CoV-2/Gamma	0.985	0.978	0.982
SARS-CoV-2/Beta	0.997	0.991	0.994
SARS-CoV-2/Alpha	0.995	0.987	0.991

continues on next page

SARS-CoV-2/Omicron	0.998	0.99	0.994
Bat_cov/BANAL-20-52	0.967	0.961	0.964
Bat_cov/BANAL-20-236	0.959	0.952	0.956
SARS-CoV-2/WIV04	1	0.993	0.997
discovered_strain_1	0.993	1	0.989
discovered_strain_2	0.997	0.989	1

Table 4.2: Similarity comparison on the nucleotide sequences for SARS-CoV-2 original strain(WIV04), two discovered strains, and other related coronaviruses. The similarity is calculated as the length of the shared sub-sequence divided by the total length of the nucleotide sequence.

The fact that the discovered strains show a more remote relationship to all the coronaviruses in the table than the SARS-CoV-2 original strain may indicate that these two coexisting strains are produced by mutations from the original strain, but they were not widely circulated and may have diminished after the successful control of the pandemic outbreak in Wuhan, China.

4.5 Sequences Comparison

The differences between discovered strain 1, discovered strain 2, and the original strain of SARS-CoV-2 are shown in Table 4.3 and Table 4.4. The information on segment lengths and frequency of reads are also included. The segment size is determined by the number of continuous matching bases of the substituted read itself and other reads among multiple samples that overlap with the substituted read's covered positions. The read frequency was also accounted for. These two statistics are essential for determining the reliability of the substitution. Our criteria for inclusion here are: either the frequency of the read is greater than one (the read appears more than once in SRR11092062) or the segment size is larger than or equal to 50.

Position	bases in original	bases in strain 1	Frequency	Segment Length	change type
348	T	C	1	156	non-synonymous
565	T	C	5	0	synonymous
1645	C	G	1	54	non-synonymous
1804	A	T	2	66	non-synonymous
2248	T	A	2	0	synonymous
2249	G	A	2	0	non-synonymous
2251	A	T	2	0	non-synonymous
2253	A	G	2	0	non-synonymous
2254	A	G	2	0	non-synonymous
2450	C	A	2	149	non-synonymous
2451	T	A	2	1	non-synonymous
2452	A	G	2	1	non-synonymous
2453	C	T	2	1	non-synonymous
2454	T	C	2	1	non-synonymous
2455	C	G	2	0	non-synonymous

continues on next page

2456	A	G	2	0	non-synonymous
2457	T	A	2	0	non-synonymous
2458	G	T	2	0	non-synonymous
2460	C	G	2	2	non-synonymous
2462	C	A	2	0	non-synonymous
2463	T	G	2	0	non-synonymous
3264	C	T	2	0	non-synonymous
3625	A	G	4	176	synonymous
3784	C	T	3	0	synonymous
3912	C	T	2	0	non-synonymous
4288	G	T	2	0	non-synonymous
5091	C	T	2	0	non-synonymous
6119	T	C	2	0	non-synonymous
6120	C	T	2	0	non-synonymous
6121	A	C	2	0	non-synonymous
6122	A	C	2	1	synonymous
6125	G	C	2	0	non-synonymous

continues on next page

6126	A	T	2	0	non-synonymous
6127	G	T	2	0	non-synonymous
6270	C	A	2	0	non-synonymous
6273	A	C	2	0	non-synonymous
6274	T	G	2	0	non-synonymous
6275	A	G	2	0	non-synonymous
6277	A	T	2	0	non-synonymous
7241	G	C	3	17	non-synonymous
7242	A	C	3	1	non-synonymous
7244	T	C	3	1	non-synonymous
7245	G	T	3	2	non-synonymous
7246	G	T	3	1	non-synonymous
7247	T	G	3	1	non-synonymous
7248	T	G	3	1	non-synonymous
7249	T	C	3	1	non-synonymous
7250	T	C	3	1	synonymous
7252	G	C	3	2	synonymous

continues on next page

7253	G	C	3	2	non-synonymous
7254	C	G	3	3	non-synonymous
7256	T	C	3	1	non-synonymous
7257	A	T	3	0	non-synonymous
7930	A	G	2	0	synonymous
7933	A	G	2	4	synonymous
7935	C	A	2	2	non-synonymous
7936	G	T	2	1	non-synonymous
7937	T	C	2	1	non-synonymous
7938	C	G	2	1	non-synonymous
8173	A	G	2	24	synonymous
8174	G	C	2	1	non-synonymous
8175	C	T	2	2	non-synonymous
8177	C	G	2	3	non-synonymous
8179	G	C	2	3	non-synonymous
8181	A	T	2	0	non-synonymous
8182	A	C	2	2	non-synonymous

continues on next page

8183	G	C	2	0	non-synonymous
8185	G	A	2	0	non-synonymous
8186	T	C	2	0	non-synonymous
9167	T	A	3	161	non-synonymous
10009	T	A	2	0	synonymous
10451	A	G	12	179	non-synonymous
10933	T	-	2	0	non-synonymous
12506	A	-	2	0	non-synonymous
12847	T	A	2	118	synonymous
13529	T	C	2	0	non-synonymous
13927	T	C	3	0	synonymous
15060	T	C	2	0	non-synonymous
15771	T	C	3	150	non-synonymous
15957	G	A	2	137	non-synonymous
16333	C	A	2	0	synonymous
16335	T	G	2	0	non-synonymous
16757	C	A	2	51	non-synonymous

continues on next page

16759	C	G	2	2	non-synonymous
16760	C	T	2	1	non-synonymous
16761	A	C	2	1	non-synonymous
16762	C	G	2	3	non-synonymous
16763	T	G	2	1	non-synonymous
16764	T	A	2	1	non-synonymous
16765	A	T	2	1	non-synonymous
16766	A	C	2	1	non-synonymous
16767	C	G	2	0	non-synonymous
17458	C	A	3	0	non-synonymous
17459	C	G	3	1	non-synonymous
17460	A	T	3	0	non-synonymous
17650	T	G	2	175	synonymous
17825	C	T	7	1	non-synonymous
18022	A	G	3	0	synonymous
18024	T	C	3	0	non-synonymous
18025	G	T	3	0	non-synonymous

continues on next page

18597	G	A	2	0	non-synonymous
19050	T	A	2	9	non-synonymous
19817	T	A	2	80	non-synonymous
19819	C	G	2	0	non-synonymous
19820	C	T	2	0	non-synonymous
19821	A	C	2	1	non-synonymous
19823	A	G	2	2	non-synonymous
19824	G	A	2	0	non-synonymous
19825	G	T	2	1	non-synonymous
19826	T	C	2	0	non-synonymous
19828	A	T	2	3	non-synonymous
19830	A	G	2	0	non-synonymous
19831	A	C	2	0	non-synonymous
19832	T	C	2	3	non-synonymous
20136	A	C	3	0	non-synonymous
21587	C	T	2	0	non-synonymous
22114	T	G	2	0	synonymous

continues on next page

22274	T	C	2	0	non-synonymous
22990	T	A	3	0	synonymous
23364	G	A	2	0	non-synonymous
23366	A	G	2	0	non-synonymous
23367	C	T	2	0	non-synonymous
23368	A	C	2	0	non-synonymous
23369	A	G	2	0	non-synonymous
23605	T	C	4	206	synonymous
23859	C	G	1	56	non-synonymous
24108	T	A	1	177	non-synonymous
24259	T	A	1	155	synonymous
24653	G	C	2	0	non-synonymous
24654	A	T	2	1	non-synonymous
24655	G	C	2	0	non-synonymous
24656	T	C	2	0	non-synonymous
24658	T	A	2	0	non-synonymous
24659	G	C	2	0	non-synonymous

continues on next page

24661	A	T	2	0	non-synonymous
24804	T	A	2	0	non-synonymous
24805	T	G	2	2	non-synonymous
24807	G	C	2	2	non-synonymous
24808	T	G	2	0	non-synonymous
24809	C	G	2	0	non-synonymous
25046	C	G	2	0	non-synonymous
25518	T	C	2	3	synonymous
25519	T	C	2	0	non-synonymous
25522	G	C	2	3	non-synonymous
25525	T	C	2	0	non-synonymous
25526	G	T	2	1	non-synonymous
25527	G	T	2	0	non-synonymous
25669	C	A	2	1	non-synonymous
25670	A	G	2	0	non-synonymous
25671	C	T	2	0	non-synonymous
25673	T	G	2	0	non-synonymous

continues on next page

25674	T	G	2	0	non-synonymous
25675	T	A	2	0	non-synonymous
25677	G	C	2	2	non-synonymous
26354	T	A	1	163	non-synonymous
26836	G	T	2	40	non-synonymous
26838	A	G	2	0	non-synonymous
26839	C	G	2	1	non-synonymous
26840	G	C	2	2	non-synonymous
26842	G	T	2	2	non-synonymous
26843	T	C	2	0	non-synonymous
26844	T	C	2	1	non-synonymous
26845	C	G	2	0	non-synonymous
26846	C	A	2	1	non-synonymous
26847	A	C	2	0	non-synonymous
26849	G	T	2	5	non-synonymous
27017	T	A	4	158	synonymous
27372	A	G	3	0	synonymous

continues on next page

27885	T	G	3	2	non-synonymous
28293	A	G	3	151	non-synonymous
28510	T	C	3	1	synonymous
28777	A	C	2	0	synonymous
29066	A	C	3	0	non-synonymous
29456	C	T	3	0	non-synonymous
29840	T	A	3	142	synonymous
29843	G	C	3	3	synonymous
29844	A	G	3	162	synonymous
29845	T	G	3	162	synonymous
29846	T	A	3	1	synonymous

Table 4.3: Verified substituted bases from the coexisting strain 1, a "-" sign means the base is deleted.

Position	**bases in original**	**bases in strain 2**	**Frequency**	**Segment Length**	**change type**
3625	A	C	3	153	synonymous

continues on next page

4291	A	T	2	0	synonymous
6160	T	A	2	0	synonymous
6161	G	A	2	0	non-synonymous
6162	A	G	2	0	non-synonymous
6163	T	A	2	11	non-synonymous
6164	G	C	2	11	non-synonymous
6165	T	C	2	11	non-synonymous
6167	G	C	2	11	non-synonymous
6169	G	T	2	11	non-synonymous
6171	C	G	2	1	non-synonymous
6172	T	C	2	11	non-synonymous
6173	A	C	2	2	non-synonymous
6175	T	C	2	3	non-synonymous
6176	G	C	2	3	non-synonymous
6177	A	G	2	3	non-synonymous
6178	T	A	2	9	non-synonymous
6179	T	C	2	4	non-synonymous

continues on next page

6180	A	T	2	3	non-synonymous
7305	T	A	2	159	non-synonymous
10451	A	C	6	214	non-synonymous
13392	G	T	2	56	non-synonymous
13950	T	A	2	50	non-synonymous
13951	A	G	2	0	non-synonymous
13953	A	C	2	2	non-synonymous
13954	T	G	2	1	non-synonymous
13955	T	G	2	2	non-synonymous
13957	C	T	2	0	non-synonymous
13958	G	C	2	1	non-synonymous
13959	C	G	2	1	non-synonymous
13960	G	T	2	1	non-synonymous
13961	T	A	2	0	non-synonymous
15882	T	A	1	176	non-synonymous
16469	C	A	2	129	non-synonymous
16470	C	A	2	150	non-synonymous

continues on next page

16471	A	G	2	150	non-synonymous
16473	T	C	2	1	non-synonymous
16474	A	G	2	1	non-synonymous
17374	A	G	2	157	synonymous
17709	A	G	4	233	non-synonymous
17886	T	C	3	153	non-synonymous
18108	A	G	2	1	non-synonymous
19150	G	A	2	145	non-synonymous
19151	C	A	2	1	non-synonymous
19152	C	G	2	1	non-synonymous
19153	A	T	2	2	non-synonymous
19155	A	G	2	2	non-synonymous
19156	C	G	2	2	non-synonymous
19159	T	C	2	1	synonymous
19160	C	G	2	2	non-synonymous
19162	G	A	2	3	non-synonymous
19163	A	G	2	3	non-synonymous

continues on next page

19165	A	C	2	2	non-synonymous
19167	A	T	2	1	non-synonymous
19168	T	G	2	3	non-synonymous
19171	A	G	2	4	synonymous
19172	C	T	2	4	non-synonymous
19173	A	T	2	1	non-synonymous
19174	G	C	2	0	non-synonymous
19175	A	T	2	1	non-synonymous
19176	T	G	2	1	non-synonymous
19177	G	T	2	0	non-synonymous
19896	T	C	2	0	non-synonymous
21486	T	A	2	0	non-synonymous
21487	A	C	2	0	non-synonymous
21488	A	T	2	3	non-synonymous
21489	A	T	2	2	non-synonymous
21638	C	A	2	143	non-synonymous
21639	C	A	2	143	non-synonymous

continues on next page

21640	T	G	2	143	non-synonymous
21641	G	T	2	143	non-synonymous
21643	A	G	2	143	non-synonymous
21644	T	G	2	143	non-synonymous
22114	T	C	2	160	synonymous
23280	C	T	1	52	non-synonymous
23281	T	C	1	12	non-synonymous
23590	T	C	2	207	synonymous
28240	T	C	2	1	non-synonymous
28242	G	T	2	1	non-synonymous
28243	T	G	2	1	non-synonymous
28245	T	A	2	1	non-synonymous
28246	T	G	2	2	non-synonymous
28248	G	T	2	3	non-synonymous
28249	A	C	2	1	non-synonymous
28251	T	G	2	1	non-synonymous
28253	C	T	2	2	non-synonymous

continues on next page

28254	A	C	2	1	non-synonymous
29177	C	T	2	0	non-synonymous
29377	T	A	2	21	synonymous
29542	A	G	2	139	synonymous
29543	G	A	2	1	synonymous
29544	A	C	2	2	synonymous
29546	C	G	2	0	synonymous
29547	A	C	2	1	synonymous
29548	C	T	2	1	synonymous
29549	A	T	2	0	synonymous
29550	C	G	2	1	synonymous
29551	A	G	2	1	synonymous
29552	A	C	2	3	synonymous
29553	G	C	2	0	synonymous
29554	G	T	2	0	synonymous
29556	A	C	2	0	synonymous
29559	T	C	2	0	synonymous

continues on next page

29560	G	T	2	0	synonymous
29561	G	T	2	3	synonymous
29724	C	A	2	50	synonymous
29725	A	G	2	4	synonymous
29727	T	C	2	2	synonymous
29728	T	G	2	1	synonymous
29729	T	G	2	0	synonymous
29730	C	A	2	1	synonymous
29731	A	T	2	2	synonymous
29733	C	G	2	2	synonymous
29734	G	T	2	1	synonymous
29737	G	C	2	4	synonymous
29739	C	A	2	2	synonymous
29740	A	T	2	1	synonymous
29741	C	G	2	1	synonymous
29742	G	T	2	0	synonymous
29745	G	T	2	2	synonymous

continues on next page

Table 4.4: Verified substituted bases from coexisting train 2, a "-" sign means the base is deleted.

4.6 Protein Binding Affinity Analysis

We used Alphafold2[25] to predict the 3D structures of the spike proteins for the two coexisting strains. These predicted 3D structures were used to investigate their docking status with human cell surface protein ACE2. The docking scores and ligand root-mean-square deviation (rmsd) of the top 10 models of the coexisting strain 1, coexisting strain 2, and the reference strain are presented in Tables 4.5, 4.6, and 4.7. The binding affinity of the spike protein of the coexisting strain 1 with human receptor ACE2 is about 4.84% smaller than that of the original strain's spike protein, while the coexisting strain 2's spike protein has a 9.78% lower binding affinity score.

The complex structure formed by the spike protein of the coexisting strain 1 and human ACE2 receptor is less stable than the original strain's spike protein. On the other hand, the one formed by the coexisting strain 2's spike protein shows increased instability compared to that of the coexisting strain 1's spike protein. This prediction result suggests that the transmissibility of both newly discovered strains is weaker than the original strain. The weaker binding affinities of the two discovered strains with ACE2 also indicate that the two discovered strains could be replaced by SARS-CoV-2 reference strain, just like Delta replaced the reference strain and Omicron replaced Delta.

Rank	1	2	3	4	5	6	7	8	9	10
Docking Score	−320.46	−287.91	−283.90	−282.99	−280.66	−276.94	−274.08	−270.94	−266.21	−265.61
Ligand rmsd (Å)	380.77	369.92	339.95	327.94	317.26	400.29	368.67	289.60	403.08	383.07

Table 4.5: HDOCK results for top 10 models of binding energy scores and rsmd for WIV04 (7CWL) interaction with ACE2 (6M1D)

Rank	1	2	3	4	5	6	7	8	9	10
Docking Score	−298.65	−278.85	−274.11	−264.91	−263.87	−262.92	−260.56	−258.40	−257.71	−256.86
Ligand rmsd (Å)	346.57	292.14	308.84	381.99	393.60	374.32	289.67	343.61	248.65	358.10

Table 4.6: HDOCK results for top 10 models of binding energy scores and rsmd for discovered strain 1 spike protein 3D structure interaction with ACE2 (6M1D)

Rank	1	2	3	4	5	6	7	8	9	10
Docking Score	−298.22	−283.19	−272.89	−268.42	−267.25	−258.48	−249.65	−247.00	−243.73	−242.99
Ligand rmsd (Å)	356.14	308.39	353.95	391.48	377.23	369.87	350.24	381.94	327.39	297.72

Table 4.7: HDOCK results for top 10 models of binding energy scores and rsmd for discovered strain 2 spike protein 3D structure interaction with ACE2 (6M1D)

The details of the nucleotide sequences of the two discovered strains from SRR11092062 and their spike proteins' amino acid sequences can be found in the supplementary information.

4.7 Case Study: Co-existing Strain Identification of SARS-CoV-2 in Wastewater Sample from California

Crits-Christoph et al. [13] sampled from sewage in San Francisco Bay Area, California, USA in mid 2020 and detected multiple viral genotypes. These viral genomes identified were shown to be more close to the SARS-CoV-2 sequences observed in California than those observed in other regions of the US. The authors submitted 11 samples to SRA, here we use the SRR12596175 sample which was collected from Marin. This sample has a large amount of SARS-CoV-2 related reads and is the only one that we were able to identify the corresponding haploflow resulting sequences. In this section, the discovered strains from the wastewater sample are called MR discovered strains to avoid confusion with discovered strains from SRR11092062.

Firstly, an alignment was performed to check whether the sequence submitted by Crits-Christoph et al. [13] is supported by the reads from SRR12596175. 25 positions throughout the whole sequence were gaps as a result of the alignment, which means it is not supported by the reads. The original SARS-CoV-2 reference sequence MN996528.1 was also not supported. It is possible that RNA degradation caused this, but this sequence may not be from a SARS-CoV-2 strain that exist in the sample. A consensus sequence was assembled using MEGAHIT 1.2.9, which contains the longest contig produced by haploflow and has dozens of bases extended from the haploflow contig at the beginning and end of the sequence. This assembled sequence could be supported by the reads according to the results of alignment. Therefore this sequence is likely to be the dominant SARS-CoV-2 strain sequence in the sample.

After obtaining the sequence of the possible dominant strain, we ran the strain identification program to detect whether within-host diversity exists. The strain identification process manage to produce two sequences both with at least 10 frequency > 1 reads. The substituted reads were verified with other 10 samples submitted under PRJNA-661613 by Crits-Christoph et al. [13] using the same method described for SRR11092062 because these samples were collected within California so they might contain same circulating SARS-CoV-2 strains. Since the protein information is not available like the SARS-CoV-2

reference genome sequence, we could only estimate the possible spike protein region by translating the entire sequence with Expasy[16] and locate the index of spike region from the possible spike protein amino acid sequence location, a consecutive region of 1000+ bases at the second half of the translated sequence. The translated spike protein sequence of the assembled contig is also almost identical to the SARS-CoV-2 reference genome sequence's spike protein except for base 623. However, Expasy identified 9 additional bases at the beginning to also belong to the MR assembled contig's possible spike protein region. MR discovered strain 1 of SRR12596175 showed several insertion/deletion sites including 2 deletions and other 2 non-synonymous changes out of 8 total changes in spike region. The total amount of changes of MR discovered strain 1 is 63. MR discovered strain 2 contains much less changes, a total of 17 changes were identified with 2 in possible spike regions. The changes of both MR discovered strains are listed in the tables 4.8,4.9. Since MR discovered strain 1 contains multiple deletion/insertion in the possible spike region, the translated amino acid sequence is not a whole chunk in this region like MR discovered strain 2, but several fragmented pieces with the longest one containing 422 amino acid bases. In light of this, the synonymity status inference is not very reliable.

Position	bases in assembled contig	bases in MR discovered strain 1	Frequency	Segment Length	change type
289	G	T	4	0	unavailable
1413	T	C	17	129	non-synonymous
1659	A	C	2	0	non-synonymous
1766	T	G	20	77	synonymous
2938	C	G	1	58	non-synonymous
3270	A	G	27	117	non-synonymous
5611	C	T	2	0	non-synonymous
6725	-	T	3	1	non-synonymous
7535	A	C	5	0	non-synonymous
7749	C	A	3	0	non-synonymous
8111	G	A	18	0	non-synonymous
8367	A	T	2	2	non-synonymous
9111	G	A	1	104	non-synonymous
10347	C	T	25	0	non-synonymous
10433	T	A	1	93	non-synonymous

continues on next page

10925	A	T	4	0	synonymous
11840	C	A	3	0	synonymous
11944	T	C	20	103	non-synonymous
12168	A	G	1	118	non-synonymous
13328	-	A	2	4	non-synonymous
14363	G	T	2	0	unavailable
14833	C	T	16	0	unavailable
15082	T	-	2	0	unavailable
15795	T	G	2	51	unavailable
16288	C	T	3	0	unavailable
16592	G	T	3	6	unavailable
16687	T	-	2	0	unavailable
17542	T	C	4	0	unavailable
18452	A	G	13	0	unavailable
18803	C	A	1	88	unavailable
18888	C	T	2	0	unavailable
19867	T	A	1	88	unavailable

continues on next page

20011	C	T	2	66	unavailable
20145	C	G	2	0	unavailable
21083	G	A	3	0	unavailable
21332	C	T	23	0	unavailable
22208	G	A	2	0	synonymous
22946	G	A	2	0	synonymous
23071	C	T	2	35	non-synonymous
23259	A	-	2	0	non-synonymous
23276	C	T	2	12	synonymous
24597	A	T	2	0	non-synonymous
24773	C	T	23	0	synonymous
25220	A	-	3	0	non-synonymous
25701	T	A	1	114	unavailable
25736	C	T	3	3	unavailable
25935	G	T	20	0	unavailable
27086	A	G	2	0	unavailable
27379	T	A	1	106	unavailable

continues on next page

27754	A	C	2	85	unavailable
27755	C	T	2	1	unavailable
27756	T	A	2	0	unavailable
27898	G	T	3	7	unavailable
27992	C	T	26	0	unavailable
28115	T	C	14	140	unavailable
28243	-	T	4	1	unavailable
28500	C	T	19	0	unavailable
28813	T	A	1	81	unavailable
28897	C	T	37	0	unavailable
29028	G	T	5	36	unavailable
29080	-	A	5	57	unavailable
29524	A	T	1	128	unavailable
29609	T	C	3	9	unavailable

Table 4.8: Verified substituted bases from the MR discovered strain 1, a "-" sign in original strain means the base is inserted in MR discovered strain 1, and the "-" sign in MR discovered strain 1 means the base is deleted.

Position	bases in assembled contig	bases in MR discovered strain 2	Frequency	Segment Length	change type
1558	A	T	1	134	non-synonymous
5598	A	G	2	0	non-synonymous
7471	T	C	2	0	non-synonymous
10724	T	A	1	163	non-synonymous
13378	G	T	2	0	non-synonymous
15099	T	A	3	0	unavailable
15350	A	T	1	140	unavailable
15855	A	T	1	133	unavailable
17156	T	C	1	89	unavailable
17483	G	A	1	138	unavailable
17589	G	T	2	0	unavailable
18476	A	G	2	0	unavailable
21277	G	A	2	0	unavailable
21622	T	A	1	105	non-synonymous
22283	A	G	2	0	non-synonymous

continues on next page

27778	G	A	2	10	unavailable
28851	T	G	2	0	unavailable

Table 4.9: Verified substituted bases from MR discovered train 2

The complete nucleotide sequences of the two MR discovered strains and the possible dominant strain assembled by Megahit from wastewater sample SRR12596175 show higher similarity to the original SARS-CoV-2 strain than to any other variants of concerns in the tables below 4.10, 4.11. For the possible spike protein sequences, MR discovered strain 1 shows a broken region so its result may not mean anything, but MR discovered strain 2's possible spike protein sequence shows a closer relationship to the original SARS-CoV-2 strain than other VOCs.

Virus name	SARS-CoV-2/WIV04	WH_strain_1	WH_strain_2	MR_contig.fa	MR_strain_1	MR_strain_2
Bat_cov/RaTG13	0.961	0.954	0.957	0.959	0.943	0.959
Bat_cov/HKU2	0.514	0.512	0.514	0.513	0.51	0.512
SARS-CoV/BJ01	0.798	0.793	0.796	0.797	0.785	0.796
MERS-CoV	0.57	0.568	0.57	0.57	0.566	0.569
SARS-CoV-2/Kappa	0.996	0.989	0.993	0.995	0.977	0.994
SARS-CoV-2/Delta	0.996	0.988	0.992	0.994	0.977	0.994
SARS-CoV-2/Lambda	0.997	0.989	0.993	0.995	0.978	0.995
SARS-CoV-2/Gamma	0.985	0.978	0.982	0.984	0.966	0.983
SARS-CoV-2/Beta	0.997	0.991	0.994	0.996	0.979	0.996
SARS-CoV-2/Alpha	0.995	0.987	0.991	0.993	0.976	0.993
SARS-CoV-2/Omicron	0.998	0.99	0.994	0.996	0.979	0.996
Bat_cov/BANAL-20-52	0.967	0.961	0.964	0.966	0.949	0.965
Bat_cov/BANAL-20-236	0.959	0.952	0.956	0.957	0.941	0.957
SARS-CoV-2/WIV04	1.0	0.993	0.997	0.998	0.98	0.997
WH_discovered_strain_1	0.993	1.0	0.989	0.991	0.974	0.991
WH_discovered_strain_2	0.997	0.989	1.0	0.995	0.977	0.994

continues on next page

MR_megahit_contig.fa	0.998	0.991	0.995	1.0	0.982	0.999
MR_strain_1	0.98	0.974	0.977	0.982	1.0	0.981
MR_strain_2	0.997	0.991	0.994	0.999	0.981	1.0

Table 4.10: Similarity comparison on the nucleotide sequences for SARS-CoV-2 original strain(WIV04), two discovered strains from SRR11092062, the wastewater assembled contig, MR discovered strains from wastewater sample and other related coronaviruses. The similarity is calculated as the length of the shared sub-sequence divided by the total length of the nucleotide sequence. WH means from sample SRR11092062 from Wuhan (the original SARS-CoV-2 reference genome sample), and MR means sample from Marin wastewater SRR12596175. MR_strain_1 and MR_strain_2 are short for MR_discovered_strain_1 and MR_discovered_strain_2.

Virus name	**WIV04/s**	**WH_strain_1/s**	**WH_strain_2/s**	**MR_megahit_contig/s**	**MR_strain_1/s**	**MR_strain_2/s**
Bat_CoV/RaTG13	0.974	0.963	0.971	0.966	0.477	0.965
Bat_CoV/HKU2	0.188	0.184	0.188	0.183	0.099	0.181
SARS-CoV/BJ01	0.764	0.754	0.763	0.756	0.331	0.756
MERS-CoV	0.327	0.323	0.327	0.326	0.199	0.326
SARS-CoV-2/Kappa	0.993	0.982	0.99	0.987	0.492	0.985
SARS-CoV-2/Delta	0.993	0.982	0.99	0.987	0.49	0.985

continues on next page

SARS-CoV-2/Lambda	0.989	0.978	0.986	0.983	0.483	0.981
SARS-CoV-2/Gamma	0.991	0.98	0.988	0.984	0.489	0.983
SARS-CoV-2/Beta	0.991	0.98	0.99	0.985	0.489	0.983
SARS-CoV-2/Alpha	0.992	0.981	0.989	0.986	0.491	0.984
SARS-CoV-2/Omicron	0.995	0.984	0.992	0.989	0.491	0.987
Bat_Cov/BANAL-20-52	0.984	0.973	0.981	0.976	0.486	0.975
Bat_Cov/BANAL-20-236	0.907	0.897	0.907	0.901	0.41	0.9
SARS-CoV-2/WIV04	1.0	0.989	0.997	0.992	0.495	0.991
WH_discovered_strain_1/s	0.989	1.0	0.986	0.981	0.494	0.98
WH_discovered_strain_2/s	0.997	0.986	1.0	0.989	0.493	0.987
MR_megahit_contig/s	0.992	0.981	0.989	1.0	0.505	0.998
MR_strain_1/s	0.498	0.498	0.496	0.505	1.0	0.503
MR_strain_2/s	0.991	0.98	0.988	0.998	0.503	1.0

continues on next page

Table 4.11: Similarity comparison on the (possible) spike protein sequences for SARS-CoV-2 original strain(WIV04), two discovered strains, the wastewater assembled contig, MR discovered strains from wastewater sample and other related coronaviruses. The similarity is calculated as the length of the shared sub-sequence divided by the total length of the nucleotide sequence. WH means from sample SRR11092062 from Wuhan (the original SARS-CoV-2 reference genome sample), and MR means sample from Marin wastewater SRR12596175. MR_strain_1 and MR_strain_2 are short for MR_discovered_strain_1 and MR_discovered_strain_2.

The phylogenetic tree constructed from the nucleotide sequences of the discovered wastewater strains 4.4 seem to belong to their own clade which has a closer relationship to Omicron variant B.A.1 than other VOCs. However, the tree constructed from spike protein sequences shows that the MR discovered strain 2 and the possible dominant strain have a closer relationship to the Alpha variant.

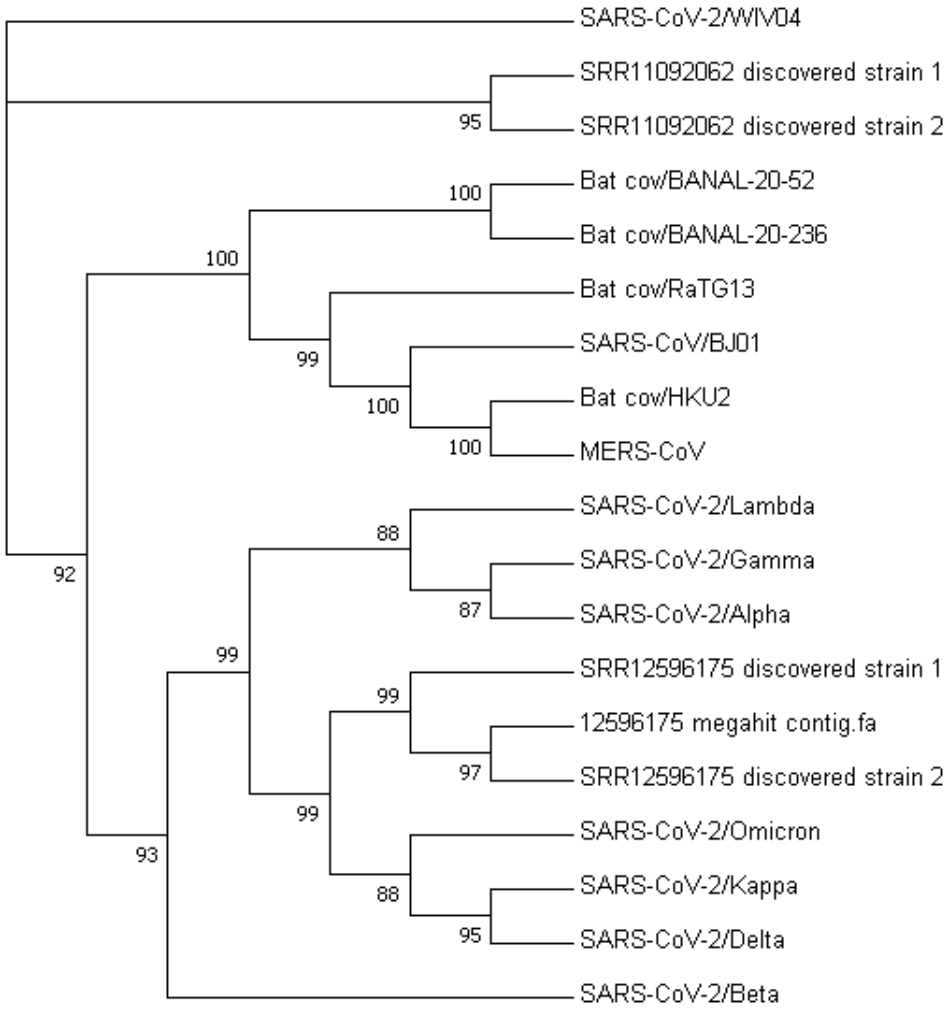

Figure 4.4: Phylogenetic tree constructed from nucleotide sequences of MR discovered strains

4.8 Foot-and-Mouth Disease Virus Animal Samples

Foot-and-mouth disease is caused by the highly contagious foot-and-mouth disease virus (FMDV), a non-enveloped RNA virus belonging to the Picornaviridae family. The FMDV is known to devastate the economy due to its ability to transmit from animals to humans through direct or indirect contact [54]. The high mutation rate exhibited by the FMDV

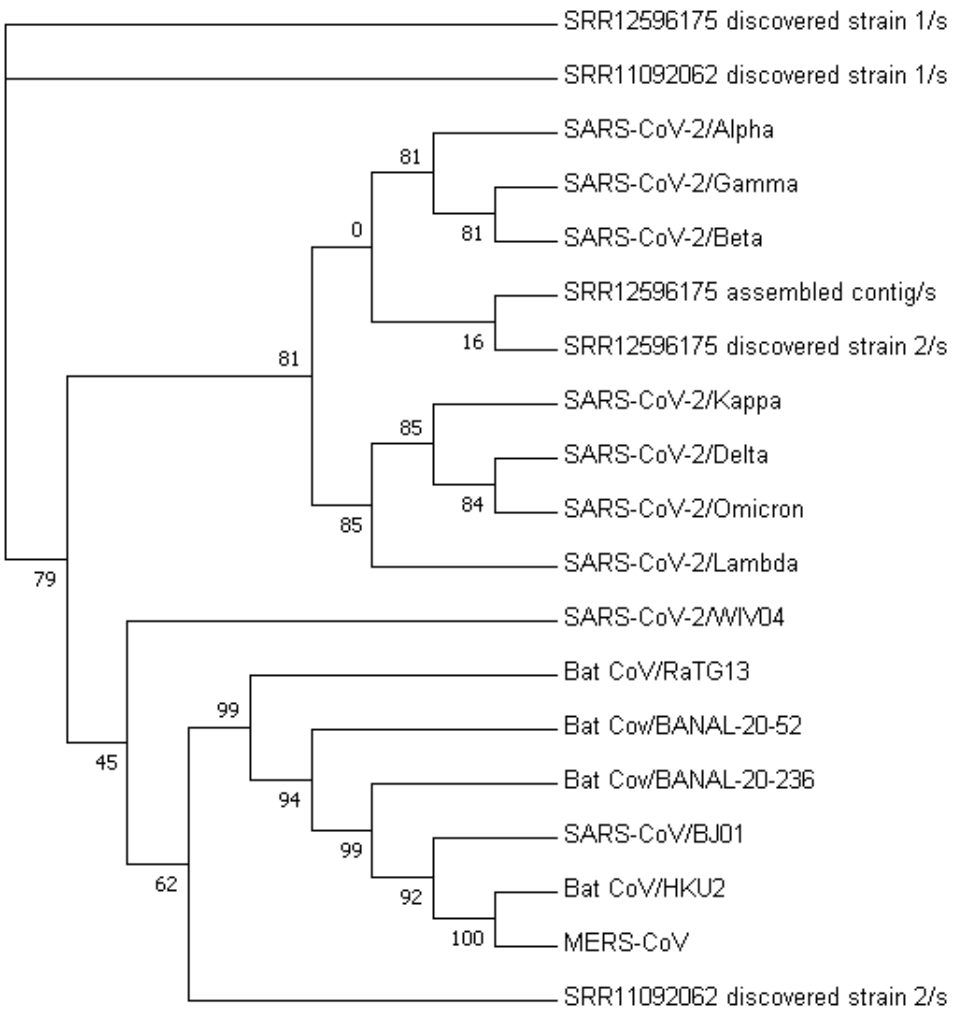

Figure 4.5: Phylogenetic tree constructed from the spike protein sequences of MR discovered strains

as an RNA virus makes it easier to identify its within-host diversity with next-generation sequencing (NGS) technologies [61].

In a study by Wright et al. [61], the FMDV was sequenced from animals using the Illumina Genome Analyzer. The researchers identified polymorphic sites from the FMDV reference genome sequence in the U.K. in 2007 [12] and the consensus sequences assembled from the samples. However, while the study provided information about the polymorphic sites, the authors did not show how those sites would form different strains of viruses, which is a crucial aspect of understanding the FMDV's diversity.

I utilized MEGAHIT v1.2.9 with default options to generate consensus sequences from the sequencing read files provided by Wright et al. [61]. My analysis revealed that 2 out of 6 read files produced identical sequences to the FMDV reference genome sequence from the UK in 2007, while the remaining files contained several mismatches. We selected the

read file ERR034193 to assemble a consensus sequence as it included all five substitutions identified by the authors, which enabled us to demonstrate our workflow on an RNA virus other than SARS-CoV-2.

After applying my strain identification process to the read file, I successfully identified multiple co-existing strains, each with over 170 changed bases from the consensus sequence. These strains shared approximately 15–20 changed bases with one another. The authors proposed over 2000 high-frequency polymorphic sites, but not all of them could be assigned to a single or two strains due to some sites having multiple possible bases. Additionally, combining sites from different strains would result in a lack of actual reads to support the sequence.

The first identified strain contained four out of the five proposed substitutions that were consistent with the reference genome sequence. Other strains generally contained one or two of the five substitutions, sometimes with different bases from both the consensus sequence and the reference genome sequence. For example, at position 3138, Strain 2 showed an A instead of the T in the reference genome or the G in the consensus sequence.

The synonymous/non-synonymous ratio of changed bases in Strain 1 was 3.56. This ratio was generally about 3.3–3.6 in all the discovered strains, except for Strain 3, which had a ratio of 4.0. This may indicate positive selection occurring in the host environment for the FMDV. The synonymity information of the three strains can be found in Tables 7.1–7.3 in the Supplementary. Here, I only report the results of the first three discovered strains, as they provide a representative sample of all the strains discovered.

In conclusion, my method demonstrated that with a higher sensitivity, it is able to detect two more reliable strains than Haploflow in the wastewater sample SRR12596175. Both MR discovered co-existing strains could be supported by the reads without gaps while Haploflow only manages to assemble the possible dominant strain that could be easily assembled by Megahit.

Chapter 5

Conclusions and future research perspectives

In this research, I have examined the mRNA vaccines' mechanism and traits as the most common used vaccine platform for SARS-CoV-2. Among all the SARS-CoV-2 vaccines, the mRNA vaccines have always been the leading ones in terms of development speed and clinical effectiveness. Both the Pfizer and Moderna mRNA vaccines demonstrate over 90% efficiency in protecting against infection of the original SARS-CoV-2 strain. The quickness of its development is a major advantage against fast evolving viruses like SARS-CoV-2 and influenza, when a slower traditional vaccine development process often finds it difficult to keep up with the mutations. The only critical information required to develop an mRNA vaccine for a new virus is the genome sequence, which can be easily acquired through NGS. Another important merit for the mRNA vaccine is its ability to be massively produced in a short time, as the production process does not involve cells or animal product. The clinical trials and real world data show its high effectiveness against SARS-CoV-2, but has strict requirements of storage and delivery. The current state of SARS-CoV-2 mRNA vaccine design could be improved with computational methods to make the antigen more effective and versatile.

A computational workflow for within-host diversity identification is designed to detect co-existing strains of RNA viruses within a sample. This method could compensate the commonly used de-Bruijn-graph-based de-novo genome assemblers like MEGAHIT by identifying low-frequency co-existing strains that are often overlooked during the merging edge process of these tools. The reference genome sequence of SARS-CoV-2 was assembled by MEGAHIT from a high coverage sample SRR11092062, and we applied the workflow to the same sample. The workflow successfully detects two other co-existing strains

of SARS-CoV-2 other than the WIV04 assembled by MEGAHIT. This is also the case for the wastewater sample SRR12596175, which two co-existing SARS-CoV-2 strains' sequences are discovered apart from a dominant strain sequence assembled by MEGAHIT. The strain identification process is more sensitive to within-host diversity than normal de-Bruijn-based assemblers is because it does not assemble a sequence from scratch, but instead works on established genome sequence along with the sequencing reads.

The process of the workflow includes extracting relevant reads, error correction, strain identification, phylogenetic analysis and protein binding analysis. We used a novel double-model error correction model that produce a reliable read set for further analysis. The identified strains' sequences could be supported by the reads in the sample and some of the variant sites also appear in other samples collected at the same location around the same date. The protein binding affinity is acquired by predicting the spike protein structure with AlphaFold2 and computing docking scores. After this complete process, a result consists of not just identified co-existing strains' sequences, but also their phylogenetic relationship with the original strain and the protein binding affinity differences. The discovered strains from SRR11092062 form more unstable complex structure with human ACE2, meaning the ability to infect is lower than the original SARS-CoV-2 strain. Theses strains also do not share any mutations with other major variant of concern.

The pre-processing step of non-human read extraction and error correction ensures the reliability of the reads used for strain identification. The identification process examines every position, which allows a better chance to spot local variant sites than working on k-mer-scale like the de-Bruijn-based assemblers. However, frequency is still used to determine the reliability of a read, as a read that shows up multiple times have a relatively low probability of containing error. When the frequency of read is 1, the read will be verified by examining whether the mismatched part of it also shows up in other samples collected along with SRR11092062, with a consecutive region of more than 50 bases. This is because we suspect the co-existing strains might be circulating in Wuhan, late 2019 as well, but other samples have inferior sequencing qualities making it more difficult to find the entire read that matches perfectly to the reference genome sequence and the identified strains. Other than that, the produced sequence can be supported by a set of perfectly matched reads that have no gaps among them, which means these reads can be concatenated together to form the sequence. Therefore, after strain identification, multiple reliable co-existing strains' sequences are produced.

The strain identification algorithm also performs better at finding co-existing strains

than Haploflow in the wastewater sample from California. It detected two possible co-existing strains of SARS-CoV-2 in the sample which Haploflow only managed to reproduce the MEGAHIT assembled contig sequence. This demonstrates its ability to compensate de-Bruijn-graph-based de-novo assemblers in identifying co-existing strains from a sequencing read library.

In the future, this workflow could also be applied to other RNA virus samples sequenced by NGS because within-host diversity is a common phenomenon in RNA virus infection. It often signals mutation and sometimes superinfection. Identifying co-existing strains and revealing the key differences in important protein regions and the binding activities will allow researchers to better infer the evolution status and circulation strains information. The identified co-existing strains of viruses could help in designing more effective and futureproof vaccines by detecting the potential dominant strains. This is because if within-host diversity comes from mutation, then the mutated strain could have stronger transmission capability, which will result in a gradual replacement of the existing strain. Designing mRNA vaccines based on the mutated strain's sequence will allow a much quicker process to protect the public against the future dominant strain. If it is a result of superinfection, a comparison of protein binding affinity among these strains also provide insights on the contagiousness of difference strains.

Chapter 6

Current Publications and Future Plan

During the first year of my study, I published a paper as the first author on Briefings in Functional Genomics.

[1] Xinhui Cai, Jiao Jiao Li, Tao Liu, Oliver Brian, Jinyan Li, Infectious disease mRNA vaccines and a review on epitope prediction for vaccine design, Briefings in Functional Genomics, Volume 20, Issue 5, September 2021, Pages 289–303, https://doi.org/10.1093/bfgp/elab027

[2] Xinhui Cai, Tian Lan, Pengyao Ping, Brian Oliver, Jinyan Li, Intra-host co-existing strains of SARS-CoV-2 reference genome uncovered by exhaustive computational search, Viruses 2023, 15, 1065. https://doi.org/10.3390/v15051065.

The results showed in this report have been published in the papers above.

Chapter 7

Appendix

Bibliography

[1] Amin Allam, Panos Kalnis, and Victor Solovyev. Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data. *Bioinformatics*, 31(21):3421–3428, 07 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv415. URL https://doi.org/10.1093/bioinformatics/btv415.

[2] Anaconda. Anaconda, 2018. URL https://www.anaconda.com/.

[3] H. M. Berman. The protein data bank. *Nucleic Acids Research*, 28:235–242, 01 2000. doi: 10.1093/nar/28.1.235.

[4] Katia J. Bruxvoort, Lina S. Sy, Lei Qian, Bradley K. Ackerson, Yi Luo, Gina S. Lee, Yun Tian, Ana Florea, Michael Aragones, Julia E. Tubert, Harpreet S. Takhar, Jennifer H. Ku, Yamuna D. Paila, Carla A. Talarico, and Hung Fu Tseng. Effectiveness of mrna-1273 against delta, mu, and other emerging variants of sars-cov-2: test negative case-control study. *BMJ*, 375:e068848, 12 2021. doi: 10.1136/bmj-2021-068848. URL https://www.bmj.com/content/375/bmj-2021-068848.

[5] Dennis R. Burton and Laura M. Walker. Rational vaccine design in the time of covid-19. *Cell Host and Microbe*, 27(5), may 2020. doi: 10.1016/j.chom.2020.04.022.

[6] Gal Cafri, Jared J. Gartner, Tal Zaks, Kristen Hopson, Noam Levin, Biman C. Paria, Maria R. Parkhurst, Rami Yossef, Frank J. Lowery, Mohammad S. Jafferji, Todd D. Prickett, Stephanie L. Goff, Christine T. McGowan, Samantha Seitter, Mackenzie L. Shindorf, Anup Parikh, Praveen D. Chatani, Paul F. Robbins, and Steven A. Rosenberg. mrna vaccine–induced neoantigen-specific t cell immunity in patients with gastrointestinal cancer. *The Journal of Clinical Investigation*, 130, 10 2020. doi: 10.1172/JCI134915. URL https://www.jci.org/articles/view/134915.

[7] CDC. Coronavirus disease 2019 (covid-19), 02 2020. URL https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html.

[8] Jasdave S. Chahal, Tao Fang, Andrew W. Woodham, Omar F. Khan, Jingjing Ling, Daniel G. Anderson, and Hidde L. Ploegh. An rna nanoparticle vaccine against zika virus elicits antibody and cd8+ t cell responses in a mouse model. *Scientific Reports*, 7, 03 2017. doi: 10.1038/s41598-017-00193-w.

[9] Jiao Chen, Yingchao Zhao, and Yanni Sun. De novo haplotype reconstruction in viral quasispecies using paired-end read guided path finding. *Bioinformatics*, 34(17): 2927–2935, 04 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty202. URL `https://doi.org/10.1093/bioinformatics/bty202`.

[10] Kizzmekia S. Corbett, Darin K. Edwards, Sarah R. Leist, Olubukola M. Abiona, Seyhan Boyoglu-Barnum, Rebecca A. Gillespie, Sunny Himansu, Alexandra Schäfer, Cynthia T. Ziwawo, Anthony T. DiPiazza, Kenneth H. Dinnon, Sayda M. Elbashir, Christine A. Shaw, Angela Woods, Ethan J. Fritch, David R. Martinez, Kevin W. Bock, Mahnaz Minai, Bianca M. Nagata, Geoffrey B. Hutchinson, Kai Wu, Carole Henry, Kapil Bahi, Dario Garcia-Dominguez, LingZhi Ma, Isabella Renzi, Wing-Pui Kong, Stephen D. Schmidt, Lingshu Wang, Yi Zhang, Emily Phung, Lauren A. Chang, Rebecca J. Loomis, Nedim Emil Altaras, Elisabeth Narayanan, Mihir Metkar, Vlad Presnyak, Cuiping Liu, Mark K. Louder, Wei Shi, Kwanyee Leung, Eun Sung Yang, Ande West, Kendra L. Gully, Laura J. Stevens, Nianshuang Wang, Daniel Wrapp, Nicole A. Doria-Rose, Guillaume Stewart-Jones, Hamilton Bennett, Gabriela S. Alvarado, Martha C. Nason, Tracy J. Ruckwardt, Jason S. McLellan, Mark R. Denison, James D. Chappell, Ian N. Moore, Kaitlyn M. Morabito, John R. Mascola, Ralph S. Baric, Andrea Carfi, and Barney S. Graham. Sars-cov-2 mrna vaccine design enabled by prototype pathogen preparedness. *Nature*, 08 2020. doi: 10.1038/s41586-020-2622-0.

[11] Kizzmekia S. Corbett, Barbara Flynn, Kathryn E. Foulds, Joseph R. Francica, Seyhan Boyoglu-Barnum, Anne P. Werner, Britta Flach, Sarah O'Connell, Kevin W. Bock, Mahnaz Minai, Bianca M. Nagata, Hanne Andersen, David R. Martinez, Amy T. Noe, Naomi Douek, Mitzi M. Donaldson, Nadesh N. Nji, Gabriela S. Alvarado, Darin K. Edwards, Dillon R. Flebbe, Evan Lamb, Nicole A. Doria-Rose, Bob C. Lin, Mark K. Louder, Sijy O'Dell, Stephen D. Schmidt, Emily Phung, Lauren A. Chang, Christina Yap, John-Paul M. Todd, Laurent Pessaint, Alex Van Ry, Shanai Browne, Jack Greenhouse, Tammy Putman-Taylor, Amanda Strasbaugh, Tracey-Ann Campbell, Anthony Cook, Alan Dodson, Katelyn Steingrebe, Wei Shi, Yi Zhang, Olubukola M. Abiona,

Lingshu Wang, Amarendra Pegu, Eun Sung Yang, Kwanyee Leung, Tongqing Zhou, I-Ting Teng, Alicia Widge, Ingelise Gordon, Laura Novik, Rebecca A. Gillespie, Rebecca J. Loomis, Juan I. Moliva, Guillaume Stewart-Jones, Sunny Himansu, Wing-Pui Kong, Martha C. Nason, Kaitlyn M. Morabito, Tracy J. Ruckwardt, Julie E. Ledgerwood, Martin R. Gaudinski, Peter D. Kwong, John R. Mascola, Andrea Carfi, Mark G. Lewis, Ralph S. Baric, Adrian McDermott, Ian N. Moore, Nancy J. Sullivan, Mario Roederer, Robert A. Seder, and Barney S. Graham. Evaluation of the mrna-1273 vaccine against sars-cov-2 in nonhuman primates. *New England Journal of Medicine*, 07 2020. doi: 10.1056/nejmoa2024671.

[12] Eleanor M. Cottam, Jemma Wadsworth, Andrew E. Shaw, Rebecca J. Rowlands, Lynnette Goatley, Sushila Maan, Narender S. Maan, Peter P. C. Mertens, Katja Ebert, Yanmin Li, Eoin D. Ryan, Nicholas Juleff, Nigel P. Ferris, John W. Wilesmith, Daniel T. Haydon, Donald P. King, David J. Paton, and Nick J. Knowles. Transmission pathways of foot-and-mouth disease virus in the united kingdom in 2007. *PLoS Pathogens*, 4:e1000050, 04 2008. doi: 10.1371/journal.ppat.1000050.

[13] Alexander Crits-Christoph, Rose S. Kantor, Matthew R. Olm, Oscar N. Whitney, Basem Al-Shayeb, Yue Clare Lou, Avi Flamholz, Lauren C. Kennedy, Hannah Greenwald, Adrian Hinkle, Jonathan Hetzel, Sara Spitzer, Jeffery Koble, Asako Tan, Fred Hyde, Gary Schroth, Scott Kuersten, Jillian F. Banfield, and Kara L. Nelson. Genome sequencing of sewage detects regionally prevalent sars-cov-2 variants. *mBio*, 12, 02 2021. doi: 10.1128/mbio.02703-20.

[14] Noa Dagan, Noam Barda, Eldad Kepten, Oren Miron, Shay Perchik, Mark A. Katz, Miguel A. Hernán, Marc Lipsitch, Ben Reis, and Ran D. Balicer. Bnt162b2 mrna covid-19 vaccine in a nationwide mass vaccination setting. *The New England Journal of Medicine*, 02 2021. doi: 10.1056/NEJMoa2101765. URL `https://pubmed.ncbi.nlm.nih.gov/33626250/`.

[15] Amy S. Espeseth, Pedro J. Cejas, Michael P. Citron, Dai Wang, Daniel J. DiStefano, Cheryl Callahan, Gregory O' Donnell, Jennifer D. Galli, Ryan Swoyer, Sinoeun Touch, Zhiyun Wen, Joseph Antonello, Lan Zhang, Jessica A. Flynn, Kara S. Cox, Daniel C. Freed, Kalpit A. Vora, Kapil Bahl, Andrew H. Latham, Jeffrey S. Smith, Marian E. Gindy, Giuseppe Ciaramella, Daria Hazuda, Christine A. Shaw, and Andrew J. Bett. Modified mrna/lipid nanoparticle-based vaccines expressing respiratory

syncytial virus f protein variants are immunogenic and protective in rodent models of rsv infection. *npj Vaccines*, 5, 02 2020. doi: 10.1038/s41541-020-0163-z.

[16] Expasy. Expasy - translate tool, 2019. URL `https://web.expasy.org/translate/`.

[17] Daniel R Feikin, Melissa M Higdon, Laith J Abu-Raddad, Nick Andrews, Rafael Araos, Yair Goldberg, Michelle J Groome, Amit Huppert, Katherine L O'Brien, Peter G Smith, Annelies Wilder-Smith, Scott Zeger, Maria Deloria Knoll, and Minal K Patel. Duration of effectiveness of vaccines against sars-cov-2 infection and covid-19 disease: results of a systematic review and meta-regression. *The Lancet*, 02 2022. doi: 10.1016/s0140-6736(22)00152-0.

[18] Robert A Feldman, Rainard Fuhr, Igor Smolenov, Amilcar (Mick) Ribeiro, Lori Panther, Mike Watson, Joseph J Senn, Mike Smith, Orn Almarsson, Hari S Pujar, Michael E Laska, James Thompson, Tal Zaks, and Giuseppe Ciaramella. mrna vaccines against h10n8 and h7n9 influenza viruses of pandemic potential are immunogenic and well tolerated in healthy adults in phase 1 randomized clinical trials. *Vaccine*, 37(25):3326 – 3334, 2019. doi: https://doi.org/10.1016/j.vaccine.2019.04.074. URL `http://www.sciencedirect.com/science/article/pii/S0264410X19305626`.

[19] Rebuma Firdessa-Fite and Rémi J. Creusot. Nanoparticles versus dendritic cells as vehicles to deliver mrna encoding multiple epitopes for immunotherapy. *Molecular Therapy - Methods & Clinical Development*, 16:50 – 62, 2020. ISSN 2329-0501. doi: https://doi.org/10.1016/j.omtm.2019.10.015. URL `http://www.sciencedirect.com/science/article/pii/S2329050119301238`.

[20] Adrian Fritz, Andreas Bremges, Zhi-Luo Deng, Till Robin Lesker, Jasper Götting, Tina Ganzenmueller, Alexander Sczyrba, Alexander Dilthey, Frank Klawonn, and Alice Carolyn McHardy. Haploflow: strain-resolved de novo assembly of viral genomes. *Genome Biology*, 22, 07 2021. doi: 10.1186/s13059-021-02426-8.

[21] Qiang Gao, Linlin Bao, Haiyan Mao, Lin Wang, Kangwei Xu, Minnan Yang, Yajing Li, Ling Zhu, Nan Wang, Zhe Lv, Hong Gao, Xiaoqin Ge, Biao Kan, Yaling Hu, Jiangning Liu, Fang Cai, Deyu Jiang, Yanhui Yin, Chengfeng Qin, Jing Li, Xuejie Gong, Xiuyu Lou, Wen Shi, Dongdong Wu, Hengming Zhang, Lang Zhu, Wei Deng, Yurong Li, Jinxing Lu, Changgui Li, Xiangxi Wang, Weidong Yin, Yanjun Zhang, and Chuan Qin. Development of an inactivated vaccine candidate for sars-cov-2. *Science*,

05 2020. doi: 10.1126/science.abc1932. URL https://science.sciencemag.org/content/early/2020/05/06/science.abc1932/tab-pdf.

[22] Ole A. W. Haabeth, Timothy R. Blake, Colin J. McKinlay, Robert M. Waymouth, Paul A. Wender, and Ronald Levy. mrna vaccination with charge-altering releasable transporters elicits human t cell responses and cures established tumors in mice. *Proceedings of the National Academy of Sciences*, 115(39):E9153–E9161, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1810002115. URL https://www.pnas.org/content/115/39/E9153.

[23] Steven R. Head, H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, 56, 02 2014. doi: 10.2144/000114133. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4351865/.

[24] Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 12 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr708. URL https://doi.org/10.1093/bioinformatics/btr708.

[25] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, page 1–11, 07 2021. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2.

[26] Kazutaka Katoh, George Asimenos, and Hiroyuki Toh. Multiple alignment of dna sequences with mafft. *Methods in Molecular Biology (Clifton, N.J.)*, 537:39–64, 2009. doi: 10.1007/978-1-59745-251-9_3. URL https://pubmed.ncbi.nlm.nih.gov/19378139/.

[27] Michael S. D. Kormann, Günther Hasenpusch, Manish K. Aneja, Gabriela Nica,

Andreas W. Flemmer, Susanne Herber-Jonat, Marceline Huppmann, Lauren E. Mays, Marta Illenyi, Andrea Schams, Matthias Griese, Iris Bittmann, Rupert Handgretinger, Dominik Hartl, Joseph Rosenecker, and Carsten Rudolph. Expression of therapeutic proteins after delivery of chemically modified mrna in mice. *Nature Biotechnology*, 29:154–157, 02 2011. doi: 10.1038/nbt.1733. URL https://pubmed.ncbi.nlm.nih.gov/21217696/.

[28] Florian Krammer. Sars-cov-2 vaccines in development. *Nature*, 09 2020. doi: 10.1038/s41586-020-2798-3.

[29] Francisco A Lagunas-Rangel, Martha E Viveros-Sandoval, and Arturo Reyes-Sandoval. Current trends in zika vaccine development. *Journal of Virus Eradication*, 3:124–127, 07 2017. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5518240/.

[30] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9:357–359, 03 2012. doi: 10.1038/nmeth.1923.

[31] Susanna K.P. Lau, Patrick C.Y. Woo, Kenneth S.M. Li, Yi Huang, Ming Wang, Carol S.F. Lam, Huifang Xu, Rongtong Guo, Kwok hung Chan, Bo jian Zheng, and Kwok yung Yuen. Complete genome sequence of bat coronavirus hku2 from chinese horseshoe bats revealed a much smaller spike gene with a different evolutionary lineage from the rest of the genome. *Virology*, 367(2):428–439, 2007. ISSN 0042-6822. doi: https://doi.org/10.1016/j.virol.2007.06.009. URL https://www.sciencedirect.com/science/article/pii/S0042682207004096.

[32] Adam S. Lauring. Within-host viral diversity: A window into viral evolution. *Annual Review of Virology*, 7(1):63–81, 2020. doi: 10.1146/annurev-virology-010320-061642. URL https://doi.org/10.1146/annurev-virology-010320-061642. PMID: 32511081.

[33] R. Leinonen, H. Sugawara, and M. Shumway. The sequence read archive. *Nucleic Acids Research*, 39:D19–D21, 11 2010. doi: 10.1093/nar/gkq1019. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013647/.

[34] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676, 01 2015.

ISSN 1367-4803. doi: 10.1093/bioinformatics/btv033. URL https://doi.org/10.1093/bioinformatics/btv033.

[35] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34 (18):3094–3100, 05 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty191. URL https://doi.org/10.1093/bioinformatics/bty191.

[36] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5):589–595, 01 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp698. URL https://doi.org/10.1093/bioinformatics/btp698.

[37] Frank Liang, Gustaf Lindgren, Ang Lin, Elizabeth A. Thompson, Sebastian Ols, Josefine Röhss, Shinu John, Kimberly Hassett, Olga Yuzhakov, Kapil Bahl, Luis A. Brito, Hugh Salter, Giuseppe Ciaramella, and Karin Loré. Efficient targeting and activation of antigen-presenting cells in vivo after modified mrna vaccine administration in rhesus macaques. *Molecular Therapy*, 25(12):2635 – 2647, 2017. ISSN 1525-0016. doi: https://doi.org/10.1016/j.ymthe.2017.08.006. URL http://www.sciencedirect.com/science/article/pii/S1525001617303659.

[38] Margaret A. Liu. A comparison of plasmid dna and mrna as vaccine technologies. *Vaccines*, 7:37, 04 2019. doi: 10.3390/vaccines7020037.

[39] Katrina A. Lythgoe, Matthew Hall, Luca Ferretti, Mariateresa de Cesare, George MacIntyre-Cockett, Amy Trebes, Monique Andersson, Newton Otecko, Emma L. Wise, Nathan Moore, Jessica Lynch, Stephen Kidd, Nicholas Cortes, Matilde Mori, Rebecca Williams, Gabrielle Vernet, Anita Justice, Angie Green, Samuel M. Nicholls, M. Azim Ansari, Lucie Abeler-Dörner, Catrin E. Moore, Timothy E. A. Peto, David W. Eyre, Robert Shaw, Peter Simmonds, David Buck, John A. Todd, on behalf of the Oxford Virus Sequencing Analysis Group (OVSG)‡, Thomas R. Connor, Shirin Ashraf, Ana da Silva Filipe, James Shepherd, Emma C. Thomson, The COVID-19 Genomics UK (COG-UK) Consortium§, David Bonsall, Christophe Fraser, and Tanya Golubchik. Sars-cov-2 within-host diversity and transmission. *Science*, 372, 04 2021. doi: 10.1126/science.abg0821. URL https://science.sciencemag.org/content/372/6539/eabg0821.

[40] Fábio Madeira, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian R N Tivey, Simon C Potter, Robert D Finn,

and Rodrigo Lopez. The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic acids research*, 47(W1):W636—W641, July 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz268. URL https://europepmc.org/articles/PMC6602479.

[41] Giulietta Maruggi, Cuiling Zhang, Junwei Li, Jeffrey B. Ulmer, and Dong Yu. mrna as a transformative technology for vaccine development to control infectious diseases. *Molecular Therapy*, 27(4):757 – 772, 2019. ISSN 1525-0016. doi: https://doi.org/10.1016/j.ymthe.2019.01.020. URL http://www.sciencedirect.com/science/article/pii/S1525001619300413.

[42] Mark J. Mulligan, Kirsten E. Lyke, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, Kathleen Neuzil, Vanessa Raabe, Ruth Bailey, Kena A. Swanson, Ping Li, Kenneth Koury, Warren Kalina, David Cooper, Camila Fontes-Garfias, Pei-Yong Shi, Özlem Türeci, Kristin R. Tompkins, Edward E. Walsh, Robert Frenck, Ann R. Falsey, Philip R. Dormitzer, William C. Gruber, Uğur Şahin, and Kathrin U. Jansen. Phase 1/2 study of covid-19 rna vaccine bnt162b1 in adults. *Nature*, page 1–8, 08 2020. doi: 10.1038/s41586-020-2639-4. URL https://www.nature.com/articles/s41586-020-2639-4.

[43] Ryan Musich, Lance Cadle-Davidson, and Michael V. Osier. Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. *Frontiers in Plant Science*, 12, 04 2021. doi: 10.3389/fpls.2021.657240.

[44] NCBI. Grch38.p13 genome assembly ncbi, 2019. URL https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39.

[45] Satish Chandra Pandey, Veni Pande, Diksha Sati, Shobha Upreti, and Mukesh Samant. Vaccination strategies to combat novel corona virus sars-cov-2. *Life Sciences*, 256:117956, 2020. ISSN 0024-3205. doi: https://doi.org/10.1016/j.lfs.2020.117956.

[46] Norbert Pardi, Michael J. Hogan, Rebecca S. Pelc, Hiromi Muramatsu, Hanne Andersen, Christina R. DeMaso, Kimberly A. Dowd, Laura L. Sutherland, Richard M. Scearce, Robert Parks, Wendeline Wagner, Alex Granados, Jack Greenhouse, Michelle Walker, Elinor Willis, Jae-Sung Yu, Charles E. McGee, Gregory D. Sempowski, Barbara L. Mui, Ying K. Tam, Yan-Jang Huang, Dana Vanlandingham, Veronica M. Holmes, Harikrishnan Balachandran, Sujata Sahu, Michelle Lifton, Stephen Higgs, Scott E. Hensley, Thomas D. Madden, Michael J. Hope, Katalin Karikó, Sampa Santra, Barney S. Graham, Mark G. Lewis, Theodore C. Pierson, Barton F. Haynes,

and Drew Weissman. Zika virus protection by a single low-dose nucleoside-modified mrna vaccination. *Nature*, 543:248–251, 02 2017. doi: 10.1038/nature21428.

[47] Norbert Pardi, Michael J. Hogan, Frederick W. Porter, and Drew Weissman. mrna vaccines — a new era in vaccinology. *Nature Reviews Drug Discovery*, 17:261–279, 01 2018. doi: 10.1038/nrd.2017.243. URL https://www.nature.com/articles/nrd.2017.243.

[48] Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J Daly, Ben Neale, Daniel G. MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 2018. doi: 10.1101/201178. URL https://www.biorxiv.org/content/early/2018/07/24/201178.

[49] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. Fasttree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5(3):1–10, 03 2010. doi: 10.1371/journal.pone.0009490. URL https://doi.org/10.1371/journal.pone.0009490.

[50] Justin M. Richner, Sunny Himansu, Kimberly A. Dowd, Scott L. Butler, Vanessa Salazar, Julie M. Fox, Justin G. Julander, William W. Tang, Sujan Shresta, Theodore C. Pierson, Giuseppe Ciaramella, and Michael S. Diamond. Modified mrna vaccines protect against zika virus infection. *Cell*, 168:1114–1125.e10, 03 2017. doi: 10.1016/j.cell.2017.02.017. URL https://www.cell.com/cell/fulltext/S0092-8674(17)30195-2.

[51] Claude Roth, Tineke Cantaert, Chloé Colas, Matthieu Prot, Isabelle Casadémont, Laurine Levillayer, Jessie Thalmensi, Pierre Langlade-Demoyen, Christiane Gerke, Kapil Bahl, Giuseppe Ciaramella, Etienne Simon-Loriere, and Anavaj Sakuntabhai. A modified mrna vaccine targeting immunodominant ns epitopes protects against dengue virus infection in hla class i transgenic mice. *Frontiers in Immunology*, 10, 06 2019. doi: 10.3389/fimmu.2019.01424. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6598640/.

[52] Melanie Schirmer, Rosalinda D'Amore, Umer Z. Ijaz, Neil Hall, and Christopher

Quince. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, 17, 03 2016. doi: 10.1186/s12859-016-0976-y.

[53] Ellen Shrock, Eric Fujimura, Tomasz Kula, Richard T. Timms, I-Hsiu Lee, Yumei Leng, Matthew L. Robinson, Brandon M. Sie, Mamie Z. Li, Yuezhou Chen, Jennifer Logue, Adam Zuiani, Denise McCulloch, Felipe J. N. Lelis, Stephanie Henson, Daniel R. Monaco, Meghan Travers, Shaghayegh Habibi, William A. Clarke, Patrizio Caturegli, Oliver Laeyendecker, Alicja Piechocka-Trocha, Jon Li, Ashok Khatri, Helen Y. Chu, Alexandra-Chloé Villani, Kyle Kays, Marcia B. Goldberg, Nir Hacohen, Michael R. Filbin, Xu G. Yu, Bruce D. Walker, Duane R. Wesemann, H. Benjamin Larman, James A. Lederer, and Stephen J. Elledge. Viral epitope profiling of covid-19 patients reveals cross-reactivity and correlates of severity. *Science*, 2020. ISSN 0036-8075. doi: 10.1126/science.abd4250. URL `https://science.sciencemag.org/content/early/2020/09/28/science.abd4250`.

[54] Carolina Stenfeldt and Jonathan Arzt. The carrier conundrum; a review of recent advances and persistent gaps regarding the carrier state of foot-and-mouth disease virus. *Pathogens*, 9:167, 02 2020. doi: 10.3390/pathogens9030167.

[55] Sharon H.X. Tan, Alex R. Cook, Derrick Heng, Benjamin Ong, David C. Lye, and Kelvin B. Tan. Effectiveness of bnt162b2 vaccine against omicron in children 5 to 11 years of age. *New England Journal of Medicine*, 387(6):525–532, 2022. doi: 10.1056/NEJMoa2203209. URL `https://doi.org/10.1056/NEJMoa2203209`. PMID: 35857701.

[56] Andreas Thess, Stefanie Grund, Barbara L Mui, Michael J Hope, Patrick Baumhof, Mariola Fotin-Mleczek, and Thomas Schlake. Sequence-engineered mrna without chemical nucleoside modifications enables an effective protein therapy in large animals. *Molecular Therapy*, 23:1456–1464, 09 2015. doi: 10.1038/mt.2015.103.

[57] Laura A. VanBlargan, Sunny Himansu, Bryant M. Foreman, Gregory D. Ebel, Theodore C. Pierson, and Michael S. Diamond. An mrna vaccine protects mice against multiple tick-transmitted flavivirus infections. *Cell Reports*, 25(12):3382 – 3392.e3, 2018. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2018.11.082. URL `http://www.sciencedirect.com/science/article/pii/S2211124718318734`.

[58] Rein Verbeke, Ine Lentacker, Stefaan C. De Smedt, and Heleen Dewitte. Three

decades of messenger rna vaccine development. *Nano Today*, 28:100766, 10 2019. doi: 10.1016/j.nantod.2019.100766.

[59] Edward E. Walsh, Robert W. Frenck, Ann R. Falsey, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, Kathleen Neuzil, Mark J. Mulligan, Ruth Bailey, Kena A. Swanson, Ping Li, Kenneth Koury, Warren Kalina, David Cooper, Camila Fontes-Garfias, Pei-Yong Shi, Özlem Türeci, Kristin R. Tompkins, Kirsten E. Lyke, Vanessa Raabe, Philip R. Dormitzer, Kathrin U. Jansen, Uğur Şahin, and William C. Gruber. Safety and immunogenicity of two rna-based covid-19 vaccine candidates. *New England Journal of Medicine*, 10 2020. doi: 10.1056/nejmoa2027906.

[60] Daniel Wrapp, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S. Graham, and Jason S. McLellan. Cryo-em structure of the 2019-ncov spike in the prefusion conformation. *Science*, 367 (6483):1260–1263, 2020. ISSN 0036-8075. doi: 10.1126/science.abb2507. URL `https://science.sciencemag.org/content/367/6483/1260`.

[61] Caroline F. Wright, Marco J. Morelli, Gaël Thébaud, Nick J. Knowles, Pawel Herzyk, David J. Paton, Daniel T. Haydon, and Donald P. King. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *Journal of Virology*, 85:2266–2275, 03 2011. doi: 10.1128/JVI.01396-10. URL `https://pubmed.ncbi.nlm.nih.gov/21159860/`.

[62] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, Ming-Li Yuan, Yu-Ling Zhang, Fa-Hui Dai, Yi Liu, Qi-Min Wang, Jiao-Jiao Zheng, Lin Xu, Edward C. Holmes, and Yong-Zhen Zhang. A new coronavirus associated with human respiratory disease in china. *Nature*, 579, 02 2020. doi: 10.1038/s41586-020-2008-3.

[63] Tiantian Xiao and Wenhao Zhou. The third generation sequencing: the advanced approach to genetic diseases. *Translational Pediatrics*, 9:163–173, 04 2020. doi: 10.21037/tp.2020.03.06.

[64] XuanZhang. Srnaec, 08 2021. URL `https://github.com/XuanrZhang/SRNAEC`.

[65] Katherine S. Xue, Louise H. Moncla, Trevor Bedford, and Jesse D. Bloom. Within-host evolution of human influenza virus. *Trends in Microbiology*, 26(9):781–793, 2018. ISSN 0966-842X. doi: https://doi.org/10.1016/j.tim.2018.02.007. URL `https://www.sciencedirect.com/science/article/pii/S0966842X1830043X`.

[66] Yumeng Yan, Huanyu Tao, Jiahua He, and Sheng-You Huang. The hdock server for integrated protein–protein docking. *Nature Protocols*, 15:1829–1852, 04 2020. doi: 10.1038/s41596-020-0312-x.

[67] Cuiling Zhang, Giulietta Maruggi, Hu Shan, and Junwei Li. Advances in mrna vaccines for infectious diseases. *Frontiers in Immunology*, 10, 03 2019. doi: 10.3389/fimmu.2019.00594.

[68] Na-Na Zhang, Xiao-Feng Li, Yong-Qiang Deng, Hui Zhao, Yi-Jiao Huang, Guan Yang, Wei-Jin Huang, Peng Gao, Chao Zhou, Rong-Rong Zhang, Yan Guo, Shi-Hui Sun, Hang Fan, Shu-Long Zu, Qi Chen, Qi He, Tian-Shu Cao, Xing-Yao Huang, Hong-Ying Qiu, Jian-Hui Nie, Yuhang Jiang, Hua-Yuan Yan, Qing Ye, Xia Zhong, Xia-Lin Xue, Zhen-Yu Zha, Dongsheng Zhou, Xiao Yang, You-Chun Wang, Bo Ying, and Cheng-Feng Qin. A thermostable mrna vaccine against covid-19. *Cell*, 182:1271–1283.e16, 09 2020. doi: 10.1016/j.cell.2020.07.024. URL `https://www.sciencedirect.com/science/article/pii/S0092867420309326`.

[69] Xuan Zhang, Yuansheng Liu, Zuguo Yu, Michael Blumenstein, Gyorgy Hutvagner, and Jinyan Li. Instance-based error correction for short reads of disease-associated genes. *BMC Bioinformatics*, 22, 06 2021. doi: 10.1186/s12859-021-04058-y.

[70] Xuan Zhang, Pengyao Ping, Gyorgy Hutvagner, Michael Blumenstein, and Jinyan Li. Aberration-corrected ultrafine analysis of miRNA reads at single-base resolution: a k-mer lattice approach. *Nucleic Acids Research*, 49(18):e106–e106, 07 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab610. URL `https://doi.org/10.1093/nar/gkab610`.

[71] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, Hui-Dong Chen, Jing Chen, Yun Luo, Hua Guo, Ren-Di Jiang, Mei-Qin Liu, Ying Chen, Xu-Rui Shen, Xi Wang, Xiao-Shuang Zheng, Kai Zhao, Quan-Jiao Chen, Fei Deng, Lin-Lin Liu, Bing Yan, Fa-Xian Zhan, Yan-Yi Wang, Geng-Fu Xiao, and Zheng-Li Shi. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579, 02 2020. doi: 10.1038/s41586-020-2012-7.

7.1 Supplemental Material

7.1.1 SRR11092062 Discovered Strain 1 Spike Protein Sequence

1 MFVFLVLLSLVSSQCVNLTT RTQLPPAYTNSFTRGVYYPD KVFRSSVLHSTQDLFLPFFS NVTWFHAIHVSGTNGTKRFD NPVLPFNDGVYFASTEKSNI 100

101 IRGWIFGTTLDSKTQSLLIV NNATNVVIKVCEFQFCNDPF LGVYYHKNNKSWMESEFRVY SSANNCTFEYVSQPFLMDLE GKQGNFKNLREFVFKNIDGY 200

201 FKIYSKHTPINLVRDLPQGF SALEPLVDLPIGINITRLQT LLALHRSYLTPGDSSSGWTA GAAAYYVGYLQPRTFLLKYN ENGTITDAVDCALDPLSETK 300

301 CTLKSFTVEKGIYQTSNFRV QPTESIVRFPNITNLCPFGE VFNATRFASVYAWNRKRISN CVADYSVLYNSASFSTFKCY GVSPTKLNDLCFTNVYADSF 400

401 VIRGDEVRQIAPGQTGKIAD YNYKLPDDFTGCVIAWNSNN LDSKVGGNYNYLYRLFRKSN LKPFERDISTEIYQAGSTPC NGVEGFNCYFPLQSYGFQPT 500

501 NGVGYQPYRVVVLSFELLHA PATVCGPKKSTNLVKNKCVN FNFNGLTGTGVLTESNKKFL PFQQFGRDIADTTDAVRDPQ TLEILDITPCSFGGVSVITP 600

601 EVDTSNQVAVLYQDVNCTEV PVAIHADQLTPTWRVYSTGS NVFQTRAGCLIGAEHVNNSY ECDIPIGAGICASYQTQTNS PRRARSVASQSIIAYTMSLG 700

701 AENSVAYSNNSIAIPTNFTI SVTTEILPVSMTKTSVDCTM YICGDSTECSNLLLQYGSFC TQLNRGLTGIAVEQDKNTQE VFAQVKQIYKTPPIKDFGGF 800

801 NFSQILPDPSKPSKRSFIED LLFNKVTLADAGFIKQYGDC LGDIAARDHICAQKFNGLTV LPPLLTDEMIAQYTSALLAG TITSGWTFGAGAALQIPFAM 900

901 QMAYRFNGIGVTQNVLYENQ KLIANQFNSAIGKIQDSLSS TASALGKLQDVVNQNAQALN TLVKQLSSNFGAISSVLNDI LSRLDKVEAEVQIDRLITGR 1000

1001 LQSLQTYVTQQLIRAAEIRA SANLAATKMSLRLLGQSKRV DFCGKGYHLMSFPQSAPHGV VFLHVTYVPAQEKNFTTAPA KSDDGKAHFPREGVFVSNGT 1100

1101 HWFVTQRNFYEPQIITTDNT FVSGNCDVVIGIVNNTVYDP LQPELDSFKEELDKYFKNHT SADVDLGDISGINASVVNIQ KEIDRLNEVAKNLNESLIDL 1200

1201 QELGKYEQYIKWPWYIWLGF IAGLIAIVMVTIMLCCMTSC CSCLKGCCSCGSCCKFDEDD SEPVLKGVKLHYT SEPVLKGVKLHYT 1273

7.1.2 Difference of SRR11092062 Discovered Strain 1 Spike Protein Sequence and SARS-CoV-2 Spike Protein Sequence

The format of EMBOSS Needle is that each three lines followed an empty line, the first in the three lines is 50 amino acid bases of SARS-CoV-2 original strain spike protein, the second line is the matching status, and the third line is 50 amino acid bases of discovered strain 1. | means the two bases at the position in the first protein sequence and the second protein sequence are the same, other signs mean the two bases are different. The figure is located at 7.1 and 7.1.

7.1.3 SRR11092062 Discovered Strain 2 Spike Protein Sequence

1 MFVFLVLLPLVSSQCVNLTT RTQLPKSDTNSFTRGVYYPD KVFRSSVLHSTQDLFLPFFS NVTWFHAIHVSGTNGTKRFD NPVLPFNDGVYFASTEKSNI 100

101 IRGWIFGTTLDSKTQSLLIV NNATNVVIKVCEFQFCNDPF LGVYYHKNNKSWMESEFRVY SSANNCTFEYVSQPFLMDLE GKQGNFKNLREFVFKNIDGY 200

201 FKIYSKHTPINLVRDLPQGF SALEPLVDLPIGINITRFQT LLALHRSYLTPGDSSSGWTA GAAAYYVGYLQPRTFLLKYN ENGTITDAVDCALDPLSETK 300

301 CTLKSFTVEKGIYQTSNFRV QPTESIVRFPNITNLCPFGE VFNATRFASVYAWNRKRISN CVADYSVLYNSASFSTFKCY GVSPTKLNDLCFTNVYADSF 400

401 VIRGDEVRQIAPGQTGKIAD YNYKLPDDFTGCVIAWNSNN LDSKVGGNYNYLYRLFRKSN LKPFERDISTEIYQAGSTPC NGVEGFNCYFPLQSYGFQPT 500

501 NGVGYQPYRVVVLSFELLHA PATVCGPKKSTNLVKNKCVN FNFNGLTGTGVLTESNKKFL PFQQFGRDIADTIDAVRDPQ TLEILDITPCSFGGVSVITP 600

601 GTNTSNQVAVLYQDVNCTEV PVAIHADQLTPTWRVYSTGS NVFQTRAGCLIGAEHVNNSY ECDIPIGAGICASYQTQTNS PRRARSVASQSIIAYTMSLG 700

701 AENSVAYSNNSIAIPTNFTI SVTTEILPVSMTKTSVDCTM YICGDSTECSNLLLQYGSFC TQLNRALTGIAVEQDKNTQE VFAQVKQIYKTPPIKDFGGF 800

801 NFSQILPDPSKPSKRSFIED LLFNKVTLADAGFIKQYGDC LGDIAARDLICAQKFNGLTV LPPLLTDEMIAQYTSALLAG TITSGWTFGAGAALQIPFAM 900

901 QMAYRFNGIGVTQNVLYENQ KLIANQFNSAIGKIQDSLSS TASALGKLQDVVNQNAQALN TLVKQLSSNFGAISSVLNDI LSRLDKVEAEVQIDRLITGR 1000

1001 LQSLQTYVTQQLIRAAEIRA SANLAATKMSECVLGQSKRV DFCGKGYHLMSFPQSAPHGV VFLHVTYVPAQEKNFTTAPA ICHDGKAHFPREGVFVSNGT 1100

1101 HWFVTQRNFYEPQIITTDNT FVSGNCDVVIGIVNNTVYDP LQPELDSFKEELDKYFKNHT SPDVDLGDISGINASVVNIQ KEIDRLNEVAKNLNESLIDL 1200

1201 QELGKYEQYIKWPWYIWLGF IAGLIAIVMVTIMLCCMTSC CSCLKGCCSCGSCCKFDEDD SEPVLKGVKLHYT 1273

7.1.4 Difference of Discovered Strain 2 Spike Protein Sequence and SARS-CoV-2 Spike Protein Sequence

The format of EMBOSS Needle is that each three lines followed an empty line, the first in the three lines is 50 amino acid bases of SARS-CoV-2 original strain spike protein, the second line is the matching status, and the third line is 50 amino acid bases of discovered strain 2. | means the two bases at the position in the first protein sequence and the second protein sequence are the same, other signs mean the two bases are different.The figure is located at 7.2 and 7.2.

7.1.5 SRR11092062 Discovered Strain 1 Nucleotide Sequence

1 ATTAAAGGTTTATACCTTCC CAGGTAACAAACCAACCAAC TTTCGATCTCTTGTAGATCT GTTCTCTAAACGAACTTTAA AATCTGTGTGGCTGTCACTC 100
101 GGCTGCATGCTTAGTGCACT CACGCAGTATAATTAATAAC TAATTACTGTCGTTGACAGG ACACGAGTAACTCGTCTATC TTCTGCAGGCTGCTTACGGT 200
201 TTCGTCCGTGTTGCAGCCGA TCATCAGCACATCTAGGTTT CGTCCGGGTGTGACCGAAAG GTAAGATGGAGAGCCTTGTC CCTGGTTTCAACGAGAAAAC 300
301 ACACGTCCAACTCAGTTTGC CTGTTTTACAGGTTCGCGAC GTGCTCGCACGTGGCTTTGG AGACTCCGTGGAGGAGGTCT TATCAGAGGCACGTCAACAT 400
401 CTTAAAGATGGCACTTGTGG CTTAGTAGAAGTTGAAAAAG GCGTTTTGCCTCAACTTGAA CAGCCCTATGTGTTCATCAA ACGTTCGGATGCTCGAACTG 500
501 CACCTCATGGTCATGTTATG GTTGAGCTGGTAGCAGAACT CGAAGGCATTCAGTACGGTC GTAGCGGTGAGACACTTGGT GTCCTTGTCCCTCATGTGGG 600
601 CGAAATACCAGTGGCTTACC GCAAGGTTCTTCTTCGTAAG AACGGTAATAAAGGAGCTGG TGGCCATAGTTACGGCGCCG ATCTAAAGTCATTTGACTTA 700
701 GGCGACGAGCTTGGCACTGA TCCTTATGAAGATTTTCAAG AAAACTGGAACACTAAACAT AGCAGTGGTGTTACCCGTGA ACTCATGCGTGAGCTTAACG 800
801 GAGGGGCATACACTCGCTAT GTCGATAACAACTTCTGTGG CCCTGATGGCTACCCTCTTG AGTGCATTAAAGACCTTCTA GCACGTGCTGGTAAAGCTTC 900
901 ATGCACTTTGTCCGAACAAC TGGACTTTATTGACACTAAG AGGGGTGTATACTGCTGCCG TGAACATGAGCATGAAATTG CTTGGTACACGGAACGTTCT 1000
1001 GAAAAGAGCTATGAATTGCA GACACCTTTTGAAATTAAAT TGGCAAAGAAATTTGACACC TTCAATGGGGAATGTCCAAA TTTTGTATTTCCCTTAAATT 1100
1101 CCATAATCAAGACTATTCAA CCAAGGGTTGAAAAGAAAAA GCTTGATGGCTTTATGGGTA GAATTCGATCTGTCTATCCA GTTGCGTCACCAAATGAATG 1200
1201 CAACCAAATGTGCCTTTCAA CTCTCATGAAGTGTGATCAT TGTGGTGAAACTTCATGGCA GACGGGCGATTTTGTTAAAG CCACTTGCGAATTTTGTGGC 1300
1301 ACTGAGAATTTGACTAAAGA AGGTGCCACTACTTGTGGTT ACTTACCCCAAAATGCTGTT GTTAAAATTTATTGTCCAGC ATGTCACAATTCAGAAGTAG 1400
1401 GACCTGAGCATAGTCTTGCC GAATACCATAATGAATCTGG CTTGAAAACCATTCTTCGTA AGGGTGGTCGCACTATTGCC TTTGGAGGCTGTGTGTTCTC 1500
1501 TTATGTTGGTTGCCATAACA AGTGTGCCTATTGGGTTCCA CGTGCTAGCGCTAACATAGG TTGTAACCATACAGGTGTTG TTGGAGAAGGTTCCGAAGGT 1600
1601 CTTAATGACAACCTTCTTGA AATACTCCAAAAAGAGAAAG TCAAGATCAATATTGTTGGT GACTTTAAACTTAATGAAGA GATCGCCATTATTTTGGCAT 1700
1701 CTTTTTCTGCTTCCACAAGT GCTTTTGTGGAAACTGTGAA AGGTTTGGATTATAAAGCAT TCAAACAAATTGTTGAATCC TGTGGTAATTTTAAAGTTAC 1800
1801 AAATGGAAAAGCTAAAAAAG GTGCCTGGAATATTGGTGAA CAGAAATCAATACTGAGTCC TCTTTATGCATTTGCATCAG AGGCTGCTCGTGTTGTACGA 1900

1901 TCAATTTTCTCCCGCACTCT TGAAACTGCTCAAAATTCTG TGCGTGTTTTACAGAAGGCC GCTATAACAATACTAGATGG AATTTCACAGTATTCACTGA 2000
2001 GACTCATTGATGCTATGATG TTCACATCTGATTTGGCTAC TAACAATCTAGTTGTAATGG CCTACATTACAGGTGGTGTT GTTCAGTTGACTTCGCAGTG 2100
2101 GCTAACTAACATCTTTGGCA CTGTTTATGAAAAACTCAAA CCCGTCCTTGATTGGCTTGA AGAGAAGTTTAAGGAAGGTG TAGAGTTTCTTAGAGACGGT 2200
2201 TGGGAAATTGTTAAATTTAT CTCAACCTGTGCTTGTGAAA TTGTCGGAAGTCGGATTGTC ACCTGTGCAAAGGAAATTAA GGAGAGTGTTCAGACATTCT 2300
2301 TTAAGCTTGTAAATAAATTT TTGGCTTTGTGTGCTGACTC TATCATTATTGGTGGAGCTA AACTTAAAGCCTTGAATTTA GGTGAAACATTTGTCACGCA 2400
2401 CTCAAAGGGATTGTACAGAA AGTGTGTTAAATCCAGAGAA GAAACTGGCAAGTCGGATCG TAGAAAAGCCCCAAAAGAAA TTATCTTCTTAGAGGGAGAA 2500
2501 ACACTTCCCACAGAAGTGTT AACAGAGGAAGTTGTCTTGA AAACTGGTGATTTACAACCA TTAGAACAACCTACTAGTGA AGCTGTTGAAGCTCCATTGG 2600
2601 TTGGTACACCAGTTTGTATT AACGGGCTTATGTTGCTCGA AATCAAAGACACAGAAAAGT ACTGTGCCCTTGCACCTAAT ATGATGGTAACAAACAATAC 2700
2701 CTTCACACTCAAAGGCGGTG CACCAACAAAGGTTACTTTT GGTGATGACACTGTGATAGA AGTGCAAGGTTACAAGAGTG TGAATATCACTTTTGAACTT 2800
2801 GATGAAAGGATTGATAAAGT ACTTAATGAGAAGTGCTCTG CCTATACAGTTGAACTCGGT ACAGAAGTAAATGAGTTCGC CTGTGTTGTGGCAGATGCTG 2900
2901 TCATAAAAACTTTGCAACCA GTATCTGAATTACTTACACC ACTGGGCATTGATTTAGATG AGTGGAGTATGGCTACATAC TACTTATTTGATGAGTCTGG 3000
3001 TGAGTTTAAATTGGCTTCAC ATATGTATTGTTCTTTCTAC CCTCCAGATGAGGATGAAGA AGAAGGTGATTGTGAAGAAG AAGAGTTTGAGCCATCAACT 3100
3101 CAATATGAGTATGGTACTGA AGATGATTACCAAGGTAAAC CTTTGGAATTTGGTGCCACT TCTGCTGCTCTTCAACCTGA AGAAGAGCAAGAAGAAGATT 3200
3201 GGTTAGATGATGATAGTCAA CAAACTGTTGGTCAACAAGA CGGCAGTGAGGACAATCAGA CAATTACTATTCAAACAATT GTTGAGGTTCAACCTCAATT 3300
3301 AGAGATGGAACTTACACCAG TTGTTCAGACTATTGAAGTG AATAGTTTTAGTGGTTATTT AAAACTTACTGACAATGTAT ACATTAAAAATGCAGACATT 3400
3401 GTGGAAGAAGCTAAAAAGGT AAAACCAACAGTGGTTGTTA ATGCAGCCAATGTTTACCTT AAACATGGAGGAGGTGTTGC AGGAGCCTTAAATAAGGCTA 3500
3501 CTAACAATGCCATGCAAGTT GAATCTGATGATTACATAGC TACTAATGGACCACTTAAAG TGGGTGGTAGTTGTGTTTTA AGCGGACACAATCTTGCTAA 3600
3601 ACACTGTCTTCATGTTGTCG GCCCGAATGTTAACAAAGGT GAAGACATTCAACTTCTTAA GAGTGCTTATGAAAATTTTA ATCAGCACGAAGTTCTACTT 3700
3701 GCACCATTATTATCAGCTGG TATTTTTGGTGCTGACCCTA TACATTCTTTAAGAGTTTGT GTAGATACTGTTCGCACAAA TGTTTACTTAGCTGTCTTTG 3800
3801 ATAAAAATCTCTATGACAAA CTTGTTTCAAGCTTTTTGGA AATGAAGAGTGAAAAGCAAG TTGAACAAAAGATCGCTGAG ATTCCTAAAGAGGAAGTTAA 3900
3901 GCCATTTATAATTGAAAGTA AACCTTCAGTTGAACAGAGA AAACAAGATGATAAGAAAAT CAAAGCTTGTGTTGAAGAAG TTACAACAACTCTGGAAGAA 4000
4001 ACTAAGTTCCTCACAGAAAA CTTGTTACTTTATATTGACA TTAATGGCAATCTTCATCCA GATTCTGCCACTCTTGTTAG TGACATTGACATCACTTTCT 4100
4101 TAAAGAAAGATGCTCCATAT ATAGTGGGTGATGTTGTTCA AGAGGGTGTTTTAACTGCTG TGGTTATACCTACTAAAAAG GCTGGTGGCACTACTGAAAT 4200
4201 GCTAGCGAAAGCTTTGAGAA AAGTGCCAACAGACAATTAT ATAACCACTTACCCGGGTCA GGGTTTAAATGGTTACACTG TAGAGGATGCAAAGACAGTG 4300
4301 CTTAAAAAGTGTAAAAGTGC CTTTTACATTCTACCATCTA TTATCTCTAATGAGAAGCAA GAAATTCTTGGAACTGTTTC TTGGAATTTGCGAGAAATGC 4400
4401 TTGCACATGCAGAAGAAACA CGCAAATTAATGCCTGTCTG TGTGGAAACTAAAGCCATAG TTTCAACTATACAGCGTAAA TATAAGGGTATTAAAATACA 4500
4501 AGAGGGTGTGGTTGATTATG GTGCTAGATTTTACTTTTAC ACCAGTAAAACAACTGTAGC GTCACTTATCAACACACTTA ACGATCTAAATGAAACTCTT 4600
4601 GTTACAATGCCACTTGGCTA TGTAACACATGGCTTAAATT TGGAAGAAGCTGCTCGGTAT ATGAGATCTCTCAAAGTGCC AGCTACAGTTTCTGTTTCTT 4700
4701 CACCTGATGCTGTTACAGCG TATAATGGTTATCTTACTTC TTCTTCTAAAACACCTGAAG AACATTTTATTGAAACCATC TCACTTGCTGGTTCCTATAA 4800
4801 AGATTGGTCCTATTCTGGAC AATCTACACAACTAGGTATA GAATTTCTTAAGAGAGGTGA TAAAAGTGTATATTACACTA GTAATCCTACCACATTCCAC 4900
4901 CTAGATGGTGAAGTTATCAC CTTTGACAATCTTAAGACAC TTCTTTCTTTGAGAGAAGTG AGGACTATTAAGGTGTTTAC AACAGTAGACAACATTAACC 5000
5001 TCCACACGCAAGTTGTGGAC ATGTCAATGACATATGGACA ACAGTTTGGTCCAACTTATT TGGATGGAGCTGATGTTACT AAAATAAAACTTCATAATTC 5100
5101 ACATGAAGGTAAAACATTTT ATGTTTTACCTAATGATGAC ACTCTACGTGTTGAGGCTTT TGAGTACTACCACACAACTG ATCCTAGTTTTCTGGGTAGG 5200
5201 TACATGTCAGCATTAAATCA CACTAAAAAGTGGAAATACC CACAAGTTAATGGTTTAACT TCTATTAAATGGGCAGATAA CAACTGTTATCTTGCCACTG 5300
5301 CATTGTTAACACTCCAACAA ATAGAGTTGAAGTTTAATCC ACCTGCTCTACAAGATGCTT ATTACAGAGCAAGGGCTGGT GAAGCTGCTAACTTTTGTGC 5400
5401 ACTTATCTTAGCCTACTGTA ATAAGACAGTAGGTGAGTTA GGTGATGTTAGAGAAACAAT GAGTTACTTGTTTCAACATG CCAATTTAGATTCTTGCAAA 5500
5501 AGAGTCTTGAACGTGGTGTG TAAAACTTGTGGACAACAGC AGACAACCCTTAAGGGTGTA GAAGCTGTTATGTACATGGG CACACTTTCTTATGAACAAT 5600
5601 TTAAGAAAGGTGTTCAGATA CCTTGTACGTGTGGTAAACA AGCTACAAAATATCTAGTAC AACAGGAGTCACCTTTTGTT ATGATGTCAGCACCACCTGC 5700
5701 TCAGTATGAACTTAAGCATG GTACATTTACTTGTGCTAGT GAGTACACTGGTAATTACCA GTGTGGTCACTATAAACATA TAACTTCTAAAGAAACTTTG 5800
5801 TATTGCATAGACGGTGCTTT ACTTACAAAGTCCTCAGAAT ACAAAGGTCCTATTACGGAT GTTTTCTACAAAGAAAACAG TTACACAACAACCATAAAAC 5900
5901 CAGTTACTTATAAATTGGAT GGTGTTGTTTGTACAGAAAT TGACCCTAAGTTGGACAATT ATTATAAGAAAGACAATTCT TATTTCACAGAGCAACCAAT 6000
6001 TGATCTTGTACCAAACCAAC CATATCCAAACGCAAGCTTC GATAATTTTAAGTTTGTATG TGATAATATCAAATTTGCTG ATGATTTAAACCAGTTAACT 6100
6101 GGTTATAAGAAACCTGCTCT CCGACTTCTTAAAGTTACAT TTTTCCCTGACTTAAATGGT GATGTGGTGGCTATTGATTA TAAACACTACACACCCTCTT 6200
6201 TTAAGAAAGGAGCTAAATTG TTACATAAACCTATTGTTTG GCATGTTAACAATGCAACTA ATAAAGCCAAGTCGGATCCA AATACCTGGTGTATACGTTG 6300
6301 TCTTTGGAGCACAAAACCAG TTGAAACATCAAATTCGTTT GATGTACTGAAGTCAGAGGA CGCGCAGGGAATGGATAATC TTGCCTGCGAAGATCTAAAA 6400
6401 CCAGTCTCTGAAGAAGTAGT GGAAAATCCTACCATACAGA AAGACGTTCTTGAGTGTAAT GTGAAAACTACCGAAGTTGT AGGAGACATTATACTTAAAC 6500
6501 CAGCAAATAATAGTTTAAAA ATTACAGAAGAGGTTGGCCA CACAGATCTAATGGCTGCTT ATGTAGACAATTCTAGTCTT ACTATTAAGAAACCTAATGA 6600
6601 ATTATCTAGAGTATTAGGTT TGAAAACCCTTGCTACTCAT GGTTTAGCTGCTGTTAATAG TGTCCCTTGGGATACTATAG CTAATTATGCTAAGCCTTTT 6700
6701 CTTAACAAAGTTGTTAGTAC AACTACTAACATAGTTACAC GGTGTTTAAACCGTGTTTGT ACTAATTATATGCCTTATTT CTTTACTTTATTGCTACAAT 6800
6801 TGTGTACTTTTACTAGAAGT ACAAATTCTAGAATTAAAGC ATCTATGCCGACTACTATAG CAAAGAATACTGTTAAGAGT GTCGGTAAATTTTGTCTAGA 6900
6901 GGCTTCATTTAATTATTTGA AGTCACCTAATTTTTCTAAA CTGATAAATATTATAATTTG GTTTTTACTATTAAGTGTTT GCCTAGGTTCTTTAATCTAC 7000
7001 TCAACCGCTGCTTTAGGTGT TTTAATGTCTAATTTAGGCA TGCCTTCTTACTGTACTGGT TACAGAGAAGGCTATTTGAA CTCTACTAATGTCACTATTG 7100
7101 CAACCTACTGTACTGGTTCT ATACCTTGTAGTGTTTGTCT TAGTGGTTTAGATTCTTTAG ACACCTATCCTTCTTTAGAA ACTATACAAATTACCATTTC 7200
7201 ATCTTTTAAATGGGATTTAA CTGCTTTTGGCTTAGTTGCA CCGCTTGGCCTCCGACTTAT TCTTTTCACTAGGTTTTTCT ATGTACTTGGATTGGCTGCA 7300
7301 ATCATGCAATTGTTTTTCAG CTATTTTGCAGTACATTTTA TTAGTAATTCTTGGCTTATG TGGTTAATAATTAATCTTGT ACAAATGGCCCCGATTTCAG 7400
7401 CTATGGTTAGAATGTACATC TTCTTTGCATCATTTTATTA TGTATGGAAAAGTTATGTGC ATGTTGTAGACGGTTGTAAT TCATCAACTTGTATGATGTG 7500

7501 TTACAAACGTAATAGAGCAA CAAGAGTCGAATGTACAACT ATTGTTAATGGTGTTAGAAG GTCCTTTTATGTCTATGCTA ATGGAGGTAAAGGCTTTTGC 7600
7601 AAACTACACAATTGGAATTG TGTTAATTGTGATACATTCT GTGCTGGTAGTACATTTATT AGTGATGAAGTTGCGAGAGA CTTGTCACTACAGTTTAAAA 7700
7701 GACCAATAAATCCTACTGAC CAGTCTTCTTACATCGTTGA TAGTGTTACAGTGAAGAATG GTTCCATCCATCTTTACTTT GATAAAGCTGGTCAAAAGAC 7800
7801 TTATGAAAGACATTCTCTCT CTCATTTTGTTAACTTAGAC AACCTGAGAGCTAATAACAC TAAAGGTTCATTGCCTATTA ATGTTATAGTTTTTGATGGT 7900
7901 AAATCAAAATGTGAAGAATC ATCTGCAAAGTCGGATCGTG TTTACTACAGTCAGCTTATG TGTCAACCTATACTGTTACT AGATCAGGCATTAGTGTCTG 8000
8001 ATGTTGGTGATAGTGCGGAA GTTGCAGTTAAAATGTTTGA TGCTTACGTTAATACGTTTT CATCAACTTTTAACGTACCA ATGGAAAAACTCAAAACACT 8100
8101 AGTTGCAACTGCAGAAGCTG AACTTGCAAAGAATGTGTCC TTAGACAATGTCTTATCTAC TTTTATTTCAGCGCTTGGCC TCCGACTTGTTGATTCAGAT 8200
8201 GTAGAAACTAAAGATGTTGT TGAATGTCTTAAATTGTCAC ATCAATCTGACATAGAAGTT ACTGGCGATAGTTGTAATAA CTATATGCTCACCTATAACA 8300
8301 AAGTTGAAAACATGACACCC CGTGACCTTGGTGCTTGTAT TGACTGTAGTGCGCGTCATA TTAATGCGCAGGTAGCAAAA AGTCACAACATTGCTTTGAT 8400
8401 ATGGAACGTTAAAGATTTCA TGTCATTGTCTGAACAACTA CGAAAACAAATACGTAGTGC TGCTAAAAAGAATAACTTAC CTTTTAAGTTGACATGTGCA 8500
8501 ACTACTAGACAAGTTGTTAA TGTTGTAACAACAAAGATAG CACTTAAGGGTGGTAAAATT GTTAATAATTGGTTGAAGCA GTTAATTAAAGTTACACTTG 8600
8601 TGTTCCTTTTTGTTGCTGCT ATTTTCTATTTAATAACACC TGTTCATGTCATGTCTAAAC ATACTGACTTTTCAAGTGAA ATCATAGGATACAAGGCTAT 8700
8701 TGATGGTGGTGTCACTCGTG ACATAGCATCTACAGATACT TGTTTTGCTAACAAACATGC TGATTTTGACACATGGTTTA GCCAGCGTGGTGGTAGTTAT 8800
8801 ACTAATGACAAAGCTTGCCC ATTGATTGCTGCAGTCATAA CAAGAGAAGTGGGTTTTGTC GTGCCTGGTTTGCCTGGCAC GATATTACGCACAACTAATG 8900
8901 GTGACTTTTTGCATTTCTTA CCTAGAGTTTTTAGTGCAGT TGGTAACATCTGTTACACAC CATCAAAACTTATAGAGTAC ACTGACTTTGCAACATCAGC 9000
9001 TTGTGTTTTGGCTGCTGAAT GTACAATTTTTAAAGATGCT TCTGGTAAGCCAGTACCATA TTGTTATGATACCAATGTAC TAGAAGGTTCTGTTGCTTAT 9100
9101 GAAAGTTTACGCCCTGACAC ACGTTATGTGCTCATGGATG GCTCTATTATTCAATTTCCT AACACCAACCTTGAAGGTTC TGTTAGAGTGGTAACAACTT 9200
9201 TTGATTCTGAGTACTGTAGG CACGGCACTTGTGAAAGATC AGAAGCTGGTGTTTGTGTAT CTACTAGTGGTAGATGGGTA CTTAACAATGATTATTACAG 9300
9301 ATCTTTACCAGGAGTTTTCT GTGGTGTAGATGCTGTAAAT TTACTTACTAATATGTTTAC ACCACTAATTCAACCTATTG GTGCTTTGGACATATCAGCA 9400
9401 TCTATAGTAGCTGGTGGTAT TGTAGCTATCGTAGTAACAT GCCTTGCCTACTATTTTATG AGGTTTAGAAGAGCTTTTGG TGAATACAGTCATGTAGTTG 9500
9501 CCTTTAATACTTTACTATTC CTTATGTCATTCACTGTACT CTGTTTAACACCAGTTTACT CATTCTTACCTGGTGTTTAT TCTGTTATTTACTTGTACTT 9600
9601 GACATTTTATCTTACTAATG ATGTTTCTTTTTTAGCACAT ATTCAGTGGATGGTTATGTT CACACCTTTAGTACCTTTCT GGATAACAATTGCTTATATC 9700
9701 ATTTGTATTTCCACAAAGCA TTTCTATTGGTTCTTTAGTA ATTACCTAAAGAGACGTGTA GTCTTTAATGGTGTTTCCTT TAGTACTTTTGAAGAAGCTG 9800
9801 CGCTGTGCACCTTTTTGTTA AATAAAGAAATGTATCTAAA GTTGCGTAGTGATGTGCTAT TACCTCTTACGCAATATAAT AGATACTTAGCTCTTTATAA 9900
9901 TAAGTACAAGTATTTTAGTG GAGCAATGGATACAACTAGC TACAGAGAAGCTGCTTGTTG TCATCTCGCAAAGGCTCTCA ATGACTTCAGTAACTCAGGT 10000
10001 TCTGATGTACTTTACCAACC ACCACAAACCTCTATCACCT CAGCTGTTTTGCAGAGTGGT TTTAGAAAAATGGCATTCCC ATCTGGTAAAGTTGAGGGTT 10100
10101 GTATGGTACAAGTAACTTGT GGTACAACTACACTTAACGG TCTTTGGCTTGATGACGTAG TTTACTGTCCAAGACATGTG ATCTGCACCTCTGAAGACAT 10200
10201 GCTTAACCCTAATTATGAAG ATTTACTCATTCGTAAGTCT AATCATAATTTCTTGGTACA GGCTGGTAATGTTCAACTCA GGGTTATTGGACATTCTATG 10300
10301 CAAAATTGTGTACTTAAGCT TAAGGTTGATACAGCCAATC CTAAGACACCTAAGTATAAG TTTGTTCGCATTCAACCAGG ACAGACTTTTTCAGTGTTAG 10400
10401 CTTGTTACAATGGTTCACCA TCTGGTGTTTACCAATGTGC TATGAGGCCCGATTTCACTA TTAAGGGTTCATTCCTTAAT GGTTCATGTGGTAGTGTTGG 10500
10501 TTTTAACATAGATTATGACT GTGTCTCTTTTTGTTACATG CACCATATGGAATTACCAAC TGGAGTTCATGCTGGCACAG ACTTAGAAGGTAACTTTTAT 10600
10601 GGACCTTTTGTTGACAGGCA AACAGCACAAGCAGCTGGTA CGGACACAACTATTACAGTT AATGTTTTAGCTTGGTTGTA CGCTGCTGTTATAAATGGAG 10700
10701 ACAGGTGGTTTCTCAATCGA TTTACCACAACTCTTAATGA CTTTAACCTTGTGGCTATGA AGTACAATTATGAACCTCTA ACACAAGACCATGTTGACAT 10800
10801 ACTAGGACCTCTTTCTGCTC AAACTGGAATTGCCGTTTTA GATATGTGTGCTTCATTAAA AGAATTACTGCAAAATGGTA TGAATGGACGTACCATATTG 10900
10901 GGTAGTGCTTTATTAGAAGA TGAATTTACACCTTTGATGT TGTTAGACAATGCTCAGGTG TTACTTTCCAAAGTGCAGTG AAAAGAACAATCAAGGGTAC 11000
11001 ACACCACTGGTTGTTACTCA CAATTTTGACTTCACTTTTA GTTTTAGTCCAGAGTACTCA ATGGTCTTTGTTCTTTTTTT TGTATGAAAATGCCTTTTTA 11100
11101 CCTTTTGCTATGGGTATTAT TGCTATGTCTGCTTTTGCAA TGATGTTTGTCAAACATAAG CATGCATTTCTCTGTTTGTT TTTGTTACCTTCTCTCTTGCCA 11200
11201 CTGTAGCTTATTTTAATATG GTCTATATGCCTGCTAGTTG GGTGATGCGTATTATGACAT GGTTGGATATGGTTGATACT AGTTTGTCTGGTTTTAAGCT 11300
11301 AAAAGACTGTGTTATGTATG CATCAGCTGTAGTGTTACTA ATCCTTATGACAGCAAGAAC TGTGTATGATGATGGTGCTA GGAGAGTGTGGACACTTATG 11400
11401 AATGTCTTGACACTCGTTTA TAAAGTTTATTATGGTAATG CTTTAGATCAAGCCATTTCC ATGTGGGCTCTTATAATCTC TGTTACTTCTAACTACTCAG 11500
11501 GTGTAGTTACAACTGTCATG TTTTTGGCCAGAGGTATTGT TTTTATGTGTGTTGAGTATT GCCCTATTTTCTTCATAACT GGTAATACACTTCAGTGTAT 11600
11601 AATGCTAGTTTATTGTTTCT TAGGCTATTTTTGTACTTGT TACTTTGGCCTCTTTTGTTT ACTCAACCGCTACTTTAGAC TGACTCTTGGTGTTTATGAT 11700
11701 TACTTAGTTTCTACACAGGA GTTTAGATATATGAATTCAC AGGGACTACTCCCACCCAAG AATAGCATAGATGCCTTCAA ACTCAACATTAAATTGTTGG 11800
11801 GTGTTGGTGGCAAACCTTGT ATCAAAGTAGCCACTGTACA GTCTAAAATGTCAGATGTAA AGTGCACATCAGTAGTCTTA CTCTCAGTTTTGCAACAACT 11900
11901 CAGAGTAGAATCATCATCTA AATTGTGGGCTCAATGTGTC CAGTTACACAATGACATTCT CTTAGCTAAAGATACTACTG AAGCCTTTGAAAAAATGGTT 12000
12001 TCACTACTTTCTGTTTTGCT TTCCATGCAGGGTGCTGTAG ACATAAACAAGCTTTGTGAA GAAATGCTGGACAACAGGGC AACCTTACAAGCTATAGCCT 12100
12101 CAGAGTTTAGTTCCCTTCCA TCATATGCAGCTTTTGCTAC TGCTCAAGAAGCTTATGAGC AGGCTGTTGCTAATGGTGAT TCTGAAGTTGTTCTTAAAAA 12200
12201 GTTGAAGAAGTCTTTGAATG TGGCTAAATCTGAATTTGAC CGTGATGCAGCCATGCAACG TAAGTTGGAAAAGATGGCTG ATCAAGCTATGACCCAAATG 12300
12301 TATAAACAGGCTAGATCTGA GGACAAGAGGGCAAAAGTTA CTAGTGCTATGCAGACAATG CTTTTCACTATGCTTAGAAA GTTGGATAATGATGCACTCA 12400
12401 ACAACATTATCAACAATGCA AGAGATGGTTGTGTTCCCTT GAACATAATACCTCTTACAA CAGCAGCCAAACTAATGGTT GTCATACCAGACTATAACAC 12500
12501 ATATAAAATACGTGTGATGG TACAACATTTACTTATGCAT CAGCATTGTGGGAAATCCAA CAGGTTGTAGATGCAGATAG TAAAATTGTTCAACTTAGTG 12600
12601 AAATTAGTATGGACAATTCA CCTAATTTAGCATGGCCTCT TATTGTAACAGCTTTAAGGG CCAATTCTGCTGTCAAATTA CAGAATAATGAGCTTAGTCC 12700
12701 TGTTGCACTACGACAGATGT CTTGTGCTGCCGGTACTACA CAAACTGCTTGCACTGATGA CAATGCGTTAGCTTACTACA ACACAACAAAGGGAGGTAGG 12800
12801 TTTGTACTTGCACTGTTATC CGATTTACAGGATTTGAAAT GGGCAAGATTCCCTAAGAGT GATGGAACTGGTACTATCTA TACAGAACTGGAACCACCTT 12900
12901 GTAGGTTTGTTACAGACACA CCTAAAGGTCCTAAAGTGAA GTATTTATACTTTATTAAAG GATTAAACAACCTAAATAGA GGTATGGTACTTGGTAGTTT 13000
13001 AGCTGCCACAGTACGTCTAC AAGCTGGTAATGCAACAGAA GTGCCTGCCAATTCAACTGT ATTATCTTTCTGTGCTTTTG CTGTAGATGCTGCTAAAGCT 13100

13101 TACAAAGATTATCTAGCTAG TGGGGGACAACCAATCACTA ATTGTGTTAAGATGTTGTGT ACACACACTGGTACTGGTCA GGCAATAACAGTTACACCGG 13200
13201 AAGCCAATATGGATCAAGAA TCCTTTGGTGGTGCATCGTG TTGTCTGTACTGCCGTTGCC ACATAGATCATCCAAATCCT AAAGGATTTTGTGACTTAAA 13300
13301 AGGTAAGTATGTACAAATAC CTACAACTTGTGCTAATGAC CCTGTGGGTTTTACACTTAA AAACACAGTCTGTACCGTCT GCGGTATGTGGAAAGGTTAT 13400
13401 GGCTGTAGTTGTGATCAACT CCGCGAACCCATGCTTCAGT CAGCTGATGCACAATCGTTT TTAAACGGGTTTGCGGTGTA AGTGCAGCCCGTCTTACACC 13500
13501 GTGCGGCACAGGCACTAGTA CTGATGCCGTATACAGGGCT TTTGACATCTACAATGATAA AGTAGCTGGTTTTGCTAAAT TCCTAAAAACTAATTGTTGT 13600
13601 CGCTTCCAAGAAAAGGACGA AGATGACAATTTAATTGATT CTTACTTTGTAGTTAAGAGA CACACTTTCTCTAACTACCA ACATGAAGAAACAATTTATA 13700
13701 ATTTACTTAAGGATTGTCCA GCTGTTGCTAAACATGACTT CTTTAAGTTTAGAATAGACG GTGACATGGTACCACATATA TCACGTCAACGTCTTACTAA 13800
13801 ATACACAATGGCAGACCTCG TCTATGCTTTAAGGCATTTT GATGAAGGTAATTGTGACAC ATTAAAAGAAATACTTGTCA CATACAATTGTTGTGATGAT 13900
13901 GATTATTTCAATAAAAAGGA CTGGCATGATTTTGTAGAAA ACCCAGATATATTACGCGTA TACGCCAACTTAGGTGAACG TGTACGCCAAGCTTTGTTAA 14000
14001 AAACAGTACAATTCTGTGAT GCCATGCGAAATGCTGGTAT TGTTGGTGTACTGACATTAG ATAATCAAGATCTCAATGGT AACTGGTATGATTTCGGTGA 14100
14101 TTTCATACAAACCACGCCAG GTAGTGGAGTTCCTGTTGTA GATTCTTATTATTCATTGTT AATGCCTATATTAACCTTGA CCAGGGCTTTAACTGCAGAG 14200
14201 TCACATGTTGACACTGACTT AACAAAGCCTTACATTAAGT GGGATTTGTTAAAATATGAC TTCACGGAAGAGAGGTTAAA ACTCTTTGACCGTTATTTTA 14300
14301 AATATTGGGATCAGACATAC CACCCAAATTGTGTTAACTG TTTGGATGACAGATGCATTC TGCATTGTGCAAACTTTAAT GTTTTATTCTCTACAGTGTT 14400
14401 CCCACCTACAAGTTTTGGAC CACTAGTGAGAAAAATATTT GTTGATGGTGTTCCATTTGT AGTTTCAACTGGATACCACT TCAGAGAGCTAGGTGTTGTA 14500
14501 CATAATCAGGATGTAAACTT ACATAGCTCTAGACTTAGTT TTAAGGAATTACTTGTGTAT GCTGCTGACCCTGCTATGCA CGCTGCTTCTGGTAATCTAT 14600
14601 TACTAGATAAACGCACTACG TGCTTTTCAGTAGCTGCACT TACTAACAATGTTGCTTTTC AAACTGTCAAACCCGGTAAT TTTAACAAAGACTTCTATGA 14700
14701 CTTTGCTGTGTCTAAGGGTT TCTTTAAGGAAGGAAGTTCT GTTGAATTAAAACACTTCTT CTTTGCTCAGGATGGTAATG CTGCTATCAGCGATTATGAC 14800
14801 TACTATCGTTATAATCTACC AACAATGTGTGATATCAGAC AACTACTATTTGTAGTTGAA GTTGTTGATAAGTACTTTGA TTGTTACGATGGTGGCTGTA 14900
14901 TTAATGCTAACCAAGTCATC GTCAACAACCTAGACAAATC AGCTGGTTTTCCATTTAATA AATGGGGTAAGGCTAGACTT TATTATGATTCAATGAGTTA 15000
15001 TGAGGATCAAGATGCACTTT TCGCATATACAAAACGTAAT GTCATCCCTACTATAACCCA AATGAATCTTAAGTATGCCA TTAGTGCAAAGAATAGAGCT 15100
15101 CGCACCGTAGCTGGTGTCTC TATCTGTAGTACTATGACCA ATAGACAGTTTCATCAAAAA TTATTGAAATCAATAGCCGC CACTAGAGGAGCTACTGTAG 15200
15201 TAATTGGAACAAGCAAATTC TATGGTGGTTGGCACAACAT GTTAAAAACTGTTTATAGTG ATGTAGAAAACCCTCACCTT ATGGGTTGGGATTATCCTAA 15300
15301 ATGTGATAGAGCCATGCCTA ACATGCTTAGAATTATGGCC TCACTTGTTCTTGCTCGCAA ACATACAACGTGTTGTAGCT TGTCACACCGTTTCTATAGA 15400
15401 TTAGCTAATGAGTGTGCTCA AGTATTGAGTGAAATGGTCA TGTGTGGCGGTTCACTATAT GTTAAACCAGGTGGAACCTC ATCAGGAGATGCCACAACTG 15500
15501 CTTATGCTAATAGTGTTTTT AACATTTGTCAAGCTGTCAC GGCCAATGTTAATGCACTTT TATCTACTGATGGTAACAAA ATTGCCGATAAGTATGTCCG 15600
15601 CAATTTACAACACAGACTTT ATGAGTGTCTCTATAGAAAT AGAGATGTTGACACAGACTT TGTGAATGAGTTTTACGCAT ATTTGCGTAAACATTTCTCA 15700
15701 ATGATGATACTCTCTGACGA TGCTGTTGTGTGTTTCAATA GCACTTATGCATCTCAAGGT CTAGTGGCCAGCATAAAGAA CTTTAAGTCAGTTCTTTATT 15800
15801 ATCAAAACAATGTTTTTATG TCTGAAGCAAAATGTTGGAC TGAGACTGACCTTACTAAAG GACCTCATGAATTTTGCTCT CAACATACAATGCTAGTTAA 15900
15901 ACAGGGTGATGATTATGTGT ACCTTCCTTACCCAGATCCA TCAAGAATCCTAGGAGCCGG CTGTTTTGTAGATGATATCG TAAAAACAGATGGTACACTT 16000
16001 ATGATTGAACGGTTCGTGTC TTTAGCTATAGATGCTTACC CACTTACTAAACATCCTAAT CAGGAGTATGCTGATGTCTT TCATTTGTACTTACAATACA 16100
16101 TAAGAAAGCTACATGATGAG TTAACAGGACACATGTTAGA CATGTATTCTGTTATGCTTA CTAATGATAACACTTCAAGG TATTGGGAACCTGAGTTTTA 16200
16201 TGAGGCTATGTACACACCGC ATACAGTCTTACAGGCTGTT GGGGCTTGTGTTCTTTGCAA TTCACAGACTTCATTAAGAT GTGGTGCTTGCATACGTAGA 16300
16301 CCATTCTTATGTTGTAAATG CTGTTACGACAAGGTCATAT CAACATCACATAAATTAGTC TTGTCTGTTAATCCGTATGT TTGCAATGCTCCAGGTTGTG 16400
16401 ATGTCACAGATGTGACTCAA CTTTACTTAGGAGGTATGAG CTATTATTGTAAATCACATA AACCACCCATTAGTTTTCCA TTGTGTGCTAATGGACAAGT 16500
16501 TTTTGGTTTATATAAAAATA CATGTGTTGGTAGCGATAAT GTTACTGACTTTAATGCAAT TGCAACATGTGACTGGACAA ATGCTGGTGATTACATTTTA 16600
16601 GCTAACACCTGTACTGAAAG ACTCAAGCTTTTTGCAGCAG AAACGCTCAAAGCTACTGAG GAGACATTTAAACTGTCTTA TGGTATTGCTACTGTACGTG 16700
16701 AAGTGCTGTCTGACAGAGAA TTACATCTTTCATGGGAAGT TGGTAAACCTAGACAAGTCG GATCGCGAAATTATGTCTTT ACTGGTTATCGTGTAACTAA 16800
16801 AAACAGTAAAGTACAAATAG GAGAGTACACCTTTGAAAAA GGTGACTATGGTGATGCTGT TGTTTACCGAGGTACAACAA CTTACAAATTAAATGTTGGT 16900
16901 GATTATTTTGTGCTGACATC ACATACAGTAATGCCATTAA GTGCACCTACACTAGTGCCA CAAGAGCACTATGTTAGAAT TACTGGCTTATACCCAACAC 17000
17001 TCAATATCTCAGATGAGTTT TCTAGCAATGTTGCAAATTA TCAAAAGGTTGGTATGCAAA AGTATTCTACACTCCAGGGA CCACCTGGTACTGGTAAGAG 17100
17101 TCATTTTGCTATTGGCCTAG CTCTCTACTACCCTTCTGCT CGCATAGTGTATACAGCTTG CTCTCATGCCGCTGTTGATG CACTATGTGAGAAGGCATTA 17200
17201 AAATATTTGCCTATAGATAA ATGTAGTAGAATTATACCTG CACGTGCTCGTGTAGAGTGT TTTGATAAATTCAAAGTGAA TTCAACATTAGAACAGTATG 17300
17301 TCTTTTGTACTGTAAATGCA TTGCCTGAGACGACAGCAGA TATAGTTGTCTTTGATGAAA TTTCAATGGCCACAAATTAT GATTTGAGTGTTGTCAATGC 17400
17401 CAGATTACGTGCTAAGCACT ATGTGTACATTGGCGACCCT GCTCAATTACCTGCAAGTCG CACATTGCTAACTAAGGGCA CACTAGAACCAGAATATTTC 17500
17501 AATTCAGTGTGTAGACTTAT GAAAACTATAGGTCCAGACA TGTTCCTCGGAACTTGTCGG CGTTGTCCTGCTGAAATTGT TGACACTGTGAGTGCTTTGG 17600
17601 TTTATGATAATAAGCTTAAA GCACATAAAGACAAATCAGC TCAATGCGTTAAAATGTTTT ATAAGGGTGTTATCACGCAT GATGTTTCATCTGCAATTAA 17700
17701 CAGGCCACAAATAGGCGTGG TAAGAGAATTCCTTACACGT AACCCTGCTTGGAGAAAAGC TGTCTTTATTTCACCTTATA ATTCACAGAATGCTGTAGCC 17800
17801 TCAAAGATTTTGGGACTACC AATTCAAACTGTTGATTCAT CACAGGGCTCAGAATATGAC TATGTCATATTCACTCAAAC CACTGAAACAGCTCACTCTT 17900
17901 GTAATGTAAACAGATTTAAT GTTGCTATTACCAGAGCAAA AGTAGGCATACTTTGCATAA TGTCTGATAGAGACCTTTAT GACAAGTTGCAATTTACAAG 18000
18001 TCTTGAAATTCCACGTAGGG ACTTGGCAACTTTACAAGCT GAAAATGTAACAGGACTCTT TAAAGATTGTAGTAAGGTAA TCACTGGGTTACATCCTACA 18100
18101 CAGGCACCTACACACCTCAG TGTTGACACTAAATTCAAAA CTGAAGGTTTATGTGTTGAC ATACCTGGCATACCTAAGGA CATGACCTATAGAAGACTCA 18200
18201 TCTCTATGATGGGTTTTAAA ATGAATTATCAAGTTAATGG TTACCCTAACATGTTTATCA CCCGCGAAGAAGCTATAAGA CATGTACGTGCATGGATTGG 18300
18301 CTTCGATGTCGAGGGGTGTC ATGCTACTAGAGAAGCTGTT GGTACCAATTTACCTTTACA GCTAGGTTTTTCTACAGGTG TTAACCTAGTTGCTGTACCT 18400
18401 ACAGGTTATGTTGATACACC TAATAATACAGATTTTTCCA GAGTTAGTGCTAAACCACCG CCTGGAGATCAATTTAAACA CCTCATACCACTTATGTACA 18500
18501 AAGGACTTCCTTGGAATGTA GTGCGTATAAAGATTGTACA AATGTTAAGTGACACACTTA AAAATCTCTCTGACAGAGTC GTATTTGTCTTATGAGCACA 18600
18601 TGGCTTTGAGTTGACATCTA TGAAGTATTTTGTGAAAATA GGACCTGAGCGCACCTGTTG TCTATGTGATAGACGTGCCA CATGCTTTTCCACTGCTTCA 18700

18701 GACACTTATGCCTGTTGGCA TCATTCTATTGGATTTGATT ACGTCTATAATCCGTTTATG ATTGATGTTCAACAATGGGG TTTTACAGGTAACCTACAAA 18800
18801 GCAACCATGATCTGTATTGT CAAGTCCATGGTAATGCACA TGTAGCTAGTTGTGATGCAA TCATGACTAGGTGTCTAGCT GTCCACGAGTGCTTTGTTAA 18900
18901 GCGTGTTGACTGGACTATTG AATATCCTATAATTGGTGAT GAACTGAAGATTAATGCGGC TTGTAGAAAGGTTCAACACA TGGTTGTTAAAGCTGCATTA 19000
19001 TTAGCAGACAAATTCCCAGT TCTTCACGACATTGGTAACC CTAAAGCAATTAAGTGTGTA CCTCAAGCTGATGTAGAATG GAAGTTCTATGATGCACAGC 19100
19101 CTTGTAGTGACAAAGCTTAT AAAATAGAAGAATTATTCTA TTCTTATGCCACACATTCTG ACAAATTCACAGATGGTGTA TGCCTATTTTGGAATTGCAA 19200
19201 TGTCGATAGATATCCTGCTA ATTCCATTGTTTGTAGATTT GACACTAGAGTGCTATCTAA CCTTAACTTGCCTGGTTGTG ATGGTGGCAGTTTGTATGTA 19300
19301 AATAAACATGCATTCCACAC ACCAGCTTTTGATAAAAGTG CTTTTGTTAATTTAAAACAA TTACCATTTTTCTATTACTC TGACAGTCCATGTGAGTCTC 19400
19401 ATGGAAAACAAGTAGTGTCA GATATAGATTATGTACCACT AAAGTCTGCTACGTGTATAA CACGTTGCAATTTAGGTGGT GCTGTCTGTAGACATCATGC 19500
19501 TAATGAGTACAGATTGTATC TCGATGCTTATAACATGATG ATCTCAGCTGGCTTTAGCTT GTGGGTTTACAAACAATTTG ATACTTATAACCTCTGGAAC 19600
19601 ACTTTTACAAGACTTCAGAG TTTAGAAAATGTGGCTTTTA ATGTTGTAAATAAGGGACAC TTTGATGGACAACAGGGTGA AGTACCAGTTTCTATCATTA 19700
19701 ATAACACTGTTTACACAAAA GTTGATGGTGTTGATGTAGA ATTGTTTGAAAATAAAACAA CATTACCTGTTAATGTAGCA TTTGAGCTTTGGGCTAAGCG 19800
19801 CAACATTAAACCAGAAGTCG GATCGTAGCCACTCAATAAT TTGGGTGTGGACATTGCTGC TAATACTGTGATCTGGGACT ACAAAAGAGATGCTCCAGCA 19900
19901 CATATATCTACTATTGGTGT TTGTTCTATGACTGACATAG CCAAGAAACCAACTGAAACG ATTTGTGCACCACTCACTGT CTTTTTTGATGGTAGAGTTG 20000
20001 ATGGTCAAGTAGACTTATTT AGAAATGCCCGTAATGGTGT TCTTATTACAGAAGGTAGTG TTAAAGGTTTACAACCATCT GTAGGTCCCAAACAAGCTAG 20100
20101 TCTTAATGGAGTCACATTAA TTGGAGAAGCCGTCAAAACA CAGTTCAATTATTATAAGAA AGTTGATGGTGTTGTCCAAC AATTACCTGAAACTTACTTT 20200
20201 ACTCAGAGTAGAAATTTACA AGAATTTAAACCCAGGAGTC AAATGGAAATTGATTTCTTA GAATTAGCTATGGATGAATT CATTGAACGGTATAAATTAG 20300
20301 AAGGCTATGCCTTCGAACAT ATCGTTTATGGAGATTTTAG TCATAGTCAGTTAGGTGGTT TACATCTACTGATTGGACTA GCTAAACGTTTTAAGGAATC 20400
20401 ACCTTTTGAATTAGAAGATT TTATTCCTATGGACAGTACA GTTAAAAACTATTTCATAAC AGATGCGCAAACAGGTTCAT CTAAGTGTGTGTGTTCTGTT 20500
20501 ATTGATTTATTACTTGATGA TTTTGTTGAAATAATAAAAT CCCAAGATTTATCTGTAGTT TCTAAGGTTGTCAAAGTGAC TATTGACTATACAGAAATTT 20600
20601 CATTTATGCTTTGGTGTAAA GATGGCCATGTAGAAACATT TTACCCAAAATTACAATCTA GTCAAGCGTGGCAACCGGGT GTTGCTATGCCTAATCTTTA 20700
20701 CAAAATGCAAAGAATGCTAT TAGAAAAGTGTGACCTTCAA AATTATGGTGATAGTGCAAC ATTACCTAAAGGCATAATGA TGAATGTCGCAAAATATACT 20800
20801 CAACTGTGTCAATATTTAAA CACATTAACATTAGCTGTAC CCTATAATATGAGAGTTATA CATTTTGGTGCTGGTTCTGA TAAAGGAGTTGCACCAGGTA 20900
20901 CAGCTGTTTTAAGACAGTGG TTGCCTACGGGTACGCTGCT TGTCGATTCAGATCTTAATG ACTTTGTCTCTGATGCAGAT TCAACTTTGATTGGTGATTG 21000
21001 TGCAACTGTACATACAGCTA ATAAATGGGATCTCATTATT AGTGATATGTACGACCCTAA GACTAAAAATGTTACAAAAG AAAATGACTCTAAAGAGGGT 21100
21101 TTTTTCACTTACATTTGTGG GTTTATACAACAAAAGCTAG CTCTTGGAGGTTCCGTGGCT ATAAAGATAACAGAACATTC TTGGAATGCTGATCTTTATA 21200
21201 AGCTCATGGGACACTTCGCA TGGTGGACAGCCTTTGTTAC TAATGTGAATGCGTCATCAT CTGAAGCATTTTTAATTGGA TGTAATTATCTTGGCAAACC 21300
21301 ACGCGAACAAATAGATGGTT ATGTCATGCATGCAAATTAC ATATTTTGGAGGAATACAAA TCCAATTCAGTTGTCTTCCT ATTCTTTATTTGACATGAGT 21400
21401 AAATTTCCCCTTAAATTAAG GGGTACTGCTGTTATGTCTT TAAAAGAAGGTCAAATCAAT GATATGATTTTATCTCTTCT TAGTAAAGGTAGACTTATAA 21500
21501 TTAGAGAAAACAACAGAGTT GTTATTTCTAGTGATGTTCT TGTTAACAACTAAACGAACA ATGTTTGTTTTTCTTGTTTT ATTGTCACTAGTCTCTAGTC 21600
21601 AGTGTGTTAATCTTACAACC AGAACTCAATTACCCCCTGC ATACACTAATTCTTTCACAC GTGGTGTTTATTACCCTGAC AAAGTTTTCAGATCCTCAGT 21700
21701 TTTACATTCAACTCAGGACT TGTTCTTACCTTTCTTTTCC AATGTTACTTGGTTCCATGC TATACATGTCTCTGGGACCA ATGGTACTAAGAGGTTTGAT 21800
21801 AACCCTGTCCTACCATTTAA TGATGGTGTTTATTTTGCTT CCACTGAGAAGTCTAACATA ATAAGAGGCTGGATTTTTGG TACTACTTTAGATTCGAAGA 21900
21901 CCCAGTCCCTACTTATTGTT AATAACGCTACTAATGTTGT TATTAAAGTCTGTGAATTTC AATTTTGTAATGATCCATTT TTGGGTGTTTATTACCACAA 22000
22001 AAACAACAAAAGTTGGATGG AAAGTGAGTTCAGAGTTTAT TCTAGTGCGAATAATTGCAC TTTTGAATATGTCTCTCAGC CTTTTCTTATGGACCTTGAA 22100
22101 GGAAAACAGGGGAATTTCAA AAATCTTAGGGAATTTGTGT TTAAGAATATTGATGGTTAT TTTAAAATATATTCTAAGCA CACGCCTATTAATTTAGTGC 22200
22201 GTGATCTCCCTCAGGGTTTT TCGGCTTTAGAACCATTGGT AGATTTGCCAATAGGTATTA ACATCACTAGGCTTCAAACT TTACTTGCTTTACATAGAAG 22300
22301 TTATTTGACTCCTGGTGATT CTTCTTCAGGTTGGACAGCT GGTGCTGCAGCTTATTATGT GGGTTATCTTCAACCTAGGA CTTTTCTATTAAAATATAAT 22400
22401 GAAAATGGAACCATTACAGA TGCTGTAGACTGTGCACTTG ACCCTCTCTCAGAAACAAAG TGTACGTTGAAATCCTTCAC TGTAGAAAAAGGAATCTATC 22500
22501 AAACTTCTAACTTTAGAGTC CAACCAACAGAATCTATTGT TAGATTTCCTAATATTACAA ACTTGTGCCCTTTTGGTGAA GTTTTTAACGCCACCAGATT 22600
22601 TGCATCTGTTTATGCTTGGA ACAGGAAGAGAATCAGCAAC TGTGTTGCTGATTATTCTGT CCTATATAATTCCGCATCAT TTTCCACTTTTAAGTGTTAT 22700
22701 GGAGTGTCTCCTACTAAATT AAATGATCTCTGCTTTACTA ATGTCTATGCAGATTCATTT GTAATTAGAGGTGATGAAGT CAGACAAATCGCTCCAGGGC 22800
22801 AAACTGGAAAGATTGCTGAT TATAATTATAAATTACCAGA TGATTTTACAGGCTGCGTTA TAGCTTGGAATTCTAACAAT CTTGATTCTAAGGTTGGTGG 22900
22901 TAATTATAATTACCTGTATA GATTGTTTAGGAAGTCTAAT CTCAAACCTTTTGAGAGAGA TATTTCAACTGAAATCTATC AGGCCGGAAGCACACCTTGT 23000
23001 AATGGTGTTGAAGGTTTTAA TTGTTACTTTCCTTTACAAT CATATGGTTTCCAACCCACT AATGGTGTTGGTTACCAACC ATACAGAGTAGTAGTACTTT 23100
23101 CTTTTGAACTTCTACATGCA CCAGCAACTGTTTGTGGACC TAAAAAGTCTACTAATTTGG TTAAAAACAAATGTGTCAAT TTCAACTTCAATGGTTTAAC 23200
23201 AGGCACAGGTGTTCTTACTG AGTCTAACAAAAAGTTTCTG CCTTTCCAACAATTTGGCAG AGACATTGCTGACACTACTG ATGCTGTCCGTGATCCACAG 23300
23301 ACACTTGAGATTCTTGACAT TACACCATGTTCTTTTGGTG GTGTCAGTGTTATAACACCA GAAGTCGATACTTCTAACCA GGTTGCTGTTCTTTATCAGG 23400
23401 ATGTTAACTGCACAGAAGTC CCTGTTGCTATTCATGCAGA TCAACTTACTCCTACTTGGC GTGTTTATTCTACAGGTTCT AATGTTTTTCAAACACGTGC 23500
23501 AGGCTGTTTAATAGGGGCTG AACATGTCAACAACTCATAT GAGTGTGACATACCCATTGG TGCAGGTATATGCGCTAGTT ATCAGACTCAGACTAATTCT 23600
23601 CCCCGGCGGGCACGTAGTGT AGCTAGTCAATCCATCATTG CCTACACTATGTCACTTGGT GCAGAAAATTCAGTTGCTTA CTCTAATAACTCTATTGCCA 23700
23701 TACCCACAAATTTTACTATT AGTGTTACCACAGAAATTCT ACCAGTGTCTATGACCAAGA CATCAGTAGATTGTACAATG TACATTTGTGGTGATTCAAC 23800
23801 TGAATGCAGCAATCTTTTGT TGCAATATGGCAGTTTTTGT ACACAATTAAACCGTGGTTT AACTGGAATAGCTGTTGAAC AAGACAAAAACACCCAAGAA 23900
23901 GTTTTTGCACAAGTCAAACA AATTTACAAAACACCACCAA TTAAAGATTTTGGTGGTTTT AATTTTTCACAAATATTACC AGATCCATCAAAACCAAGCA 24000
24001 AGAGGTCATTTATTGAAGAT CTACTTTTCAACAAAGTGAC ACTTGCAGATGCTGGCTTCA TCAAACAATATGGTGATTGC CTTGGTGATATTGCTGCTAG 24100
24101 AGACCACATTTGTGCACAAA AGTTTAACGGCCTTACTGTT TTGCCACCTTTGCTCACAGA TGAAATGATTGCTCAATACA CTTCTGCACTGTTAGCGGGT 24200
24201 ACAATCACTTCTGGTTGGAC CTTTGGTGCAGGTGCTGCAT TACAAATACCATTTGCAATG CAAATGGCTTATAGGTTTAA TGGTATTGGAGTTACACAGA 24300

24301 ATGTTCTCTATGAGAACCAA AAATTGATTGCCAACCAATT TAATAGTGCTATTGGCAAAA TTCAAGACTCACTTTCTTCC ACAGCAAGTGCACTTGGAAA 24400
24401 ACTTCAAGATGTGGTCAACC AAAATGCACAAGCTTTAAAC ACGCTTGTTAAACAACTTAG CTCCAATTTTGGTGCAATTT CAAGTGTTTTAAATGATATC 24500
24501 CTTTCACGTCTTGACAAAGT TGAGGCTGAAGTGCAAATTG ATAGGTTGATCACAGGCAGA CTTCAAAGTTTGCAGACATA TGTGACTCAACAATTAATTA 24600
24601 GAGCTGCAGAAATCAGAGCT TCTGCTAATCTTGCTGCTAC TAAAATGTCACTCCGACTTC TTGGACAATCAAAAAGAGTT GATTTTTGTGGAAAGGGCTA 24700
24701 TCATCTTATGTCCTTCCCTC AGTCAGCACCTCATGGTGTA GTCTTCTTGCATGTGACTTA TGTCCCTGCACAAGAAAAGA ACTTCACAACTGCTCCTGCC 24800
24801 AAGTCGGATGATGGAAAAGC ACACTTTCCTCGTGAAGGTG TCTTTGTTTCAAATGGCACA CACTGGTTTGTAACACAAAG GAATTTTTATGAACCACAAA 24900
24901 TCATTACTACAGACAACACA TTTGTGTCTGGTAACTGTGA TGTTGTAATAGGAATTGTCA ACAACACAGTTTATGATCCT TTGCAACCTGAATTAGACTC 25000
25001 ATTCAAGGAGGAGTTAGATA AATATTTTAAGAATCATACA TCAGCAGATGTTGATTTAGG TGACATCTCTGGCATTAATG CTTCAGTTGTAAACATTCAA 25100
25101 AAAGAAATTGACCGCCTCAA TGAGGTTGCCAAGAATTTAA ATGAATCTCTCATCGATCTC CAAGAACTTGGAAAGTATGA GCAGTATATAAAATGGCCAT 25200
25201 GGTACATTTGGCTAGGTTTT ATAGCTGGCTTGATTGCCAT AGTAATGGTGACAATTATGC TTTGCTGTATGACCAGTTGC TGTAGTTGTCTCAAGGGCTG 25300
25301 TTGTTCTTGTGGATCCTGCT GCAAATTTGATGAAGACGAC TCTGAGCCAGTGCTCAAAGG AGTCAAATTACATTACACAT AAACGAACTTATGGATTTGT 25400
25401 TTATGAGAATCTTCACAATT GGAACTGTAACTTTGAAGCA AGGTGAAATCAAGGATGCTA CTCCTTCAGATTTTGTTCGC GCTACTGCAACGATACCGAT 25500
25501 ACAAGCCTCACTCCCCCTCC GACTTCTTATTGTTGGCGTT GCACTTCTTGCTGTTTTTCA GAGCGCTTCCAAAATCATAA CCCTCAAAAAGAGATGGCAA 25600
25601 CTAGCACTCTCCAAGGGTGT TCACTTTGTTTGCAACTTGC TGTTGTTGTTTGTAACAGTT TACTCAAGTCGGATCCTCGT TGCTGCTGGCCTTGAAGCCC 25700
25701 CTTTTCTCTATCTTTATGCT TTAGTCTACTTCTTGCAGAG TATAAACTTTGTAAGAATAA TAATGAGGCTTTGGCTTTGC TGGAAATGCCGTTCCAAAAA 25800
25801 CCCATTACTTTATGATGCCA ACTATTTTCTTTGCTGGCAT ACTAATTGTTACGACTATTG TATACCTTACAATAGTGTAA CTTCTTCAATTGTCATTACT 25900
25901 TCAGGTGATGGCACAACAAG TCCTATTTCTGAACATGACT ACCAGATTGGTGGTTATACT GAAAAATGGGAATCTGGAGT AAAAGACTGTGTTGTATTAC 26000
26001 ACAGTTACTTCACTTCAGAC TATTACCAGCTGTACTCAAC TCAATTGAGTACAGACACTG GTGTTGAACATGTTACCTTC TTCATCTACAATAAAATTGT 26100
26101 TGATGAGCCTGAAGAACATG TCCAAATTCACACAATCGAC GGTTCATCCGGAGTTGTTAA TCCAGTAATGGAACCAATTT ATGATGAACCGACGACGACT 26200
26201 ACTAGCGTGCCTTTGTAAGC ACAAGCTGATGAGTACGAAC TTATGTACTCATTCGTTTCG GAAGAGACAGGTACGTTAAT AGTTAATAGCGTACTTCTTT 26300
26301 TTCTTGCTTTCGTGGTATTC TTGCTAGTTACACTAGCCAT CCTTACTGCGCATCGATTGT GTGCGTACTGCTGCAATATT GTTAACGTGAGTCTTGTAAA 26400
26401 ACCTTCTTTTTACGTTTACT CTCGTGTTAAAAATCTGAAT TCTTCTAGAGTTCCTGATCT TCTGGTCTAAACGAACTAAA TATTATATTAGTTTTTCTGT 26500
26501 TTGGAACTTTAATTTTAGCC ATGGCAGATTCCAACGGTAC TATTACCGTTGAAGAGCTTA AAAAGCTCCTTGAACAATGG AACCTAGTAATAGGTTTCCT 26600
26601 ATTCCTTACATGGATTTGTC TTCTACAATTTGCCTATGCC AACAGGAATAGGTTTTTGTA TATAATTAAGTTAATTTTCC TCTGGCTGTTATGGCCAGTA 26700
26701 ACTTTAGCTTGTTTTGTGCT TGCTGCTGTTTACAGAATAA ATTGGATCACCGGTGGAATT GCTATCGCAATGGCTTGTCT TGTAGGCTTGATGTGGCTCA 26800
26801 GCTACTTCATTGCTTCTTTC AGACTGTTTGCGCTTGGCCT CCGACTTTGGTCATTCAATC CAGAAACTAACATTCTTCTC AACGTGCCACTCCATGGCAC 26900
26901 TATTCTGACCAGACCGCTTC TAGAAAGTGAACTCGTAATC GGAGCTGTGATCCTTCGTGG ACATCTTCGTATTGCTGGAC ACCATCTAGGACGCTGTGAC 27000
27001 ATCAAGGACCTGCCAAAAGA AATCACTGTTGCTACATCAC GAACGCTTTCTTATTACAAA TTGGGAGCTTCGCAGCGTGT AGCAGGTGACTCAGGTTTTG 27100
27101 CTGCATACAGTCGCTACAGG ATTGGCAACTATAAATTAAA CACAGACCATTCCAGTAGCA GTGACAATATTGCTTTGCTT GTACAGTAAGTGACAACAGA 27200
27201 TGTTTCATCTCGTTGACTTT CAGGTTACTATAGCAGAGAT ATTACTAATTATTATGAGGA CTTTTAAAGTTTCCATTTGG AATCTTGATTACATCATAAA 27300
27301 CCTCATAATTAAAAATTTAT CTAAGTCACTAACTGAGAAT AAATATTCTCAATTAGATGA AGAGCAACCGATGGAGATTG ATTAAACGAACATGAAAATT 27400
27401 ATTCTTTTCTTGGCACTGAT AACACTCGCTACTTGTGAGC TTTATCACTACCAAGAGTGT GTTAGAGGTACAACAGTACT TTTAAAAGAACCTTGCTCTT 27500
27501 CTGGAACATACGAGGGCAAT TCACCATTTCATCCTCTAGC TGATAACAAATTTGCACTGA CTTGCTTTAGCACTCAATTT GCTTTTGCTTGTCCTGACGG 27600
27601 CGTAAAACACGTCTATCAGT TACGTGCCAGATCAGTTTCA CCTAAACTGTTCATCAGACA AGAGGAAGTTCAAGAACTTT ACTCTCCAATTTTTCTTATT 27700
27701 GTTGCGGCAATAGTGTTTAT AACACTTTGCTTCACACTCA AAAGAAAGACAGAATGATTG AACTTTCATTAATTGACTTC TATTTGTGCTTTTTAGCCTT 27800
27801 TCTGCTATTCCTTGTTTTAA TTATGCTTATTATCTTTTGG TTCTCACTTGAACTGCAAGA TCATAATGAAACTTGTCACG CCGAAACGAACATGAAATTT 27900
27901 CTTGTTTTCTTAGGAATCAT CACAACTGTAGCTGCATTTC ACCAAGAATGTAGTTTACAG TCATGTACTCAACATCAACC ATATGTAGTTGATGACCCGT 28000
28001 GTCCTATTCACTTCTATTCT AAATGGTATATTAGAGTAGG AGCTAGAAAATCAGCACCTT TAATTGAATTGTGCGTGGAT GAGGCTGGTTCTAAATCACC 28100
28101 CATTCAGTACATCGATATCG GTAATTATACAGTTTCCTGT TTACCTTTTACAATTAATTG CCAGGAACCTAAATTGGGTA GTCTTGTAGTGCGTTGTTCG 28200
28201 TTCTATGAAGACTTTTTAGA GTATCATGACGTTCGTGTTG TTTTAGATTTCATCTAAACG AACAAACTAAAATGTCTGAT AATGGACCCCGAAATCAGCG 28300
28301 AAATGCACCCCGCATTACGT TTGGTGGACCCTCAGATTCA ACTGGCAGTAACCAGAATGG AGAACGCAGTGGGGCGCGAT CAAAACAACGTCGGCCCCAA
28400
28401 GGTTTACCCAATAATACTGC GTCTTGGTTCACCGCTCTCA CTCAACATGGCAAGGAAGAC CTTAAATTCCCTCGAGGACA AGGCGTTCCAATTAACACCA 28500
28501 ATAGCAGCCCAGATGACCAA ATTGGCTACTACCGAAGAGC TACCAGACGAATTCGTGGTG GTGACGGTAAAATGAAAGAT CTCAGTCCAAGATGGTATTT 28600
28601 CTACTACCTAGGAACTGGGC CAGAAGCTGGACTTCCCTAT GGTGCTAACAAAGACGGCAT CATATGGGTTGCAACTGAGG GAGCCTTGAATACACCAAAA 28700
28701 GATCACATTGGCACCCGCAA TCCTGCTAACAATGCTGCAA TCGTGCTACAACTTCCTCAA GGAACAACATTGCCCAAAGG CTTCTACGCAGAAGGGAGCA 28800
28801 GAGGCGGCAGTCAAGCCTCT TCTCGTTCCTCATCACGTAG TCGCAACAGTTCAAGAAATT CAACTCCAGGCAGCAGTAGG GGAACTTCTCCTGCTAGAAT 28900
28901 GGCTGGCAATGGCGGTGATG CTGCTCTTGCTTTGCTGCTG CTTGACAGATTGAACCAGCT TGAGAGCAAAATGTCTGGTA AAGGCCAACAACAACAAGGC
29000
29001 CAAACTGTCACTAAGAAATC TGCTGCTGAGGCTTCTAAGA AGCCTCGGCAAAAACGTACT GCCCCTAAAGCATACAATGT AACACAAGCTTTCGGCAGAC 29100
29101 GTGGTCCAGAACAAACCCAA GGAAATTTTGGGGACCAGGA ACTAATCAGACAAGGAACTG ATTACAAACATTGGCCGCAA ATTGCACAATTTGCCCCCAG 29200
29201 CGCTTCAGCGTTCTTCGGAA TGTCGCGCATTGGCATGGAA GTCACACCTTCGGGAACGTG GTTGACCTACACAGGTGCCA TCAAATTGGATGACAAAGAT
29300
29301 CCAAATTTCAAAGATCAAGT CATTTTGCTGAATAAGCATA TTGACGCATACAAAACATTC CCACCAACAGAGCCTAAAAA GGACAAAAAGAAGAAGGCTG 29400
29401 ATGAAACTCAAGCCTTACCG CAGAGACAGAAGAAACAGCA AACTGTGACTCTTTTTCCTG CTGCAGATTTGGATGATTTC TCCAAACAATTGCAACAATC 29500
29501 CATGAGCAGTGCTGACTCAA CTCAGGCCTAAACTCATGCA GACCACACAAGGCAGATGGG CTATATAAACGTTTTCGCTT TTCCGTTTACGATATATAGT 29600

29601 CTACTCTTGTGCAGAATGAA TTCTCGTAACTACATAGCAC AAGTAGATGTAGTTAACTTT AATCTCACATAGCAATCTTT AATCAGTGTGTAACATTAGG 29700
29701 GAGGACTTGAAAGAGCCACC ACATTTTCACCGAGGCCACG CGGAGTACGATCGAGTGTAC AGTGAACAATGCTAGGGAGA GCTGCCTATATGGAAGAGCC 29800
29801 CTAATGTGTAAAATTAATTT TAGTAGTGCTATCCCCAAGT CGGATTAATAGCTTCTTAGG AGAATGACAAAAAAAAAAAA AAAAAAAAA 29889

7.1.6 SRR11092062 Discovered Strain 2 Nucleotide Sequence

1 ATTAAAGGTTTATACCTTCC CAGGTAACAAACCAACCAAC TTTCGATCTCTTGTAGATCT GTTCTCTAAACGAACTTTAA AATCTGTGTGGCTGTCACTC 100

101 GGCTGCATGCTTAGTGCACT CACGCAGTATAATTAATAAC TAATTACTGTCGTTGACAGG ACACGAGTAACTCGTCTATC TTCTGCAGGCTGCTTACGGT 200

201 TTCGTCCGTGTTGCAGCCGA TCATCAGCACATCTAGGTTT CGTCCGGGTGTGACCGAAAG GTAAGATGGAGAGCCTTGTC CCTGGTTTCAACGAGAAAAC 300

301 ACACGTCCAACTCAGTTTGC CTGTTTTACAGGTTCGCGAC GTGCTCGTACGTGGCTTTGG AGACTCCGTGGAGGAGGTCT TATCAGAGGCACGTCAACAT 400

401 CTTAAAGATGGCACTTGTGG CTTAGTAGAAGTTGAAAAAG GCGTTTTGCCTCAACTTGAA CAGCCCTATGTGTTCATCAA ACGTTCGGATGCTCGAACTG 500

501 CACCTCATGGTCATGTTATG GTTGAGCTGGTAGCAGAACT CGAAGGCATTCAGTACGGTC GTAGTGGTGAGACACTTGGT GTCCTTGTCCCTCATGTGGG 600

601 CGAAATACCAGTGGCTTACC GCAAGGTTCTTCTTCGTAAG AACGGTAATAAAGGAGCTGG TGGCCATAGTTACGGCGCCG ATCTAAAGTCATTTGACTTA 700

701 GGCGACGAGCTTGGCACTGA TCCTTATGAAGATTTTCAAG AAAACTGGAACACTAAACAT AGCAGTGGTGTTACCCGTGA ACTCATGCGTGAGCTTAACG 800

801 GAGGGGCATACACTCGCTAT GTCGATAACAACTTCTGTGG CCCTGATGGCTACCCTCTTG AGTGCATTAAAGACCTTCTA GCACGTGCTGGTAAAGCTTC 900

901 ATGCACTTTGTCCGAACAAC TGGACTTTATTGACACTAAG AGGGGTGTATACTGCTGCCG TGAACATGAGCATGAAATTG CTTGGTACACGGAACGTTCT 1000

1001 GAAAAGAGCTATGAATTGCA GACACCTTTTGAAATTAAAT TGGCAAAGAAATTTGACACC TTCAATGGGGAATGTCCAAA TTTTGTATTTCCCTTAAATT 1100

1101 CCATAATCAAGACTATTCAA CCAAGGGTTGAAAAGAAAAA GCTTGATGGCTTTATGGGTA GAATTCGATCTGTCTATCCA GTTGCGTCACCAAATGAATG 1200

1201 CAACCAAATGTGCCTTTCAA CTCTCATGAAGTGTGATCAT TGTGGTGAAACTTCATGGCA GACGGGCGATTTTGTTAAAG CCACTTGCGAATTTTGTGGC 1300

1301 ACTGAGAATTTGACTAAAGA AGGTGCCACTACTTGTGGTT ACTTACCCCAAAATGCTGTT GTTAAAATTTATTGTCCAGC ATGTCACAATTCAGAAGTAG 1400

1401 GACCTGAGCATAGTCTTGCC GAATACCATAATGAATCTGG CTTGAAAACCATTCTTCGTA AGGGTGGTCGCACTATTGCC TTTGGAGGCTGTGTGTTCTC 1500

1501 TTATGTTGGTTGCCATAACA AGTGTGCCTATTGGGTTCCA CGTGCTAGCGCTAACATAGG TTGTAACCATACAGGTGTTG TTGGAGAAGGTTCCGAAGGT 1600

1601 CTTAATGACAACCTTCTTGA AATACTCCAAAAAGAGAAAG TCAACATCAATATTGTTGGT GACTTTAAACTTAATGAAGA GATCGCCATTATTTTGGCAT 1700

1701 CTTTTTCTGCTTCCACAAGT GCTTTTGTGGAAACTGTGAA AGGTTTGGATTATAAAGCAT TCAAACAAATTGTTGAATCC TGTGGTAATTTTAAAGTTAC 1800

1801 AAAAGGAAAAGCTAAAAAAG GTGCCTGGAATATTGGTGAA CAGAAATCAATACTGAGTCC TCTTTATGCATTTGCATCAG AGGCTGCTCGTGTTGTACGA 1900

1901 TCAATTTTCTCCCGCACTCT TGAAACTGCTCAAAATTCTG TGCGTGTTTTACAGAAGGCC GCTATAACAATACTAGATGG AATTTCACAGTATTCACTGA 2000

2001 GACTCATTGATGCTATGATG TTCACATCTGATTTGGCTAC TAACAATCTAGTTGTAATGG CCTACATTACAGGTGGTGTT GTTCAGTTGACTTCGCAGTG 2100

2101 GCTAACTAACATCTTTGGCA CTGTTTATGAAAAACTCAAA CCCGTCCTTGATTGGCTTGA AGAGAAGTTTAAGGAAGGTG TAGAGTTTCTTAGAGACGGT 2200

2201 TGGGAAATTGTTAAATTTAT CTCAACCTGTGCTTGTGAAA TTGTCGGTGGACAAATTGTC ACCTGTGCAAAGGAAATTAA GGAGAGTGTTCAGACATTCT 2300

2301 TTAAGCTTGTAAATAAATTT TTGGCTTTGTGTGCTGACTC TATCATTATTGGTGGAGCTA AACTTAAAGCCTTGAATTTA GGTGAAACATTTGTCACGCA 2400

2401 CTCAAAGGGATTGTACAGAA AGTGTGTTAAATCCAGAGAA GAAACTGGCCTACTCATGCC TCTAAAAGCCCCAAAAGAAA TTATCTTCTTAGAGGGAGAA 2500

2501 ACACTTCCCACAGAAGTGTT AACAGAGGAAGTTGTCTTGA AAACTGGTGATTTACAACCA TTAGAACAACCTACTAGTGA AGCTGTTGAAGCTCCATTGG 2600

2601 TTGGTACACCAGTTTGTATT AACGGGCTTATGTTGCTCGA AATCAAAGACACAGAAAAGT ACTGTGCCCTTGCACCTAAT ATGATGGTAACAAACAATAC 2700

2701 CTTCACACTCAAAGGCGGTG CACCAACAAAGGTTACTTTT GGTGATGACACTGTGATAGA AGTGCAAGGTTACAAGAGTG TGAATATCACTTTTGAACTT 2800

2801 GATGAAAGGATTGATAAAGT ACTTAATGAGAAGTGCTCTG CCTATACAGTTGAACTCGGT ACAGAAGTAAATGAGTTCGC CTGTGTTGTGGCAGATGCTG 2900

2901 TCATAAAAACTTTGCAACCA GTATCTGAATTACTTACACC ACTGGGCATTGATTTAGATG AGTGGAGTATGGCTACATAC TACTTATTTGATGAGTCTGG 3000

3001 TGAGTTTAAATTGGCTTCAC ATATGTATTGTTCTTTCTAC CCTCCAGATGAGGATGAAGA AGAAGGTGATTGTGAAGAAG AAGAGTTTGAGCCATCAACT 3100

3101 CAATATGAGTATGGTACTGA AGATGATTACCAAGGTAAAC CTTTGGAATTTGGTGCCACT TCTGCTGCTCTTCAACCTGA AGAAGAGCAAGAAGAAGATT 3200

3201 GGTTAGATGATGATAGTCAA CAAACTGTTGGTCAACAAGA CGGCAGTGAGGACAATCAGA CAACTACTATTCAAACAATT GTTGAGGTTCAACCTCAATT 3300

3301 AGAGATGGAACTTACACCAG TTGTTCAGACTATTGAAGTG AATAGTTTTAGTGGTTATTT AAAACTTACTGACAATGTAT ACATTAAAAATGCAGACATT 3400

3401 GTGGAAGAAGCTAAAAAGGT AAAACCAACAGTGGTTGTTA ATGCAGCCAATGTTTACCTT AAACATGGAGGAGGTGTTGC AGGAGCCTTAAATAAGGCTA 3500

3501 CTAACAATGCCATGCAAGTT GAATCTGATGATTACATAGC TACTAATGGACCACTTAAAG TGGGTGGTAGTTGTGTTTTA AGCGGACACAATCTTGCTAA 3600

3601 ACACTGTCTTCATGTTGTCG GCCCCAATGTTAACAAAGGT GAAGACATTCAACTTCTTAA GAGTGCTTATGAAAATTTTA ATCAGCACGAAGTTCTACTT 3700

3701 GCACCATTATTATCAGCTGG TATTTTTGGTGCTGACCCTA TACATTCTTTAAGAGTTTGT GTAGATACTGTTCGCACAAA TGTCTACTTAGCTGTCTTTG 3800

3801 ATAAAAATCTCTATGACAAA CTTGTTTCAAGCTTTTTGGA AATGAAGAGTGAAAAGCAAG TTGAACAAAAGATCGCTGAG ATTCCTAAAGAGGAAGTTAA 3900

3901 GCCATTTATAACTGAAAGTA AACCTTCAGTTGAACAGAGA AAACAAGATGATAAGAAAAT CAAAGCTTGTGTTGAAGAAG TTACAACAACTCTGGAAGAA 4000

4001 ACTAAGTTCCTCACAGAAAA CTTGTTACTTTATATTGACA TTAATGGCAATCTTCATCCA GATTCTGCCACTCTTGTTAG TGACATTGACATCACTTTCT 4100

4101 TAAAGAAAGATGCTCCATAT ATAGTGGGTGATGTTGTTCA AGAGGGTGTTTTAACTGCTG TGGTTATACCTACTAAAAAG GCTGGTGGCACTACTGAAAT 4200

4201 GCTAGCGAAAGCTTTGAGAA AAGTGCCAACAGACAATTAT ATAACCACTTACCCGGGTCA GGGTTTAAATGGTTACACTG TAGAGGAGGCTAAGACAGTG 4300

4301 CTTAAAAAGTGTAAAAGTGC CTTTTACATTCTACCATCTA TTATCTCTAATGAGAAGCAA GAAATTCTTGGAACTGTTTC TTGGAATTTGCGAGAAATGC 4400

4401 TTGCACATGCAGAAGAAACA CGCAAATTAATGCCTGTCTG TGTGGAAACTAAAGCCATAG TTTCAACTATACAGCGTAAA TATAAGGGTATTAAAATACA 4500

4501 AGAGGGTGTGGTTGATTATG GTGCTAGATTTTACTTTTAC ACCAGTAAAACAACTGTAGC GTCACTTATCAACACACTTA ACGATCTAAATGAAACTCTT 4600

4601 GTTACAATGCCACTTGGCTA TGTAACACATGGCTTAAATT TGGAAGAAGCTGCTCGGTAT ATGAGATCTCTCAAAGTGCC AGCTACAGTTTCTGTTTCTT 4700

4701 CACCTGATGCTGTTACAGCG TATAATGGTTATCTTACTTC TTCTTCTAAAACACCTGAAG AACATTTTATTGAAACCATC TCACTTGCTGGTTCCTATAA 4800

4801 AGATTGGTCCTATTCTGGAC AATCTACACAACTAGGTATA GAATTTCTTAAGAGAGGTGA TAAAAGTGTATATTACACTA GTAATCCTACCACATTCCAC 4900

4901 CTAGATGGTGAAGTTATCAC CTTTGACAATCTTAAGACAC TTCTTTCTTTGAGAGAAGTG AGGACTATTAAGGTGTTTAC AACAGTAGACAACATTAACC 5000

5001 TCCACACGCAAGTTGTGGAC ATGTCAATGACATATGGACA ACAGTTTGGTCCAACTTATT TGGATGGAGCTGATGTTACT AAAATAAAACCTCATAATTC 5100

5101 ACATGAAGGTAAAACATTTT ATGTTTTACCTAATGATGAC ACTCTACGTGTTGAGGCTTT TGAGTACTACCACACAACTG ATCCTAGTTTTCTGGGTAGG 5200

5201 TACATGTCAGCATTAAATCA CACTAAAAAGTGGAAATACC CACAAGTTAATGGTTTAACT TCTATTAAATGGGCAGATAA CAACTGTTATCTTGCCACTG 5300

5301 CATTGTTAACACTCCAACAA ATAGAGTTGAAGTTTAATCC ACCTGCTCTACAAGATGCTT ATTACAGAGCAAGGGCTGGT GAAGCTGCTAACTTTTGTGC 5400

5401 ACTTATCTTAGCCTACTGTA ATAAGACAGTAGGTGAGTTA GGTGATGTTAGAGAAACAAT GAGTTACTTGTTTCAACATG CCAATTTAGATTCTTGCAAA 5500

5501 AGAGTCTTGAACGTGGTGTG TAAAACTTGTGGACAACAGC AGACAACCCTTAAGGGTGTA GAAGCTGTTATGTACATGGG CACACTTTCTTATGAACAAT 5600

5601 TTAAGAAAGGTGTTCAGATA CCTTGTACGTGTGGTAAACA AGCTACAAAATATCTAGTAC AACAGGAGTCACCTTTTGTT ATGATGTCAGCACCACCTGC 5700

5701 TCAGTATGAACTTAAGCATG GTACATTTACTTGTGCTAGT GAGTACACTGGTAATTACCA GTGTGGTCACTATAAACATA TAACTTCTAAAGAAACTTTG 5800

5801 TATTGCATAGACGGTGCTTT ACTTACAAAGTCCTCAGAAT ACAAAGGTCCTATTACGGAT GTTTTCTACAAAGAAAACAG TTACACAACAACCATAAAAC 5900

5901 CAGTTACTTATAAATTGGAT GGTGTTGTTTGTACAGAAAT TGACCCTAAGTTGGACAATT ATTATAAGAAAGACAATTCT TATTTCACAGAGCAACCAAT 6000

6001 TGATCTTGTACCAAACCAAC CATATCCAAACGCAAGCTTC GATAATTTTAAGTTTGTATG TGATAATATCAAATTTGCTG ATGATTTAAACCAGTTAACT 6100

6101 GGTTATAAGAAACCTGCTTC AAGAGAGCTTAAAGTTACAT TTTTCCCTGACTTAAATGGA AGACCGCTTGGCCTCCGACT TAAACACTACACACCCTCTT 6200

6201 TTAAGAAAGGAGCTAAATTG TTACATAAACCTATTGTTTG GCATGTTAACAATGCAACTA ATAAAGCCACGTATAAACCA AATACCTGGTGTATACGTTG 6300

6301 TCTTTGGAGCACAAAACCAG TTGAAACATCAAATTCGTTT GATGTACTGAAGTCAGAGGA CGCGCAGGGAATGGATAATC TTGCCTGCGAAGATCTAAAA 6400

6401 CCAGTCTCTGAAGAAGTAGT GGAAAATCCTACCATACAGA AAGACGTTCTTGAGTGTAAT GTGAAAACTACCGAAGTTGT AGGAGACATTATACTTAAAC 6500

6501 CAGCAAATAATAGTTTAAAA ATTACAGAAGAGGTTGGCCA CACAGATCTAATGGCTGCTT ATGTAGACAATTCTAGTCTT ACTATTAAGAAACCTAATGA 6600

6601 ATTATCTAGAGTATTAGGTT TGAAAACCCTTGCTACTCAT GGTTTAGCTGCTGTTAATAG TGTCCCTTGGGATACTATAG CTAATTATGCTAAGCCTTTT 6700

6701 CTTAACAAAGTTGTTAGTAC AACTACTAACATAGTTACAC GGTGTTTAAACCGTGTTTGT ACTAATTATATGCCTTATTT CTTTACTTTATTGCTACAAT 6800

6801 TGTGTACTTTTACTAGAAGT ACAAATTCTAGAATTAAAGC ATCTATGCCGACTACTATAG CAAAGAATACTGTTAAGAGT GTCGGTAAATTTTGTCTAGA 6900

6901 GGCTTCATTTAATTATTTGA AGTCACCTAATTTTTCTAAA CTGATAAATATTATAATTTG GTTTTTACTATTAAGTGTTT GCCTAGGTTCTTTAATCTAC 7000

7001 TCAACCGCTGCTTTAGGTGT TTTAATGTCTAATTTAGGCA TGCCTTCTTACTGTACTGGT TACAGAGAAGGCTATTTGAA CTCTACTAATGTCACTATTG 7100

7101 CAACCTACTGTACTGGTTCT ATACCTTGTAGTGTTTGTCT TAGTGGTTTAGATTCTTTAG ACACCTATCCTTCTTTAGAA ACTATACAAATTACCATTTC 7200

7201 ATCTTTTAAATGGGATTTAA CTGCTTTTGGCTTAGTTGCA GAGTGGTTTTTGGCATATAT TCTTTTCACTAGGTTTTTCT ATGTACTTGGATTGGCTGCA 7300

7301 ATCAAGCAATTGTTTTTCAG CTATTTTGCAGTACATTTTA TTAGTAATTCTTGGCTTATG TGGTTAATAATTAATCTTGT ACAAATGGCCCCGATTTCAG 7400

7401 CTATGGTTAGAATGTACATC TTCTTTGCATCATTTTATTA TGTATGGAAAAGTTATGTGC ATGTTGTAGACGGTTGTAAT TCATCAACTTGTATGATGTG 7500

7501 TTACAAACGTAATAGAGCAA CAAGAGTCGAATGTACAACT ATTGTTAATGGTGTTAGAAG GTCCTTTTATGTCTATGCTA ATGGAGGTAAAGGCTTTTGC 7600

7601 AAACTACACAATTGGAATTG TGTTAATTGTGATACATTCT GTGCTGGTAGTACATTTATT AGTGATGAAGTTGCGAGAGA CTTGTCACTACAGTTTAAAA 7700

7701 GACCAATAAATCCTACTGAC CAGTCTTCTTACATCGTTGA TAGTGTTACAGTGAAGAATG GTTCCATCCATCTTTACTTT GATAAAGCTGGTCAAAAGAC 7800

7801 TTATGAAAGACATTCTCTCT CTCATTTTGTTAACTTAGAC AACCTGAGAGCTAATAACAC TAAAGGTTCATTGCCTATTA ATGTTATAGTTTTTGATGGT 7900

7901 AAATCAAAATGTGAAGAATC ATCTGCAAAATCAGCGTCTG TTTACTACAGTCAGCTTATG TGTCAACCTATACTGTTACT AGATCAGGCATTAGTGTCTG 8000

8001 ATGTTGGTGATAGTGCGGAA GTTGCAGTTAAAATGTTTGA TGCTTACGTTAATACGTTTT CATCAACTTTTAACGTACCA ATGGAAAAACTCAAAACACT 8100

8101 AGTTGCAACTGCAGAAGCTG AACTTGCAAAGAATGTGTCC TTAGACAATGTCTTATCTAC TTTTATTTCAGCAGCTCGGC AAGGGTTTGTTGATTCAGAT 8200

8201 GTAGAAACTAAAGATGTTGT TGAATGTCTTAAATTGTCAC ATCAATCTGACATAGAAGTT ACTGGCGATAGTTGTAATAA CTATATGCTCACCTATAACA 8300

8301 AAGTTGAAAACATGACACCC CGTGACCTTGGTGCTTGTAT TGACTGTAGTGCGCGTCATA TTAATGCGCAGGTAGCAAAA AGTCACAACATTGCTTTGAT 8400

8401 ATGGAACGTTAAAGATTTCA TGTCATTGTCTGAACAACTA CGAAAACAAATACGTAGTGC TGCTAAAAAGAATAACTTAC CTTTTAAGTTGACATGTGCA 8500

8501 ACTACTAGACAAGTTGTTAA TGTTGTAACAACAAAGATAG CACTTAAGGGTGGTAAAATT GTTAATAATTGGTTGAAGCA GTTAATTAAAGTTACACTTG 8600

8601 TGTTCCTTTTTGTTGCTGCT ATTTTCTATTTAATAACACC TGTTCATGTCATGTCTAAAC ATACTGACTTTTCAAGTGAA ATCATAGGATACAAGGCTAT 8700

8701 TGATGGTGGTGTCACTCGTG ACATAGCATCTACAGATACT TGTTTTGCTAACAAACATGC TGATTTTGACACATGGTTTA GCCAGCGTGGTGGTAGTTAT 8800

8801 ACTAATGACAAAGCTTGCCC ATTGATTGCTGCAGTCATAA CAAGAGAAGTGGGTTTTGTC GTGCCTGGTTTGCCTGGCAC GATATTACGCACAACTAATG 8900

8901 GTGACTTTTTGCATTTCTTA CCTAGAGTTTTTAGTGCAGT TGGTAACATCTGTTACACAC CATCAAAACTTATAGAGTAC ACTGACTTTGCAACATCAGC 9000

9001 TTGTGTTTTGGCTGCTGAAT GTACAATTTTTAAAGATGCT TCTGGTAAGCCAGTACCATA TTGTTATGATACCAATGTAC TAGAAGGTTCTGTTGCTTAT 9100

9101 GAAAGTTTACGCCCTGACAC ACGTTATGTGCTCATGGATG GCTCTATTATTCAATTTCCT AACACCTACCTTGAAGGTTC TGTTAGAGTGGTAACAACTT 9200

9201 TTGATTCTGAGTACTGTAGG CACGGCACTTGTGAAAGATC AGAAGCTGGTGTTTGTGTAT CTACTAGTGGTAGATGGGTA CTTAACAATGATTATTACAG 9300

9301 ATCTTTACCAGGAGTTTTCT GTGGTGTAGATGCTGTAAAT TTACTTACTAATATGTTTAC ACCACTAATTCAACCTATTG GTGCTTTGGACATATCAGCA 9400

9401 TCTATAGTAGCTGGTGGTAT TGTAGCTATCGTAGTAACAT GCCTTGCCTACTATTTTATG AGGTTTAGAAGAGCTTTTGG TGAATACAGTCATGTAGTTG 9500

9501 CCTTTAATACTTTACTATTC CTTATGTCATTCACTGTACT CTGTTTAACACCAGTTTACT CATTCTTACCTGGTGTTTAT TCTGTTATTTACTTGTACTT 9600

9601 GACATTTTATCTTACTAATG ATGTTTCTTTTTTAGCACAT ATTCAGTGGATGGTTATGTT CACACCTTTAGTACCTTTCT GGATAACAATTGCTTATATC 9700

9701 ATTTGTATTTCCACAAAGCA TTTCTATTGGTTCTTTAGTA ATTACCTAAAGAGACGTGTA GTCTTTAATGGTGTTTCCTT TAGTACTTTTGAAGAAGCTG 9800

9801 CGCTGTGCACCTTTTTGTTA AATAAAGAAATGTATCTAAA GTTGCGTAGTGATGTGCTAT TACCTCTTACGCAATATAAT AGATACTTAGCTCTTTATAA 9900

9901 TAAGTACAAGTATTTTAGTG GAGCAATGGATACAACTAGC TACAGAGAAGCTGCTTGTTG TCATCTCGCAAAGGCTCTCA ATGACTTCAGTAACTCAGGT 10000

10001 TCTGATGTTCTTTACCAACC ACCACAAACCTCTATCACCT CAGCTGTTTTGCAGAGTGGT TTTAGAAAAATGGCATTCCC ATCTGGTAAAGTTGAGGGTT 10100

10101 GTATGGTACAAGTAACTTGT GGTACAACTACACTTAACGG TCTTTGGCTTGATGACGTAG TTTACTGTCCAAGACATGTG ATCTGCACCTCTGAAGACAT 10200

10201 GCTTAACCCTAATTATGAAG ATTTACTCATTCGTAAGTCT AATCATAATTTCTTGGTACA GGCTGGTAATGTTCAACTCA GGGTTATTGGACATTCTATG 10300

10301 CAAAATTGTGTACTTAAGCT TAAGGTTGATACAGCCAATC CTAAGACACCTAAGTATAAG TTTGTTCGCATTCAACCAGG ACAGACTTTTTCAGTGTTAG 10400

10401 CTTGTTACAATGGTTCACCA TCTGGTGTTTACCAATGTGC TATGAGGCCCCATTTCACTA TTAAGGGTTCATTCCTTAAT GGTTCATGTGGTAGTGTTGG 10500

10501 TTTTAACATAGATTATGACT GTGTCTCTTTTTGTTACATG CACCATATGGAATTACCAAC TGGAGTTCATGCTGGCACAG ACTTAGAAGGTAACTTTTAT 10600

10601 GGACCTTTTGTTGACAGGCA AACAGCACAAGCAGCTGGTA CGGACACAACTATTACAGTT AATGTTTTAGCTTGGTTGTA CGCTGCTGTTATAAATGGAG 10700

10701 ACAGGTGGTTTCTCAATCGA TTTACCACAACTCTTAATGA CTTTAACCTTGTGGCTATGA AGTACAATTATGAACCTCTA ACACAAGACCATGTTGACAT 10800

10801 ACTAGGACCTCTTTCTGCTC AAACTGGAATTGCCGTTTTA GATATGTGTGCTTCATTAAA AGAATTACTGCAAAATGGTA TGAATGGACGTACCATATTG 10900

10901 GGTAGTGCTTTATTAGAAGA TGAATTTACACCTTTTGATG TTGTTAGACAATGCTCAGGT GTTACTTTCCAAAGTGCAGT GAAAAGAACAATCAAGGGTA 11000

11001 CACACCACTGGTTGTTACTC ACAATTTTGACTTCACTTTT AGTTTTAGTCCAGAGTACTC AATGGTCTTTGTTCTTTTTT TTGTATGAAAATGCCTTTTT 11100

11101 ACCTTTTGCTATGGGTATTA TTGCTATGTCTGCTTTTGCA ATGATGTTTGTCAAACATAA GCATGCATTTCTCTGTTTGT TTTTGTTACCTTCTCTTGCC 11200

11201 ACTGTAGCTTATTTTAATAT GGTCTATATGCCTGCTAGTT GGGTGATGCGTATTATGACA TGGTTGGATATGGTTGATAC TAGTTTGTCTGGTTTTAAGC 11300

11301 TAAAAGACTGTGTTATGTAT GCATCAGCTGTAGTGTTACT AATCCTTATGACAGCAAGAA CTGTGTATGATGATGGTGCT AGGAGAGTGTGGACACTTAT 11400

11401 GAATGTCTTGACACTCGTTT ATAAAGTTTATTATGGTAAT GCTTTAGATCAAGCCATTTC CATGTGGGCTCTTATAATCT CTGTTACTTCTAACTACTCA 11500

11501 GGTGTAGTTACAACTGTCAT GTTTTTGGCCAGAGGTATTG TTTTTATGTGTGTTGAGTAT TGCCCTATTTTCTTCATAAC TGGTAATACACTTCAGTGTA 11600

11601 TAATGCTAGTTTATTGTTTC TTAGGCTATTTTTGTACTTG TTACTTTGGCCTCTTTTGTT TACTCAACCGCTACTTTAGA CTGACTCTTGGTGTTTATGA 11700

11701 TTACTTAGTTTCTACACAGG AGTTTAGATATATGAATTCA CAGGGACTACTCCCACCCAA GAATAGCATAGATGCCTTCA AACTCAACATTAAATTGTTG 11800

11801 GGTGTTGGTGGCAAACCTTG TATCAAAGTAGCCACTGTAC AGTCTAAAATGTCAGATGTA AAGTGCACATCAGTAGTCTT ACTCTCAGTTTTGCAACAAC 11900

11901 TCAGAGTAGAATCATCATCT AAATTGTGGGCTCAATGTGT CCAGTTACACAATGACATTC TCTTAGCTAAAGATACTACT GAAGCCTTTGAAAAAATGGT 12000

12001 TTCACTACTTTCTGTTTTGC TTTCCATGCAGGGTGCTGTA GACATAAACAAGCTTTGTGA AGAAATGCTGGACAACAGGG CAACCTTACAAGCTATAGCC 12100

12101 TCAGAGTTTAGTTCCCTTCC ATCATATGCAGCTTTTGCTA CTGCTCAAGAAGCTTATGAG CAGGCTGTTGCTAATGGTGA TTCTGAAGTTGTTCTTAAAA 12200

12201 AGTTGAAGAAGTCTTTGAAT GTGGCTAAATCTGAATTTGA CCGTGATGCAGCCATGCAAC GTAAGTTGGAAAAGATGGCT GATCAAGCTATGACCCAAAT 12300

12301 GTATAAACAGGCTAGATCTG AGGACAAGAGGGCAAAAGTT ACTAGTGCTATGCAGACAAT GCTTTTCACTATGCTTAGAA AGTTGGATAATGATGCACTC 12400

12401 AACAACATTATCAACAATGC AAGAGATGGTTGTGTTCCCT TGAACATAATACCTCTTACA ACAGCAGCCAAACTAATGGT TGTCATACCAGACTATAACA 12500

12501 CATATAAAAATACGTGTGAT GGTACAACATTTACTTATGC ATCAGCATTGTGGGAAATCC AACAGGTTGTAGATGCAGAT AGTAAAATTGTTCAACTTAG 12600

12601 TGAAATTAGTATGGACAATT CACCTAATTTAGCATGGCCT CTTATTGTAACAGCTTTAAG GGCCAATTCTGCTGTCAAAT TACAGAATAATGAGCTTAGT 12700

12701 CCTGTTGCACTACGACAGAT GTCTTGTGCTGCCGGTACTA CACAAACTGCTTGCACTGAT GACAATGCGTTAGCTTACTA CAACACAACAAAGGGAGGTA 12800

12801 GGTTTGTACTTGCACTGTTA TCCGATTTACAGGATTTGAA ATGGGCTAGATTCCCTAAGA GTGATGGAACTGGTACTATC TATACAGAACTGGAACCACC 12900

12901 TTGTAGGTTTGTTACAGACA CACCTAAAGGTCCTAAAGTG AAGTATTTATACTTTATTAA AGGATTAAACAACCTAAATA GAGGTATGGTACTTGGTAGT 13000

13001 TTAGCTGCCACAGTACGTCT ACAAGCTGGTAATGCAACAG AAGTGCCTGCCAATTCAACT GTATTATCTTTCTGTGCTTT TGCTGTAGATGCTGCTAAAG 13100

13101 CTTACAAAGATTATCTAGCT AGTGGGGGACAACCAATCAC TAATTGTGTTAAGATGTTGT GTACACACACTGGTACTGGT CAGGCAATAACAGTTACACC 13200

13201 GGAAGCCAATATGGATCAAG AATCCTTTGGTGGTGCATCG TGTTGTCTGTACTGCCGTTG CCACATAGATCATCCAAATC CTAAAGGATTTTGTGACTTA 13300

13301 AAAGGTAAGTATGTACAAAT ACCTACAACTTGTGCTAATG ACCCTGTGGGTTTTACACTT AAAAACACAGTCTGTACCGT CTGCGGTATGTTGAAAGGTT 13400

13401 ATGGCTGTAGTTGTGATCAA CTCCGCGAACCCATGCTTCA GTCAGCTGATGCACAATCGT TTTTAAACGGGTTTGCGGTG TAAGTGCAGCCCGTCTTACA 13500

13501 CCGTGCGGCACAGGCACTAG TACTGATGTCGTATACAGGG CTTTTGACATCTACAATGAT AAAGTAGCTGGTTTTGCTAA ATTCCTAAAAACTAATTGTT 13600

13601 GTCGCTTCCAAGAAAAGGAC GAAGATGACAATTTAATTGA TTCTTACTTTGTAGTTAAGA GACACACTTTCTCTAACTAC CAACATGAAGAAACAATTTA 13700

13701 TAATTTACTTAAGGATTGTC CAGCTGTTGCTAAACATGAC TTCTTTAAGTTTAGAATAGA CGGTGACATGGTACCACATA TATCACGTCAACGTCTTACT 13800

13801 AAATACACAATGGCAGACCT CGTCTATGCTTTAAGGCATT TTGATGAAGGTAATTGTGAC ACATTAAAAGAAATACTTGT CACATACAATTGTTGTGATG 13900

13901 ATGATTATTTCAATAAAAAG GACTGGTATGATTTTGTAGA AAACCCAGAAGTCGGATCGT AATACGCCAACTTAGGTGAA CGTGTACGCCAAGCTTTGTT 14000

14001 AAAAACAGTACAATTCTGTG ATGCCATGCGAAATGCTGGT ATTGTTGGTGTACTGACATT AGATAATCAAGATCTCAATG GTAACTGGTATGATTTCGGT 14100

14101 GATTTCATACAAACCACGCC AGGTAGTGGAGTTCCTGTTG TAGATTCTTATTATTCATTG TTAATGCCTATATTAACCTT GACCAGGGCTTTAACTGCAG 14200

14201 AGTCACATGTTGACACTGAC TTAACAAAGCCTTACATTAA GTGGGATTTGTTAAAATATG ACTTCACGGAAGAGAGGTTA AAACTCTTTGACCGTTATTT 14300

14301 TAAATATTGGGATCAGACAT ACCACCCAAATTGTGTTAAC TGTTTGGATGACAGATGCAT TCTGCATTGTGCAAACTTTA ATGTTTTATTCTCTACAGTG 14400

14401 TTCCCACCTACAAGTTTTGG ACCACTAGTGAGAAAAATAT TTGTTGATGGTGTTCCATTT GTAGTTTCAACTGGATACCA CTTCAGAGAGCTAGGTGTTG 14500

14501 TACATAATCAGGATGTAAAC TTACATAGCTCTAGACTTAG TTTTAAGGAATTACTTGTGT ATGCTGCTGACCCTGCTATG CACGCTGCTTCTGGTAATCT 14600

14601 ATTACTAGATAAACGCACTA CGTGCTTTTCAGTAGCTGCA CTTACTAACAATGTTGCTTT TCAAACTGTCAAACCCGGTA ATTTTAACAAAGACTTCTAT 14700

14701 GACTTTGCTGTGTCTAAGGG TTTCTTTAAGGAAGGAAGTT CTGTTGAATTAAAACACTTC TTCTTTGCTCAGGATGGTAA TGCTGCTATCAGCGATTATG 14800

14801 ACTACTATCGTTATAATCTA CCAACAATGTGTGATATCAG ACAACTACTATTTGTAGTTG AAGTTGTTGATAAGTACTTT GATTGTTACGATGGTGGCTG 14900

14901 TATTAATGCTAACCAAGTCA TCGTCAACAACCTAGACAAA TCAGCTGGTTTTCCATTTAA TAAATGGGGTAAGGCTAGAC TTTATTATGATTCAATGAGT 15000

15001 TATGAGGATCAAGATGCACT TTTCGCATATACAAAACGTA ATGTCATCCCTACTATAACT CAAATGAATCTTAAGTATGC CATTAGTGCAAAGAATAGAG 15100

15101 CTCGCACCGTAGCTGGTGTC TCTATCTGTAGTACTATGAC CAATAGACAGTTTCATCAAA AATTATTGAAATCAATAGCC GCCACTAGAGGAGCTACTGT 15200

15201 AGTAATTGGAACAAGCAAAT TCTATGGTGGTTGGCACAAC ATGTTAAAAACTGTTTATAG TGATGTAGAAAACCCTCACC TTATGGGTTGGGATTATCCT 15300

15301 AAATGTGATAGAGCCATGCC TAACATGCTTAGAATTATGG CCTCACTTGTTCTTGCTCGC AAACATACAACGTGTTGTAG CTTGTCACACCGTTTCTATA 15400

15401 GATTAGCTAATGAGTGTGCT CAAGTATTGAGTGAAATGGT CATGTGTGGCGGTTCACTAT ATGTTAAACCAGGTGGAACC TCATCAGGAGATGCCACAAC 15500

15501 TGCTTATGCTAATAGTGTTT TTAACATTTGTCAAGCTGTC ACGGCCAATGTTAATGCACT TTTATCTACTGATGGTAACA AAATTGCCGATAAGTATGTC 15600

15601 CGCAATTTACAACACAGACT TTATGAGTGTCTCTATAGAA ATAGAGATGTTGACACAGAC TTTGTGAATGAGTTTTACGC ATATTTGCGTAAACATTTCT 15700

15701 CAATGATGATACTCTCTGAC GATGCTGTTGTGTGTTTCAA TAGCACTTATGCATCTCAAG GTCTAGTGGCTAGCATAAAG AACTTTAAGTCAGTTCTTTA 15800

15801 TTATCAAAACAATGTTTTTA TGTCTGAAGCAAAATGTTGG ACTGAGACTGACCTTACTAA AGGACCTCATGAATTTTGCT CACAACATACAATGCTAGTT 15900

15901 AAACAGGGTGATGATTATGT GTACCTTCCTTACCCAGATC CATCAAGAATCCTAGGGGCC GGCTGTTTTGTAGATGATAT CGTAAAAACAGATGGTACAC 16000

16001 TTATGATTGAACGGTTCGTG TCTTTAGCTATAGATGCTTA CCCACTTACTAAACATCCTA ATCAGGAGTATGCTGATGTC TTTCATTTGTACTTACAATA 16100

16101 CATAAGAAAGCTACATGATG AGTTAACAGGACACATGTTA GACATGTATTCTGTTATGCT TACTAATGATAACACTTCAA GGTATTGGGAACCTGAGTTT 16200

16201 TATGAGGCTATGTACACACC GCATACAGTCTTACAGGCTG TTGGGGCTTGTGTTCTTTGC AATTCACAGACTTCATTAAG ATGTGGTGCTTGCATACGTA 16300

16301 GACCATTCTTATGTTGTAAA TGCTGTTACGACCATGTCAT ATCAACATCACATAAATTAG TCTTGTCTGTTAATCCGTAT GTTTGCAATGCTCCAGGTTG 16400

16401 TGATGTCACAGATGTGACTC AACTTTACTTAGGAGGTATG AGCTATTATTGTAAATCACA TAAACCACAAGTCGGTTTTC CATTGTGTGCTAATGGACAA 16500

16501 GTTTTTGGTTTATATAAAAA TACATGTGTTGGTAGCGATA ATGTTACTGACTTTAATGCA ATTGCAACATGTGACTGGAC AAATGCTGGTGATTACATTT 16600

16601 TAGCTAACACCTGTACTGAA AGACTCAAGCTTTTTGCAGC AGAAACGCTCAAAGCTACTG AGGAGACATTTAAACTGTCT TATGGTATTGCTACTGTACG 16700

16701 TGAAGTGCTGTCTGACAGAG AATTACATCTTTCATGGGAA GTTGGTAAACCTAGACCACC ACTTAACCGAAATTATGTCT TTACTGGTTATCGTGTAACT 16800

16801 AAAAACAGTAAAGTACAAAT AGGAGAGTACACCTTTGAAA AAGGTGACTATGGTGATGCT GTTGTTTACCGAGGTACAAC AACTTACAAATTAAATGTTG 16900

16901 GTGATTATTTTGTGCTGACA TCACATACAGTAATGCCATT AAGTGCACCTACACTAGTGC CACAAGAGCACTATGTTAGA ATTACTGGCTTATACCCAAC 17000

17001 ACTCAATATCTCAGATGAGT TTTCTAGCAATGTTGCAAAT TATCAAAAGGTTGGTATGCA AAAGTATTCTACACTCCAGG GACCACCTGGTACTGGTAAG 17100

17101 AGTCATTTTGCTATTGGCCT AGCTCTCTACTACCCTTCTG CTCGCATAGTGTATACAGCT TGCTCTCATGCCGCTGTTGA TGCACTATGTGAGAAGGCAT 17200

17201 TAAAATATTTGCCTATAGAT AAATGTAGTAGAATTATACC TGCACGTGCTCGTGTAGAGT GTTTTGATAAATTCAAAGTG AATTCAACATTAGAACAGTA 17300

17301 TGTCTTTTGTACTGTAAATG CATTGCCTGAGACGACAGCA GATATAGTTGTCTTTGATGA AATTTCAATGGCCGCAAATT ATGATTTGAGTGTTGTCAAT 17400

17401 GCCAGATTACGTGCTAAGCA CTATGTGTACATTGGCGACC CTGCTCAATTACCTGCACCA CGCACATTGCTAACTAAGGG CACACTAGAACCAGAATATT 17500

17501 TCAATTCAGTGTGTAGACTT ATGAAAACTATAGGTCCAGA CATGTTCCTCGGAACTTGTC GGCGTTGTCCTGCTGAAATT GTTGACACTGTGAGTGCTTT 17600

17601 GGTTTATGATAATAAGCTTA AAGCACATAAAGACAAATCA GCTCAATGCTTTAAAATGTT TTATAAGGGTGTTATCACGC ATGATGTTTCATCTGCAATT 17700

17701 AACAGGCCGCAAATAGGCGT GGTAAGAGAATTCCTTACAC GTAACCCTGCTTGGAGAAAA GCTGTCTTTATTTCACCTTA TAATTCACAGAATGCTGTAG 17800

17801 CCTCAAAGATTTTGGGACTA CCAACTCAAACTGTTGATTC ATCACAGGGCTCAGAATATG ACTATGTCATATTCACTCAA ACCACCGAAACAGCTCACTC 17900

17901 TTGTAATGTAAACAGATTTA ATGTTGCTATTACCAGAGCA AAAGTAGGCATACTTTGCAT AATGTCTGATAGAGACCTTT ATGACAAGTTGCAATTTACA 18000

18001 AGTCTTGAAATTCCACGTAG GAATGTGGCAACTTTACAAG CTGAAAATGTAACAGGACTC TTTAAAGATTGTAGTAAGGT AATCACTGGGTTACATCCTA 18100

18101 CACAGGCGCCTACACACCTC AGTGTTGACACTAAATTCAA AACTGAAGGTTTATGTGTTG ACATACCTGGCATACCTAAG GACATGACCTATAGAAGACT 18200

18201 CATCTCTATGATGGGTTTTA AAATGAATTATCAAGTTAAT GGTTACCCTAACATGTTTAT CACCCGCGAAGAAGCTATAA GACATGTACGTGCATGGATT 18300

18301 GGCTTCGATGTCGAGGGGTG TCATGCTACTAGAGAAGCTG TTGGTACCAATTTACCTTTA CAGCTAGGTTTTTCTACAGG TGTTAACCTAGTTGCTGTAC 18400

18401 CTACAGGTTATGTTGATACA CCTAATAATACAGATTTTTC CAGAGTTAGTGCTAAACCAC CGCCTGGAGATCAATTTAAA CACCTCATACCACTTATGTA 18500

18501 CAAAGGACTTCCTTGGAATG TAGTGCGTATAAAGATTGTA CAAATGTTAAGTGACACACT TAAAAATCTCTCTGACAGAG TCGTATTTGTCTTATGGGCA 18600

18601 CATGGCTTTGAGTTGACATC TATGAAGTATTTTGTGAAAA TAGGACCTGAGCGCACCTGT TGTCTATGTGATAGACGTGC CACATGCTTTTCCACTGCTT 18700

18701 CAGACACTTATGCCTGTTGG CATCATTCTATTGGATTTGA TTACGTCTATAATCCGTTTA TGATTGATGTTCAACAATGG GGTTTTACAGGTAACCTACA 18800

18801 AAGCAACCATGATCTGTATT GTCAAGTCCATGGTAATGCA CATGTAGCTAGTTGTGATGC AATCATGACTAGGTGTCTAG CTGTCCACGAGTGCTTTGTT 18900

18901 AAGCGTGTTGACTGGACTAT TGAATATCCTATAATTGGTG ATGAACTGAAGATTAATGCG GCTTGTAGAAAGGTTCAACA CATGGTTGTTAAAGCTGCAT 19000

19001 TATTAGCAGACAAATTCCCA GTTCTTCACGACATTGGTAA CCCTAAAGCTATTAAGTGTG TACCTCAAGCTGATGTAGAA TGGAAGTTCTATGATGCACA 19100

19101 GCCTTGTAGTGACAAAGCTT ATAAAATAGAAGAATTATTC TATTCTTATAAGTCGGATCG TAGCCATGTCGTTCTGTGTG TATGCCTATTTTGGAATTGC 19200

19201 AATGTCGATAGATATCCTGC TAATTCCATTGTTTGTAGAT TTGACACTAGAGTGCTATCT AACCTTAACTTGCCTGGTTG TGATGGTGGCAGTTTGTATG 19300

19301 TAAATAAACATGCATTCCAC ACACCAGCTTTTGATAAAAG TGCTTTTGTTAATTTAAAAC AATTACCATTTTTCTATTAC TCTGACAGTCCATGTGAGTC 19400

19401 TCATGGAAAACAAGTAGTGT CAGATATAGATTATGTACCA CTAAAGTCTGCTACGTGTAT AACACGTTGCAATTTAGGTG GTGCTGTCTGTAGACATCAT 19500

19501 GCTAATGAGTACAGATTGTA TCTCGATGCTTATAACATGA TGATCTCAGCTGGCTTTAGC TTGTGGGTTTACAAACAATT TGATACTTATAACCTCTGGA 19600

19601 ACACTTTTACAAGACTTCAG AGTTTAGAAAATGTGGCTTT TAATGTTGTAAATAAGGGAC ACTTTGATGGACAACAGGGT GAAGTACCAGTTTCTATCAT 19700

19701 TAATAACACTGTTTACACAA AAGTTGATGGTGTTGATGTA GAATTGTTTGAAAATAAAAC AACATTACCTGTTAATGTAG CATTTGAGCTTTGGGCTAAG 19800

19801 CGCAACATTAAACCAGTACC AGAGGTGAAAATACTCAATA ATTTGGGTGTGGACATTGCT GCTAATACTGTGATCTGGGA CTACAAAAGAGATGCCCCAG 19900

19901 CACATATATCTACTATTGGT GTTTGTTCTATGACTGACAT AGCCAAGAAACCAACTGAAA CGATTTGTGCACCACTCACT GTCTTTTTTGATGGTAGAGT 20000

20001 TGATGGTCAAGTAGACTTAT TTAGAAATGCCCGTAATGGT GTTCTTATTACAGAAGGTAG TGTTAAAGGTTTACAACCAT CTGTAGGTCCCAAACAAGCT 20100

20101 AGTCTTAATGGAGTCACATT AATTGGAGAAGCCGTAAAAA CACAGTTCAATTATTATAAG AAAGTTGATGGTGTTGTCCA ACAATTACCTGAAACTTACT 20200

20201 TTACTCAGAGTAGAAATTTA CAAGAATTTAAACCCAGGAG TCAAATGGAAATTGATTTCT TAGAATTAGCTATGGATGAA TTCATTGAACGGTATAAATT 20300

20301 AGAAGGCTATGCCTTCGAAC ATATCGTTTATGGAGATTTT AGTCATAGTCAGTTAGGTGG TTTACATCTACTGATTGGAC TAGCTAAACGTTTTAAGGAA 20400

20401 TCACCTTTTGAATTAGAAGA TTTTATTCCTATGGACAGTA CAGTTAAAAACTATTTCATA ACAGATGCGCAAACAGGTTC ATCTAAGTGTGTGTGTTCTG 20500

20501 TTATTGATTTATTACTTGAT GATTTTGTTGAAATAATAAA ATCCCAAGATTTATCTGTAG TTTCTAAGGTTGTCAAAGTG ACTATTGACTATACAGAAAT 20600

20601 TTCATTTATGCTTTGGTGTA AAGATGGCCATGTAGAAACA TTTTACCCAAAATTACAATC TAGTCAAGCGTGGCAACCGG GTGTTGCTATGCCTAATCTT 20700

20701 TACAAAATGCAAAGAATGCT ATTAGAAAAGTGTGACCTTC AAAATTATGGTGATAGTGCA ACATTACCTAAAGGCATAAT GATGAATGTCGCAAAATATA 20800

20801 CTCAACTGTGTCAATATTTA AACACATTAACATTAGCTGT ACCCTATAATATGAGAGTTA TACATTTTGGTGCTGGTTCT GATAAAGGAGTTGCACCAGG 20900

20901 TACAGCTGTTTTAAGACAGT GGTTGCCTACGGGTACGCTG CTTGTCGATTCAGATCTTAA TGACTTTGTCTCTGATGCAG ATTCAACTTTGATTGGTGAT 21000

21001 TGTGCAACTGTACATACAGC TAATAAATGGGATCTCATTA TTAGTGATATGTACGACCCT AAGACTAAAAATGTTACAAA AGAAAATGACTCTAAAGAGG 21100

21101 GTTTTTTCACTTACATTTGT GGGTTTATACAACAAAAGCT AGCTCTTGGAGGTTCCGTGG CTATAAAGATAACAGAACAT TCTTGGAATGCTGATCTTTA 21200

21201 TAAGCTCATGGGACACTTCG CATGGTGGACAGCCTTTGTT ACTAATGTGAATGCGTCATC ATCTGAAGCATTTTTAATTG GATGTAATTATCTTGGCAAA 21300

21301 CCACGCGAACAAATAGATGG TTATGTCATGCATGCAAATT ACATATTTTGGAGGAATACA AATCCAATTCAGTTGTCTTC CTATTCTTTATTTGACATGA 21400

21401 GTAAATTTCCCCTTAAATTA AGGGGTACTGCTGTTATGTC TTTAAAAGAAGGTCAAATCA ATGATATGATTTTATCTCTT CTTAGACTTGGTAGACTTAT 21500

21501 AATTAGAGAAAACAACAGAG TTGTTATTTCTAGTGATGTT CTTGTTAACAACTAAACGAA CAATGTTTGTTTTTCTTGTT TTATTGCCACTAGTCTCTAG 21600

21601 TCAGTGTGTTAATCTTACAA CCAGAACTCAATTACCCAAG TCGGACACTAATTCTTTCAC ACGTGGTGTTTATTACCCTG ACAAAGTTTTCAGATCCTCA 21700

21701 GTTTTACATTCAACTCAGGA CTTGTTCTTACCTTTCTTTT CCAATGTTACTTGGTTCCAT GCTATACATGTCTCTGGGAC CAATGGTACTAAGAGGTTTG 21800

21801 ATAACCCTGTCCTACCATTT AATGATGGTGTTTATTTTGC TTCCACTGAGAAGTCTAACA TAATAAGAGGCTGGATTTTT GGTACTACTTTAGATTCGAA 21900

21901 GACCCAGTCCCTACTTATTG TTAATAACGCTACTAATGTT GTTATTAAAGTCTGTGAATT TCAATTTTGTAATGATCCAT TTTTGGGTGTTTATTACCAC 22000

22001 AAAAACAACAAAAGTTGGAT GGAAAGTGAGTTCAGAGTTT ATTCTAGTGCGAATAATTGC ACTTTTGAATATGTCTCTCA GCCTTTTCTTATGGACCTTG 22100

22101 AAGGAAAACAGGGCAATTTC AAAAATCTTAGGGAATTTGT GTTTAAGAATATTGATGGTT ATTTTAAAATATATTCTAAG CACACGCCTATTAATTTAGT 22200

22201 GCGTGATCTCCCTCAGGGTT TTTCGGCTTTAGAACCATTG GTAGATTTGCCAATAGGTAT TAACATCACTAGGTTTCAAA CTTTACTTGCTTTACATAGA 22300

22301 AGTTATTTGACTCCTGGTGA TTCTTCTTCAGGTTGGACAG CTGGTGCTGCAGCTTATTAT GTGGGTTATCTTCAACCTAG GACTTTTCTATTAAAATATA 22400

22401 ATGAAAATGGAACCATTACA GATGCTGTAGACTGTGCACT TGACCCTCTCTCAGAAACAA AGTGTACGTTGAAATCCTTC ACTGTAGAAAAAGGAATCTA 22500

22501 TCAAACTTCTAACTTTAGAG TCCAACCAACAGAATCTATT GTTAGATTTCCTAATATTAC AAACTTGTGCCCTTTTGGTG AAGTTTTTAACGCCACCAGA 22600

22601 TTTGCATCTGTTTATGCTTG GAACAGGAAGAGAATCAGCA ACTGTGTTGCTGATTATTCT GTCCTATATAATTCCGCATC ATTTTCCACTTTTAAGTGTT 22700

22701 ATGGAGTGTCTCCTACTAAA TTAAATGATCTCTGCTTTAC TAATGTCTATGCAGATTCAT TTGTAATTAGAGGTGATGAA GTCAGACAAATCGCTCCAGG 22800

22801 GCAAACTGGAAAGATTGCTG ATTATAATTATAAATTACCA GATGATTTTACAGGCTGCGT TATAGCTTGGAATTCTAACA ATCTTGATTCTAAGGTTGGT 22900

22901 GGTAATTATAATTACCTGTA TAGATTGTTTAGGAAGTCTA ATCTCAAACCTTTTGAGAGA GATATTTCAACTGAAATCTA TCAGGCCGGTAGCACACCTT 23000

23001 GTAATGGTGTTGAAGGTTTT AATTGTTACTTTCCTTTACA ATCATATGGTTTCCAACCCA CTAATGGTGTTGGTTACCAA CCATACAGAGTAGTAGTACT 23100

23101 TTCTTTTGAACTTCTACATG CACCAGCAACTGTTTGTGGA CCTAAAAAGTCTACTAATTT GGTTAAAAACAAATGTGTCA ATTTCAACTTCAATGGTTTA 23200

23201 ACAGGCACAGGTGTTCTTAC TGAGTCTAACAAAAAGTTTC TGCCTTTCCAACAATTTGGC AGAGACATTGCTGACACTAT CGATGCTGTCCGTGATCCAC 23300

23301 AGACACTTGAGATTCTTGAC ATTACACCATGTTCTTTTGG TGGTGTCAGTGTTATAACAC CAGGAACAAATACTTCTAAC CAGGTTGCTGTTCTTTATCA 23400

23401 GGATGTTAACTGCACAGAAG TCCCTGTTGCTATTCATGCA GATCAACTTACTCCTACTTG GCGTGTTTATTCTACAGGTT CTAATGTTTTTCAAACACGT 23500

23501 GCAGGCTGTTTAATAGGGGC TGAACATGTCAACAACTCAT ATGAGTGTGACATACCCATT GGTGCAGGTATATGCGCTAG TTATCAGACCCAGACTAATT 23600

23601 CTCCTCGGCGGGCACGTAGT GTAGCTAGTCAATCCATCAT TGCCTACACTATGTCACTTG GTGCAGAAAATTCAGTTGCT TACTCTAATAACTCTATTGC 23700

23701 CATACCCACAAATTTTACTA TTAGTGTTACCACAGAAATT CTACCAGTGTCTATGACCAA GACATCAGTAGATTGTACAA TGTACATTTGTGGTGATTCA 23800

23801 ACTGAATGCAGCAATCTTTT GTTGCAATATGGCAGTTTTT GTACACAATTAAACCGTGCT TTAACTGGAATAGCTGTTGA ACAAGACAAAAACACCCAAG 23900

23901 AAGTTTTTGCACAAGTCAAA CAAATTTACAAAACACCACC AATTAAAGATTTTGGTGGTT TTAATTTTTCACAAATATTA CCAGATCCATCAAAACCAAG 24000

24001 CAAGAGGTCATTTATTGAAG ATCTACTTTTCAACAAAGTG ACACTTGCAGATGCTGGCTT CATCAAACAATATGGTGATT GCCTTGGTGATATTGCTGCT 24100

24101 AGAGACCTCATTTGTGCACA AAAGTTTAACGGCCTTACTG TTTTGCCACCTTTGCTCACA GATGAAATGATTGCTCAATA CACTTCTGCACTGTTAGCGG 24200

24201 GTACAATCACTTCTGGTTGG ACCTTTGGTGCAGGTGCTGC ATTACAAATACCATTTGCTA TGCAAATGGCTTATAGGTTT AATGGTATTGGAGTTACACA 24300

24301 GAATGTTCTCTATGAGAACC AAAAATTGATTGCCAACCAA TTTAATAGTGCTATTGGCAA AATTCAAGACTCACTTTCTT CCACAGCAAGTGCACTTGGA 24400

24401 AAACTTCAAGATGTGGTCAA CCAAAATGCACAAGCTTTAA ACACGCTTGTTAAACAACTT AGCTCCAATTTTGGTGCAAT TTCAAGTGTTTTAAATGATA 24500

24501 TCCTTTCACGTCTTGACAAA GTTGAGGCTGAAGTGCAAAT TGATAGGTTGATCACAGGCA GACTTCAAAGTTTGCAGACA TATGTGACTCAACAATTAAT 24600

24601 TAGAGCTGCAGAAATCAGAG CTTCTGCTAATCTTGCTGCT ACTAAAATGTCAGAGTGTGT ACTTGGACAATCAAAAAGAG TTGATTTTTGTGGAAAGGGC 24700

24701 TATCATCTTATGTCCTTCCC TCAGTCAGCACCTCATGGTG TAGTCTTCTTGCATGTGACT TATGTCCCTGCACAAGAAAA GAACTTCACAACTGCTCCTG 24800

24801 CCATTTGTCATGATGGAAAA GCACACTTTCCTCGTGAAGG TGTCTTTGTTTCAAATGGCA CACACTGGTTTGTAACACAA AGGAATTTTTATGAACCACA 24900

24901 AATCATTACTACAGACAACA CATTTGTGTCTGGTAACTGT GATGTTGTAATAGGAATTGT CAACAACACAGTTTATGATC CTTTGCAACCTGAATTAGAC 25000

25001 TCATTCAAGGAGGAGTTAGA TAAATATTTTAAGAATCATA CATCACCAGATGTTGATTTA GGTGACATCTCTGGCATTAA TGCTTCAGTTGTAAACATTC 25100

25101 AAAAAGAAATTGACCGCCTC AATGAGGTTGCCAAGAATTT AAATGAATCTCTCATCGATC TCCAAGAACTTGGAAAGTAT GAGCAGTATATAAAATGGCC 25200

25201 ATGGTACATTTGGCTAGGTT TTATAGCTGGCTTGATTGCC ATAGTAATGGTGACAATTAT GCTTTGCTGTATGACCAGTT GCTGTAGTTGTCTCAAGGGC 25300

25301 TGTTGTTCTTGTGGATCCTG CTGCAAATTTGATGAAGACG ACTCTGAGCCAGTGCTCAAA GGAGTCAAATTACATTACAC ATAAACGAACTTATGGATTT 25400

25401 GTTTATGAGAATCTTCACAA TTGGAACTGTAACTTTGAAG CAAGGTGAAATCAAGGATGC TACTCCTTCAGATTTTGTTC GCGCTACTGCAACGATACCG 25500

25501 ATACAAGCCTCACTCCCTTT CGGATGGCTTATTGTTGGCG TTGCACTTCTTGCTGTTTTT CAGAGCGCTTCCAAAATCAT AACCCTCAAAAAGAGATGGC 25600

25601 AACTAGCACTCTCCAAGGGT GTTCACTTTGTTTGCAACTT GCTGTTGTTGTTTGTAACAG TTTACTCACACCTTTTGCTC GTTGCTGCTGGCCTTGAAGC 25700

25701 CCCTTTTCTCTATCTTTATG CTTTAGTCTACTTCTTGCAG AGTATAAACTTTGTAAGAAT AATAATGAGGCTTTGGCTTT GCTGGAAATGCCGTTCCAAA 25800

25801 AACCCATTACTTTATGATGC CAACTATTTTCTTTGCTGGC ATACTAATTGTTACGACTAT TGTATACCTTACAATAGTGT AACTTCTTCAATTGTCATTA 25900

25901 CTTCAGGTGATGGCACAACA AGTCCTATTTCTGAACATGA CTACCAGATTGGTGGTTATA CTGAAAAATGGGAATCTGGA GTAAAAGACTGTGTTGTATT 26000

26001 ACACAGTTACTTCACTTCAG ACTATTACCAGCTGTACTCA ACTCAATTGAGTACAGACAC TGGTGTTGAACATGTTACCT TCTTCATCTACAATAAAATT 26100

26101 GTTGATGAGCCTGAAGAACA TGTCCAAATTCACACAATCG ACGGTTCATCCGGAGTTGTT AATCCAGTAATGGAACCAAT TTATGATGAACCGACGACGA 26200

26201 CTACTAGCGTGCCTTTGTAA GCACAAGCTGATGAGTACGA ACTTATGTACTCATTCGTTT CGGAAGAGACAGGTACGTTA ATAGTTAATAGCGTACTTCT 26300

26301 TTTTCTTGCTTTCGTGGTAT TCTTGCTAGTTACACTAGCC ATCCTTACTGCGCTTCGATT GTGTGCGTACTGCTGCAATA TTGTTAACGTGAGTCTTGTA 26400

26401 AAACCTTCTTTTTACGTTTA CTCTCGTGTTAAAAATCTGA ATTCTTCTAGAGTTCCTGAT CTTCTGGTCTAAACGAACTA AATATTATATTAGTTTTTCT 26500

26501 GTTTGGAACTTTAATTTTAG CCATGGCAGATTCCAACGGT ACTATTACCGTTGAAGAGCT TAAAAAGCTCCTTGAACAAT GGAACCTAGTAATAGGTTTC 26600

26601 CTATTCCTTACATGGATTTG TCTTCTACAATTTGCCTATG CCAACAGGAATAGGTTTTTG TATATAATTAAGTTAATTTT CCTCTGGCTGTTATGGCCAG 26700

26701 TAACTTTAGCTTGTTTTGTG CTTGCTGCTGTTTACAGAAT AAATTGGATCACCGGTGGAA TTGCTATCGCAATGGCTTGT CTTGTAGGCTTGATGTGGCT 26800

26801 CAGCTACTTCATTGCTTCTT TCAGACTGTTTGCGCGTACG CGTTCCATGTGGTCATTCAA TCCAGAAACTAACATTCTTC TCAACGTGCCACTCCATGGC 26900

26901 ACTATTCTGACCAGACCGCT TCTAGAAAGTGAACTCGTAA TCGGAGCTGTGATCCTTCGT GGACATCTTCGTATTGCTGG ACACCATCTAGGACGCTGTG 27000

27001 ACATCAAGGACCTGCCTAAA GAAATCACTGTTGCTACATC ACGAACGCTTTCTTATTACA AATTGGGAGCTTCGCAGCGT GTAGCAGGTGACTCAGGTTT 27100

27101 TGCTGCATACAGTCGCTACA GGATTGGCAACTATAAATTA AACACAGACCATTCCAGTAG CAGTGACAATATTGCTTTGC TTGTACAGTAAGTGACAACA 27200

27201 GATGTTTCATCTCGTTGACT TTCAGGTTACTATAGCAGAG ATATTACTAATTATTATGAG GACTTTTAAAGTTTCCATTT GGAATCTTGATTACATCATA 27300

27301 AACCTCATAATTAAAAATTT ATCTAAGTCACTAACTGAGA ATAAATATTCTCAATTAGAT GAAGAGCAACCAATGGAGAT TGATTAAACGAACATGAAAA 27400

27401 TTATTCTTTTCTTGGCACTG ATAACACTCGCTACTTGTGA GCTTTATCACTACCAAGAGT GTGTTAGAGGTACAACAGTA CTTTTAAAAGAACCTTGCTC 27500

27501 TTCTGGAACATACGAGGGCA ATTCACCATTTCATCCTCTA GCTGATAACAAATTTGCACT GACTTGCTTTAGCACTCAAT TTGCTTTTGCTTGTCCTGAC 27600

27601 GGCGTAAAACACGTCTATCA GTTACGTGCCAGATCAGTTT CACCTAAACTGTTCATCAGA CAAGAGGAAGTTCAAGAACT TTACTCTCCAATTTTTCTTA 27700

27701 TTGTTGCGGCAATAGTGTTT ATAACACTTTGCTTCACACT CAAAAGAAAGACAGAATGAT TGAACTTTCATTAATTGACT TCTATTTGTGCTTTTTAGCC 27800

27801 TTTCTGCTATTCCTTGTTTT AATTATGCTTATTATCTTTT GGTTCTCACTTGAACTGCAA GATCATAATGAAACTTGTCA CGCCTAAACGAACATGAAAT 27900

27901 TTCTTGTTTTCTTAGGAATC ATCACAACTGTAGCTGCATT TCACCAAGAATGTAGTTTAC AGTCATGTACTCAACATCAA CCATATGTAGTTGATGACCC 28000

28001 GTGTCCTATTCACTTCTATT CTAAATGGTATATTAGAGTA GGAGCTAGAAAATCAGCACC TTTAATTGAATTGTGCGTGG ATGAGGCTGGTTCTAAATCA 28100

28101 CCCATTCAGTACATCGATAT CGGTAATTATACAGTTTCCT GTTTACCTTTTACAATTAAT TGCCAGGAACCTAAATTGGG TAGTCTTGTAGTGCGTTGTT 28200

28201 CGTTCTATGAAGACTTTTTA GAGTATCATGACGTTCGTGC TTGTAGATCTGTTCTCTAAA CGAACAAACTAAAATGTCTG ATAATGGACCCCAAAATCAG 28300

28301 CGAAATGCACCCCGCATTAC GTTTGGTGGACCCTCAGATT CAACTGGCAGTAACCAGAAT GGAGAACGCAGTGGGGCGCG ATCAAAACAACGTCGGCCCC

28400

28401 AAGGTTTACCCAATAATACT GCGTCTTGGTTCACCGCTCT CACTCAACATGGCAAGGAAG ACCTTAAATTCCCTCGAGGA CAAGGCGTTCCAATTAACAC 28500

28501 CAATAGCAGTCCAGATGACC AAATTGGCTACTACCGAAGA GCTACCAGACGAATTCGTGG TGGTGACGGTAAAATGAAAG ATCTCAGTCCAAGATGGTAT 28600

28601 TTCTACTACCTAGGAACTGG GCCAGAAGCTGGACTTCCCT ATGGTGCTAACAAAGACGGC ATCATATGGGTTGCAACTGA GGGAGCCTTGAATACACCAA 28700

28701 AAGATCACATTGGCACCCGC AATCCTGCTAACAATGCTGC AATCGTGCTACAACTTCCTC AAGGAACAACATTGCCAAAA GGCTTCTACGCAGAAGGGAG 28800

28801 CAGAGGCGGCAGTCAAGCCT CTTCTCGTTCCTCATCACGT AGTCGCAACAGTTCAAGAAA TTCAACTCCAGGCAGCAGTA GGGGAACTTCTCCTGCTAGA

28900

28901 ATGGCTGGCAATGGCGGTGA TGCTGCTCTTGCTTTGCTGC TGCTTGACAGATTGAACCAG CTTGAGAGCAAAATGTCTGG TAAAGGCCAACAACAACAAG

29000

29001 GCCAAACTGTCACTAAGAAA TCTGCTGCTGAGGCTTCTAA GAAGCCTCGGCAAAAACGTA CTGCCACTAAAGCATACAAT GTAACACAAGCTTTCGGCAG 29100

29101 ACGTGGTCCAGAACAAACCC AAGGAAATTTTGGGGACCAG GAACTAATCAGACAAGGAAC TGATTACAAACATTGGTCGC AAATTGCACAATTTGCCCCC 29200

29201 AGCGCTTCAGCGTTCTTCGG AATGTCGCGCATTGGCATGG AAGTCACACCTTCGGGAACG TGGTTGACCTACACAGGTGC CATCAAATTGGATGACAAAG

29300

29301 ATCCAAATTTCAAAGATCAA GTCATTTTGCTGAATAAGCA TATTGACGCATACAAAACAT TCCCACCAACAGAGCCAAAA AAGGACAAAAAGAAGAAGGC 29400

29401 TGATGAAACTCAAGCCTTAC CGCAGAGACAGAAGAAACAG CAAACTGTGACTCTTCTTCC TGCTGCAGATTTGGATGATT TCTCCAAACAATTGCAACAA 29500

29501 TCCATGAGCAGTGCTGACTC AACTCAGGCCTAAACTCATG CGACCGCTTGGCCTCCGACT TGCTATATAAACGTTTTCGC TTTTCCGTTTACGATATATA 29600

29601 GTCTACTCTTGTGCAGAATG AATTCTCGTAACTACATAGC ACAAGTAGATGTAGTTAACT TTAATCTCACATAGCAATCT TTAATCAGTGTGTAACATTA 29700

29701 GGGAGGACTTGAAAGAGCCA CCAAGTCGGATCGTAGCCAT GTCGTAGTACGATCGAGTGT ACAGTGAACAATGCTAGGGA GAGCTGCCTATATGGAAGAG

29800

29801 CCCTAATGTGTAAAATTAAT TTTAGTAGTGCTATCCCCAT GTGATTTTAATAGCTTCTTA GGAGAATGACAAAAAAAAAA AAAAAAAAAAA 29891

7.1.7 SRR12596175 Assembled Contig Spike Protein Sequence

1 MFLLTTKRTMFVFLVLLPLV SSQCVNLTTRTQLPPAYTNS FTRGVYYPDKVFRSSVLHST QDLFLPFFSNVTWFHAIHVS GTNGTKRFDNPVLPFNDGVY 100

101 FASTEKSNIIRGWIFGTTLD SKTQSLLIVNNATNVVIKVC EFQFCNDPFLGVYYHKNNKS WMESEFRVYSSANNCTFEYV SQPFLMDLEGKQGNFKNLRE 200

201 FVFKNIDGYFKIYSKHTPIN LVRDLPQGFSALEPLVDLPI GINITRFQTLLALHRSYLTP GDSSSGWTAGAAAYYVGYLQ PRTFLLKYNENGTITDAVDC 300

301 ALDPLSETKCTLKSFTVEKG IYQTSNFRVQPTESIVRFPN ITNLCPFGEVFNATRFASVY AWNRKRISNCVADYSVLYNS ASFSTFKCYGVSPTKLNDLC 400

401 FTNVYADSFVIRGDEVRQIA PGQTGKIADYNYKLPDDFTG CVIAWNSNNLDSKVGGNYNY LYRLFRKSNLKPFERDISTE IYQAGSTPCNGVEGFNCYFP 500

501 LQSYGFQPTNGVGYQPYRVV VLSFELLHAPATVCGPKKST NLVKNKCVNFNFNGLTGTGV LTESNKKFLPFQQFGRDIAD TTDAVRDPQTLEILDITPCS 600

601 FGGVSVITPGTNTSNQVAVL YQGVNCTEVPVAIHADQLTP TWRVYSTGSNVFQTRAGCLI GAEHVNNSYECDIPIGAGIC ASYQTQTNSPRRARSVASQS 700

701 IIAYTMSLGAENSVAYSNNS IAIPTNFTISVTTEILPVSM TKTSVDCTMYICGDSTECSN LLLQYGSFCTQLNRALTGIA VEQDKNTQEVFAQVKQIYKT 800

801 PPIKDFGGFNFSQILPDPSK PSKRSFIEDLLFNKVTLADA GFIKQYGDCLGDIAARDLIC AQKFNGLTVLPPLLTDEMIA QYTSALLAGTITSGWTFGAG 900

901 AALQIPFAMQMAYRFNGIGV TQNVLYENQKLIANQFNSAI GKIQDSLSSTASALGKLQDV VNQNAQALNTLVKQLSSNFG AISSVLNDILSRLDKVEAEV 1000

1001 QIDRLITGRLQSLQTYVTQQ LIRAAEIRASANLAATKMSE CVLGQSKRVDFCGKGYHLMS FPQSAPHGVVFLHVTYVPAQ EKNFTTAPAICHDGKAHFPR 1100

1101 EGVFVSNGTHWFVTQRNFYE PQIITTDNTFVSGNCDVVIG IVNNTVYDPLQPELDSFKEE LDKYFKNHTSPDVDLGDISG INASVVNIQKEIDRLNEVAK 1200

1201 NLNESLIDLQELGKYEQYIK WPWYIWLGFIAGLIAIVMVT IMLCCMTSCCSCLKGCCSCG SCCKFDEDDSEPVLKGVKLH YT 1282

7.1.8 SRR12596175 Discovered Strain 1 Possible Spike Protein Sequence

1 MFLLTTIRTMFVFLVLLPLV SSQCVNLKSRTQLPPAYTNS FTRGVYYPGKVFRSSVLHST QDLFLPFFSNVTWFHAIHVS GTIGTKRFDNPVLPFNDGVY 100

101 FASTEKSIIIRGWIFGTTLD SKTQSLLIVNNSTNVVIKVC EFQFCNDPFLGVYYHKNNKS WMESEFRVYSSANNCTFE-V SQPFLMDLEGKQGNFKNLRE 200

201 FVFKNIDGYFKIYSKHTPIN LVRDLPQGFSALEPLVDLPI GINITRFQIILSLHRSYLTP GDSSSGWTAGAAAYYVGYLQ PRSFLLKYNENGTITDAVDC 300

301 ALDPLSGTKCTLKSFTVEKG IYQTSNFRVQPTESIVRFPN IKYLCPFGEVFNATRFASVY AWNRKRISNCVADYSVLYNS VPFSTFKCYGVSPTKLNDLC 400

401 FTNVYADSFVIRGDEVRQIA PGQTGKIADYNYKLPDDFTG CVIAWNSNNLDSKVGGNYNY LYRLFRKSNLKPFERDISTE IYQAGSTPCNGVEGFNCYFL 500

501 LQLYGFQPTNGVGYQPYRVV VLSFELLHAPATVCASKKST NLVKNKCVNFNFNGLTGTGV LTESNKSFCLFNNLAETLLT LLMLSEVHRHLRFLTLHHVL 600

601 LVVSVL-RQEQILLTRLLFF IRVSTAQKSLLLFMQINLLL LGVFILQVLKFLKHVQAV– GLNMSTTHMSVTYPLVQVYA LVIRLSLILLGGHVV-LVNP 700

701 SLHSLCHLVQKIQLLTLITL LPYPQILLLVLPQKFYQCL- PRHQ-IVQCTFGMIQLNAAI YCCNMAVFVHN-TVL-LE-L LNKTKTPKKFVHKSNKFTKH 800

801 HQLKILVVLIFHK-FQIHQN QARGHLLKIYFSTK-HLQLL ASSNNMVIALVILLLEPSFV HKSLTALLFCHLCSQIR-LL NTLLHC-RVLSLLVGPLVQV 900

901 LHYKYHVLCKWLIGLMVLEL HRMFSMRTKN-LPTNLIVLL AKFKTHFLPQQANLENFKMW STKMHKL-TRLLNNLAPILV QFQVF-MISFHVLTKLRLKC 1000

1001 KLIG-SQADFNVCRHM-LNN -LELQKSELLLILLLLNCQS VYLDNQKELIFVERAIILCP SLSQHLMV-FSCM-LMSLHK KRTSQLLLPFVMME-HTFLV 1100

1101 KVSLFQMAHTGL-HKGIFMN HKSLLQTTHLCLVTVML–E VSTTQFMILCNLN-THSRRS -INILRIIHHQMLI-VTSLA LMLQL-TYKKKLTASMRLQ- 1200

1201 I-MNLSSISKNLESMSSI-M AMVHLARFYSWLDCHSNGDN YALLYDQLL-LSQGLLFLGI LLQI–RRL-ASAQRSQITL HI 1282

Difference of SRR11092062 Discovered Strain 1 Spike Protein Sequence and SARS-CoV-2 Spike Protein Sequence

The format of EMBOSS Needle is that each three lines followed an empty line, the first in the three lines is 50 amino acid bases of SARS-CoV-2 original strain spike protein, the second line is the matching status, and the third line is 50 amino acid bases of discovered strain 1. | means the two bases at the position in the first protein sequence and the

second protein sequence are the same, other signs mean the two bases are different. The figure is located at 7.1 and 7.1.

7.1.9 SRR12596175 Discovered Strain 2 Possible Spike Protein Sequence

1 MFLLTTKRTMFVFLVLLPLD SSQCVNLTTRTQLPPAYTNS FTRGVYYPDKVFRSSVLHST QDLFLPFFSNVTWFHAIHVS GTNGTKRFDNPVLPFNDGVY 100

101 FASTEKSNIIRGWIFGTTLD SKTQSLLIVNNATNVVIKVC EFQFCNDPFLGVYYHKNNKS WMESEFRVYSSANNCTFEYV SQPFLMDLEGKQGNFKNLRE 200

201 FVFKNIDGYFKIYSKHTPIN LVRDLPQGFSALEPLVDLPM GINITRFQTLLALHRSYLTP GDSSSGWTAGAAAYYVGYLQ PRTFLLKYNENGTITDAVDC 300

301 ALDPLSETKCTLKSFTVEKG IYQTSNFRVQPTESIVRFPN ITNLCPFGEVFNATRFASVY AWNRKRISNCVADYSVLYNS ASFSTFKCYGVSPTKLNDLC 400

401 FTNVYADSFVIRGDEVRQIA PGQTGKIADYNYKLPDDFTG CVIAWNSNNLDSKVGGNYNY LYRLFRKSNLKPFERDISTE IYQAGSTPCNGVEGFNCYFP 500

501 LQSYGFQPTNGVGYQPYRVV VLSFELLHAPATVCGPKKST NLVKNKCVNFNFNGLTGTGV LTESNKKFLPFQQFGRDIAD TTDAVRDPQTLEILDITPCS 600

601 FGGVSVITPGTNTSNQVAVL YQGVNCTEVPVAIHADQLTP TWRVYSTGSNVFQTRAGCLI GAEHVNNSYECDIPIGAGIC ASYQTQTNSPRRARSVASQS 700

701 IIAYTMSLGAENSVAYSNNS IAIPTNFTISVTTEILPVSM TKTSVDCTMYICGDSTECSN LLLQYGSFCTQLNRALTGIA VEQDKNTQEVFAQVKQIYKT 800

801 PPIKDFGGFNFSQILPDPSK PSKRSFIEDLLFNKVTLADA GFIKQYGDCLGDIAARDLIC AQKFNGLTVLPPLLTDEMIA QYTSALLAGTITSGWTFGAG 900

901 AALQIPFAMQMAYRFNGIGV TQNVLYENQKLIANQFNSAI GKIQDSLSSTASALGKLQDV VNQNAQALNTLVKQLSSNFG AISSVLNDILSRLDKVEAEV 1000

1001 QIDRLITGRLQSLQTYVTQQ LIRAAEIRASANLAATKMSE CVLGQSKRVDFCGKGYHLMS FPQSAPHGVVFLHVTYVPAQ EKNFTTAPAICHDGKAHFPR 1100

1101 EGVFVSNGTHWFVTQRNFYE PQIITTDNTFVSGNCDVVIG IVNNTVYDPLQPELDSFKEE LDKYFKNHTSPDVDLGDISG INASVVNIQKEIDRLNEVAK 1200

1201 NLNESLIDLQELGKYEQYIK WPWYIWLGFIAGLIAIVMVT IMLCCMTSCCSCLKGCCSCG SCCKFDEDDSEPVLKGVKLH YT 1282

Difference of Discovered Strain 2 Spike Protein Sequence and SARS-CoV-2 Spike Protein Sequence

The format of EMBOSS Needle is that each three lines followed an empty line, the first in the three lines is 50 amino acid bases of SARS-CoV-2 original strain spike protein, the second line is the matching status, and the third line is 50 amino acid bases of discovered strain 2. | means the two bases at the position in the first protein sequence and the second protein sequence are the same, other signs mean the two bases are different.The figure is located at 7.2 and 7.2.

7.1.10 FMDV sample sequence differences

Table 7.1: Substituted bases of strain 1 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.56. There are 178 changes in total.

Position	**bases in assembled contig**	**bases in FMDV discovered strain 1**	**Frequency**	**change type**
64	A	C	2	unavailable
158	G	A	2	unavailable
233	T	C	3	unavailable
243	A	G	21	unavailable
299	C	A	2	unavailable
413	A	G	81	unavailable
426	T	G	2	unavailable
455	A	C	11	unavailable
538	T	G	10	unavailable
562	A	G	2	unavailable
593	A	G	2	unavailable
645	T	C	4	unavailable
767	T	A	20	non-synonymous
799	A	C	5	non-synonymous
853	T	C	13	non-synonymous
863	A	C	63	non-synonymous

continues on next page

Table 7.1: Substituted bases of strain 1 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.56. There are 178 changes in total.

942	C	T	5	synonymous
973	A	G	2	non-synonymous
1009	C	T	2	non-synonymous
1063	A	C	115	non-synonymous
1123	T	G	3	non-synonymous
1175	T	G	5	non-synonymous
1189	A	C	23	non-synonymous
1262	T	G	2	non-synonymous
1342	T	A	8	non-synonymous
1394	T	A	31	non-synonymous
1398	C	A	2	non-synonymous
1419	T	C	4	synonymous
1480	T	C	8	non-synonymous
1483	A	C	56	non-synonymous
1563	T	G	7	synonymous
1622	G	A	3	non-synonymous
1639	A	C	27	non-synonymous
1640	A	C	49	non-synonymous
1690	T	A	7	non-synonymous
1710	T	C	2	synonymous
1765	T	G	2	non-synonymous

continues on next page

Table 7.1: Substituted bases of strain 1 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.56. There are 178 changes in total.

1820	A	C	49	non-synonymous
1901	A	G	53	non-synonymous
1912	A	C	33	non-synonymous
1984	A	G	27	non-synonymous
2034	A	G	2	synonymous
2081	A	C	44	non-synonymous
2135	T	G	130	non-synonymous
2142	T	C	26	synonymous
2157	T	C	4	synonymous
2220	T	G	3	synonymous
2282	T	G	116	non-synonymous
2312	T	C	2	non-synonymous
2330	A	C	15	non-synonymous
2405	G	A	453	non-synonymous
2417	A	G	528	non-synonymous
2468	T	C	2	non-synonymous
2486	C	T	2	non-synonymous
2539	T	A	4	non-synonymous
2605	A	G	81	non-synonymous
2689	T	G	9	non-synonymous
2714	T	C	49	non-synonymous

continues on next page

Table 7.1: Substituted bases of strain 1 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.56. There are 178 changes in total.

2720	A	C	96	non-synonymous
2788	A	G	808	non-synonymous
2856	T	G	102	synonymous
2872	A	G	109	non-synonymous
2932	A	C	303	non-synonymous
2993	T	C	3	non-synonymous
3046	A	G	87	non-synonymous
3103	T	C	4	non-synonymous
3120	T	C	2	synonymous
3199	A	C	19	non-synonymous
3266	A	C	128	non-synonymous
3320	T	G	2	non-synonymous
3338	A	G	5	non-synonymous
3400	A	G	163	non-synonymous
3475	G	A	2	non-synonymous
3524	T	C	119	non-synonymous
3564	A	G	92	synonymous
3644	A	C	32	non-synonymous
3645	G	C	18	non-synonymous
3671	A	C	3	non-synonymous
3725	T	G	72	non-synonymous

continues on next page

Table 7.1: Substituted bases of strain 1 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.56. There are 178 changes in total.

3759	A	C	1302	non-synonymous
3848	T	A	524	non-synonymous
3886	A	C	23	non-synonymous
3947	T	C	8	non-synonymous
3963	A	C	11	synonymous
4037	C	T	10	non-synonymous
4056	A	C	58	synonymous
4106	T	G	2390	non-synonymous
4169	T	G	66	non-synonymous
4170	C	G	39	non-synonymous
4171	A	C	8	non-synonymous
4223	A	C	12	non-synonymous
4276	T	G	421	non-synonymous
4292	A	C	384	non-synonymous
4381	T	A	237	non-synonymous
4419	A	G	3	synonymous
4441	A	C	123	non-synonymous
4450	A	C	706	non-synonymous
4510	A	C	230	non-synonymous
4565	A	T	6	non-synonymous
4629	T	A	49	synonymous

continues on next page

Table 7.1: Substituted bases of strain 1 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.56. There are 178 changes in total.

4663	A	C	33	non-synonymous
4735	G	T	3226	non-synonymous
4788	C	T	2709	synonymous
4868	T	G	670	non-synonymous
4896	G	T	282	non-synonymous
4946	T	G	124	non-synonymous
4983	C	T	5	synonymous
5033	A	T	3	non-synonymous
5085	C	T	28	synonymous
5118	A	C	6	non-synonymous
5177	T	C	5	non-synonymous
5228	T	C	3	non-synonymous
5278	T	G	5	non-synonymous
5299	A	C	4	non-synonymous
5394	T	C	28	synonymous
5419	T	C	22	non-synonymous
5469	C	T	2	synonymous
5478	A	C	64	synonymous
5544	T	C	4	synonymous
5552	A	G	3	non-synonymous
5621	T	G	12	non-synonymous

continues on next page

Table 7.1: Substituted bases of strain 1 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.56. There are 178 changes in total.

5624	A	G	2	non-synonymous
5625	G	A	6	synonymous
5679	T	G	8	synonymous
5720	C	A	664	non-synonymous
5780	T	G	8	non-synonymous
5832	C	A	3	non-synonymous
5846	A	C	6	non-synonymous
5851	A	G	44	non-synonymous
5860	A	C	308	non-synonymous
5922	A	T	4	synonymous
5923	G	T	4	non-synonymous
5945	A	C	36	non-synonymous
5996	G	A	8	non-synonymous
6010	A	C	24	non-synonymous
6066	T	G	16	synonymous
6112	A	C	169	non-synonymous
6165	A	G	553	synonymous
6236	T	G	8	non-synonymous
6270	T	C	2	synonymous
6338	T	G	9	non-synonymous
6396	T	G	43	synonymous

continues on next page

Table 7.1: Substituted bases of strain 1 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.56. There are 178 changes in total.

6399	G	T	51	synonymous
6400	T	A	78	non-synonymous
6459	T	G	130	synonymous
6479	T	A	24	non-synonymous
6543	T	G	22	non-synonymous
6561	A	C	70	synonymous
6624	T	G	14	synonymous
6629	A	G	10	non-synonymous
6680	T	C	50	non-synonymous
6683	A	G	4	non-synonymous
6740	T	G	226	non-synonymous
6760	A	C	2	non-synonymous
6793	A	C	71	non-synonymous
6852	T	C	3	synonymous
6857	A	G	21	non-synonymous
6930	C	A	4	non-synonymous
6953	C	T	4	non-synonymous
7004	T	C	5839	non-synonymous
7091	T	G	136	non-synonymous
7148	T	G	17	non-synonymous
7197	A	G	324	synonymous

continues on next page

Table 7.1: Substituted bases of strain 1 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.56. There are 178 changes in total.

7254	C	A	5	non-synonymous
7262	A	C	56	non-synonymous
7344	T	G	49	non-synonymous
7350	T	C	125	synonymous
7401	C	A	3	synonymous
7405	A	C	16	non-synonymous
7461	C	A	2	synonymous
7478	A	C	187	non-synonymous
7531	A	C	30	non-synonymous
7612	T	C	110	non-synonymous
7671	A	G	3	synonymous
7708	G	T	2	non-synonymous
7777	T	G	52	unavailable
7828	-	A	33	unavailable
7829	-	A	12	unavailable

Table 7.2: Substituted bases of strain 2 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.68. There are 166 changes in total.

Position	**bases in assembled contig**	**bases in FMDV discovered strain 2**	**Frequency**	**change type**
93	A	C	2	unavailable
165	A	G	2	unavailable
188	G	A	2	unavailable
225	A	G	2	unavailable
258	A	C	20	unavailable
323	A	G	2	unavailable
394	T	G	203	unavailable
413	A	G	25	unavailable
473	T	C	10	unavailable
495	G	A	3	unavailable
548	A	G	2	unavailable
607	A	G	2	unavailable
768	C	A	4	synonymous
804	A	G	37	synonymous
853	T	C	24	non-synonymous
910	T	C	2	non-synonymous

continues on next page

Table 7.2: Substituted bases of strain 2 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.68. There are 166 changes in total.

936	G	A	2	non-synonymous
974	T	C	2	non-synonymous
1026	T	C	2	synonymous
1063	A	C	4	non-synonymous
1075	A	C	34	non-synonymous
1131	T	C	2	synonymous
1201	T	G	13	non-synonymous
1229	A	T	2	non-synonymous
1347	T	C	5	synonymous
1402	T	A	18	non-synonymous
1440	G	A	6	synonymous
1495	T	C	9	non-synonymous
1498	A	C	17	non-synonymous
1575	T	C	2	synonymous
1603	A	C	3	non-synonymous
1654	T	C	10	non-synonymous
1693	G	A	2	non-synonymous
1795	A	C	9	non-synonymous
1865	T	C	10	non-synonymous
1867	A	C	8	non-synonymous
1919	T	G	130	non-synonymous

continues on next page

Table 7.2: Substituted bases of strain 2 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.68. There are 166 changes in total.

1942	T	G	2	non-synonymous
1985	A	C	2	non-synonymous
2067	C	T	55	synonymous
2142	T	C	35	synonymous
2163	T	G	2	synonymous
2214	C	T	2	synonymous
2274	T	G	49	synonymous
2338	C	G	10	non-synonymous
2339	G	C	4	non-synonymous
2404	C	T	298	non-synonymous
2417	A	G	237	non-synonymous
2456	T	G	5	non-synonymous
2492	A	G	2	non-synonymous
2522	C	T	3	non-synonymous
2558	T	C	2	non-synonymous
2577	A	G	3	synonymous
2662	T	C	13	non-synonymous
2706	T	C	2	synonymous
2746	A	C	73	non-synonymous
2799	T	A	2	non-synonymous
2841	A	T	15	synonymous

continues on next page

Table 7.2: Substituted bases of strain 2 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.68. There are 166 changes in total.

2892	T	G	17	synonymous
2930	T	C	54	non-synonymous
2932	A	C	54	non-synonymous
3023	T	G	3	non-synonymous
3063	G	A	3	synonymous
3107	C	T	2	non-synonymous
3157	A	C	8	non-synonymous
3235	A	C	26	non-synonymous
3294	T	G	5	synonymous
3296	A	G	2	non-synonymous
3297	G	A	2	non-synonymous
3362	T	C	3	non-synonymous
3409	A	G	6	non-synonymous
3482	A	C	4	non-synonymous
3557	T	C	125	non-synonymous
3566	A	G	5	non-synonymous
3617	T	C	18	non-synonymous
3668	A	G	3	non-synonymous
3680	A	G	3	non-synonymous
3744	T	G	92	synonymous
3768	C	A	384	non-synonymous

continues on next page

Table 7.2: Substituted bases of strain 2 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.68. There are 166 changes in total.

3867	T	G	357	synonymous
3895	G	T	4	non-synonymous
3968	T	C	9	non-synonymous
3972	G	T	6	synonymous
4057	C	A	7	non-synonymous
4106	T	G	853	non-synonymous
4111	C	A	853	non-synonymous
4171	A	C	8	non-synonymous
4172	C	T	2	non-synonymous
4177	A	G	5	non-synonymous
4234	T	A	16	non-synonymous
4250	A	G	101	non-synonymous
4310	T	C	21	non-synonymous
4364	T	G	138	non-synonymous
4381	T	A	12	non-synonymous
4419	A	C	53	synonymous
4478	A	C	185	non-synonymous
4549	G	A	18	non-synonymous
4594	T	C	10	non-synonymous
4649	T	A	7	non-synonymous
4684	A	C	38	non-synonymous

continues on next page

Table 7.2: Substituted bases of strain 2 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.68. There are 166 changes in total.

4735	G	T	140	non-synonymous
4762	T	G	140	non-synonymous
4807	A	C	7	non-synonymous
4868	T	G	250	non-synonymous
4871	A	G	250	non-synonymous
4899	C	A	545	non-synonymous
4958	T	G	22	non-synonymous
4984	C	T	3	non-synonymous
5066	C	T	3	non-synonymous
5115	C	T	7	synonymous
5172	A	G	7	synonymous
5258	T	G	4	non-synonymous
5325	G	A	4	synonymous
5335	A	G	2	non-synonymous
5394	T	G	12	synonymous
5423	A	G	10	non-synonymous
5475	A	C	42	synonymous
5528	A	G	7	non-synonymous
5591	T	C	15	non-synonymous
5604	A	C	7	synonymous
5676	T	G	3	non-synonymous

continues on next page

Table 7.2: Substituted bases of strain 2 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.68. There are 166 changes in total.

5685	A	C	82	synonymous
5744	A	G	24	non-synonymous
5801	T	C	4	non-synonymous
5843	T	G	7	non-synonymous
5851	A	G	90	non-synonymous
5907	A	G	5	synonymous
5947	C	T	8	non-synonymous
6026	T	G	12	non-synonymous
6099	T	C	326	synonymous
6173	T	G	117	non-synonymous
6185	A	C	16	non-synonymous
6243	C	A	8	non-synonymous
6252	C	T	2	synonymous
6298	G	A	3	non-synonymous
6328	A	T	9	non-synonymous
6398	T	G	22	non-synonymous
6408	C	G	30	synonymous
6409	G	A	72	non-synonymous
6459	T	C	53	synonymous
6500	A	G	24	non-synonymous
6559	T	C	16	non-synonymous

continues on next page

Table 7.2: Substituted bases of strain 2 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.68. There are 166 changes in total.

6569	A	C	54	non-synonymous
6668	T	C	60	non-synonymous
6752	T	C	229	non-synonymous
6793	A	G	35	non-synonymous
6844	A	G	26	non-synonymous
6896	T	C	10	non-synonymous
6960	T	G	88	synonymous
6994	A	G	10	non-synonymous
7031	A	G	10	non-synonymous
7129	T	G	91	non-synonymous
7197	A	G	6	synonymous
7248	C	A	4	non-synonymous
7253	A	C	29	non-synonymous
7350	T	C	1116	synonymous
7399	A	C	60	non-synonymous
7445	A	C	59	non-synonymous
7499	C	A	3	non-synonymous
7512	A	C	23	non-synonymous
7607	T	C	131	non-synonymous
7652	A	G	2	non-synonymous
7695	T	C	2	synonymous

continues on next page

Table 7.2: Substituted bases of strain 2 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 3.68. There are 166 changes in total.

7726	T	C	4	unavailable
7784	T	C	200	unavailable
7790	G	T	180	unavailable

Table 7.3: Substituted bases of strain 3 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 4.0. There are 170 changes in total.

Position	**bases in assembled contig**	**bases in FMDV discovered strain 3**	**Frequency**	**change type**
246	A	C	15	unavailable
376	A	G	2	unavailable
436	T	C	15	unavailable
451	A	C	19	unavailable
521	T	G	5	unavailable
588	C	T	2	unavailable
649	A	G	2	unavailable
727	A	G	2	non-synonymous

continues on next page

Table 7.3: Substituted bases of strain 3 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 4.0. There are 170 changes in total.

785	T	C	4	non-synonymous
792	A	C	8	synonymous
844	A	C	23	non-synonymous
918	A	G	2	synonymous
961	A	G	4	non-synonymous
1030	C	T	2	non-synonymous
1061	A	C	16	non-synonymous
1113	T	G	2	synonymous
1149	T	C	2	synonymous
1204	T	G	4	non-synonymous
1253	A	G	4	non-synonymous
1319	G	T	2	non-synonymous
1372	T	A	18	non-synonymous
1409	T	C	2	non-synonymous
1461	T	G	5	non-synonymous
1489	A	C	12	non-synonymous
1543	G	A	2	non-synonymous
1606	A	C	3	non-synonymous
1648	A	C	9	non-synonymous
1778	A	C	5	non-synonymous
1830	A	C	4	non-synonymous

continues on next page

Table 7.3: Substituted bases of strain 3 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 4.0. There are 170 changes in total.

1878	T	C	138	synonymous
1942	T	G	34	non-synonymous
2016	T	C	2	synonymous
2051	C	T	2	non-synonymous
2071	G	T	14	non-synonymous
2132	T	G	4	non-synonymous
2220	T	C	4	synonymous
2279	T	C	15	non-synonymous
2331	C	-	58	non-synonymous
2398	T	G	9	non-synonymous
2456	T	G	11	non-synonymous
2503	A	C	2	non-synonymous
2552	T	C	4	non-synonymous
2581	A	C	3	non-synonymous
2643	A	C	10	synonymous
2706	T	C	2	synonymous
2714	T	C	5	non-synonymous
2720	A	C	13	non-synonymous
2725	T	C	42	non-synonymous
2850	T	G	26	non-synonymous
2856	T	G	3	synonymous

continues on next page

Table 7.3: Substituted bases of strain 3 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 4.0. There are 170 changes in total.

2867	A	G	12	non-synonymous
2924	A	T	8	non-synonymous
2932	A	C	8	non-synonymous
2976	G	A	2	synonymous
3029	C	A	3	non-synonymous
3061	C	T	2	non-synonymous
3122	A	G	2	non-synonymous
3155	T	C	2	non-synonymous
3202	A	C	3	non-synonymous
3256	A	C	12	non-synonymous
3319	G	A	2	non-synonymous
3350	T	C	2	non-synonymous
3401	T	C	4	non-synonymous
3433	C	A	3	synonymous
3509	A	G	110	non-synonymous
3559	G	A	8	non-synonymous
3560	C	T	2	non-synonymous
3573	G	T	2	non-synonymous
3626	T	A	10	non-synonymous
3627	G	T	2	synonymous
3678	T	C	3	synonymous

continues on next page

Table 7.3: Substituted bases of strain 3 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 4.0. There are 170 changes in total.

3716	C	T	2	non-synonymous
3766	T	A	10	non-synonymous
3767	A	C	37	non-synonymous
3823	T	G	275	non-synonymous
3836	T	G	13	non-synonymous
3839	T	C	2	non-synonymous
3889	T	C	13	non-synonymous
3936	G	A	4	synonymous
3990	A	G	85	synonymous
4059	T	G	7	synonymous
4106	T	G	51	non-synonymous
4111	C	A	51	non-synonymous
4112	T	A	51	non-synonymous
4114	A	T	12	non-synonymous
4187	T	G	11	non-synonymous
4258	G	C	575	non-synonymous
4287	C	G	375	non-synonymous
4373	G	C	53	non-synonymous
4381	T	A	6	non-synonymous
4389	G	T	3	non-synonymous
4441	A	C	428	non-synonymous

continues on next page

Table 7.3: Substituted bases of strain 3 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 4.0. There are 170 changes in total.

4494	C	A	3	non-synonymous
4510	A	G	83	non-synonymous
4571	C	T	3	non-synonymous
4614	A	G	7	synonymous
4660	A	C	16	non-synonymous
4726	T	G	274	non-synonymous
4785	C	A	5	non-synonymous
4797	A	C	8	synonymous
4868	T	G	32	non-synonymous
4869	T	G	32	non-synonymous
4908	A	C	139	non-synonymous
4976	T	G	12	non-synonymous
5015	C	T	2	non-synonymous
5076	C	A	3	non-synonymous
5085	C	T	8	synonymous
5137	T	C	6	non-synonymous
5181	G	T	3	non-synonymous
5265	T	C	4	synonymous
5316	T	C	3	synonymous
5362	A	C	10	non-synonymous
5412	C	A	3	non-synonymous

continues on next page

Table 7.3: Substituted bases of strain 3 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 4.0. There are 170 changes in total.

5417	A	C	2	non-synonymous
5463	A	C	39	synonymous
5520	A	C	4	synonymous
5584	G	A	4	non-synonymous
5612	A	G	4	non-synonymous
5668	A	G	4	non-synonymous
5689	A	C	38	non-synonymous
5745	C	G	22	non-synonymous
5808	T	C	2	synonymous
5810	A	C	23	non-synonymous
5860	A	C	76	non-synonymous
5904	A	G	2	synonymous
5924	A	G	7	non-synonymous
6009	C	T	57	synonymous
6064	G	C	3	non-synonymous
6081	C	G	4	non-synonymous
6082	A	G	4	non-synonymous
6145	A	C	100	non-synonymous
6196	A	C	5	non-synonymous
6252	C	T	13	synonymous
6325	G	A	3	non-synonymous

continues on next page

Table 7.3: Substituted bases of strain 3 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 4.0. There are 170 changes in total.

6344	A	G	2	non-synonymous
6396	T	C	8	synonymous
6404	A	C	49	non-synonymous
6467	T	C	14	non-synonymous
6541	T	C	6	non-synonymous
6554	A	C	48	non-synonymous
6604	A	C	2	non-synonymous
6654	T	A	5	non-synonymous
6657	A	C	26	synonymous
6707	T	C	14	non-synonymous
6758	T	C	87	non-synonymous
6830	T	A	6	non-synonymous
6831	A	G	13	synonymous
6884	A	G	16	non-synonymous
6956	T	G	14	non-synonymous
6998	A	C	5	non-synonymous
7056	C	A	3	non-synonymous
7085	A	G	8	non-synonymous
7136	T	C	27	non-synonymous
7163	C	T	5	non-synonymous
7227	C	A	4	synonymous

continues on next page

Table 7.3: Substituted bases of strain 3 from FMDV sample. The consensus sequence doesn't contain the first 350 bases in the reference genome sequence EU448639.1, so the positions are adjusted when determining the synonymity. The non-synonymous/synonymous ratio is 4.0. There are 170 changes in total.

7262	A	C	31	non-synonymous
7306	A	C	9	non-synonymous
7370	T	C	15	non-synonymous
7379	A	C	16	non-synonymous
7432	T	C	3	non-synonymous
7462	A	C	27	non-synonymous
7512	A	G	6	synonymous
7563	T	G	91	synonymous
7627	T	A	54	non-synonymous
7648	T	C	2	non-synonymous
7685	T	C	2	non-synonymous
7733	G	C	2	unavailable
7734	A	G	2	unavailable
7783	C	T	267	unavailable
7784	T	C	267	unavailable

7.1.11 SRR12596175 Assembled Contig Nucleotide Sequence

1 GTGCTATCCCCATGTGATTT TTTTTTTGTCATTCTCCTAA GAAGCTATTAAAATCACATG GGGATAGCACTACTAAAATT AATTTTACACATTAGGGCTC 100

101 TTCCATATAGGCAGCTCTCC CTAGCATTGTTCACTGTACA CTCGATCGTACTCCGCGTGG CCTCGGTGAAAATGTGGTGG CTCTTTCAAGTCCTCCCTAA 200

201 TGTTACACACTGATTAAAGA TTGCTATGTGAGATTAAAGT TAACTACATCTACTTGTGCT ATGTAGTTACGAGAATTCAT TCTGCACAAGAGTAGACTAT 300

301 ATATCGTAAACGGAAAAGCG AAAACGTTTATATAGCCCAT CTGCCTTGTGTGGTCTGCAT GAGTTTAGGCCTGAGTTGAG TCAGCACTGCTCATGGATTG 400

401 TTGCAATTGTTTGGAGAAAT CATCCAAATCTGCAGCAGGA AGAAGAGTCACAGTTTGCTG TTTCTTCTGTCTCTGCGGTA AGGCTTGAGTTTCATCAGCC 500

501 TTCTTCTTTTTGTCCTTTTT AGGCTCTGTTGGTGGGAATG TTTTGTATGCGTCAATATGC TTATTCAGCAAAATGACTTG ATCTTTGAAATTTGGATCTT 600

601 TGTCATCCAATTTGATGGCA CCTGTGTAGGTCAACCACGT TCCCGAAGGTGTGACTTCCA TGCCAATGCGCGACATTCCG AAGAACGCTGAAGCGCTGGG 700

701 GGCAAATTGTGCAATTTGCG GCCAATGTTTGTAATCAGTT CCTTGTCTGATTAGTTCCTG GTCCCCAAAATTTCCTTGGG TTTGTTCTGGACCACGTCTG 800

801 CCGAAAGCTTGTGTTACATT GTATGCTTTAGTGGCAGTAC GTTTTTGCCGAGGCTTCTTA GAAGCCTCAGCAGCAGATTT CTTAGTGACAGTTTGGCCTT 900

901 GTTGTTGTTGGCCTTTACCA GACATTTTGCTCTCAAGCTG GTTCAATCTGTCAAGCAGCA GCAAAGCAAGAGCAGCATCA CCGCCATTGCCAGCCATTCT 1000

1001 AGCAGGAGAAGTTCCCCTAC TGCTGCCTGGAGTTGAATTT CTTGAACTGTTGCGACTACG TGATGAGGAACGAGAAGAGG CTTGACTGCCGCCTCTGCTC 1100

1101 CCTTCTGCGTAGAAGCCTTT TGGCAATGTTGTTCCTTGAG GAAGTTGTAGCACGATTGCA GCATTGTTAGCAGGATTGCG GGTGCCAATGTGATCTTTTG 1200

1201 GTGTATTCAAGGCTCCCTCA GTTGCAACCCATATGATGCC GTCTTTGTTAGCACCATAGG GAAGTCCAGCTTCTGGCCCA GTTCCTAGGTAGTAGAAATA 1300

1301 CCATCTTGGACTGAGATCTT TCATTTTACCGTCACCACCA CGAATTCGTCTGGTAGCTCT TCGGTAGTAGCCAATTTGGT CATCTGGACTGCTATTGGTG 1400

1401 TTAATTGGAACGCCTTGTCC TCGAGGGAATTTAAGGTCTT CCTTGCCATGTTGAGTGAGA GCGGTGAACCAAGACGCAGT ATTATTGGGTAAACCTTGGG 1500

1501 GCCGACGTTGTTTTGATCGC GCCCCACTGCGTTCTCCATT CTGGTTACTGCCAGTTGAAT CTGAGGGTCCACCAAACGTA ATGCGGGGTGCATTTCGCTG 1600

1601 ATTTTGGGGTCCATTATCAG ACATTTTAGTTTGTTCGTTT AGATGAAATCTAAAACAACA CGAACGTCATGATACTCTAA AAAGTCTTCATAGAACGAAC 1700

1701 AACGCACTACAAGACTACCC AATTTAGGTTCCTGGCAATT AATTGTAAAAGGTAAACAGG AAACTGTATAATTACCGATA TCGATGTACTGAATGGGTGA 1800

1801 TTTAGAACCAACCTCATCCA CGCACAATTCAATTAAAGGT GCTGATTTTCTAGCTCCTAC TCTAATATACCATTTAGAAT AGAAGTGAATAGGACACGGG 1900

1901 TCATCAACTACATATGGTTG ATGTTGAGTACATGACTGTA AACTACATTCTTGGTGAAAT GCAGCTACAGTTGTGATGAT TCCTAAGAAAACAAGAAATT 2000

2001 TCATGTTCGTTTAGGCGTGA CAAGTTTCATTATGATCTTG CAGTTCAAGTGAGAACCAAA AGATAATAAGCATAATTAAA ACAAGGAATAGCAGAAAGGC 2100

2101 TAAAAAGCACAAATAGAAGT CAATTAATGAAAGTTCAATC ATTCTGTCTTTCTTTTGAGT GTGAAGCAAAGTGTTATAAA CACTATTGCCGCAACAATAA 2200

2201 GAAAAATTGGAGAGTAAAGT TCTTGAACTTCCTCTTGTCT GATGAACAGTTTAGGTGAAA CTGATCTGGCACGTAACTGA TAGACGTGTTTTACGCCGTC 2300

2301 AGGACAAGCAAAAGCAAATT GAGTGCTAAAGCAAGTCAGT GCAAATTTGTTATCAGCTAG AGGATGAAATGGTGAATTGC CCTCGTATGTTCCAGAAGAG 2400

2401 CAAGGTTCTTTTAAAAGTAC TGTTGTACCTCTAACACACT CTTGGTAGTGATAAAGCTCA CAAGTAGCGAGTGTTATCAG TGCCAAGAAAAGAATAATTT 2500

2501 TCATGTTCGTTTAATCAATC TCCATTGGTTGCTCTTCATC TAATTGAGAATATTTATTCT CAGTTAGTGACTTAGATAAA TTTTTAATTATGAGGTTTAT 2600

2601 GATGTAATCAAGATTCCAAA TGGAAACTTTAAAAGTCCTC ATAATAATTAGTAATATCTC TGCTATAGTAACCTGAAAGT CAACGAGATGAAACATCTGT 2700

2701 TGTCACTTACTGTACAAGCA AAGCAATATTGTCACTGCTA CTGGAATGGTCTGTGTTTAA TTTATAGTTGCCAATCCTGT AGCGACTGTATGCAGCAAAA 2800

2801 CCTGAGTCACCTGCTACACG CTGCGAAGCTCCCAATTTGT AATAAGAAAGCGTTCGTGAT GTAGCAACAGTGATTTCTTT AGGCAGGTCCTTGATGTCAC 2900

2901 AGCGTCCTAGATGGTGTCCA GCAATACGAAGATGTCCACG AAGGATCACAGCTCCGATTA CGAGTTCACTTTCTAGAAGC GGTCTGGTCAGAATAGTGCC 3000

3001 ATGGAGTGGCACGTTGAGAA GAATGTTAGTTTCTGGATTG AATGACCACATGGAACGCGT ACGCGCAAACAGTCTGAAAG AAGCAATGAAGTAGCTGAGC 3100

3101 CACATCAAGCCTACAAGACA AGCCATTGCGATAGCAATTC CACCGGTGATCCAATTTATT CTGTAAACAGCAGCAAGCAC AAAACAAGCTAAAGTTACTG 3200

3201 GCCATAACAGCCAGAGGAAA ATTAACTTAATTATATACAA AAACCTATTCCTGTTGGCAT AGGCAAATTGTAGAAGACAA ATCCATGTAAGGAATAGGAA 3300

3301 ACCTATTACTAGGTTCCATT GTTCAAGGAGCTTTTTAAGC TCTTCAACGGTAATAGTACC GTTGGAATCTGCCATGGCTA AAATTAAAGTTCCAAACAGA 3400

3401 AAAACTAATATAATATTTAG TTCGTTTAGACCAGAAGATC AGGAACTCTAGAAGAATTCA GATTTTTAACACGAGAGTAA ACGTAAAAAGAAGGTTTTAC 3500

3501 AAGACTCACGTTAACAATAT TGCAGCAGTACGCACACAAT CGAAGCGCAGTAAGGATGGC TAGTGTAACTAGCAAGAATA CCACGAAAGCAAGAAAAAGA 3600

3601 AGTACGCTATTAACTATTAA CGTACCTGTCTCTTCCGAAA CGAATGAGTACATAAGTTCG TACTCATCAGCTTGTGCTTA CAAAGGCACGCTAGTAGTCG 3700

3701 TCGTCGGTTCATCATAAATT GGTTCCATTACTGGATTAAC AACTCCGGATGAACCGTCGA TTGTGTGAATTTGGACATGT TCTTCAGGCTCATCAACAAT 3800

3801 TTTATTGTAGATGAAGAAGG TAACATGTTCAACACCAGTG TCTGTACTCAATTGAGTTGA GTACAGCTGGTAATAGTCTG AAGTGAAGTAACTGTGTAAT 3900

3901 ACAACACAGTCTTTTACTCC AGATTCCCATTTTTCAGTAT AACCACCAATCTGGTAGTCA TGTTCAGAAATAGGACTTGT TGTGCCATCACCTGAAGTAA 4000

4001 TGACAATTGAAGAAGTTACA CTATTGTAAGGTATACAATA GTCGTAACAATTAGTATGCC AGCAAAGAAAATAGTTGGCA TCATAAAGTAATGGGTTTTT 4100

4101 GGAACGGCATTTCCAGCAAA GCCAAAGCCTCATTATTATT CTTACAAAGTTTATACTCTG CAAGAAGTAGACTAAAGCAT AAAGATAGAGAAAAGGGGCT 4200

4201 TCAAGGCCAGCAGCAACGAG CAAAAGGTGTGAGTAAACTG TTACAAACAACAACAGCAAG TTGCAAACAAAGTGAACACC CTTGGAGAGTGCTAGTTGCC 4300

4301 ATCTCTTTTTGAGGGTTATG ATTTTGGAAGCGCTATGAAA AACAGCAAGAAGTGCAACGC CAACAATAAGCCATCCGAAA GGGAGTGAGGCTTGTATCGG 4400

4401 TATCGTTGCAGTAGCGCGAA CAAAATCTGAAGGAGTAGCA TCCTTGATTTCACCTTGCTT CAAAGTTACAGTTCCAATTG TGAAGATTCTCATAAACAAA 4500

4501 TCCATAAGTTCGTTTATGTG TAATGTAATTTGACTCCTTT GAGCACTGGCTCAGAGTCGT CTTCATCAAATTTGCAGCAG GATCCACAAGAACAACAGCC 4600

4601 CTTGAGACAACTACAGCAAC TGGTCATACAGCAAAGCATA ATTGTCACCATTACTATGGC AATCAAGCCAGCTATAAAAC CTAGCCAAATGTACCATGGC 4700

4701 CATTTTATATACTGCTCATA CTTTCCAAGTTCTTGGAGAT CGATGAGAGATTCATTTAAA TTCTTGGCAACCTCATTGAG GCGGTCAATTTCTTTTTGAA 4800

4801 TGTTTACAACTGAAGCATTA ATGCCAGAGATGTCACCTAA ATCAACATCTGGTGATGTAT GATTCTTAAAATATTTATCT AACTCCTCCTTGAATGAGTC 4900

4901 TAATTCAGGTTGCAAAGGAT CATAAACTGTGTTGTTGACA ATTCCTATTACAACATCACA GTTACCAGACACAAATGTGT TGTCTGTAGTAATGATTTGT 5000

5001 GGTTCATAAAAATTCCTTTG TGTTACAAACCAGTGTGTGC CATTTGAAACAAAGACACCT TCACGAGGAAAGTGTGCTTT TCCATCATGACAAATGGCAG 5100

5101 GAGCAGTTGTGAAGTTCTTT TCTTGTGCAGGGACATAAGT CACATGCAAGAAGACTACAC CATGAGGTGCTGACTGAGGG AAGGACATAAGATGATAGCC 5200

5201 CTTTCCACAAAAATCAACTC TTTTTGATTGTCCAAGTACA CACTCTGACATTTTAGTAGC AGCAAGATTAGCAGAAGCTC TGATTTCTGCAGCTCTAATT 5300

5301 AATTGTTGAGTCACATATGT CTGCAAACTTTGAAGTCTGC CTGTGATCAACCTATCAATT TGCACTTCAGCCTCAACTTT GTCAAGACGTGAAAGGATAT 5400

5401 CATTTAAAACACTTGAAATT GCACCAAAATTGGAGCTAAG TTGTTTAACAAGCGTGTTTA AAGCTTGTGCATTTTGGTTG ACCACATCTTGAAGTTTTCC 5500

5501 AAGTGCACTTGCTGTGGAAG AAAGTGAGTCTTGAATTTTG CCAATAGCACTATTAAATTG GTTGGCAATCAATTTTTGGT TCTCATAGAGAACATTCTGT 5600

5601 GTAACTCCAATACCATTAAA CCTATAAGCCATTTGCATAG CAAATGGTATTTGTAATGCA GCACCTGCACCAAAGGTCCA ACCAGAAGTGATTGTACCCG 5700

5701 CTAACAGTGCAGAAGTGTAT TGAGCAATCATTTCATCTGT GAGCAAAGGTGGCAAAACAG TAAGGCCGTTAAACTTTTGT GCACAAATGAGGTCTCTAGC 5800

5801 AGCAATATCACCAAGGCAAT CACCATATTGTTTGATGAAG CCAGCATCTGCAAGTGTCAC TTTGTTGAAAAGTAGATCTT CAATAAATGACCTCTTGCTT 5900

5901 GGTTTTGATGGATCTGGTAA TATTTGTGAAAAATTAAAAC CACCAAAATCTTTAATTGGT GGTGTTTTGTAAATTTGTTT GACTTGTGCAAAAACTTCTT 6000

6001 GGGTGTTTTTGTCTTGTTCA ACAGCTATTCCAGTTAAAGC ACGGTTTAATTGTGTACAAA AACTGCCATATTGCAACAAA AGATTGCTGCATTCAGTTGA 6100

6101 ATCACCACAAATGTACATTG TACAATCTACTGATGTCTTG GTCATAGACACTGGTAGAAT TTCTGTGGTAACACTAATAG TAAAATTTGTGGGTATGGCA 6200

6201 ATAGAGTTATTAGAGTAAGC AACTGAATTTTCTGCACCAA GTGACATAGTGTAGGCAATG ATGGATTGACTAGCTACACT ACGTGCCCGCCGAGGAGAAT 6300

6301 TAGTCTGAGTCTGATAACTA GCGCATATACCTGCACCAAT GGGTATGTCACACTCATATG AGTTGTTGACATGTTCAGCC CCTATTAAACAGCCTGCACG 6400

6401 TGTTTGAAAAACATTAGAAC CTGTAGAATAAACACGCCAA GTAGGAGTAAGTTGATCTGC ATGAATAGCAACAGGGACTT CTGTGCAGTTAACACCCTGA 6500

6501 TAAAGAACAGCAACCTGGTT AGAAGTATTTGTTCCTGGTG TTATAACACTGACACCACCA AAAGAACATGGTGTAATGTC AAGAATCTCAAGTGTCTGTG 6600

6601 GATCACGGACAGCATCAGTA GTGTCAGCAATGTCTCTGCC AAATTGTTGGAAAGGCAGAA ACTTTTTGTTAGACTCAGTA AGAACACCTGTGCCTGTTAA 6700

6701 ACCATTGAAGTTGAAATTGA CACATTTGTTTTTAACCAAA TTAGTAGACTTTTTAGGTCC ACAAACAGTTGCTGGTGCAT GTAGAAGTTCAAAAGAAAGT 6800

6801 ACTACTACTCTGTATGGTTG GTAACCAACACCATTAGTGG GTTGGAAACCATATGATTGT AAAGGAAAGTAACAATTAAA ACCTTCAACACCATTACAAG 6900

6901 GTGTGCTACCGGCCTGATAG ATTTCAGTTGAAATATCTCT CTCAAAAGGTTTGAGATTAG ACTTCCTAAACAATCTATAC AGGTAATTATAATTACCACC 7000

7001 AACCTTAGAATCAAGATTGT TAGAATTCCAAGCTATAACG CAGCCTGTAAAATCATCTGG TAATTTATAATTATAATCAG CAATCTTTCCAGTTTGCCCT 7100

7101 GGAGCGATTTGTCTGACTTC ATCACCTCTAATTACAAATG AATCTGCATAGACATTAGTA AAGCAGAGATCATTTAATTT AGTAGGAGACACTCCATAAC 7200

7201 ACTTAAAAGTGGAAAATGAT GCGGAATTATATAGGACAGA ATAATCAGCAACACAGTTGC TGATTCTCTTCCTGTTCCAA GCATAAACAGATGCAAATCT 7300

7301 GGTGGCGTTAAAAACTTCAC CAAAAGGGCACAAGTTTGTA ATATTAGGAAATCTAACAAT AGATTCTGTTGGTTGGACTC TAAAGTTAGAAGTTTGATAG 7400

7401 ATTCCTTTTTCTACAGTGAA GGATTTCAACGTACACTTTG TTTCTGAGAGAGGGTCAAGT GCACAGTCTACAGCATCTGT AATGGTTCCATTTTCATTAT 7500

7501 ATTTTAATAGAAAAGTCCTA GGTTGAAGATAACCCACATA ATAAGCTGCAGCACCAGCTG TCCAACCTGAAGAAGAATCA CCAGGAGTCAAATAACTTCT 7600

7601 ATGTAAAGCAAGTAAAGTTT GAAACCTAGTGATGTTAATA CCTATTGGCAAATCTACCAA TGGTTCTAAAGCCGAAAAAC CCTGAGGGAGATCACGCACT 7700

7701 AAATTAATAGGCGTGTGCTT AGAATATATTTTAAAATAAC CATCAATATTCTTAAACACA AATTCCCTAAGATTTTTGAA ATTACCCTGTTTTCCTTCAA 7800

7801 GGTCCATAAGAAAAGGCTGA GAGACATATTCAAAAGTGCA ATTATTCGCACTAGAATAAA CTCTGAACTCACTTTCCATC CAACTTTTGTTGTTTTTGTG 7900

7901 GTAATAAACACCCAAAAATG GATCATTACAAAATTGAAAT TCACAGACTTTAATAACAAC ATTAGTAGCGTTATTAACAA TAAGTAGGGACTGGGTCTTC 8000

8001 GAATCTAAAGTAGTACCAAA AATCCAGCCTCTTATTATGT TAGACTTCTCAGTGGAAGCA AAATAAACACCATCATTAAA TGGTAGGACAGGGTTATCAA 8100

8101 ACCTCTTAGTACCATTGGTC CCAGAGACATGTATAGCATG GAACCAAGTAACATTGGAAA AGAAAGGTAAGAACAAGTCC TGAGTTGAATGTAAAACTGA 8200

8201 GGATCTGAAAACTTTGTCAG GGTAATAAACACCACGTGTG AAAGAATTAGTGTATGCAGG GGGTAATTGAGTTCTGGTTG TAAGATTAACACACTGACTA 8300

8301 GAGACTAGTGGCAATAAAAC AAGAAAAACAAACATTGTTC GTTTAGTTGTTAACAAGAAC ATCACTAGAAATAACAACTC TGTTGTTTTCTCTAATTATA 8400

8401 AGTCTACCTTTACTAAGAAG AGATAAAATCATATCATTGA TTTGACCTTCTTTTAAAGAC ATAACAGCAGTACCCCTTAA TTTAAGGGGAAATTTACTCA 8500

8501 TGTCAAATAAAGAATAGGAA GACAACTGAATTGGATTTGT ATTCCTCCAAAATATGTAAT TTGCATGCATGACATAACCA TCTATTTGTTCGCGTGGTTT 8600

8601 GCCAAGATAATTACATCCAA TTAAAAATGCTTCAGATGAT GACGCATTCACATTAGTAAC AAAGGCTGTCCACCATGCGA AGTGTCCCATGAGCTTATAA 8700

8701 AGATCAGCATTCCAAGAATG TTCTGTTATCTTTATAGCCA CGGAACCTCCAAGAGCTAGC TTTTGTTGTATAAACCCACA AATGTAAGTGAAAAAACCCT 8800

8801 CTTTAGAGTCATTTTCTTTT GTAACATTTTTAGTCTTAGG GTCGTACATATCACTAATAA TGAGATCCCATTTATTAGCT GTATGTACAGTTGCACAATC 8900

8901 ACCAATCAAAGTTGAATCTG CATCAGAGACAAAGTCATTA AGATCTGAATCGACAAGCAG CGTACCCGTAGGCAACCACT GTCTTAAAACAGCTGTACCT 9000

9001 GGTGCAACTCCTTTATCAGA ACCAGCACCAAAATGTATAA CTCTCATATTATAGGGTACA GCTAATGTTAATGTGTTTAA ATATTGACACAGTTGAGTAT 9100

9101 ATTTTGCGACATTCATCATT ATGCCTTTAGGTAATGTTGC ACTATCACCATAATTTTGAA GGTCACACTTTTCTAATAGC ATTCTTTGCATTTTGTAAAG 9200

9201 ATTAGGCATAGCAACACCCG GTTGCCACGCTTGACTAGAT TGTAATTTTGGGTAAAATGT TTCTACATGGCCATCTTTAC ACCAAAGCATAAATGAAATT 9300

9301 TCTGTATAGTCAATAGTCAC TTTGACAACCTTAGAAACTA CAGATAAATCTTGGGATTTT ATTATTTCAACAAAATCATC AAGTAATAAATCAATAACAG 9400

9401 AACACACACACTTAGATGAA CCTGTTTGCGCATCTGTTAT GAAATAGTTTTTAACTGTAC TGTCCATAGGAATAAAATCT TCTAATTCAAAAGGTGATTC 9500

9501 CTTAAAACGTTTAGCTAGTC CAATCAGTAGATGTAAACCA CCTAACTGACTATGACTAAA ATCTCCATAAACGATATGTT CGAAGGCATAGCCTTCTAAT 9600

9601 TTATACCGTTCAATGAATTC ATCCATAGCTAATTCTAAGA AATCAATTTCCATTTGACTC CTGGGTTTAAATTCTTGTAA ATTTCTACTCTGAGTAAAGT 9700

9701 AAGTTTCAGGTAATTGTTGG ACAACACCATCAACTTTCTT ATAATAATTGAACTGTGTTT TTACGGCTTCTCCAATTAAT GTGACTCCATTAAGACTAGC 9800

9801 TTGTTTGGGACCTACAGATG GTTGTAAACCTTTAACACTA CCTTCTGTAATAAGAACACC ATTACGGGCATTTCTAAATA AGTCTACTTGACCATCAACT 9900

9901 CTACCATCAAAAAAGACAGT GAGTGGTGCACAAATCGTTT CAGTTGGTTTCTTGGCTATG TCAGTCATAGAACAAACACC AATAGTAGATATATGTGCTG 10000

10001 GAGCATCTCTTTTGTAGTCC CAGATCACAGTATTAGCAGC AATGTCCACACCCAAATTAT TGAGTATTTTCACCTCTGGT ACTGGTTTAATGTTGCGCTT 10100

10101 AGCCCAAAGCTCAAATGCTA CATTAACAGGTAATGTTGTT TTATTTTCAAACAATTCTAC ATCAACACCATCAACTTTTG TGTAAACAGTGTTATTAATG 10200

10201 ATAGAAACTGGTACTTCACC CTGTTGTCCATCAAAGTGTC CCTTATTTACAACATTAAAA GCCACATTTTCTAAACTCTG AAGTCTTGTAAAAGTGTTCC 10300

10301 AGAGGTTATAAGTATCAAAT TGTTTGTAAACCCACAAGCT AAAGCCAGCTGAGATCATCA TGTTATAAGCATCGAGATAC AATCTGTACTCATTAGCATG 10400

10401 ATGTCTACAGACAGCACCAC CTAAATTGCAACGTGTTATA CACGTAGCAGACTTTAGTGG TACATAATCTATATCTGACA CTACTTGTTTTCCATGAGAC 10500

10501 TCACATGGACTGTCAGAGTA ATAGAAAAATGGTAATTGTT TTAAATTAACAAAAGCACTT TTATCAAAAGCTGGTGTGTG GAATGCATGTTTATTTACAT 10600

10601 ACAAACTGCCACCATCACAA CCAGGCAAGTTAAGGTTAGA TAGCACTCTAGTGTCAAATC TACAAACAATGGAATTAGCA GGATATCTATCGACATTGCA 10700

10701 ATTCCAAAATAGGCATACAC CATCTGTGAATTTGTCAGAA TGTGTGGCATAAGAATAGAA TAATTCTTCTATTTTATAAG CTTTGTCACTACAAGGCTGT 10800

10801 GCATCATAGAACTTCCATTC TACATCAGCTTGAGGTACAC ACTTAATAGCTTTAGGGTTA CCAATGTCGTGAAGAACTGG GAATTTGTCTGCTAATAATG 10900

10901 CAGCTTTAACAACCATGTGT TGAACCTTTCTACAAGCCGC ATTAATCTTCAGTTCATCAC CAATTATAGGATATTCAATA GTCCAGTCAACACGCTTAAC 11000

11001 AAAGCACTCGTGGACAGCTA GACACCTAGTCATGATTGCA TCACAACTAGCTACATGTGC ATTACCATGGACTTGACAAT ACAGATCATGGTTGCTTTGT 11100

11101 AGGTTACCTGTAAAACCCCA TTGTTGAACATCAATCATAA ACGGATTATAGACGTAATCA AATCCAATAGAATGATGCCA ACAGGCATAAGTGTCTGAAG 11200

11201 CAGTGGAAAAGCATGTGGCA CGTCTATCACATAGACAACA GGTGCGCTCAGGTCCTATTT TCACAAAATACTTCATAGAT GTCAACTCAAAGCCATGTGC 11300

11301 CCATAAGACAAATACGACTC TGTCAGAGAGATTTTTAAGT GTGTCACTTAACATTTGTAC AATCTTTATACGCACTACAT TCCAAGGAAGTCCTTTGTAC 11400

11401 ATAAGTGGTATGAGGTGTTT AAATTGATCTCCAGGCGGTG GTTTAGCACTAACTCTGGAA AAATCTGTATTATTAGGTGT ATCAACATAACCTGTAGGTA 11500

11501 CAGCAACTAGGTTAACACCT GTAGAAAAACCTAGCTGTAA AGGTAAATTGGTACCAACAG CTTCTCTAGTAGCATGACAC CCCTCGACATCGAAGCCAAT 11600

11601 CCATGCACGTACATGTCTTA TAGCTTCTTCGCGGGTGATA AACATGTTAGGGTAACCATT AACTTGATAATTCATTTTAA AACCCATCATAGAGATGAGT 11700

11701 CTTCTATAGGTCATGTCCTT AGGTATGCCAGGTATGTCAA CACATAAACCTTCAGTTTTG AATTTAGTGTCAACACTGAG GTGTGTAGGTGCCTGTGTAG 11800

11801 GATGTAACCCAGTGATTACC TTACTACAATCTTTAAAGAG TCCTGTTACATTTTCAGCTT GTAAAGTTGCCACATTCCTA CGTGGAATTTCAAGACTTGT 11900

11901 AAATTGCAACTTGTCATAAA GGTCTCTATCAGACATTATG CAAAGTATGCCTACTTTTGC TCTGGTAATAGCAACATTAA ATCTGTTTACATTACAAGAG 12000

12001 TGAGCTGTTTCAGTGGTTTG AGTGAATATGACATAGTCAT ATTCTGAGCCCTGTGATGAA TCAACAGTTTGAGTTGGTAG TCCCAAAATCTTTGAGGCTA 12100

12101 CAGCATTCTGTGAATTATAA GGTGAAATAAAGACAGCTTT TCTCCAAGCAGGGTTACGTG TAAGGAATTCTCTTACCACG CCTATTTGTGGCCTGTTAAT 12200

12201 TGCAGATGAAACATCATGCG TGATAACACCCTTATAAAAC ATTTTAAAGCATTGAGCTGA TTTGTCTTTATGTGCTTTAA GCTTATTATCATAAACCAAA 12300

12301 GCACTCACAGTGTCAACAAT TTCAGCAGGACAACGCCGAC AAGTTCCGAGGAACATGTCT GGACCTATAGTTTTCATAAG TCTACACACTGAATTGAAAT 12400

12401 ATTCTGGTTCTAGTGTGCCC TTAGTTAGCAATGTGCGTGG TGCAGGTAATTGAGCAGGGT CGCCAATGTACACATAGTGC TTAGCACGTAATCTGGCATT 12500

12501 GACAACACTCAAATCATAAT TTGTGGCCATTGAAATTTCA TCAAAGACAACTATATCTGC TGTCGTCTCAGGCAATGCAT TTACAGTACAAAAGACATAC 12600

12601 TGTTCTAATGTTGAATTCAC TTTGAATTTATCAAAACACT CTACACGAGCACGTGCAGGT ATAATTCTACTACATTTATC TATAGGCAAATATTTTAATG 12700

12701 CCTTCTCACATAGTGCATCA ACAGCGGCATGAGAGCAAGC TGTATACACTATGCGAGCAG AAGGGTAGTAGAGAGCTAGG CCAATAGCAAAATGACTCTT 12800

12801 ACCAGTACCAGGTGGTCCCT GGAGTGTAGAATACTTTTGC ATACCAACCTTTTGATAATT TGCAACATTGCTAGAAAACT CATCTGAGATATTGAGTGTT 12900

12901 GGGTATAAGCCAGTAATTCT AACATAGTGCTCTTGTGGCA CTAGTGTAGGTGCACTTAAT GGCATTACTGTATGTGATGT CAGCACAAAATAATCACCAA 13000

13001 CATTTAATTTGTAAGTTGTT GTACCTCGGTAAACAACAGC ATCACCATAGTCACCTTTTT CAAAGGTGTACTCTCCTATT TGTACTTTACTGTTTTTAGT 13100

13101 TACACGATAACCAGTAAAGA CATAATTTCGGTTAAGTGGT GGTCTAGGTTTACCAACTTC CCATGAAAGATGTAATTCTC TGTCAGACAGCACTTCACGT 13200

13201 ACAGTAGCAATACCATAAGA CAGTTTAAATGTCTCCTCAG TAGCTTTGAGCGTTTCTGCT GCAAAAAGCTTGAGTCTTTC AGTACAGGTGTTAGCTAAAA 13300

13301 TGTAATCACCAGCATTTGTC CAGTCACATGTTGCAATTGC ATTAAAGTCAGTAACATTAT CGCTACCAACACATGTATTT TTATATAAACCAAAAACTTG 13400

13401 TCCATTAGCACACAATGGAA AACTAATGGGTGGTTTATGT GATTTACAATAATAGCTCAT ACCTCCTAAGTAAAGTTGAG TCACATCTGTGACATCACAA 13500

13501 CCTGGAGCATTGCAAACATA CGGATTAACAGACAAGACTA ATTTATGTGATGTTGATATG ACATGGTCGTAACAGCATTT ACAACATAAGAATGGTCTAC 13600

13601 GTATGCAAGCACCACATCTT AATGAAGTCTGTGAATTGCA AAGAACACAAGCCCCAACAG CCTGTAAGACTGTATGCGGT GTGTACATAGCCTCATAAAA 13700

13701 CTCAGGTTCCCAATACCTTG AAGTGTTATCATTAGTAAGC ATAACAGAATACATGTCTAA CATGTGTCCTGTTAACTCAT CATGTAGCTTTCTTATGTAT 13800

13801 TGTAAGTACAAATGAAAGAC ATCAGCATACTCCTGATTAG GATGTTTAGTAAGTGGGTAA GCATCTATAGCTAAAGACAC GAACCGTTCAATCATAAGTG 13900

13901 TACCATCTGTTTTTACGATA TCATCTACAAAACAGCCGGC CCCTAGGATTCTTGATGGAT CTGGGTAAGGAAGGTACACA TAATCATCACCCTGTTTAAC 14000

14001 TAGCATTGTATGTTGAGAGC AAAATTCATGAGGTCCTTTA GTAAGGTCAGTCTCAGTCCA ACATTTTGCTTCAGACATAA AAACATTGTTTTGATAATAA 14100

14101 AGAACTGACTTAAAGTTCTT TATGCTAGCCACTAGACCTT GAGATGCATAAGTGCTATTG AAACACACAACAGCATCGTC AGAGAGTATCATCATTGAGA 14200

14201 AATGTTTACGCAAATATGCG TAAAACTCATTCACAAAGTC TGTGTCAACATCTCTATTTC TATAGAGACACTCATAAAGT CTGTGTTGTAAATTGCGGAC 14300

14301 ATACTTATCGGCAATTTTGT TACCATCAGTAGATAAAAGT GCATTAACATTGGCCGTGAC AGCTTGACAAATGTTAAAAA CACTATTAGCATAAGCAGTT 14400

14401 GTGGCATCTCCTGATGAGGT TCCACCTGGTTTAACATATA GTGAACCGCCACACATGACC ATTTCACTCAATACTTGAGC ACACTCATTAGCTAATCTAT 14500

14501 AGAAACGGTGTGACAAGCTA CAACACGTTGTATGTTTGCG AGCAAGAACAAGTGAGGCCA TAATTCTAAGCATGTTAGGC ATGGCTCTATCACATTTAGG 14600

14601 ATAATCCCAACCCATAAGGT GAGGGTTTTCTACATCACTA TAAACAGTTTTTAACATGTT GTGCCAACCACCATAGAATT TGCTTGTTCCAATTACTACA 14700

14701 GTAGCTCCTCTAGTGGCGGC TATTGATTTCAATAATTTTT GATGAAACTGTCTATTGGTC ATAGTACTACAGATAGAGAC ACCAGCTACGGTGCGAGCTC 14800

14801 TATTCTTTGCACTAATGGCA TACTTAAGATTCATTTGAGT TATAGTAGGGATGACATTAC GTTTTGTATATGCGAAAAGT GCATCTTGATCCTCATAACT 14900

14901 CATTGAATCATAATAAAGTC TAGCCTTACCCCATTTATTA AATGGAAAACCAGCTGATTT GTCTAGGTTGTTGACGATGA CTTGGTTAGCATTAATACAG 15000

15001 CCACCATCGTAACAATCAAA GTACTTATCAACAACTTCAA CTACAAATAGTAGTTGTCTG ATATCACACATTGTTGGTAG ATTATAACGATAGTAGTCAT 15100

15101 AATCGCTGATAGCAGCATTA CCATCCTGAGCAAAGAAGAA GTGTTTTAATTCAACAGAAC TTCCTTCCTTAAAGAAACCC TTAGACACAGCAAAGTCATA 15200

15201 GAAGTCTTTGTTAAAATTAC CGGGTTTGACAGTTTGAAAA GCAACATTGTTAGTAAGTGC AGCTACTGAAAAGCACGTAG TGCGTTTATCTAGTAATAGA 15300

15301 TTACCAGAAGCAGCGTGCAT AGCAGGGTCAGCAGCATACA CAAGTAATTCCTTAAAACTA AGTCTAGAGCTATGTAAGTT TACATCCTGATTATGTACAA 15400

15401 CACCTAGCTCTCTGAAGTGG TATCCAGTTGAAACTACAAA TGGAACACCATCAACAAATA TTTTTCTCACTAGTGGTCCA AAACTTGTAAGTGGGAACAC 15500

15501 TGTAGAGAATAAAACATTAA AGTTTGCACAATGCAGAATG CATCTGTCATCCAAACAGTT AACACAATTTGGGTGGTATG TCTGATCCCAATATTTAAAA 15600

15601 TAACGGTCAAAGAGTTTTAA CCTCTCTTCCGTGAAGTCAT ATTTTAACAAATCCCACTTA ATGTAAGGCTTTGTTAAGTC AGTGTCAACATGTGACTCTG 15700

15701 CAGTTAAAGCCCTGGTCAAG GTTAATATAGGCATTAACAA TGAATAATAAGAATCTACAA CAGGAACTCCACTACCTGGC GTGGTTTGTATGAAATCACC 15800

15801 GAAATCATACCAGTTACCAT TGAGATCTTGATTATCTAAT GTCAGTACACCAACAATACC AGCATTTCGCATGGCATCAC AGAATTGTACTGTTTTTAAC 15900

15901 AAAGCTTGGCGTACACGTTC ACCTAAGTTGGCGTATACGC GTAATATATCTGGGTTTTCT ACAAAATCATACCAGTCCTT TTTATTGAAATAATCATCAT 16000

16001 CACAACAATTGTATGTGACA AGTATTTCTTTTAATGTGTC ACAATTACCTTCATCAAAAT GCCTTAAAGCATAGACGAGG TCTGCCATTGTGTATTTAGT 16100

16101 AAGACGTTGACGTGATATAT GTGGTACCATGTCACCGTCT ATTCTAAACTTAAAGAAGTC ATGTTTAGCAACAGCTGGAC AATCCTTAAGTAAATTATAA 16200

16201 ATTGTTTCTTCATGTTGGTA GTTAGAGAAAGTGTGTCTCT TAACTACAAAGTAAGAATCA ATTAAATTGTCATCTTCGTC CTTTTCTTGGAAGCGACAAC 16300

16301 AATTAGTTTTTAGGAATTTA GCAAAACCAGCTACTTTATC ATTGTAGATGTCAAAAGCCC TGTATACGACATCAGTACTA GTGCCTGTGCCGCACGGTGT 16400

16401 AAGACGGGCTGCACTTACAC CGCAAACCCGTTTAAAAACG ATTGTGCATCAGCTGACTGA AGCATGGGTTCGCGGAGTTG ATCACAACTACAGCCATAAC 16500

16501 CTTTCCACATACCGCAGACG GTACAGACTGTGTTTTTAAG TGTAAAACCCACAGGGTCAT TAGCACAAGTTGTAGGTATT TGTACATACTTACCTTTTAA 16600

16601 GTCACAAAATCCTTTAGGAT TTGGATGATCTATGTGGCAA CGGCAGTACAGACAACACGA TGCACCACCAAAGGATTCTT GATCCATATTGGCTTCCGGT 16700

16701 GTAACTGTTATTGCCTGACC AGTACCAGTGTGTGTACACA ACATCTTAACACAATTAGTG ATTGGTTGTCCCCCACTAGC TAGATAATCTTTGTAAGCTT 16800

16801 TAGCAGCATCTACAGCAAAA GCACAGAAAGATAATACAGT TGAATTGGCAGGCACTTCTG TTGCATTACCAGCTTGTAGA CGTACTGTGGCAGCTAAACT 16900

16901 ACCAAGTACCATACCTCTAT TTAGGTTGTTTAATCCTTTA ATAAAGTATAAATACTTCAC TTTAGGACCTTTAGGTGTGT CTGTAACAAACCTACAAGGT 17000

17001 GGTTCCAGTTCTGTATAGAT AGTACCAGTTCCATCACTCT TAGGGAATCTAGCCCATTTC AAATCCTGTAAATCGGATAA CAGTGCAAGTACAAACCTAC 17100

17101 CTCCCTTTGTTGTGTTGTAG TAAGCTAACGCATTGTCATC AGTGCAAGCAGTTTGTGTAG TACCGGCAGCACAAGACATC TGTCGTAGTGCAACAGGACT 17200

17201 AAGCTCATTATTCTGTAATT TGACAGCAGAATTGGCCCTT AAAGCTGTTACAATAAGAGG CCATGCTAAATTAGGTGAAT TGTCCATACTAATTTCACTA 17300

17301 AGTTGAACAATTTTACTATC TGCATCTACAACCTGTTGGA TTTCCCACAATGCTGATGCA TAAGTAAATGTTGTACCATC ACACGTATTTTTATATGTGT 17400

17401 TATAGTCTGGTATGACAACC ATTAGTTTGGCTGCTGTTGT AAGAGGTATTATGTTCAAGG GAACACAACCATCTCTTGCA TTGTTGATAATGTTGTTGAG 17500

17501 TGCATCATTATCCAACTTTC TAAGCATAGTGAAAAGCATT GTCTGCATAGCACTAGTAAC TTTTGCCCTCTTGTCCTCAG ATCTAGCCTGTTTATACATT 17600

17601 TGGGTCATAGCTTGATCAGC CATCTTTTCCAACTTACGTT GCATGGCTGCATCACGGTCA AATTCAGATTTAGCCACATT CAAAGACTTCTTCAACTTTT 17700

17701 TAAGAACAACTTCAGAATCA CCATTAGCAACAGCCTGCTC ATAAGCTTCTTGAGCAGTAG CAAAAGCTGCATATGATGGA AGGGAACTAAACTCTGAGGC 17800

17801 TATAGCTTGTAAGGTTGCCC TGTTGTCCAGCATTTCTTCA CAAAGCTTGTTTATGTCTAC AGCACCCTGCATGGAAAGCA AAACAGAAAGTAGTGAAACC 17900

17901 ATTTTTTCAAAGGCTTCAGT AGTATCTTTAGCTAAGAGAA TGTCATTGTGTAACTGGACA CATTGAGCCCACAATTTAGA TAATGATTCTACTCTGAGTT 18000

18001 GTTGCAAAACTGAGAGTAAG ACTACTGATGTGCACTTTAC ATCTGACATTTTAGACTGTA CAGTGGCTACTTTGATACAA GGTTTGCCACCAACACCCAA 18100

18101 CAATTTAATGTTGAGTTTGA AGGCATCTATGCTATTCTTG GGTGGGAGTAGTCCCTGTGA ATTCATATATCTAAACTCCT GTGTAGAAACTAAGTAATCA 18200

18201 TAAACACCAAGAGTCAGTCT AAAGTAGCGGTTGAGTAAAC AAAAGAGGCCAAAGTAACAA GTACAAAAATAGCCTAAGAA ACAATAAACTAGCATTATAC 18300

18301 ACTGAAGTGTATTACCAGTT ATGAAGAAAATAGGGCAATA CTCAACACACATAAAAACAA TACCTCTGGCCAAAAACATG ACAGTTGTAACTACACCTGA 18400

18401 GTAGTTAGAAGTAACAGAGA TTATAAGAGCCCACATGGAA ATGGCTTGATCTAAAGCATT ACCATAATAAACTTTATAAA CGAGTGTCAAGACATTCATA 18500

18501 AGTGTCCACACTCTCCTAGC ACCATCATCATACACAGTTC TTGCTGTCATAAGGATTAGT AACACTACAGCTGATGCATA CATAACACAGTCTTTTAGCT 18600

18601 TAAAACCAGACAAACTAGTA TCAACCATATCCAACCATGT CATAATACGCATCACCCAAC TAGCAGGCATATAGACCATA TTAAAATAAGCTACAGTGGC 18700

18701 AAGAGAAGGTAACAAAAACA AACAGAGAAATGCATGCTTA TGTTTGACAAACATCATTGC AAAAGCAGACATAGCAATAA TACCCATAGCAAAAGGTAAA 18800

18801 AAGGCATTTTCATACAAAAA AAAGAACAAAGACCATTGAG TACTCTGGACTAAAACTAAA AGTGAAGTCAAAATTGTGAG TAACAACCAGTGGTGTGTAC 18900

18901 CCTTGATTGTTCTTTTCACT GCACTTTGGAAAGTAACACC TGAGCATTGTCTAACAACAT CAAAAGGTGTAAATTCATCT TCTAATAAAGCACTACCCAA 19000

19001 TATGGTACGTCCATTCATAC CATTTTGCAGTAATTCTTTT AATGAAGCACACATATCTAA AACGGCAATTCCAGTTTGAG CAGAAAGAGGTCCTAGTATG 19100

19101 TCAACATGGTCTTGTGTTAG AGGTTCATAATTGTACTTCA TAGCCACAAGGTTAAAGTCA TTAAGAGTTGTGGTAAATCG ATTGAGAAACCACCTGTCTC 19200

19201 CATTTATAACAGCAGCGTAC AACCAAGCTAAAACATTAAC TGTAATAGTTGTGTCCGTAC CAGCTGCTTGTGCTGTTTGC CTGTCAACAAAAGGTCCATA 19300

19301 AAAGTTACCTTCTAAGTCTG TGCCAGCATGAACTCCAGTT GGTAATTCCATATGGTGCAT GTAACAAAAAGAGACACAGT CATAATCTATGTTAAAACCA 19400

19401 ACACTACCACATGAACCATT AAGGAATGAACCCTTAATAG TGAAATTGGGCCTCATAGCA CATTGGTAAACACCAGATGG TGAACCATTGTAACAAGCTA 19500

19501 ACACTGAAAAAGTCTGTCCT GGTTGAATGCGAACAAACTT ATACTTAGGTGTCTTAGGAT TGGCTGTATCAACCTTAAGC TTAAGTACACAATTTTGCAT 19600

19601 AGAATGTCCAATAACCCTGA GTTGAACATTACCAGCCTGT ACCAAGAAATTATGATTAGA CTTACGAATGAGTAAATCTT CATAATTAGGGTTAAGCATG 19700

19701 TCTTCAGAGGTGCAGATCAC ATGTCTTGGACAGTAAACTA CGTCATCAAGCCAAAGACCG TTAAGTGTAGTTGTACCACA AGTTACTTGTACCATACAAC 19800

19801 CCTCAACTTTACCAGATGGG AATGCCATTTTTCTAAAACC ACTCTGCAAAACAGCTGAGG TGATAGAGGTTTGTGGTGGT TGGTAAAGAACATCAGAACC 19900

19901 TGAGTTACTGAAGTCATTGA GAGCCTTTGCGAGATGACAA CAAGCAGCTTCTCTGTAGCT AGTTGTATCCATTGCTCCAC TAAAATACTTGTACTTATTA 20000

20001 TAAAGAGCTAAGTATCTATT ATATTGCGTAAGAGGTAATA GCACATCACTACGCAACTTT AGATACATTTCTTTATTTAA CAAAAAGGTGCACAGCGCAG 20100

20101 CTTCTTCAAAAGTACTAAAG GAAACACCATTAAAGACTAC ACGTCTCTTTAGGTAATTAC TAAAGAACCAATAGAAATGC TTTGTGGAAATACAAATGAT 20200

20201 ATAAGCAATTGTTATCCAGA AAGGTACTAAAGGTGTGAAC ATAACCATCCACTGAATATG TGCTAAAAAAGAAACATCAT TAGTAAGATAAAATGTCAAG 20300

20301 TACAAGTAAATAACAGAATA AACACCAGGTAAGAATGAGT AAACTGGTGTTAAACAGAGT ACAGTGAATGACATAAGGAA TAGTAAAGTATTAAAGGCAA 20400

20401 CTACATGACTGTATTCACCA AAAGCTCTTCTAAACCTCAT AAAATAGTAGGCAAGGCATG TTACTACGATAGCTACAATA CCACCAGCTACTATAGATGC 20500

20501 TGATATGTCCAAAGCACCAA TAGGTTGAATTAGTGGTGTA AACATATTAGTAAGTAAATT TACAGCATCTACACCACAGA AAACTCCTGGTAAAGATCTG 20600

20601 TAATAATCATTGTTAAGTAC CCATCTACCACTAGTAGATA CACAAACACCAGCTTCTGAT CTTTCACAAGTGCCGTGCCT ACAGTACTCAGAATCAAAAG 20700

20701 TTGTTACCACTCTAACAGAA CCTTCAAGGTAGGTGTTAGG AAATTGAATAATAGAGCCAT CCATGAGCACATAACGTGTG TCAGGGCGTAAACTTTCATA 20800

20801 AGCAACAGAACCTTCTAGTA CATTGGTATCATAACAATAT GGTACTGGCTTACCAGAAGC ATCTTTAAAAATTGTACATT CAGCAGCCAAAACACAAGCT 20900

20901 GATGTTGCAAAGTCAGTGTA CTCTATAAGTTTTGATGGTG TGTAACAGATGTTACCAACT GCACTAAAAACTCTAGGTAA GAAATGCAAAAAGTCACCAT 21000

21001 TAGTTGTGCGTAATATCGTG CCAGGCAAACCAGGCACGAC AAAACCCACTTCTCTTGTTA TGACTGCAGCAATCAATGGG CAAGCTTTGTCATTAGTATA 21100

21101 ACTACCACCACGCTGGCTAA ACCATGTGTCAAAATCAGCA TGTTTGTTAGCAAAACAAGT ATCTGTAGATGCTATGTCAC GAGTGACACCACCATCAATA 21200

21201 GCCTTGTATCCTATGATTTC ACTTGAAAAGTCAGTATGTT TAGACATGACATGAACAGGT GTTATTAAATAGAAAATAGC AGCAACAAAAAGGAACACAA 21300

21301 GTGTAACTTTAATTAACTGC TTCAACCAATTATTAACAAT TTTACCACCCTTAAGTGCTA TCTTTGTTGTTACAACATTA ACAACTTGTCTAGTAGTTGC 21400

21401 ACATGTCAACTTAAAAGGTA AGTTATTCTTTTTAGCAGCA CTACGTATTTGTTTTCGTAG TTGTTCAGACAATGACATGA AATCTTTAACGTTCCATATC 21500

21501 AAAGCAATGTTGTGACTTTT TGCTACCTGCGCATTAATAT GACGCGCACTACAGTCAATA CAAGCACCAAGGTCACGGGG TGTCATGTTTTCAACTTTGT 21600

21601 TATAGGTGAGCATATAGTTA TTACAACTATCGCCAGTAAC TTCTATGTCAGATTGATGTG ACAATTTAAGACATTCAACA ACATCTTTAGTTTCTACATC 21700

21701 TGAATCAACAAACCCTTGCC GAGCTGCTGAAATAAAAGTA GATAAGACATTGTCTAAGGA CACATTCTTTGCAAGTTCAG CTTCTGCAGTTGCAACTAGT 21800

21801 GTTTTGAGTTTTTCCATTGG TACGTTAAAAGTTGATGAAA ACGTATTAACGTAAGCATCA AACATTTTAACTGCAACTTC CGCACTATCACCAACATCAG 21900

21901 ACACTAATGCCTGATCTAGT AACAGTATAGGTTGACACAT AAGCTGACTGTAGTAAACAG ACGCTGATTTTGCAGATGAT TCTTCACATTTTGATTTACC 22000

22001 ATCAAAAACTATAACATTAA TAGGCAATGAACCTTTAGTG TTATTAGCTCTCAGGTTGTC TAAGTTAACAAAATGAGAGA GAGAATGTCTTTCATAAGTC 22100

22101 TTTTGACCAGCTTTATCAAA GTAAAGATGGATGGAACCAT TCTTCACTGTAACACTATCA ACGATGTAAGAAGACTGGTC AGTAGGATTTATTGGTCTTT 22200

22201 TAAACTGTAGTGACAAGTCT CTCGCAACTTCATCACTAAT AAATGTACTACCAGCACAGA ATGTATCACAATTAACACAA TTCCAATTGTGTAGTTTGCA 22300

22301 AAAGCCTTTACCTCCATTAG CATAGACATAAAAGGACCTT CTAACACCATTAACAATAGT TGTACATTCGACTCTTGTTG CTCTATTACGTTTGTAACAC 22400

22401 ATCATACAAGTTGATGAATT ACAACCGTCTACAACATGCA CATAACTTTTCCATACATAA TAAAATGATGCAAAGAAGAT GTACATTCTAACCATAGCTG 22500

22501 AAATCGGGGCCATTTGTACA AGATTAATTATTAACCACAT AAGCCAAGAATTACTAATAA AATGTACTGCAAAATAGCTG AAAAACAATTGCATGATTGC 22600

22601 AGCCAATCCAAGTACATAGA AAAACCTAGTGAAAAGAATA TATGCCAAAAACCACTCTGC AACTAAGCCAAAAGCAGTTA AATCCCATTTAAAAGATGAA 22700

22701 ATGGTAATTTGTATAGTTTC TAAAGAAGGATAGGTGTCTA AAGAATCTAAACCACTAAGA CAAACACTACAAGGTATAGA ACCAGTACAGTAGGTTGCAA 22800

22801 TAGTGACATTAGTAGAGTTC AAATAGCCTTCTCTGTAACC AGTACAGTAAGAAGGCATGC CTAAATTAGACATTAAAACA CCTAAAGCAGCGGTTGAGTA 22900

22901 GATTAAAGAACCTAGGCAAA CACTTAATAGTAAAAACCAA ATTATAATATTTATCAGTTT AGAAAAATTAGGTGACTTCA AATAATTAAATGAAGCCTCT 23000

23001 AGACAAAATTTACCGACACT CTTAACAGTATTCTTTGCTA TAGTAGTCGGCATAGATGCT TTAATTCTAGAATTTGTACT TCTAGTAAAAGTACACAATT 23100

23101 GTAGCAATAAAGTAAAGAAA TAAGGCATATAATTAGTACA AACACGGTTTAAACACCGTG TAACTATGTTAGTAGTTGTA CTAACAACTTTGTTAAGAAA 23200

23201 AGGCTTAGCATAATTAGCTA TAGTATCCCAAGGGACACTA TTAACAGCAGCTAAACCATG AGTAGCAAGGGTTTTCAAAC CTAATACTCTAGATAATTCA 23300

23301 TTAGGTTTCTTAATAGTAAG ACTAGAATTGTCTACATAAG CAGCCATTAGATCTGTGTGG CCAACCTCTTCTGTAATTTT TAAACTATTATTTGCTGGTT 23400

23401 TAAGTATAATGTCTCCTACA ACTTCGGTAGTTTTCACATT ACACTCAAGAACGTCTTTCT GTATGGTAGGATTTTCCACT ACTTCTTCAGAGACTGGTTT 23500

23501 TAGATCTTCGCAGGCAAGAT TATCCATTCCCTGCGCGTCC TCTGACTTCAGTACATCAAA CGAATTTGATGTTTCAACTG GTTTTGTGCTCCAAAGACAA 23600

23601 CGTATACACCAGGTATTTGG TTTATACGTGGCTTTATTAG TTGCATTGTTAACATGCCAA ACAATAGGTTTATGTAACAA TTTAGCTCCTTTCTTAAAAG 23700

23701 AGGGTGTGTAGTGTTTATAA TCAATAGCCACCACATCACC ATTTAAGTCAGGGAAAAATG TAACTTTAAGCTCTCTTGAA GCAGGTTTCTTATAACCAGT 23800

23801 TAACTGGTTTAAATCATCAG CAAATTTGATATTATCACAT ACAAACTTAAAATTATCGAA GCTTGCGTTTGGATATGGTT GGTTTGGTACAAGATCAATT 23900

23901 GGTTGCTCTGTGAAATAAGA ATTGTCTTTCTTATAATAAT TGTCCAACTTAGGGTCAATT TCTGTACAAACAACACCATC CAATTTATAAGTAACTGGTT 24000

24001 TTATGGTTGTTGTGTAACTG TTTTCTTTGTAGAAAACATC CGTAATAGGACCTTTGTATT CTGAGGACTTTGTAAGTAAA GCACCGTCTATGCAATACAA 24100

24101 AGTTTCTTTAGAAGTTATAT GTTTATAGTGACCACACTGG TAATTACCAGTGTACTCACT AGCACAAGTAAATGTACCAT GCTTAAGTTCATACTGAGCA 24200

24201 GGTGGTGCTGACATCATAAC AAAAGGTGACTCCTGTTGTA CTAGATATTTTGTAGCTTGT TTACCACACGTACAAGGTAT CTGAACACCTTTCTTAAATT 24300

24301 GTTCATAAGAAAGTGTGCCC ATGTACATAACAGCTTCTAC ACCCTTAAGGGTTGTCTGCT GTTGTCCACAAGTTTTACAC ACCACGTTCAAGACTCTTTT 24400

24401 GCAAGAATCTAAATTGGCAT GTTGAAACAAGTAACTCATT GTTTCTCTAACATCACCTAA CTCACCTACTGTCTTATTAC AGTAGGCTAAGATAAGTGCA 24500

24501 CAAAAGTTAGCAGCTTCACC AGCCCTTGCTCTGTAATAAG CATCTTGTAGAGCAGGTGGA TTAAACTTCAACTCTATTTG TTGGAGTGTTAACAATGCAG 24600

24601 TGGCAAGATAACAGTTGTTA TCTGCCCATTTAATAGAAGT TAAACCATTAACTTGTGGGT ATTTCCACTTTTTAGTGTGA TTTAATGCTGACATGTACCT 24700

24701 ACCCAGAAAACTAGGATCAG TTGTGTGGTAGTACTCAAAA GCCTCAACACGTAGAGTGTC ATCATTAGGTAAAACATAAA ATGTTTTACCTTCATGTGAA 24800

24801 TTATGAGGTTTTATTTTAGT AACATCAGCTCCATCCAAAT AAGTTGGACCAAACTGTTGT CCATATGTCATTGACATGTC CACAACTTGCGTGTGGAGGT 24900

24901 TAATGTTGTCTACTGTTGTA AACACCTTAATAGTCCTCAC TTCTCTCAAAGAAAGAAGTG TCTTAAGATTGTCAAAGGTG ATAACTTCACCATCTAGGTG 25000

25001 GAATGTGGTAGGATTACTAG TGTAATATACACTTTTATCA CCTCTCTTAAGAAATTCTAT ACCTAGTTGTGTAGATTGTC CAGAATAGGACCAATCTTTA 25100

25101 TAGGAACCAGCAAGTGAGAT GGTTTCAATAAAATGTTCTT CAGGTGTTTTAGAAGAAGAA GTAAGATAACCATTATACGC TGTAACAGCATCAGGTGAAG 25200

25201 AAACAGAAACTGTAGCTGGC ACTTTGAGAGATCTCATATA CCGAGCAGCTTCTTCCAAAT TTAAGCCATGTGTTACATAG CCAAGTGGCATTGTAACAAG 25300

25301 AGTTTCATTTAGATCGTTAA GTGTGTTGATAAGTGACGCT ACAGTTGTTTTACTGGTGTA AAAGTAAAATCTAGCACCAT AATCAACCACACCCTCTTGT 25400

25401 ATTTTAATACCCTTATATTT ACGCTGTATAGTTGAAACTA TGGCTTTAGTTTCCACACAG ACAGGCATTAATTTGCGTGT TTCTTCTGCATGTGCAAGCA 25500

25501 TTTCTCGCAAATTCCAAGAA ACAGTTCCAAGAATTTCTTG CTTCTCATTAGAGATAATAG ATGGTAGAATGTAAAAGGCA CTTTTACACTTTTTAAGCAC 25600

25601 TGTCTTTGCCTCCTCTACAG TGTAACCATTTAAACCCTGA CCCGGGTAAGTGGTTATATA ATTGTCTGTTGGCACTTTTC TCAAAGCTTTCGCTAGCATT 25700

25701 TCAGTAGTGCCACCAGCCTT TTTAGTAGGTATAACCACAG CAGTTAAAACACCCTCTTGA ACAACATCACCCACTATATA TGGAGCATCTTTCTTTAAGA 25800

25801 AAGTGATGTCAATGTCACTA ACAAGAGTGGCAGAATCTGG ATGAAGATTGCCATTAATGT CAATATAAAGTAACAAGTTT TCTGTGAGGAACTTAGTTTC 25900

25901 TTCCAGAGTTGTTGTAACTT CTTCAACACAAGCTTTGATT TTCTTATCATCTTGTTTTCT CTGTTCAACTGAAGGTTTAC TTTCAGTTATAAATGGCTTA 26000

26001 ACTTCCTCTTTAGGAATCTC AGCGATCTTTTGTTCAACTT GCTTTTCACTCTTCATTTCC AAAAAGCTTGAAACAAGTTT GTCATAGAGATTTTTATCAA 26100

26101 AGACAGCTAAGTAGACATTT GTGCGAACAGTATCTACACA AACTCTTAAAGAATGTATAG GGTCAGCACCAAAAATACCA GCTGATAATAATGGTGCAAG 26200

26201 TAGAACTTCGTGCTGATTAA AATTTTCATAAGCACTCTTA AGAAGTTGAATGTCTTCACC TTTGTTAACATTTGGGCCGA CAACATGAAGACAGTGTTTA 26300

26301 GCAAGATTGTGTCCGCTTAA AACACAACTACCACCCACTT TAAGTGGTCCATTAGTAGCT ATGTAATCATCAGATTCAAC TTGCATGGCATTGTTAGTAG 26400

26401 CCTTATTTAAGGCTCCTGCA ACACCTCCTCCATGTTTAAG GTAAACATTGGCTGCATTAA CAACCACTGTTGGTTTTACC TTTTTAGCTTCTTCCACAAT 26500

26501 GTCTGCATTTTTAATGTATA CATTGTCAGTAAGTTTTAAA TAACCACTAAAACTATTCAC TTCAATAGTCTGAACAACTG GTGTAAGTTCCATCTCTAAT 26600

26601 TGAGGTTGAACCTCAACAAT TGTTTGAATAGTAGTTGTCT GATTGTCCTCACTGCTGTCT TGTTGACCAACAGTTTGTTG ACTATCATCATCTAACCAAT 26700

26701 CTTCTTCTTGCTCTTCTTCA GGTTGAAGAGCAGCAGAAGT GGCACCAAATTCCAAAGGTT TACCTTGGTAATCATCTTCA GTACCATACTCATATTGAGT 26800

26801 TGATGGCTCAAACTCTTCTT CTTCACAATCACCTTCTTCT TCATCCTCATCTGGAGGGTA AAAAGAACAATACATATGTG AAGCCAATTTAAACTCACCA 26900

26901 GACTCATCAAATAAGTAGTA TGTAGCCATACTCCACTCAT CTAAATCAATGCCCAGTGGT GTAAGTAATTCAGATACTGG TTGCAAAGTTTTTATGACAG 27000

27001 CATCTGCCACAACACAGGCG AACTCATTTACTTCTGTACC GAGTTCAACTGTATAGGCAG AGCACTTCTCATTAAGTACT TTATCAATCCTTTCATCAAG 27100

27101 TTCAAAAGTGATATTCACAC TCTTGTAACCTTGCACTTCT ATCACAGTGTCATCACCAAA AGTAACCTTTGTTGGTGCAC CGCCTTTGAGTGTGAAGGTA 27200

27201 TTGTTTGTTACCATCATATT AGGTGCAAGGGCACAGTACT TTTCTGTGTCTTTGATTTCG AGCAACATAAGCCCGTTAAT ACAAACTGGTGTACCAACCA 27300

27301 ATGGAGCTTCAACAGCTTCA CTAGTAGGTTGTTCTAATGG TTGTAAATCACCAGTTTTCA AGACAACTTCCTCTGTTAAC ACTTCTGTGGGAAGTGTTTC 27400

27401 TCCCTCTAAGAAGATAATTT CTTTTGGGGCTTTTAGAGGC ATGAGTAGGCCAGTTTCTTC TCTGGATTTAACACACTTTC TGTACAATCCCTTTGAGTGC 27500

27501 GTGACAAATGTTTCACCTAA ATTCAAGGCTTTAAGTTTAG CTCCACCAATAATGATAGAG TCAGCACACAAAGCCAAAAA TTTATTTACAAGCTTAAAGA 27600

27601 ATGTCTGAACACTCTCCTTA ATTTCCTTTGCACAGGTGAC AATTTGTCCACCGACAATTT CACAAGCACAGGTTGAGATA AATTTAACAATTTCCCAACC 27700

27701 GTCTCTAAGAAACTCTACAC CTTCCTTAAACTTCTCTTCA AGCCAATCAAGGACGGGTTT GAGTTTTTCATAAACAGTGC CAAAGATGTTAGTTAGCCAC 27800

27801 TGCGAAGTCAACTGAACAAC ACCACCTGTAATGTAGGCCA TTACAACTAGATTGTTAGTA GCCAAATCAGATGTGAACAT CATAGCATCAATGAGTCTCA 27900

27901 GTGAATACTGTGAAATTCCA TCTAGTATTGTTATAGCGGC CTTCTGTAAAACACGCACAG AATTTTGAGCAGTTTCAAGA GTGCGGGAGAAAATTGATCG 28000

28001 TACAACACGAGCAGCCTCTG ATGCAAATGCATAAAGAGGA CTCAGTATTGATTTCTGTTC ACCAATATTCCAGGCACCTT TTTTAGCTTTTCCTTTTGTA 28100

28101 ACTTTAAAATTACCACAGGA TTCAACAATTTGTTTGAATG CTTTATAATCCAAACCTTTA ACAGTTTCCACAAAAGCACT TGTGGAAGCAGAAAAAGATG 28200

28201 CCAAAATAATGGCGATCTCT TCATTAAGTTTAAAGTCACC AACAATATTGATGTTGACTT TCTCTTTTTGGAGTATTTCA AGAAGGTTGTCATTAAGACC 28300

28301 TTCGGAACCTTCTCCAACAA CACCTGTATGGTTACAACCT ATGTTAGCGCTAGCACGTGG AACCCAATAGGCACACTTGT TATGGCAACCAACATAAGAG 28400

28401 AACACACAGCCTCCAAAGGC AATAGTGCGACCACCCTTAC GAAGAATGGTTTTCAAGCCA GATTCATTATGGTATTCGGC AAGACTATGCTCAGGTCCTA 28500

28501 CTTCTGAATTGTAACATGCT GGACAATAAATTTTAACAAC AGCATTTTGGGGTAAGTAAC CACAAGTAGTGGCACCTTCT TTAGTCAAATTCTCAGTGCC 28600

28601 ACAAAATTCGCAAGTGGCTT TAACAAAATCGCCCGTCTGC CATGAAGTTTCACCACAATG ATCACACTTCATGAGAGTTG AAAGGCACATTTGGTTGCAT 28700

28701 TCATTTGGTGACGCAACTGG ATAGACAGATCGAATTCTAC CCATAAAGCCATCAAGCTTT TTCTTTTCAACCCTTGGTTG AATAGTCTTGATTATGGAAT 28800

28801 TTAAGGGAAATACAAAATTT GGACATTCCCCATTGAAGAT GTCAAATTTCTTTGCCAATT TAATTTCAAAAGGTGTCTGC AATTCATAGCTCTTTTCAGA 28900

28901 ACGTTCCGTGTACCAAGCAA TTTCATGCTCATGTTCACGG CAGCAGTATACACCCCTCTT AGTGTCAATAAAGTCCAGTT GTTCGGACAAAGTGCATGAA 29000

29001 GCTTTACCAGCACGTGCTAG AAGGTCTTTAATGCACTCAA GAGGGTAGCCATCAGGGCCA CAGAAGTTGTTATCGACATA GCGAGTGTATGCCCCTCCGT 29100

29101 TAAGCTCACGCATGAGTTCA CGGGTAACACCACTGCTATG TTTAGTGTTCCAGTTTTCTT GAAAATCTTCATAAGGATCA GTGCCAAGCTCGTCGCCTAA 29200

29201 GTCAAATGACTTTAGATCGG CGCCGTAACTATGGCCACCA GCTCCTTTATTACCGTTCTT ACGAAGAAGAACCTTGCGGT AAGCCACTGGTATTTCGCCC 29300

29301 ACATGAGGGACAAGGACACC AAGTGTCTCACCACTACGAC CGTACTGAATGCCTTCGAGT TCTGCTACCAGCTCAACCAT AACATGACCATGAGGTGCAG 29400

29401 TTCGAGCATCCGAACGTTTG ATGAACACATAGGGCTGTTC AAGTTGAGGCAAAACGCCTT TTTCAACTTCTACTAAGCCA CAAGTGCCATCTTTAAGATG 29500

29501 TTGACGTGCCTCTGATAAGA CCTCCTCCACGGAGTCTCCA AAGCCACGTACGAGCACGTC GCGAACCTGTAAAACAGGCA AACTGAGTTGGACGTGTGTT

29600

29601 TTCTCGTTGAAACCAGGGAC AAGGCTCTCCATCTTACCTT TCGGTCACACCCGGACAAAA CCTAGATGTGCTGATGATCG GCTGCAACACGGACGAAACC 29700

29701 GTAAGCAGCCTGCAGAAGAT AGACGAGTTACTCGTGTCCT GTCAACGACAGTAATTAGTT ATTAATTATACTGCGTGAGT GCACTAAGCATGCAGCCGAG 29800

29801 TGACAGCCACACAGATTTTA AAGTTCGTTTAGAGAACAGA TCTACAAGAGATCGAAAGTT GGTTGGTTTATACCTTCCCA GGTAACAAACCAACCAACTT 29900

29901 TCGATCTCTTGTAGATCTGT TCTCT 29925

7.1.12 SRR12596175 Discovered Strain 1 Nucleotide Sequence

1 AGAGAACAGATCTACAAGAG ATCGAAAGTTGGTTGGTTTG TTACCTGGGAAGGTATAAAC CAACCAACTTTCGATCTCTT GTAGATCTGTTCTCTAAACG 100

101 AACTTTAAAATCTGTGTGGC TGTCACTCGGCTGCATGCTT AGTGCACTCACGCAGTATAA TTAATAACTAATTACTGTCG TTGACAGGACACGAGTAACT 200

201 CGTCTATCTTCTGCAGGCTG CTTACGGTTTCGTCCGTGTT GCAGCCGATCATCAGCACAT CTAGGTTTTGTCCGGGTGTG ACCGAAAGTTAAGATGGAGA 300

301 GCCTTGTCCCTGGTTTCAAC GAGAAAACACACGTCCAACT CAGTTTGCCTGTTTTACAGG TTCGCGACGTGCTCGTACGT GGCTTTGGAGACTCCGTGGA 400

401 GGAGGTCTTATCAGAGGCAC GTCAACATCTTAAAGATGGC ACTTGTGGCTTAGTAGAAGT TGAAAAAGGCGTTTTGCCTC AACTTGAACAGCCCTATGTG 500

501 TTCATCAAACGTTCGGATGC TCGAACTGCACCTCATGGTC ATGTTATGGTTGAGCTGGTA GCAGAACTCGAAGGCATTCA GTACGGTCGTAGTGGTGAGA 600

601 CACTTGGTGTCCTTGTCCCT CATGTGGGCGAAATACCAGT GGCTTACCGCAAGGTTCTTC TTCGTAAGAACGGTAATAAA GGAGCTGGTGGCCATAGTTA 700

701 CGGCGCCGATCTAAAGTCAT TTGACTTAGGCGACGAGCTT GGCACTGATCCTTATGAAGA TTTTCAAGAAAACTGGAACA CTAAACATAGCAGTGGTGTT 800

801 ACCCGTGAACTCATGCGTGA GCTTAACGGAGGGGCATACA CTCGCTATGTCGATAACAAC TTCTGTGGCCCTGATGGCTA CCCTCTTGAGTGCATTAAAG 900

901 ACCTTCTAGCACGTGCTGGT AAAGCTTCATGCACTTTGTC CGAACAACTGGACTTTATTG ACACTAAGAGGGGTGTATAC TGCTGCCGTGAACATGAGCA 1000

1001 TGAAATTGCTTGGTACACGG AACGTTCTGAAAAGAGCTAT GAATTGCAGACACCTTTTGA AATTAAATTGGCAAAGAAAT TTGACATCTTCAATGGGGAA 1100

1101 TGTCCAAATTTTGTATTTCC CTTAAATTCCATAATCAAGA CTATTCAACCAAGGGTTGAA AAGAAAAAGCTTGATGGCTT TATGGGTAGAATTCGATCTG 1200

1201 TCTATCCAGTTGCGTCACCA AATGAATGCAACCAAATGTG CCTTTCAACTCTCATGAAGT GTGATCATTGTGGTGAAACT TCATGGCAGACGGGCGATTT 1300

1301 TGTTAAAGCCACTTGCGAAT TTTGTGGCACTGAGAATTTG ACTAAAGAAGGTGCCACTAC TTGTGGTTACTTACCCCAAA ATGCTGTTGTTAAAATTTAT 1400

1401 TGTCCAGCATGTCACAATTC AGAAGTAGGACCTGAGCATA GTCTTGCCGAATACCATAAT GAATCTGGCTTGAAAACCAT TCTTCGTAAGGGTGGTCGCA 1500

1501 CTATTGCCTTTGGAGGCTGT GTGTTCTCTTATGTTGGTTG CCATAACAAGTGTGCCTATT GGGTTCCACGTGCTAGCGCT AACATAGGTTGTAACCATAC 1600

1601 AGGTGTTGTTGGAGAAGGTT CCGAAGGTCTTAATGACAAC CTTCTTGAAATACTCCAACA AGAGAAAGTCAACATCAATA TTGTTGGTGACTTTAAACTT 1700

1701 AATGAAGAGATCGCCATTAT TTTGGCATCTTTTTCTGCTT CCACAAGTGCTTTTGTGGAA ACTGTGAAAGGTTTGGATTA TAAAGCATTCAAACAAATTG 1800

1801 TTGAATCCTGTGGTAATTTT AAAGTTACAAAAGGAAAAGC TAAAAAAGGTGCCTGGAATA TTGGTGAACAGAAATCAATA CTGAGTCCTCTTTATGCATT 1900

1901 TGCATCAGAGGCTGCTCGTG TTGTACGATCAATTTTCTCC CGCACTCTTGAAACTGCTCA AAATTCTGTGCGTGTTTTAC AGAAGGCCGCTATAACAATA 2000

2001 CTAGATGGAATTTCACAGTA TTCACTGAGACTCATTGATG CTATGATGTTCACATCTGAT TTGGCTACTAACAATCTAGT TGTAATGGCCTACATTACAG 2100

2101 GTGGTGTTGTTCAGTTGACT TCGCAGTGGCTAACTAACAT CTTTGGCACTGTTTATGAAA AACTCAAACCCGTCCTTGAT TGGCTTGAAGAGAAGTTTAA 2200

2201 GGAAGGTGTAGAGTTTCTTA GAGACGGTTGGGAAATTGTT AAATTTATCTCAACCTGTGC TTGTGAAATTGTCGGTGGAC AAATTGTCACCTGTGCAAAG 2300

2301 GAAATTAAGGAGAGTGTTCA GACATTCTTTAAGCTTGTAA ATAAATTTTTGGCTTTGTGT GCTGACTCTATCATTATTGG TGGAGCTAAACTTAAAGCCT 2400

2401 TGAATTTAGGTGAAACATTT GTCACGCACTCAAAGGGATT GTACAGAAAGTGTGTTAAAT CCAGAGAAGAAACTGGCCTA CTCATGCCTCTAAAAGCCCC 2500

2501 AAAAGAAATTATCTTCTTAG AGGGAGAAACACTTCCCACA GAAGTGTTAACAGAGGAAGT TGTCTTGAAAACTGGTGATT TACAACCATTAGAACAACCT 2600

2601 ACTAGTGAAGCTGTTGAAGC TCCATTGGTTGGTACACCAG TTTGTATTAACGGGCTTATG TTGCTCGAAATCAAAGACAC AGAAAAGTACTGTGCCCTTG 2700

2701 CACCTAATATGATGGTAACA AACAATACCTTCACACTCAA AGGCGGTGCACCAACAAAGG TTACTTTTGGTGATGACACT GTGATAGAAGTGCAAGGTTA 2800

2801 CAAGAGTGTGAATATCACTT TTGAACTTGATGAAAGGATT GATAAAGTACTTAATGAGAA GTGCTCTGCCTATACAGTTG AACTCGGTACAGAAGTAAAT 2900

2901 GAGTTCGCCTGTGTTGTGGC AGATGCTGTCATAAAAAGTT TGCAACCAGTATCTGAATTA CTTACACCACTGGGCATTGA TTTAGATGAGTGGAGTATGG 3000

3001 CTACATACTACTTATTTGAT GAGTCTGGTGAGTTTAAATT GGCTTCACATATGTATTGTT CTTTTTACCCTCCAGATGAG GATGAAGAAGAAGGTGATTG 3100

3101 TGAAGAAGAAGAGTTTGAGC CATCAACTCAATATGAGTAT GGTACTGAAGATGATTACCA AGGTAAACCTTTGGAATTTG GTGCCACTTCTGCTGCTCTT 3200

3201 CAACCTGAAGAAGAGCAAGA AGAAGATTGGTTAGATGATG ATAGTCAACAAACTGTTGGT CAACAAGACGGCAGTGAGGA CAATCAGACAACTACTATTC 3300

3301 AAACAATTGTTGAGGTTCAA CCTCAATTAGAGATGGAACT TACACCAGTTGTTCAGACTA TTGAAGTGAATAGTTTTAGT GGTTATTTAAAACTTACTGA 3400

3401 CAATGTATACATTAAAAATG CAGACATTGTGGAAGAAGCT AAAAAGGTAAAACCAACAGT GGTTGTTAATGCAGCCAATG TTTACCTTAAACATGGAGGA 3500

3501 GGTGTTGCAGGAGCCTTAAA TAAGGCTACTAACAATGCCA TGCAAGTTGAATCTGATGAT TACATAGCTACTAATGGACC ACTTAAAGTGGGTGGTAGTT 3600

3601 GTGTTTTAAGCGGACACAAT CTTGCTAAACACTGTCTTCA TGTTGTCGGCCCAAATGTTA ACAAAGGTGAAGACATTCAA CTTCTTAAGAGTGCTTATGA 3700

3701 AAATTTTAATCAGCACGAAG TTCTACTTGCACCATTATTA TCAGCTGGTATTTTTGGTGC TGACCCTATACATTCTTTAA GAGTTTGTGTAGATACTGTT 3800

3801 CGCACAAATGTCTACTTAGC TGTCTTTGATAAAAATCTCT ATGACAAACTTGTTTCAAGC TTTTTGGAAATGAAGAGTGA AAAGCAAGTTGAACAAAAGA 3900

3901 TCGCTGAGATTCCTAAAGAG GAAGTTAAGCCATTTATAAC TGAAAGTAAACCTTCAGTTG AACAGAGAAAACAAGATGAT AAGAAAATCAAAGCTTGTGT 4000

4001 TGAAGAAGTTACAACAACTC TGGAAGAAACTAAGTTCCTC ACAGAAAACTTGTTACTTTA TATTGACATTAATGGCAATC TTCATCCAGATTCTGCCACT 4100

4101 CTTGTTAGTGACATTGACAT CACTTTCTTAAAGAAAGATG CTCCATATATAGTGGGTGAT GTTGTTCAAGAGGGTGTTTT AACTGCTGTGGTTATACCTA 4200

4201 CTAAAAAGGCTGGTGGCACT ACTGAAATGCTAGCGAAAGC TTTGAGAAAAGTGCCAACAG ACAATTATATAACCACTTAC CCGGGTCAGGGTTTAAATGG 4300

4301 TTACACTGTAGAGGAGGCAA AGACAGTGCTTAAAAAGTGT AAAAGTGCCTTTTACATTCT ACCATCTATTATCTCTAATG AGAAGCAAGAAATTCTTGGA 4400

4401 ACTGTTTCTTGGAATTTGCG AGAAATGCTTGCACATGCAG AAGAAACACGCAAATTAATG CCTGTCTGTGTGGAAACTAA AGCCATAGTTTCAACTATAC 4500

4501 AGCGTAAATATAAGGGTATT AAAATACAAGAGGGTGTGGT TGATTATGGTGCTAGATTTT ACTTTTACACCAGTAAAACA ACTGTAGCGTCACTTATCAA 4600

4601 CACACTTAACGATCTAAATG AAACTCTTGTTACAATGCCA CTTGGCTATGTAACACATGG CTTAAATTTGGAAGAAGCTG CTCGGTATATGAGATCTCTC 4700

4701 AAAGTGCCAGCTACAGTTTC TGTTTCTTCACCTGATGCTG TTACAGCGTATAATGGTTAT CTTACTTCTTCTTCTAAAAC ACCTGAAGAACATTTTATTG 4800

4801 AAACCATCTCACTTGCTGGT TCCTATAAAGATTGGTCCTA TTCTGGACAATCTACACAAC TAGGTATAGAATTTCTTAAG AGAGGTGATAAAAGTGTATA 4900

4901 TTACACTAGTAATCCTACCA CATTCCACCTAGATGGTGAA GTTATCACCTTTGACAATCT TAAGACACTTCTTTCTTTGA GAGAAGTGAGGACTATTAAG 5000

5001 GTGTTTACAACAGTAGACAA CATTAACCTCCACACGCAAG TTGTGGACATGTCAATGACA TATGGACAACAGTTTGGTCC AACTTATTTGGATGGAGCTG 5100

5101 ATGTTACTAAAATAAAACCT CATAATTCACATGAAGGTAA AACATTTTATGTTTTACCTA ATGATGACACTCTACGTGTT GAGGCTTTTGAGTACTACCA 5200

5201 CACAACTGATCCTAGTTTTC TGGGTAGGTACATGTCAGCA TTAAATCACACTAAAAAGTG GAAATACCCACAAGTTAATG GTTTAACTTCTATTAAATGG 5300

5301 GCAGATAACAACTGTTATCT TGCCACTGCATTGTTAACAC TCCAACAAATAGAGTTGAAG TTTAATCCACCTGCTCTACA AGATGCTTATTACAGAGCAA 5400

5401 GGGCTGGTGAAGCTGCTAAC TTTTGTGCACTTATCTTAGC CTACTGTAATAAGACAGTAG GTGAGTTAGGTGATGTTAGA GAAACAATGAGTTACTTGTT 5500

5501 TCAACATGCCAATTTAGATT CTTGCAAAAGAGTCTTGAAC GTGGTGTGTAAAACTTGTGG ACAACAGCAGACAACCCTTA AGGGTGTAGAAGCTGTTATG 5600

5601 TACATGGGCATACTTTCTTA TGAACAATTTAAGAAAGGTG TTCAGATACCTTGTACGTGT GGTAAACAAGCTACAAAATA TCTAGTACAACAGGAGTCAC 5700

5701 CTTTTGTTATGATGTCAGCA CCACCTGCTCAGTATGAACT TAAGCATGGTACATTTACTT GTGCTAGTGAGTACACTGGT AATTACCAGTGTGGTCACTA 5800

5801 TAAACATATAACTTCTAAAG AAACTTTGTATTGCATAGAC GGTGCTTTACTTACAAAGTC CTCAGAATACAAAGGTCCTA TTACGGATGTTTTCTACAAA 5900

5901 GAAAACAGTTACACAACAAC CATAAAACCAGTTACTTATA AATTGGATGGTGTTGTTTGT ACAGAAATTGACCCTAAGTT GGACAATTATTATAAGAAAG 6000

6001 ACAATTCTTATTTCACAGAG CAACCAATTGATCTTGTACC AAACCAACCATATCCAAACG CAAGCTTCGATAATTTTAAG TTTGTATGTGATAATATCAA 6100

6101 ATTTGCTGATGATTTAAACC AGTTAACTGGTTATAAGAAA CCTGCTTCAAGAGAGCTTAA AGTTACATTTTTCCCTGACT TAAATGGTGATGTGGTGGCT 6200

6201 ATTGATTATAAACACTACAC ACCCTCTTTTAAGAAAGGAG CTAAATTGTTACATAAACCT ATTGTTTGGCATGTTAACAA TGCAACTAATAAAGCCACGT 6300

6301 ATAAACCAAATACCTGGTGT ATACGTTGTCTTTGGAGCAC AAAACCAGTTGAAACATCAA ATTCGTTTGATGTACTGAAG TCAGAGGACGCGCAGGGAAT 6400

6401 GGATAATCTTGCCTGCGAAG ATCTAAAACCAGTCTCTGAA GAAGTAGTGGAAAATCCTAC CATACAGAAAGACGTTCTTG AGTGTAATGTGAAAACTACC 6500

6501 GAAGTTGTAGGAGACATTAT ACTTAAACCAGCAAATAATA GTTTAAAAATTACAGAAGAG GTTGGCCACACAGATCTAAT GGCTGCTTATGTAGACAATT 6600

6601 CTAGTCTTACTATTAAGAAA CCTAATGAATTATCTAGAGT ATTAGGTTTGAAAACCCTTG CTACTCATGGTTTAGCTGCT GTTAATAGTGTCCCTTGGGA 6700

6701 TACTATAGCTAATTATGCTA AGCCTTTTTCTTAACAAAGT TGTTAGTACAACTACTAACA TAGTTACACGGTGTTTAAAC CGTGTTTGTACTAATTATAT 6800

6801 GCCTTATTTCTTTACTTTAT TGCTACAATAGTGTACTTTT ACTAGAAGTACAAATTCTAG AATTAAAGCATCTATGCCGA CTACTATAGCAAAGAATACT 6900

6901 GTTAAGAGTGTCGGTAAATT TTGTCTAGAGGCTCGATTTA ATTATTTGAAGTCACCTAAT TTTTCTAAACTGATAAATAT TATAATTTGGTTTTTACTAT 7000

7001 TAAGTGTTTGCCTAGGTTCT TTAATCTACTCAACCGCTGC TTTAGTAGTTTTAATGTCTA ATTTAGGCATGCCTTCTTAC TGTACTGGTTACAGAGAAGG 7100

7101 CTATTTGAACTCTACTAATG TCACTATTGCAACTTACTGT ACTGGTTCTATACCTTGTAG TGTTTGTCTATGTGGTTTAG ATTCTTTAGACACTAACACT 7200

7201 TCTTTAGAAACTATACAAAT TACCATTTCATCTTTTAAAT GGGATTTAACTGCTTTTGGC TTAGTTGCAGAGTGGTTTTT GGCATATATTCTTTTCACTA 7300

7301 GGTTTTTCTATGTACTTGGA TTGGCTGCAATCATGCAATT GTTTTTCAGCTTCTTTGCAG TACATTTTATTAGTAATTCT TGGCTTATGTGGTTAATAAT 7400

7401 TAATCTTGTCAAAATGGCCC CGATTTCAGCTATGGTTAGA ATGTACATCTTCTTTGCATC ATTTTATTATGTATGGAAAA GTTATGTGCATGTTGTAGAC 7500

7501 GGTTGTAATCAATCAACTTG TATGATGTGTTACAACCGTA ATAGAGCAACAAGAGTCGAA TGTACAACTATTGTTAATGG TGTTAGAAGGTCCTTTTATG 7600

7601 TCTATGCTAATGGAGGATAA GGCTTTTGCAAACTACACAA TTGGAATTGTGTTAATTGTG ATACATTCTGTGCTGGTAGT ACATTTATTAGTGATGAAGT 7700

7701 TGCGAGAGACTTGTCACTAC GTTTTAAAAGACCAATAAAT CCTACTGACAAGTCTTCTTA CATCGTTGATAGTGTTACAG TGAAGAATGGTTCCATCCAT 7800

7801 CTTTACTTTGATAAAGCTGG TCAAAAGACTTATGAAAGAC ATTCTCTCTCTCATTTTGTT AACTTAGACAACCTGAGAGC TAATAACACTAAAGGTTCTA 7900

7901 TGCCTATTAATGTTATAGTT TTTGATGGTAATCAAAAATG TGAAGAATCATCTGCAAAAT CAGCGTCTGTTTACTACAGT CAGCTTATGTGTCAACCTAT 8000

8001 ATAGTTACTAGATCAGGCAT TAGTGTCTGATGTTGGTGAT AGTGCGGAAGTTGCAGTTAA AATGTTTGATGCTTACGTTA ATACGTTTTCATCAACTTTT 8100

8101 AACGTACCAATAGAAAAACT CAAAACACTAGTTGCAACTG CAGAAGCTGAACTTGCAAAG AATGTGTCCTTAGACAATGT CTTATCTATGTTTATTTCAG 8200

8201 CAGCTCGGCAAGGGTTTGTT GATTCAGATGTAGAAACTAA AGATGTTGTTGAATGTCTTA AATTGCTACATCAATCTGAC ATAGAAGTTACTGGCGATAG 8300

8301 TTGTAATAACTATATGCTCA CCTATAACAAAGTTGAAAAC ATGACACCCCGTGACCTTGG TGCTTGTTTTGACTGTAGTG CGCGTCATATTAATGCGCAG 8400

8401 GTAGCAAAAAGTCACAACAT TGCTTTGATATGGAACGTTA AAGATTTCATTCCATTGTCT GAACAACTACGAAAACAAAT ACGTAGTGCTGCTAAAAAGA 8500

8501 ATAACTTACCTTTTAAGTTG ACATGTGCAACTACTAGACA AGTTGTTAATGTTGTAAATA CAAAGATAGCACTTAAGGGT GGTAAAATTGTTAATAATTG 8600

8601 GTTGAAGCAGTTAATTAAAG TTACACTTGTGTTCTTTTTT GTTGCTGCTATTTTCTATTT AATAAATCCTGTTCATGTCA TGTCTAAACATACTGACTTT 8700

8701 TCAAGTGAAATCATAGGATA CAAGGCTATTGATGGTGGTG TCACTCGTGACATAGCATCT ACAGATACTAGTTTTGCTAA CAAACATGCTGATTTTGACA 8800

8801 CATGGTTTAGCCAGCGTGGT GGTAGTTATACTAATGACAA AGCTTGCCCATTGATTGCTG CAGTCATAACAAGAGAAGTG GTTTTTGTCGTGCCTGGTTT 8900

8901 GCCTGGCACGATATTACGCA CAACTAATGGTACCTTTTTG CATTTCTTACCTAGAGTTTT TAGTGCAGTTGGTAACATCT GTTACACACCATCAAAACTA 9000

9001 TTAGAGTACACTGACTTTGC AACATCAGCTTGTGTTTTGG CTGCTGAATGTACAATTTTT AAAGATGCTTCTGGTAAGCC AGTACATTATTGTTATGATA 9100

9101 CCAATGTACTAAAAGGTTCT GTTGCTTATGAAAGTTTACG CCCTGACACACGTTATGTGC TCATGGATGGCTCTATTATT CAATTTCCTACAACTAACCT 9200

9201 TGAAGGTTCTGTAAGAGGCG TAACAACTTTTGATTCTGAG TACTGTAGGCACGGCACTTG TGAAAGATCAGAAGCTGGTG TTTGTGTATCTACTAGTGGT 9300

9301 AGATGGGTACTTAACAATGA TTATTACAGATCTTTACCGT GAGTTTTCTGTGGTGTAGAT GCTGTAAATTTACTTACTAA TATGTTTACACCACTAATTC 9400

9401 AACCTATTGGTGCTTTGGAC ATATCAGCATTAATAGTAGC TGGTGGTATTGTAGCTATCG TAGTAACATGCCTTGCCTAC TATTTTATGAGGTTTAGAAG 9500

9501 AGCTTTTGGTGAATACAGTC ATGTGCTTGCCTTTAATACT TTACTATTCCTTATGTCATT CACTTCACTCTGTTTAACAC CAGTTTACTCATTCTTACCT 9600

9601 GGTGTTTATTCTGTTATTTA CTTGTACTTGACATTTTATC TTACATATGATGTTTCTTTT TTAGCACATATTCAGTGGAT GGTTATGTTCACACCTTTAG 9700

9701 TACCTTTCTGGATAACAATT GCTTATATCATTTGTATTTC ATCATAGCATTTCTATTGGT TCTTTAGTAATTACCTAAAG AGACGTGTAGTCTTTAATGG 9800

9801 TGTTCCCTTTAGTACTTTTG AAGAAGCTGCGCTGTGCACC TTTTTGTTAAATAAAGAAGT GTATCTAAAGTTGCGTAGTG ATGTGCTATTACCTCTTACG 9900

9901 CAATATAATAGATACTTAGC TCTTTATAATAAGTACAAGT ATTGTAGTGGAGCAATGGAT ACAACTAGCTACAGAGAAGC TGCTTGTTGTCATCTCGCAA 10000

10001 AGGCTCTCAATGACTTCAGT AACTCAGGTTCTGATGTTCT TTACCAACCACCACAAACCT CTATCACCTCAGCTGTTTTG CAGAGGTGTTTTAGAAAAAT 10100

10101 GGCATTCCCATCTGGTAAAG TTGAGGGTTGTATGGTACAA GTAACTTGTGGTACAACTAC ACTTAACGGTCTTTGGCTTG TAGACGTAGTTTACTGTCCA 10200

10201 AGACATGTGATCTGCACCTC TAGAGACATGCTTAACCCTA ATTATGAAGATTTACTCATT CGTAAGTCTAATCATAATTT CTTGGTACAGGCTGGTAATG 10300

10301 TTCAACTCAGGGTTATTGGA CATTCTATGCAAAATTGTGT ACTTAAGTTTAAGGTTGATA CAGCCAATCCTAAGACACCT AAGTATAAGTTTGTTCGCAT 10400

10401 TCAACCAGGACAGACTTTTT CAGTGTTAGCTTGATACAAT GGTTCACCATCTGGTGTTCA CCAATGTGCTATGAGGCCCA ATTTCACTATTAAGGGTTCA 10500

10501 TTCCTTAATGGTTCTGGTGG TAGTGTTGGTTTTAACATAG ATTATGACTGTGTCTCTTTT TGTTACATGCACCATATGGA ATTACCAACTGGAGTTCATG 10600

10601 CTGGCACAGACTTAGAAGGT AACTTTTAGTGACCTTTTGT TGACAGGCAAACAGCACAAG CAGCTGGTACGGACACAACT ATTACGATTAATGTTTTAGC 10700

10701 TTGGTTGTACGCTGCTTCTA TAAATGGAGACAGGTGGTTT CTCAATCGATTTACCACAAC TCTTAATGACTTTAACCTTG TGGCTATGAAGTACAATTAT 10800

10801 GAACCTCTAACACAAGACCA TGTTAGCATACTAGGACCTC TTTCTGCTCAAACTGGAATT GCCGTTTTAGATATGTGTGC TTCATTAAAAGAATTACTGC 10900

10901 AAAATGGTATGAATGGACGT ACCATTTTGGGTAGTGCTTT ATTAGAAGATGAATTTACAC CTTTTGATGTTGTTAGACAA TGCTCAGGTGTTACTTTCCA 11000

11001 AAGTGCAGTGAAAAGAACAA TCAAGGGTCGACACCACTGG TTGTTACTCACAATTTTGAC TTCACTTTTAGTTTTAGTCC AGAGTACTCAATGGTCTTTG 11100

11101 TTCTTTTTTTTTGTATGAAAA TGCCTATATACCTTTTGCTA TGGTAATTATTGCTATGTCT GCTTTTGCAATGATGTTTGT CAAACATAAGCATGCATTTC 11200

11201 TCTGTTTGTTTTTGTTACCT TCTTATGCCACTGTAGCTTA TTTTAATATGGTCTATATGC CTGCTAGTAGGGTGATGCGT ATTATGACATGGTTGGATAT 11300

11301 GGTTGATACTAGTTTGTCTG GTTTTAAGCTAAAAGACTGT GTTATGTATGCATCAGCTGT AGTGTTACTAATCCTTATGA CAGCAAGAACTGTGTATGAT 11400

11401 GATGGTGCTAGGAAGGTGTG GACACTTATGAATGTCTTGA CACTCGTTTATAAAGTTTTA TATGGTAATGCTTTAGATCA AGCCATTTCCTCGTGGGCTC 11500

11501 TTATAATCTCTGTTACTTCT AACTACTCAGGTGTAGTTAC AACGATCATGTTTTTGGCCA GAGGTATTGTTTTTATGTGT GTTGAGTATTGCCCTATTTT 11600

11601 CTTCATAACTGGTAATACAC TTCATAGTATAATGTAAGTT TATTGTTTCTTAGGCTATTT TTGTACTTGTTACTTTGGCC TCTTTTGTTTACTCAACCGC 11700

11701 TATCTTAGACTGACTCTTGG TGTTTATGATTACTTAGTTT CTACACAGGAGTTTAGATAT ATGAATTCACAGGGACTACT CCCACCCAAGAATAGCATAG 11800

11801 ATGCCTTCAAACTCAACATT AAATTGTTGGGTGTTGGTGG AAAACCTTGTATCAAAGTAG CCACTGTACAGTCTAAAATG TCAGATGTAAAGTGCACATC 11900

11901 AGTAGTCTTACTCTCAGTTT TGCAACAACTCAGAGTAGAA TCATCATCTAAATTGTGGGC TCAATGTGTCCAGTTACACA ATGACATTCTCTTAGCTAAA 12000

12001 GATACTACTGAAGCCTTTGA AAAAATGGTTTCACTACTTT CTGTCTTGCTTTCCATGCAG GGTGCTGTAGACATAAACAA GCTTTGTGAAGAAATGCTGG 12100

12101 ACAACAGGGCAACCTTACAA GCTATAGCCTCAGAGTTTAG TTCCCTTCCATCATATGCAG CTTTTGCTGCTGCTCAAGAA GCTTATGAGCAGGCTGTTGC 12200

12201 TAATGGTGATTCTGAAGTTG TTCTTAAAATGTTGAAGAAG TCTTTGAATGTGGCTAAATC TGAATTTGACCGTGATGCAG CCATGCAACGTAAGTTGGAA 12300

12301 AAGATGGCTACTCAAGCTAT GACCCAAATGTATAAACAGG CTAGATCTGAGGACAAGAGG GCAAAAGTTACTAGTGCTAT GCAATCAATGCTTTTCACTA 12400

12401 TGCTTAGAAAGTTGGATAAT GATGCACTCAACAACATTAT CAACAATGCAAGAGATGTAT GTGTTCCCTTGAACATAATA CCTCTTACAACAGCAGCCAA 12500

12501 ACTAATGGTTGTCATACCAG ACTATAACACATATAAAAAT ACGTGTGATGTCACAACATT TACTTATGCATCAGCATTGT GGGAAATCCAACAGGTTGTA 12600

12601 GATGCAGATAGTAAAATTGT TCAACTTAGGAAAATTAGTA TGGACAATTCACCTAATTTA GCATGGCCTCTTATTGTAAC AGCTTTAAGGGCCAATTCTG 12700

12701 CTGTCAAATTACAGAATAAT GAGCTTAGCACTGTTGCACT ACGACAGATGTCTTGTGCTG CCGGTACTACACAAACTGCT TGCACTGATGACAATGCGTT 12800

12801 AGCTTACTACAACACAACAA AGGGAGTTAGGTTTGTACTT GCACTGTTATCCGATTTACA GGATTTGAAATGGGCTAGAT TCCCTAAGAGTGATGGACCT 12900

12901 GGTACTATCTATACAGAACT GGAACCACCTTGTAGGTTTG TTACAGACACACCTAAAGGT CCTAAAGTGAAGTATTTATA CTTTATAAAAGGATTAAACA 13000

13001 ACCTAAATAGAGGTATGGTA CTTGGTAGTTTAGCTGCCAC AGTACGTCTACAAGCTGGTA ATGCAACAGAAGTGCCTGCT AATTCAACTGTATTATCTTT 13100

13101 CTGTGCTTTTGCTGTAGATG CTGCTAAAGCTTACAAAGAT TATCTAGCTAGTGGGGGACA ACCAATCACTAATTGTGTTA AACTGTTGTGTACACACACT 13200

13201 GGTACTGGTCAGGCATATAC AGTTACACCGGAAGCCAATA TGGATCAAGAATCCTTTGGT GGTGCATCGTGTTGTCTGTA CTGCCGTTGCCACATGTATC 13300

13301 ATCCAAATCCTAAAGGATTT TGTGACTTAAAAAGGTAAGT ATGTACAAATACCTACAACT TGTGCTAATGACCCTGTGGG TTTTACACTTAAAAACACAG 13400

13401 TCTGTACTGACTGCGGTATG TGGAAAGGTTATGGCTGTAG TTGTGATCAACTCCGCGAAC CCATGCTTCAGTCAGCTGAT GCACAATCGTTTTTAAACGG 13500

13501 GTTTGCGGTGTAAGTGCAGC CTGCCTTACACCGTGCGGCA CAGGCACTAGTACTGATGTC GTATACAGGGCTTTTGACAT CTACAATGATAAAATTGCTG 13600

13601 GTTTTGCTAAATTCCTAAAA ACTAATTGTTGTCGCTTCCA AGTAAAGGACGAAGATGACA ATTTAATTGATTCTTACTTT GTAGTTAAGAGACACACTTT 13700

13701 CTCTAACTACCAACATGAAG AAACAATTTATAATTTACTT AAGTAATGTCCAGCTGTTGC TAAACATGACTTCTTTAAGT TTAGAATAGACGGTGACATG 13800

13801 GTACCATACATATCACGTCA ACGTCTTACTAAATACACAA TGGCAGACCTCGTCTATGCT TTAAGGCATTTTGATGAAGG TAATTGTGACACATTAAAAG 13900

13901 AATTACTTGTCACATACAAT TGTTGTGATGATGATTATTT CAATAAAAAGGACTGGTATG ATTTTGTAGAAAGCCCAGAT ATATTACGCGTATACGCCAA 14000

14001 CTTAGGTGAACGTGTACGCC AAGCTTTGTTAAAAACAGTA AATTTCTGTGATGCCATGCA ATATGCTGGTATTGTTGGTA TTCTGACATTAGAATTTCAA 14100

14101 GATCTCAATGGTAACTGGTA TGATTTCGGTGATTTCATAC AAACCACGCCAGGTAGTGGA GTTCCTGTTGTAGATTCTTA TTATACTTTGTTAATGCCTA 14200

14201 TATTAACCTTGAACCGGGCT TTAACTGCAGAGTCACATGT TGACACTGACTTAACAAAGC CTTACATTAAGTGGGATTAG TTAAAATATGACTTCACGGA 14300

14301 AGAGAGGTTAAAACTCTTTG ACCGTTATTTTAAATATTGG GATCAGACATACCACCCAAA TTGTTTTAACTGTTTGGATG ACAGATGCATTCTGCATTGT 14400

14401 GCAAACTTTAATGTTTTATT CTCTACAGTGTTCCCACTTA CAAGTTTTGGACCACTAGTG AGAAAAATATTTGTTGATTG CGTTCCATTTGTAGTTTCAA 14500

14501 CTGGATACCACTTCAGAGAG CTAGGTGTTGTACATAATCA GGATGTAAACTTACATAGCT CTAGACTTAGTTTTAGGTAA TTACTTGTGTATGCTGCTGA 14600

14601 CCCTGCTATGCACGCTGCTT CTGGTAATCTATTACTAGAT AAACGCACTACGTGCTTTTC AGTAGCTGCACTTACTAACA GTCTTGCTTTTCAAACTGTC 14700

14701 AAACCCGGTAATTTTAACAA AGACTTCTATGACTTTGCTG TGTCTAAGGGTTTCTTTAAG GAAGGAAGTTCTGTTGAATT AAAACACTTCTTCTTTGCTC 14800

14801 AGGATGGTAATGCTGCTATC AGCGATTATGACTATTATCG TTATAATCTACCAACAATGT GTGATATCAGACAACTACTA TTTGTAGTTGAAGTTGTTGA 14900

14901 TAAGTTCCTTGATTGTTACG ATGGTGGCTGTATTAATGCT AACCAAGTCATCGTCAACAA CCTAGACAAATCAGCTGGTT TTCCATTTAATAAAGGTGGT 15000

15001 AAGGCTAGACTTTATTATGA TTCAATGAGTTATGGGTATC AAGATGCACTTTTCGCATAT ACAAAACGTAATGTCATCCC TAC-ATAACTCAAATGAATC 15100

15101 TTAAGTATGCCATTTGCGCA AAGAATAGAGCTCGCACCGT AGCTGGTGTCTCTATCTGTA GTTCAATGACCAATAGACAT TATCCTAAAAAAATTTTGAA 15200

15201 ATCAATAGCCGCCACTAGAG GAGCTACTGTAGTAATTGGA ACAAGCAAATTCTATGGTGG TTGGCACAACATGTTAAAAA CTGTTTATAGTGATGTAGAA 15300

15301 AACCCTCACCTTATGGGTTG GGATTATTCCAAATGTGATA GAGCCATGCCTAACATGCTT AGAATTATGGCCTCACTTGT TCTTGCTCGCAAACATACAA 15400

15401 CGTGTTGTAGCTTGTCACAG CGTTTCTATAGATTAGCTAA TGAGTGTGCTCAAGTTTAGA GTGAAATGGTTAAGTGTGGC GGTTCACTATATGTTAAACC 15500

15501 AGGTGGAACCTCATCAGGAG ACGACACAACTGCTTATGCT AATAGTGTTTTTAACATTTG TCAAGCTGTCACGGCCAATG TTAATGCACTTTTATCTACT 15600

15601 GATGGTAACAAAATTGCCGA TAAGTATGTCCGCAATTTAC AACACACATTTTATGAGTGT CTCTATAGAAATAGAGATGT TGACACAGACTTTGTGAATG 15700

15701 AGTTTTACGCATATTTGCGT AAACATTTCTCAATGATGAT ACTCTCTGACGATGCTGTTG TGTGTTTCAATAGCACTTAT GCATCTCAAGGTCTAGGGGC 15800

15801 TAGCATAAAGAACTTTAAGT CAGTTCTTTATTATCAAAAC AATGTTTTTATGTCTGAAGC AAAATGTTGGACTGAGACTG ACCTTACTAAAGGACCTCGT 15900

15901 TAATTTTGCTCTCAACATAC AATGCTAGTTAAACAGGGTG ATGATTATGTGTCCTTTCCT TACCCAGATCCATCAAGAAT CCTAGGGGCCGGCTGTTTTG 16000

16001 TAGATGATATCGTAAAAACA GATGGTACATTAATGATTGA ACGGTTCGTGTCTTTAGCTA TAGATGCTTACCCACTTACT AAACATCCTAATCAGGAGTA 16100

16101 TGCTGATGTCTTTCATTTGT ACTTACAATACATAAGAAAG CTACATGATGAGTAATCAGG ACACATGTTAGACATGTATT CTGTTATGCTTACTAATGAT 16200

16201 AACACTTCAAGGTATTGAGT ACCTGAGTTTTATGAGGCTA TGTACACACCGCATACAGTC TTACAGGCTGTTGGGGCTTG TGTTCTTTGTAATTCACAGA 16300

16301 CTTCATTAAGATGTGGTGCT TGCATACGTAGACCATTCTT ATGTTGTAAATGGTTTTACG ACCATGTCATATCAACATCA CATAAATTACTATTGTCTGT 16400

16401 TAATCCGTATGTTTGCAATG CTCCAGGTTGTGATGTCACA GATGTGACTCAACTTTACTT AGGAGGTATGAGCTATTATT GTAAATCACATAAAACTCCG 16500

16501 ATTAGTTTTCCATTGTGTTC TAATGGACAAGTTTTTGGTT TATATAAAAATACATGTGTT GGTAGCGATAATGTTACTGA CTTTAATGCAATTTCAACAT 16600

16601 GTGACTGGACAAATGCTGGT GATTACATTTTAGCTAACAC CTGTACTGAACGACTCAAGC TTTTTGCAGCAGAAACGCTC AAAGCTAC-GAGGAGACATT 16700

16701 TAAACTGTCTTATGGTATTG CTACTGTACGTGAAGTTCAG TCCGACAGAGAATTACATCT TTCATGGGAAGTTGGTAAAC CTAGACCACCACTTAACCGA 16800

16801 AATTATGTCTTTACTGGTTA TCGTGTAACTAAAAACAGTA AAGTAAATATAGGAGAGTAC ACCTTTGAAAAAGGTGACTA TGGTGATTCCGTTGTTTACC 16900

16901 GAGGTACAACAACTTACAAA TTAAATGTGGTTGATTATTT TGTGCTGACATCACATACAG TAATGCCATTAAGTGCACCT ACACTAGTGCCACAAGAGCA 17000

17001 CTATGTTAGAATTTCAGGCT TATACCCAACACTCAATATC TCAGATGAGTTTTCTAGCAA TGTTGCAAATTATCAAAAGG TGGTTATGCAAAAGTATTCT 17100

17101 ACACTCCAGGGACCACCTGG TACTGGTAAGAGTCATTTTG CTATTGGCCTAGCTCTCAAC TACCCTTCTGCTCGCATAGT GTATACAGCTTGCTCTCATG 17200

17201 CCGCTGTTGATGCACTATGT GAGAAGGCATTAAAATATTT GTCAATAGATAAATGTAGTA GAATTATACCTGCACGTGCT CGTGTAGAGTGTTTAGATAA 17300

17301 ATTCAATGAGAATTCAACAT TAGAACAGTATGTCTTTTGT ACTGTAAATGCATTGCCTGA GACGACAGCAGATATAGTTG TCTTTGATGAATTATCAATG 17400

17401 GCCACAAATTATGATTTGAG TGTTGTCAATGCCAGATTAC GTGCTAAGCACTATGTGTAC ATTGGCGACCCTGCTCAATT ACCTGCACCACGCACATTGC 17500

17501 TAACTAAGGGCACACTAGAA CCAGAATATTTCAATTCAGT GTGCAGACTTATGAAAACTA TAGGTCCAGACATGTTCCTC GGAACTTGTCGGCGTTGTCC 17600

17601 TGCTGAAATTGTTGACACTG TGAGTGCTTTGTTCTATGAT AATAAGCTTAAAGCACATAA AGACAAATCAGCTCAATGCT TTAAAATGTTTTAAAAGTGA 17700

17701 TTAATCACGCATGATGTTTC ATCTGCAATTAACAGGCCAC AAATAGGCGTGGTAAGAGAA TTCCTTACACGTAACCCTGC TTGGAGAAAAGCTGTCTTTA 17800

17801 TTTCACCTTATAATTCACAG AATTCAGTAGCCTCAAAGAT TTTGGGACTACCAACTCAAA CTGTTGATTCATCACAGGGC TCAGAATATGACTATGACAT 17900

17901 ATTCACTCAAACCACTGAAA CAGCTCACTCTTGTAATGTA AACAGATTTAATGTTGCTAT TACCAGAGCAAAAGTAGGCA TACTTTGTACAATGTCTGAT 18000

18001 AGAGACCTTTATGACAAGTT GCAATTTACAAGTCTTGAAA TTCCACGTAGGAATGTGGCA ACTTTACAAGCTGAAAATGT AACAGGACTCTTTAAAGATT 18100

18101 GTAGTAAGGTAATCACTGGG TTACATCCTACACAGGCACC TACACCCCTCAGTGTTGACA CTAAATTCAAAACTGAAGGT TTATGTGTTGACATACCTGG 18200

18201 CATACCTAAGGACATGACCT ATAGAAGACTCATCTCTATG ATGGGTTTTAAAATGAATTC TAAAGTTAATGGTTACCCTA ACATGTTTATCACCCGCGAA 18300

18301 GAAGCTATAAGACATGTACG TGCATGGATTGGCTTCGATC TTGAGGGGTGTCATGCTACT AGAGAAGCTGTTGGTACCAA TTTACCTTTACAGCTAGGTT 18400

18401 TTTCTACAGGTGTTAACCTA GTTGCTGTACCTACAGGTTA TGTTGATACACCTGATAATA CAGATTTTTCCAGAGTTAGT GCTAAACCACCGCCTGGAGA 18500

18501 TCAATTTAAACACCTCATAC CACTTATGTACAAAGGACTT CCTTGGTACGTAGTGCGTAT AAAGATTGTACAAATGTTAA GTGACACACTTAAAAATCTC 18600

18601 TCTGACAGAGTCATGTGTCT GTTATGGGCACATGGCTTTG AGTTGACATCTATGAAGTAT TTTGTGAAAATAGGACCTGA GCGCACCTGTTGTCTATGTG 18700

18701 ATAGACGTGCCACATTCCTT TCCACTGCTTCAGACACTTA TGCCTGTTGGCATCATTCTA TTGGATTTGATTACGTCTAT AATCCGTTTATGATTGATGT 18800

18801 TCAAAAATGGGGTTTTACAG GTAACCTACAAAGCAACCAT GATCTGTATTGTCAAGTCCA TGGTAATGCACATGTAGCTA GTTGTGATGTAATCATGACT 18900

18901 AGGTGTCTAGCTGTCCACGA GTGCTTTGTTAAGCGTGTTG ACTGGACTATTGAATATCCT ATAATTGGAGTTGAACTGAA GATTAATGCGGCTTGTAGAA 19000

19001 AGGTTCAACACATGGTTGTA ATAGCTGCATTATTAGCAGA CAAATTCCCAGTTCTTCACG ACATTGGTAACCCTAAAGCT ATTAAGTGTGTACCTCAAGC 19100

19101 TGATGTAGAATAGTAGTTCT ATGATGCACAGCCTTGTAGT GACAAAGCTTATAAAATAGA AGAATTATTCTATTCTTATG CCACACATTCTGACAAATTC 19200

19201 ACAGATGGTGTATGTCAATT TTGGAATTGCAATGTCGATA GATATCCTGCTAATTCCATT GTTAGTAGATTTGACACTAG AGTGCTATCTAACCTTAACT 19300

19301 TGCCTGGTTGTGATGGTGGC AGTTTGTATGTAAATAAACA TGCATTCCACACACCAGCTT GTAATAAAAGTGCTTTTGTT AATTTAAAACAATTACCATT 19400

19401 TTTCTATTACTCTGACAGTC CATGTGAGTCTCATGGAAAA CAAGTAGTGTCAGATATAGA TTATGTACCACTAAGGTCTG CTACGTGTATAACACGTTGC 19500

19501 AATTTAGGTGGTGCTGTCTG TAGACATCATGCTAATGAGT ACAGATTGTATCTCGATGCT TATAACATGATGATCTCAGC TGGTTATAGCTTGTGGGTTT 19600

19601 ACAAACAATTTGATACTTAT AACCTCTAGGACACTTTTAC AAGACTTCAGAGTTTAGAAA ATGTGGCTTTTAATGTTGTA AATAAGGGACACTTTGATGG 19700

19701 ACAACGGTGTGAAGTACCAG TTTCTATCATTAATAACACT GTTTACACAAAATTAGATGG TGTTGATGTAGAATTGTGTA AAAATAAAACAACAATTCCT 19800

19801 GTTAATGTAGCATTTGAGCT TTGGGCTAAGCGCAACATTA AACCAGTACCAGAGGTGAAA ATACTCAAAAATTTGGGTGT GGACATTGCTGCTAATACTG 19900

19901 TGATCTGGGACTACAAAAGA GATGCTCCAGCACATATATC TACTATTGGTGTTAGTTCTA TGACTGACATAGCCAAGAAA CCAACTGAAACGATTTGTGC 20000

20001 ACCACTCACTGTTTTTTTTG ATGGTAGAGTTGATGGTCAA GTAGACTTATTTAGAAATGC CCGTAATGGTGTTCTTATTA CAGAAGGTAGTGTTAAAGGT 20100

20101 TTACAACCATCTGTAGGTCC CAAACAAGCTAGTCTTAATG GAGTCAGATTAATTGGAGAA GCCGTAAAAACACAGTTCAA TTATTATAAGAAAGTTGATG 20200

20201 GTGTTGTCCAACAATTACCT GAAACTTACTTTACACTGAG TAGAAATTTACAAGAATTAA AACCCAGGAGTCAAATGGAA ATTGATTTCTTAGAATTAGC 20300

20301 TATGGATGAATTCATTGAAC GATTTAAATTAGAAGGCTAT GCCTTCGAACATATCGTTTA TGGAGATTTTAGTCATAGTC AGTTAGGTGGTTTACATCTA 20400

20401 CTGATTGGACTAGCTAAACG TTTTAAGAATTCACCTTTTG AATTAGAAGATTTTATTCCT ATGGACAGTACAGTTAAAAA CTATTTCATAACAGATGCGA 20500

20501 ACACAGGTTCATCTAAGTGT GTGTGTTCTGTTATTGATTT ATTACTTGATTAATTTGTTG AAATAATAAAATCCCAAGAT TTATCTGTAGTTTCTAAGGT 20600

20601 TGTCAAAGTGACTATTGACT ATACGGAAATTTCATTTATG CTTTGGTGTAAAGATCGTCA TGTAGAAACATTTTACCCAA AATTACAATCTAGTCAAGCG 20700

20701 TGGCAACCGGGTGTTGCTAT GCCTAATCTTTACAAAACGA AAAGAATGCTATTAGAAAAG TGTGACCTTCAAAATTATGG TGATAGTGCAACATTACCTA 20800

20801 ATGGCATAATGATGAATGTC GCAAAATATACTCAACTGTG TCAATATTTAAACACATTAA CATTAGCTGTACCCTATAAT ATGAGAGTTATACTTGTTGG 20900

20901 TGCTGGTTCTGATAAAGGAG TTGCACCAGGTACAGCTGTT TTAAGACAGTGGTTGCCTAC GGGTACGCTGCTTGTCGTTC CAGATCTTAATGACTTTGTC 21000

21001 TCTGATGCAGATTCAACTTT GATTGGTGATTGTGCAACTG TACATACAGCTAATAAATGG GATCTCATTATTAGTGATAT GTACAACCCTAAGACTAAAA 21100

21101 ATGTTACAAAAGAAAATGAC TCTAAAGAGGGTTTTTTCAC TTACATTTGTGGGTTTATAA ATCAATAGCTAGCTCTTGGA GGTTCCGTGGCTATAAGGAT 21200

21201 AACAGAACATTCTTGGAATG CTGATCTTTATAAGCTCATG GGACACTTCGCATGGTGGAC AGCCTTTGTTACTAATGTGA ATGCGTCATCATCTGAAGCA 21300

21301 TTTTTAATTGGATGTAATTA TCTTGGCAAACCATGCGAAC AAATAGATGGTTATGTCATG CATGCAAATTACATATTTTG GAGGAATACAAATCCAATTC 21400

21401 AGTTGTCTTCCTATTCTTTA TTTGACATGAGTAAATTTCC CCTTAAATTAAGGGGTACTG CTTTCATGTCTTTAAAAGAA GGTCAAATCAATGATATGAT 21500

21501 TTTATCTCTTCTTAGTAAAG GTAGACTTATAATTAGAGAA AACAACAGAGTTGTTATCTA TAGTGATGTTCTTGTTAACA ACAATACGAACAATGTTTGT 21600

21601 TTTTCTTGTTTTATTGCCAC TAGTCTCTAGTCAGTGTGTT AATCTTAAATCCAGAACTCA ATTACCCCCTGCATACACTA ATTCTTTCACACGTGGTGTT 21700

21701 TATTACCCAGGCAAAGTTTT CAGATCCTCAGTTTTACATT CAACTCAGGACTTGTTCTTA CCTTTCTTTTCCAATGTTAC TTGGTTCCATGCTATACATG 21800

21801 TCTCTGGGACAATTGGTACT AAGAGGTTTGATAACCCTGT CCTACCATTTAATGATGGTG TTTATTTTGCTTCCACTGAG AAGTCAATCATAATAAGAGG 21900

21901 CTGGATTTTTGGTACTACTT TAGATTCGAAGACCCAGTCC CTACTTATTGTTAATAACTC AACTAATGTTGTTATTAAAG TCTGTGAATTTCAATTTTGT 22000

22001 AATGATCCATTTTTGGGTGT TTATTACCACAAAAACAACA AAAGTTGGATGGAAAGTGAG TTCAGAGTTTATTCTAGTGC GAATAATTGCACTTTTGAAT 22100

22101 AGGTCTCTCAGCCTTTTCTT ATGGACCTTGAAGGAAAACA GGGTAATTTCAAAAATCTTA GGGAATTTGTGTTTAAGAAT ATTGATGGTTATTTTAAAAT 22200

22201 ATATTCTAAACACACGCCTA TTAATTTAGTGCGTGATCTC CCTCAGGGTTTTTCGGCTTT AGAACCATTGGTAGATTTGC CAATAGGTATTAACATCACT 22300

22301 AGGTTTCAAATTATACTTTC ATTACATAGAAGTTATTTGA CTCCTGGTGATTCTTCTTCA GGTTGGACAGCTGGTGCTGC AGCTTATTATGTGGGTTATC 22400

22401 TTCAACCTAGGTCATTTCTA TTAAAATATAATGAAAATGG AACCATTACAGATGCTGTAG ACTGTGCACTTGACCCTCTC TCAGGAACAAAGTGTACGTT 22500

22501 GAAATCCTTCACTGTAGAAA AAGGAATCTATCAAACTTCT AACTTTAGAGTCCAACCAAC AGAATCTATTGTTAGATTTC CTAATATTAAATACTTGTGC 22600

22601 CCTTTTGGTGAAGTTTTTAA CGCCACCAGATTTGCATCTG TTTATGCTTGGAACAGGAAG AGAATCAGCAACTGTGTTGC TGATTATTCTGTCCTATATA 22700

22701 ATTCCGTACCATTTTCCACT TTTAAGTGTTATGGAGTGTC TCCTACTAAATTAAATGATC TCTGCTTTACTAATGTCTAT GCAGATTCATTTGTAATTAG 22800

22801 AGGTGATGAAGTCAGACAAA TCGCTCCAGGACAAACTGGA AAGATTGCTGATTATAATTA TAAATTACCAGATGATTTTA CAGGCTGCGTTATAGCTTGG 22900

22901 AATTCTAACAATCTTGATTC TAAGGTTGGTGGTAATTATA ATTACCTATATAGATTGTTT AGGAAGTCTAATCTCAAACC TTTTGAGAGAGATATTTCAA 23000

23001 CTGAAATCTATCAGGCCGGT AGCACACCTTGTAATGGTGT TGAAGGTTTTAATTGTTACT TCCTTTTACAATTATATGGT TTCCAACCCACTAATGGTGT 23100

23101 TGGTTACCAACCATACAGAG TAGTAGTACTTTCTTTTGAA CTTCTACATGCACCAGCAAC TGTTTGTGCATCTAAAAAGT CTACTAATTTGGTTAAAAAC 23200

23201 AAATGTGTCAATTTCAACTT CAATGGTTTAACAGGCACAG GTGTTCTTACTGAGTCTAAC -AAAAGTTTCTGCCTTTTCA ACAATTTGGCAGAGACATTG 23300

23301 CTGACACTACTGATGCTGTC CGAGGTCCACAGACACTTGA GATTCTTGACATTACACCAT GTTCTTTTGGTGGTGTCAGT GTTATAACGCCAGGAACAAA 23400

23401 TACTTCTAACCAGGTTGCTG TTCTTTATCAGGGTTTCAAC TGCACAGAAGTCCCTGTTGC TATTCATGCAGATCAACTTA CTCCTACTTGGCGTGTTTAT 23500

23501 TCTACAGGTTCTTAAGTTTC TAAAACACGTGCAGGCTGTT TAATAGGGGCTGAACATGTC AACAACTCATATGAGTGTGA CATACCCATTGGTGCAGGTA 23600

23601 TATGCGCTAGTTATCAGACT CAGTCTAATTCTCCTCGGCG GGCACGTAGTGTAGCTAGTC AATCCATCATTGCATTCACT ATGTCACTTGGTGCAGAAAA 23700

23701 TTCAGTTGCTTACTCTAATA ACTCTATTGCCATACCCACA AATTTTACTATTAGTGTTAC CACAGAAATTCTACCAGTGT CTATGACCAAGACATCAGTA 23800

23801 GATTGTACAATGTACATTTG GGATGATTCAACTGAATGCA GCAATTTATTGTTGCAATAT GGCAGTTTTTGTACACAATT AAACCGTGCTTTAACTGGAA 23900

23901 TAGCTGTTGAACAAGACAAA AACACCCAAGAAGTTTGTAC ACAAGTCAAACAAATTTACA AAACACCACCAATTAAAGAT TTTGGTGGTTTTAATTTTTC 24000

24001 ACAAATAATTCCAGATCCAT CAAAACCAAGCAAGAGGTCA TTTATTGAAGATCTACTTTT CAACAAAGTGACACTTGCAG TTGCTGGCTTCATCAAACAA 24100

24101 TATGGTGATTGCCTTGGTGA TATTGCTGCTAGAGCCCTCA TTTGTGCACAAAAGTTTAAC GGCCTTACTGTTTTGCCACC TTTGCTCACAGATAAGATGA 24200

24201 TTGCTCAATACACTTCTGCA CTGTTAGCGGGTACTATCAC TTCTGGTTGGACCTTTGGTG CAGGTGCTGCATTACAAATA CCATGTACTATGCAAATGGC 24300

24301 TTATAGGTTTAATGGTATTG GAGTTACACAGAATGTTCTC TATGAGAACCAAAAATTGAT TGCCAACCAATTTAATAGTG CTATTGGCAAAATTCAAGAC 24400

24401 TCACTTTCTTCCACAGCAAG CGAACTTGGAAAACTTCAAG ATGTGGTCAACCAAAATGCA CAAGCTTTAAACACGCTTGT TAAACAACTTAGCTCCAATT 24500

24501 TTGGTGCAATTTCAAGTGTT TTAAATGATATCCTTTCACG TCTTGACAAAGTTGAGGCTG AAGTGCAAATTGATAGGTTG ATCACAGGCAGACTTCAATG 24600

24601 TTTGCAGACATATGTGACTC AACAATTAATTAGAGCTGCA GAAATCAGAGCTTCTGCTAA TCTTGCTGCTACTAAATTGT CAGAGTGTGTACTTGGACAA 24700

24701 TCAAAAAGAGTTGATTTTTG TGGAAAGGGCTATCATCTTA TGTCCTTCCCTCAGTCAGCA CCTCATGGTGTAGTTTTCTT GCATGTGACTTATGTCCCTG 24800

24801 CACAAGAAAAGAACTTCACA ACTGCTCCTGCCATTTGTCA TGATGGAATAGCACACTTTC CTCGTGAAGGTGTCTTTGTT TCAAATGGCACACACTGGTT 24900

24901 TGTAACACAAAGGAATTTTT ATGAACCACAAATCATTACT ACAGACAACACATTTGTGTC TGGTAACTGTGATGTTGTAA TAGGAAGTCTCAACAACACA 25000

25001 GTTTATGATCCTTTGCAACC TGAATTAGACTCATTCAAGG AGGAGTTAGATAAATATTTT AAGAATCATACATCACCAGA TGTTGATTTAGGTGACATCT 25100

25101 CTGGCATTAATGCTTCAGTT GTAAACTTACAAAAAGAAAT TGACCGCCTCAATGAGGTTG CAATGAATTTAAATGAATCT CTCATCGATCTCCAAGAACT 25200

25201 TGGAAAGTATGAGCAGTATA T-AAATGGCCATGGTACATT TGGCTAGGTTTTATAGCTGG CTTGATTGCCATAGTAATGG TGACAATTATGCTTTGCTGT 25300

25301 ATGACCAGTTGCTGTAGTTG TCTCAAGGGCTGTTGTTCTT GGGTATCCTGCTGCAAATTT GATGAAGACGACTCTGAGCC AGTGCTCAAAGGAGTCAAAT

25400

25401 TACATTACACATAAACGAAC TTATGGATTTGTTTATGAGA ATCCTCACAATTGGAACTGT AACTTTGAAGCAAGGTGAAA TCAAGGATGCTACTCCTTCA 25500

25501 GATTTTGTTCGCTCCACTGC AACGATACCGATACAAGCCT CACTCCCTTTCGGATGGCTT ATTGTTGGCGTTGCACTTCT TGCTGTTTTTCATAGCGCTT 25600

25601 ACTAAATCATAACCCTCAAA AAGAGATGGCAACTAGCACT CTCCAAGGGTGTTCACTTTG TTTGCAACTTGCTGTTGTTG TTTGTAACAGTTTACTCACA 25700

25701 CCATTTGCTCGTTGCTGCTG GCCTTGAAGCCCCTTTTTTC TATCTTTATGCTTTAGTCTA CTTCTTGCAGAGTATAAATT ATGTAAGAATAATAATGAGT 25800

25801 CATTGGCTTTGCTGGAAATG CTGATCCAAAAACCCATTAC TTTATGATGACTACTATTTT CTTTGCTGGCATACTAATTG TTACGACTATTGTATACCTT 25900

25901 ACAATAGTGTAACTTCTTCA ATTGTCATTACTTCAGTTGA TGGCACAACAAGTCCTATTT CTGAACATGACTACCAGATT GGTGGTTATACTGAAAAATG 26000

26001 GGAATCTGGAGTAAAAGACT GTGTTGTATTACACAGTTAC TTCAATTCAGACTATTACCA GCTGTACTCAACTCAATTGA GTACAGACACTGGTGTTGAA 26100

26101 CATGTTACCTTCTTCATCTA CAATAATATTGTTGATGAGC CTGAAGAACATGTCCAAATT CACACAATCGACGGTTCATC CGGAGTTGTTACTGCAGTAG 26200

26201 TAGCAGCAATTTATGATGAA CCGACGACGACTACTAGCGC GACTTTGTAAGCACAAGCTG ATGAGTGCTAACTTATGTAC TCATTCGTTTCGGAAGAGAC 26300

26301 AGGTACGTTAATAGTTAATA GCGTACTTCCTTTTCTTGCT TTCGTGGTATTTTAGCTAGT TACACTAGCCATCCTTACTG CGCTTCGATTGTGTGCGTAC 26400

26401 TGCTGCAATATTGTTAACGA GTGTCTTGTAAAACCTTCTT TTTACGTTTACTCTCGTGTT AAAAATCTGAATTCTTCTAG AGTTCCTGATCTTCTGGTCT 26500

26501 AAACGAACTAAATACTATAT TAGTTTTTCTGTTTGGAACT TTAATTTTAGCCATGGCAGA TTCCAACGGTACTATTAGCC TTGAAGAGTTAAAAAAGCTC 26600

26601 CGTCAACAATGGAACCTAGT AATAGGTTTCCTATTCCTTA CATGGATTTGTCTTCTACAA TTTGCCTATGCCAACAGGAA TGGTTTTTTGTATATAATTA 26700

26701 AGTTAATTTTCCTCTGGCTG TTATGGCCAGTAACTTTAGC TTGTTTTGTGCTTGCTGCTG TTTACAGAATAAATTGGATC AGCCGTGGAATTGCTATCGC 26800

26801 AATGGCTTGTCTTGTAGGCT TGATGTGGCTCAGCTACTTC ATTGCTTCTTTCAGACTGTT TGCGCGTACGTGATCCATGT GGTCATTCAATCCAGAAACT 26900

26901 AACATTCTTCTCAACGTGCC ACTCCATGGCACTATTCTGA CCAGCCAGCTTCTAGAAAGT GAACTCGTAATCGGAGCTGT GATCCTTCGTGGACATCTTC 27000

27001 GTATTGCTGGACACCATCTA GGACGCTGTGACATCAAGGA CCTGCCTAAAGAAATCACTG TTGCTACATCACGAACGCTT TCTTATTGCAAATTGGGAGC 27100

27101 TTCGCAGCGTGTAGCAGGTG ACTCAGGTTTTGCTGCATAC AGTCGCTACAGGATTGGCAA CTATAAATTAAACACAGACC ATCCTAGTAGCAGTGACAAT 27200

27201 ATTGCTTTGCTTGTACAGTA AGTGACAACAGATGTTTCAT CTCGTTGACTTTCAGGTTAC TATAGCAGAGATATTACTAA TTATTATGAGGACTTTTAAA 27300

27301 TTATCCATTTGGAATCTTGA TTACATCATAAACCTCATAA TTAAAAATTTATCTAAGTCA CTAACTGAGAATAAATATTC ACAATTAGATGAAGAGCAAC 27400

27401 CAATGGATAATGATTAAACG AACATGAAAATTATTCTTTT CTTGGCACTGATAACACTCG CTACTTGTGAGCTTTATCAC TACCAAGAGTGTGTTAGAGG 27500

27501 TACAACAGTACTTATTAAAG AACCTTGCTCTTCTGGAACA TACGAGGGCAATTCACCATT TCATCCTCTAGCTGATAACA AATTCGAACTGACTTGCTTT 27600

27601 AGCACTCAATTTGCTTTTGC TTGTCCTGACGGCGTAAAAC ACGTCTATCAGTTACGTGCC AGATCAGTCTTACCTAAACT GTTCATCAGACAAGAGGAAG 27700

27701 TTCAAGAACTTTACTCTCCA ATTTTTCTTATTGTTGCGGC AATAGTGTTTATAACCTATT GCTTCACACTCAAAAGAAAG ACAGAATGATTGAACTTTCA 27800

27801 TTAATTGACTTCTATTTGTG CTTTTTAGCCTTTCTGCTAT TCCTTGTTTTAATTACGATT ATTATCTTTTGGTTCTCACT TGAACTGCAAGATCATAATT 27900

27901 AAACTTGTCACGCCTAAACG AACATGAAATTTCTTGTTTT CTTAGGAATCATCACAACTG TAGCTGCATTTCACCAAGAA TGTAGTTTACAGTTATGTAC 28000

28001 TCAACATCAACCATATGTAG TTGATGACCCGTGTCCTATT CACTTCTATTCTAAATGGTA TATTAGAGTAGGAGCTAGAA AATCAGCACCTTTAATTGAA 28100

28101 TTGTGCGTGGATGAGGCTGG TTCTAAATCACCCATTCAGT ACATCGATATCGGTAATTAT ACAGTTTCCTGTTTACCTTT TACAATTAATTGCCAGGAAC 28200

28201 CTAAATTGGGTAGTCTTGTA GTGCGTTGTTCGTTCTATGA AGACTTTTTTAGAGTATCAT GACGTTCGTGTTGTTTTAGA TTTCATCTAAACGAACAAAC 28300

28301 TAAAATGTCTGATAATGCAC TCCAAAATCAGCGAAATGCA CCCCGCATTACGTTTGGTGG ACCCTCAGATTCAACTGGCA GTAACCAGAAAGGCGAACGC 28400

28401 AGTGGGGCGCGATCAAAACA ACGTCGGCCCCAAGGTTTAC CCAATAATACTGCGTCTTGG TTCACCGCTCTCACTCAACA TGGCAAGGAAGACCTTAAAT 28500

28501 TCTCTCGAGGACAAGGCGTT CCAATTAACACCAATAGCAG TCCAGATGACCAAATTGGCT ACTACCGAAGAGCTACCAGA CGAATTCGTGGTGGTGACGG 28600

28601 TAAAATGAATGAACTCAGTC CAAGATGGTATTTCTACTAC CTAGGAACTGGGCCAGAAGC TGGACTTCCCTATGGTGCTA ACAAAGACGGCATCATATGG 28700

28701 GTTGCAACTGAGGGAGCCTT GACTATACCAAAAGATCACA TTGGCACCCGCAATCCTGCT AACAATGCTGCAATCGTGCA ACTACTTCCTCAAGGAACAA 28800

28801 CATTGCCAAAAGGCTACTAC GCAGAAGGGAGCAGAGGCGG CAGTCAAGCCTCTTCTCGTT CCTCATCACGTAGTCGCAAC AGTTCAAGAAATTCAACTCT 28900

28901 AGGCAGCAGTAGGGGAACTT CTCCTGCTAGAATGGCTGGC AATGGCGGTGATGCTGCTCT TGCTTTGCTGCTGCTTAACT GATTGAACCAGCTTGAGAGC

29000

29001 AAAATGTCTGGTAAAGGCCA ACAACAACAATGCCAAACTG TCACTAAGAAATCTGCTGCT GAGGCTTCTAAGAAGCCTCG GCAAAAAACGTACTGCCACT 29100

29101 AAAGCATACAATGTAACACA AGCTTTCGGCAGACGTGGTC CAGAACAAACCCAAGGAAAT GTTGTGGACCAGGAACTAAT CAGACAAGGAACTGATTACA 29200

29201 ATCATAGGCCGCAAATTGCA CAATTTGCCCCCAGCGCTTC AGCGTTCTTCGGAATGTCGC GCATTGGCATGGAAGTCACA CCTTCGGGAACGTGGTTGAC 29300

29301 CTAGACTCCTGAGATCAAAA TGGTTGACAAAGATCCAAAA TTCTACGATTAAGTCATTTT GCTGAATAAACATTTTGACG CATACAAAACATTCCCACCA 29400

29401 ACAGAGCCTAAAAAGGACAA AAAGAAGAAGGCTGATGAAA CTCAAGCCTTACCGCAGAGA CGGAAAAAACAGCAAACTGT GACTCTTCTTCCTGCTGCAG

29500

29501 ATTTGGATGATTTCTCCAAA CAATTGCTACAATCCATGAG CAGTGCTGACTCAACTCAGG CCTAAACTCATGCAGACCAC ACAAGGCAGATGGGCTATAT 29600

29601 AAACGTTTTCGCCTTTCCGT TTACGATATATAGTCTACTC TTGTGCAGAATGAATTCTCG TAACTACATAGCACAAGTAG ATGTAGTTAACTTTAATCTC 29700

29701 ACATAGCAATATTTGATCAG TGTGTAACATTAGGGAGGAC TTGAAAGAGCCACCACATTT TCAGCGACGCCACGCGGAGT ACGATCGAGTGTACAGTGAA 29800

29801 CAATGCTAGGGAGAGCTGCC TATAAGGATGAGCCCTAATG TGTAAAATTAATTTTAGTAG TGCTATCCCCATGTGATTTT AATAGCTTCTTAGGAGAATG 29900

29901 ACAAAAAAAAAAATCACATGG GGATAGCAC 29929

7.1.13 SRR12596175 Discovered Strain 2 Nucleotide Sequence

1 AGAGAACAGATCTACAAGAG ATCGAAAGTTGGTTGGTTTG TTACCTGGGAAGGTATAAAC CAACCAACTTTCGATCTCTT GTAGATCTGTTCTCTAAACG 100

101 AACTTTAAAATCTGTGTGGC TGTCACTCGGCTGCATGCTT AGTGCACTCACGCAGTATAA TTAATAACTAATTACTGTCG TTGACAGGACACGAGTAACT 200

201 CGTCTATCTTCTGCAGGCTG CTTACGGTTTCGTCCGTGTT GCAGCCGATCATCAGCACAT CTAGGTTTTGTCCGGGTGTG ACCGAAAGGTAAGATGGAGA 300

301 GCCTTGTCCCTGGTTTCAAC GAGAAAACACACGTCCAACT CAGTTTGCCTGTTTTACAGG TTCGCGACGTGCTCGTACGT GGCTTTGGAGACTCCGTGGA 400

401 GGAGGTCTTATCAGAGGCAC GTCAACATCTTAAAGATGGC ACTTGTGGCTTAGTAGAAGT TGAAAAAGGCGTTTTGCCTC AACTTGAACAGCCCTATGTG 500

501 TTCATCAAACGTTCGGATGC TCGAACTGCACCTCATGGTC ATGTTATGGTTGAGCTGGTA GCAGAACTCGAAGGCATTCA GTACGGTCGTAGTGGTGAGA 600

601 CACTTGGTGTCCTTGTCCCT CATGTGGGCGAAATACCAGT GGCTTACCGCAAGGTTCTTC TTCGTAAGAACGGTAATAAA GGAGCTGGTGGCCATAGTTA 700

701 CGGCGCCGATCTAAAGTCAT TTGACTTAGGCGACGAGCTT GGCACTGATCCTTATGAAGA TTTTCAAGAAAACTGGAACA CTAAACATAGCAGTGGTGTT 800

801 ACCCGTGAACTCATGCGTGA GCTTAACGGAGGGGCATACA CTCGCTATGTCGATAACAAC TTCTGTGGCCCTGATGGCTA CCCTCTTGAGTGCATTAAAG 900

901 ACCTTCTAGCACGTGCTGGT AAAGCTTCATGCACTTTGTC CGAACAACTGGACTTTATTG ACACTAAGAGGGGTGTATAC TGCTGCCGTGAACATGAGCA 1000

1001 TGAAATTGCTTGGTACACGG AACGTTCTGAAAAGAGCTAT GAATTGCAGACACCTTTTGA AATTAAATTGGCAAAGAAAT TTGACATCTTCAATGGGGAA 1100

1101 TGTCCAAATTTTGTATTTCC CTTAAATTCCATAATCAAGA CTATTCAACCAAGGGTTGAA AAGAAAAAGCTTGATGGCTT TATGGGTAGAATTCGATCTG 1200

1201 TCTATCCAGTTGCGTCACCA AATGAATGCAACCAAATGTG CCTTTCAACTCTCATGAAGT GTGATCATTGTGGTGAAACT TCATGGCAGACGGGCGATTT 1300

1301 TGTTAAAGCCACTTGCGAAT TTTGTGGCACTGAGAATTTG ACTAAAGAAGGTGCCACTAC TTGTGGTTACTTACCCCAAA ATGCTGTTGTTAAAATTTAT 1400

1401 TGTCCAGCATGTTACAATTC AGAAGTAGGACCTGAGCATA GTCTTGCCGAATACCATAAT GAATCTGGCTTGAAAACCAT TCTTCGTAAGGGTGGTCGCA 1500

1501 CTATTGCCTTTGGAGGCTGT GTGTTCTCTTATGTTGGTTG CCATAACAAGTGTGCCTTTT GGGTTCCACGTGCTAGCGCT AACATAGGTTGTAACCATAC 1600

1601 AGGTGTTGTTGGAGAAGGTT CCGAAGGTCTTAATGACAAC CTTCTTGAAATACTCCAAAA AGAGAAAGTCAACATCAATA TTGTTGGTGACTTTAAACTT 1700

1701 AATGAAGAGATCGCCATTAT TTTGGCATCTTTTTCTGCTT CCACAAGTGCTTTTGTGGAA ACTGTTAAAGGTTTGGATTA TAAAGCATTCAAACAAATTG 1800

1801 TTGAATCCTGTGGTAATTTT AAAGTTACAAAAGGAAAAGC TAAAAAAGGTGCCTGGAATA TTGGTGAACAGAAATCAATA CTGAGTCCTCTTTATGCATT 1900

1901 TGCATCAGAGGCTGCTCGTG TTGTACGATCAATTTTCTCC CGCACTCTTGAAACTGCTCA AAATTCTGTGCGTGTTTTAC AGAAGGCCGCTATAACAATA 2000

2001 CTAGATGGAATTTCACAGTA TTCACTGAGACTCATTGATG CTATGATGTTCACATCTGAT TTGGCTACTAACAATCTAGT TGTAATGGCCTACATTACAG 2100

2101 GTGGTGTTGTTCAGTTGACT TCGCAGTGGCTAACTAACAT CTTTGGCACTGTTTATGAAA AACTCAAACCCGTCCTTGAT TGGCTTGAAGAGAAGTTTAA 2200

2201 GGAAGGTGTAGAGTTTCTTA GAGACGGTTGGGAAATTGTT AAATTTATCTCAACCTGTGC TTGTGAAATTGTCGGTGGAC AAATTGTCACCTGTGCAAAG 2300

2301 GAAATTAAGGAGAGTGTTCA GACATTCTTTAAGCTTGTAA ATAAATTTTTGGCTTTGTGT GCTGACTCTATCATTATTGG TGGAGCTAAACTTAAAGCCT 2400

2401 TGAATTTAGGTGAAACATTT GTCACGCACTCAAAGGGATT GTACAGAAAGTGTGTTAAAT CCAGAGAAGAAACTGGCCTA CTCATGCCTCTAAAAGCCCC 2500

2501 AAAAGAAATTATCTTCTTAG AGGGAGAAACACTTCCCACA GAAGTGTTAACAGAGGAAGT TGTCTTGAAAACTGGTGATT TACAACCATTAGAACAACCT 2600

2601 ACTAGTGAAGCTGTTGAAGC TCCATTGGTTGGTACACCAG TTTGTATTAACGGGCTTATG TTGCTCGAAATCAAAGACAC AGAAAAGTACTGTGCCCTTG 2700

2701 CACCTAATATGATGGTAACA AACAATACCTTCACACTCAA AGGCGGTGCACCAACAAAGG TTACTTTTGGTGATGACACT GTGATAGAAGTGCAAGGTTA 2800

2801 CAAGAGTGTGAATATCACTT TTGAACTTGATGAAAGGATT GATAAAGTACTTAATGAGAA GTGCTCTGCCTATACAGTTG AACTCGGTACAGAAGTAAAT 2900

2901 GAGTTCGCCTGTGTTGTGGC AGATGCTGTCATAAAAACTT TGCAACCAGTATCTGAATTA CTTACACCACTGGGCATTGA TTTAGATGAGTGGAGTATGG 3000

3001 CTACATACTACTTATTTGAT GAGTCTGGTGAGTTTAAATT GGCTTCACATATGTATTGTT CTTTTTACCCTCCAGATGAG GATGAAGAAGAAGGTGATTG 3100

3101 TGAAGAAGAAGAGTTTGAGC CATCAACTCAATATGAGTAT GGTACTGAAGATGATTACCA AGGTAAACCTTTGGAATTTG GTGCCACTTCTGCTGCTCTT 3200

3201 CAACCTGAAGAAGAGCAAGA AGAAGATTGGTTAGATGATG ATAGTCAACAAACTGTTGGT CAACAAGACAGCAGTGAGGA CAATCAGACAACTACTATTC 3300

3301 AAACAATTGTTGAGGTTCAA CCTCAATTAGAGATGGAACT TACACCAGTTGTTCAGACTA TTGAAGTGAATAGTTTTAGT GGTTATTTAAAACTTACTGA 3400

3401 CAATGTATACATTAAAAATG CAGACATTGTGGAAGAAGCT AAAAAGGTAAAACCAACAGT GGTTGTTAATGCAGCCAATG TTTACCTTAAACATGGAGGA 3500

3501 GGTGTTGCAGGAGCCTTAAA TAAGGCTACTAACAATGCCA TGCAAGTTGAATCTGATGAT TACATAGCTACTAATGGACC ACTTAAAGTGGGTGGTAGTT 3600

3601 GTGTTTTAAGCGGACACAAT CTTGCTAAACACTGTCTTCA TGTTGTCGGCCCAAATGTTA ACAAAGGTGAAGACATTCAA CTTCTTAAGAGTGCTTATGA 3700

3701 AAATTTTAATCAGCACGAAG TTCTACTTGCACCATTATTA TCAGCTGGTATTTTTGGTGC TGACCCTATACATTCTTTAA GAGTTTGTGTAGATACTGTT 3800

3801 CGCACAAATGTCTACTTAGC TGTCTTTGATAAAAATCTCT ATGACAAACTTGTTTCAAGC TTTTTGGAAATGAAGAGTGA AAAGCAAGTTGAACAAAAGA 3900

3901 TCGCTGAGATTCCTAAAGAG GAAGTTAAGCCATTTATAAC TGAAAGTAAACCTTCAGTTG AACAGAGAAAACAAGATGAT AAGAAAATCAAAGCTTGTGT 4000

4001 TGAAGAAGTTACAACAACTC TGGAAGAAACTAAGTTCCTC ACAGAAAACTTGTTACTTTA TATTGACATTAATGGCAATC TTCATCCAGATTCTGCCACT 4100

4101 CTTGTTAGTGACATTGACAT CACTTTCTTAAAGAAAGATG CTCCATATATAGTGGGTGAT GTTGTTCAAGAGGGTGTTTT AACTGCTGTGGTTATACCTA 4200

4201 CTAAAAAGGCTGGTGGCACT ACTGAAATGCTAGCGAAAGC TTTGAGAAAAGTGCCAACAG ACAATTATATAACCACTTAC CCGGGTCAGGGTTTAAATGG 4300

4301 TTACACTGTAGAGGAGGCAA AGACAGTGCTTAAAAAGTGT AAAAGTGCCTTTTACATTCT ACCATCTATTATCTCTAATG AGAAGCAAGAAATTCTTGGA 4400

4401 ACTGTTTCTTGGAATTTGCG AGAAATGCTTGCACATGCAG AAGAAACACGCAAATTAATG CCTGTCTGTGTGGAAACTAA AGCCATAGTTTCAACTATAC 4500

4501 AGCGTAAATATAAGGGTATT AAAATACAAGAGGGTGTGGT TGATTATGGTGCTAGATTTT ACTTTTACACCAGTAAAACA ACTGTAGCGTCACTTATCAA 4600

4601 CACACTTAACGATCTAAATG AAACTCTTGTTACAATGCCA CTTGGCTATGTAACACATGG CTTAAATTTGGAAGAAGCTG CTCGGTATATGAGATCTCTC 4700

4701 AAAGTGCCAGCTACAGTTTC TGTTTCTTCACCTGATGCTG TTACAGCGTATAATGGTTAT CTTACTTCTTCTTCTAAAAC ACCTGAAGAACATTTTATTG 4800

4801 AAACCATCTCACTTGCTGGT TCCTATAAAGATTGGTCCTA TTCTGGACAATCTACACAAC TAGGTATAGAATTTCTTAAG AGAGGTGATAAAAGTGTATA 4900

4901 TTACACTAGTAATCCTACCA CATTCCACCTAGATGGTGAA GTTATCACCTTTGACAATCT TAAGACACTTCTTTCTTTGA GAGAAGTGAGGACTATTAAG 5000

5001 GTGTTTACAACAGTAGACAA CATTAACCTCCACACGCAAG TTGTGGACATGTCAATGACA TATGGACAACAGTTTGGTCC AACTTATTTGGATGGAGCTG 5100

5101 ATGTTACTAAAATAAAACCT CATAATTCACATGAAGGTAA AACATTTTATGTTTTACCTA ATGATGACACTCTACGTGTT GAGGCTTTTGAGTACTACCA 5200

5201 CACAACTGATCCTAGTTTTC TGGGTAGGTACATGTCAGCA TTAAATCACACTAAAAAGTG GAAATACCCACAAGTTAATG GTTTAACTTCTATTAAATGG 5300

5301 GCAGATAACAACTGTTATCT TGCCACTGCATTGTTAACAC TCCAACAAATAGAGTTGAAG TTTAATCCACCTGCTCTACA AGATGCTTATTACAGAGCAA 5400

5401 GGGCTGGTGAAGCTGCTAAC TTTTGTGCACTTATCTTAGC CTACTGTAATAAGACAGTAG GTGAGTTAGGTGATGTTAGA GAAACAATGAGTTACTTGTT 5500

5501 TCAACATGCCAATTTAGATT CTTGCAAAAGAGTCTTGAAC GTGGTGTGTAAAACTTGTGG ACAACAGCAGACAACCCTTA AGGGTGTAGAAGCTGTTGTG 5600

5601 TACATGGGCACACTTTCTTA TGAACAATTTAAGAAAGGTG TTCAGATACCTTGTACGTGT GGTAAACAAGCTACAAAATA TCTAGTACAACAGGAGTCAC 5700

5701 CTTTTGTTATGATGTCAGCA CCACCTGCTCAGTATGAACT TAAGCATGGTACATTTACTT GTGCTAGTGAGTACACTGGT AATTACCAGTGTGGTCACTA 5800

5801 TAAACATATAACTTCTAAAG AAACTTTGTATTGCATAGAC GGTGCTTTACTTACAAAGTC CTCAGAATACAAAGGTCCTA TTACGGATGTTTTCTACAAA 5900

5901 GAAAACAGTTACACAACAAC CATAAAACCAGTTACTTATA AATTGGATGGTGTTGTTTGT ACAGAAATTGACCCTAAGTT GGACAATTATTATAAGAAAG 6000

6001 ACAATTCTTATTTCACAGAG CAACCAATTGATCTTGTACC AAACCAACCATATCCAAACG CAAGCTTCGATAATTTTAAG TTTGTATGTGATAATATCAA 6100

6101 ATTTGCTGATGATTTAAACC AGTTAACTGGTTATAAGAAA CCTGCTTCAAGAGAGCTTAA AGTTACATTTTTCCCTGACT TAAATGGTGATGTGGTGGCT 6200

6201 ATTGATTATAAACACTACAC ACCCTCTTTTAAGAAAGGAG CTAAATTGTTACATAAACCT ATTGTTTGGCATGTTAACAA TGCAACTAATAAAGCCACGT 6300

6301 ATAAACCAAATACCTGGTGT ATACGTTGTCTTTGGAGCAC AAAACCAGTTGAAACATCAA ATTCGTTTGATGTACTGAAG TCAGAGGACGCGCAGGGAAT 6400

6401 GGATAATCTTGCCTGCGAAG ATCTAAAACCAGTCTCTGAA GAAGTAGTGGAAAATCCTAC CATACAGAAAGACGTTCTTG AGTGTAATGTGAAAACTACC 6500

6501 GAAGTTGTAGGAGACATTAT ACTTAAACCAGCAAATAATA GTTTAAAAATTACAGAAGAG GTTGGCCACACAGATCTAAT GGCTGCTTATGTAGACAATT 6600

6601 CTAGTCTTACTATTAAGAAA CCTAATGAATTATCTAGAGT ATTAGGTTTGAAAACCCTTG CTACTCATGGTTTAGCTGCT GTTAATAGTGTCCCTTGGGA 6700

6701 TACTATAGCTAATTATGCTA AGCCTTTTCTTAACAAAGTT GTTAGTACAACTACTAACAT AGTTACACGGTGTTTAAACC GTGTTTGTACTAATTATATG 6800

6801 CCTTATTTCTTTACTTTATT GCTACAATTGTGTACTTTTA CTAGAAGTACAAATTCTAGA ATTAAAGCATCTATGCCGAC TACTATAGCAAAGAATACTG 6900

6901 TTAAGAGTGTCGGTAAATTT TGTCTAGAGGCTTCATTTAA TTATTTGAAGTCACCTAATT TTTCTAAACTGATAAATATT ATAATTTGGTTTTTACTATT 7000

7001 AAGTGTTTGCCTAGGTTCTT TAATCTACTCAACCGCTGCT TTAGGTGTTTTAATGTCTAA TTTAGGCATGCCTTCTTACT GTACTGGTTACAGAGAAGGC 7100

7101 TATTTGAACTCTACTAATGT CACTATTGCAACCTACTGTA CTGGTTCTATACCTTGTAGT GTTTGTCTTAGTGGTTTAGA TTCTTTAGACACCTATCCTT 7200

7201 CTTTAGAAACTATACAAATT ACCATTTCATCTTTTAAATG GGATTTAACTGCTTTTGGCT TAGTTGCAGAGTGGTTTTTG GCATATATTCTTTTCACTAG 7300

7301 GTTTTTCTATGTACTTGGAT TGGCTGCAATCATGCAATTG TTTTTCAGCTATTTTGCAGT ACATTTTATTAGTAATTCTT GGCTTATGTGGTTAATAATT 7400

7401 AATCTTGTACAAATGGCCCC GATTTCAGCTATGGTTAGAA TGTACATCTTCTTTGCATCA TTTTATTATGCATGGAAAAG TTATGTGCATGTTGTAGACG 7500

7501 GTTGTAATTCATCAACTTGT ATGATGTGTTACAAACGTAA TAGAGCAACAAGAGTCGAAT GTACAACTATTGTTAATGGT GTTAGAAGGTCCTTTTATGT 7600

7601 CTATGCTAATGGAGGTAAAG GCTTTTGCAAACTACACAAT TGGAATTGTGTTAATTGTGA TACATTCTGTGCTGGTAGTA CATTTATTAGTGATGAAGTT 7700

7701 GCGAGAGACTTGTCACTACA GTTTAAAAGACCAATAAATC CTACTGACCAGTCTTCTTAC ATCGTTGATAGTGTTACAGT GAAGAATGGTTCCATCCATC 7800

7801 TTTACTTTGATAAAGCTGGT CAAAAGACTTATGAAAGACA TTCTCTCTCTCATTTTGTTA ACTTAGACAACCTGAGAGCT AATAACACTAAAGGTTCATT 7900

7901 GCCTATTAATGTTATAGTTT TTGATGGTAAATCAAAATGT GAAGAATCATCTGCAAAATC AGCGTCTGTTTACTACAGTC AGCTTATGTGTCAACCTATA 8000

8001 CTGTTACTAGATCAGGCATT AGTGTCTGATGTTGGTGATA GTGCGGAAGTTGCAGTTAAA ATGTTTGATGCTTACGTTAA TACGTTTTCATCAACTTTTA 8100

8101 ACGTACCAATGGAAAAACTC AAAACACTAGTTGCAACTGC AGAAGCTGAACTTGCAAAGA ATGTGTCCTTAGACAATGTC TTATCTACTTTTATTTCAGC 8200

8201 AGCTCGGCAAGGGTTTGTTG ATTCAGATGTAGAAACTAAA GATGTTGTTGAATGTCTTAA ATTGTCACATCAATCTGACA TAGAAGTTACTGGCGATAGT 8300

8301 TGTAATAACTATATGCTCAC CTATAACAAAGTTGAAAACA TGACACCCCGTGACCTTGGT GCTTGTATTGACTGTAGTGC GCGTCATATTAATGCGCAGG 8400

8401 TAGCAAAAAGTCACAACATT GCTTTGATATGGAACGTTAA AGATTTCATGTCATTGTCTG AACAACTACGAAAACAAATA CGTAGTGCTGCTAAAAAGAA 8500

8501 TAACTTACCTTTTAAGTTGA CATGTGCAACTACTAGACAA GTTGTTAATGTTGTAACAAC AAAGATAGCACTTAAGGGTG GTAAAATTGTTAATAATTGG 8600

8601 TTGAAGCAGTTAATTAAAGT TACACTTGTGTTCCTTTTTG TTGCTGCTATTTTCTATTTA ATAACACCTGTTCATGTCAT GTCTAAACATACTGACTTTT 8700

8701 CAAGTGAAATCATAGGATAC AAGGCTATTGATGGTGGTGT CACTCGTGACATAGCATCTA CAGATACTTGTTTTGCTAAC AAACATGCTGATTTTGACAC 8800

8801 ATGGTTTAGCCAGCGTGGTG GTAGTTATACTAATGACAAA GCTTGCCCATTGATTGCTGC AGTCATAACAAGAGAAGTGG GTTTTGTCGTGCCTGGTTTG 8900

8901 CCTGGCACGATATTACGCAC AACTAATGGTGACTTTTTGC ATTTCTTACCTAGAGTTTTT AGTGCAGTTGGTAACATCTG TTACACACCATCAAAACTTA 9000

9001 TAGAGTACACTGACTTTGCA ACATCAGCTTGTGTTTTGGC TGCTGAATGTACAATTTTTA AAGATGCTTCTGGTAAGCCA GTACCATATTGTTATGATAC 9100

9101 CAATGTACTAGAAGGTTCTG TTGCTTATGAAAGTTTACGC CCTGACACACGTTATGTGCT CATGGATGGCTCTATTATTC AATTTCCTAACACCTACCTT 9200

9201 GAAGGTTCTGTTAGAGTGGT AACAACTTTTGATTCTGAGT ACTGTAGGCACGGCACTTGT GAAAGATCAGAAGCTGGTGT TTGTGTATCTACTAGTGGTA 9300

9301 GATGGGTACTTAACAATGAT TATTACAGATCTTTACCAGG AGTTTTCTGTGGTGTAGATG CTGTAAATTTACTTACTAAT ATGTTTACACCACTAATTCA 9400

9401 ACCTATTGGTGCTTTGGACA TATCAGCATCTATAGTAGCT GGTGGTATTGTAGCTATCGT AGTAACATGCCTTGCCTACT ATTTTATGAGGTTTAGAAGA 9500

9501 GCTTTTGGTGAATACAGTCA TGTAGTTGCCTTTAATACTT TACTATTCCTTATGTCATTC ACTGTACTCTGTTTAACACC AGTTTACTCATTCTTACCTG 9600

9601 GTGTTTATTCTGTTATTTAC TTGTACTTGACATTTTATCT TACTAATGATGTTTCTTTTT TAGCACATATTCAGTGGATG GTTATGTTCACACCTTTAGT 9700

9701 ACCTTTCTGGATAACAATTG CTTATATCATTTGTATTTCC ACAAAGCATTTCTATTGGTT CTTTAGTAATTACCTAAAGA GACGTGTAGTCTTTAATGGT 9800

9801 GTTTCCTTTAGTACTTTTGA AGAAGCTGCGCTGTGCACCT TTTTGTTAAATAAAGAAATG TATCTAAAGTTGCGTAGTGA TGTGCTATTACCTCTTACGC 9900

9901 AATATAATAGATACTTAGCT CTTTATAATAAGTACAAGTA TTTTAGTGGAGCAATGGATA CAACTAGCTACAGAGAAGCT GCTTGTTGTCATCTCGCAAA 10000

10001 GGCTCTCAATGACTTCAGTA ACTCAGGTTCTGATGTTCTT TACCAACCACCACAAACCTC TATCACCTCAGCTGTTTTGC AGAGTGGTTTTAGAAAAATG 10100

10101 GCATTCCCATCTGGTAAAGT TGAGGGTTGTATGGTACAAG TAACTTGTGGTACAACTACA CTTAACGGTCTTTGGCTTGA TGACGTAGTTTACTGTCCAA 10200

10201 GACATGTGATCTGCACCTCT GAAGACATGCTTAACCCTAA TTATGAAGATTTACTCATTC GTAAGTCTAATCATAATTTC TTGGTACAGGCTGGTAATGT 10300

10301 TCAACTCAGGGTTATTGGAC ATTCTATGCAAAATTGTGTA CTTAAGCTTAAGGTTGATAC AGCCAATCCTAAGACACCTA AGTATAAGTTTGTTCGCATT 10400

10401 CAACCAGGACAGACTTTTTC AGTGTTAGCTTGTTACAATG GTTCACCATCTGGTGTTTAC CAATGTGCTATGAGGCCCAA TTTCACTATTAAGGGTTCAT 10500

10501 TCCTTAATGGTTCATGTGGT AGTGTTGGTTTTAACATAGA TTATGACTGTGTCTCTTTTT GTTACATGCACCATATGGAA TTACCAACTGGAGTTCATGC 10600

10601 TGGCACAGACTTAGAAGGTA ACTTTTATGGACCTTTTGTT GACAGGCAAACAGCACAAGC AGCTGGTACGGACACAACTA TTACAGTTAATGTTTTAGCT 10700

10701 TGGTTGTACGCTGCTGTTAT AAAAGGAGACAGGTGGTTTC TCAATCGATTTACCACAACT CTTAATGACTTTAACCTTGT GGCTATGAAGTACAATTATG 10800

10801 AACCTCTAACACAAGACCAT GTTGACATACTAGGACCTCT TTCTGCTCAAACTGGAATTG CCGTTTTAGATATGTGTGCT TCATTAAAAGAATTACTGCA 10900

10901 AAATGGTATGAATGGACGTA CCATATTGGGTAGTGCTTTA TTAGAAGATGAATTTACACC TTTTGATGTTGTTAGACAAT GCTCAGGTGTTACTTTCCAA 11000

11001 AGTGCAGTGAAAAGAACAAT CAAGGGTACACACCACTGGT TGTTACTCACAATTTTGACT TCACTTTTAGTTTTAGTCCA GAGTACTCAATGGTCTTTGT 11100

11101 TCTTTTTTTTTGTATGAAAAT GCCTTTTTACCTTTTGCTAT GGGTATTATTGCTATGTCTG CTTTTGCAATGATGTTTGTC AAACATAAGCATGCATTTCT 11200

11201 CTGTTTGTTTTTGTTACCTT CTCTTGCCACTGTAGCTTAT TTTAATATGGTCTATATGCC TGCTAGTTGGGTGATGCGTA TTATGACATGGTTGGATATG 11300

11301 GTTGATACTAGTTTGTCTGG TTTTAAGCTAAAAGACTGTG TTATGTATGCATCAGCTGTA GTGTTACTAATCCTTATGAC AGCAAGAACTGTGTATGATG 11400

11401 ATGGTGCTAGGAGAGTGTGG ACACTTATGAATGTCTTGAC ACTCGTTTATAAAGTTTATT ATGGTAATGCTTTAGATCAA GCCATTTCCATGTGGGCTCT 11500

11501 TATAATCTCTGTTACTTCTA ACTACTCAGGTGTAGTTACA ACTGTCATGTTTTTGGCCAG AGGTATTGTTTTTATGTGTG TTGAGTATTGCCCTATTTTC 11600

11601 TTCATAACTGGTAATACACT TCAGTGTATAATGCTAGTTT ATTGTTTCTTAGGCTATTTT TGTACTTGTTACTTTGGCCT CTTTTGTTTACTCAACCGCT 11700

11701 ACTTTAGACTGACTCTTGGT GTTTATGATTACTTAGTTTC TACACAGGAGTTTAGATATA TGAATTCACAGGGACTACTC CCACCCAAGAATAGCATAGA 11800

11801 TGCCTTCAAACTCAACATTA AATTGTTGGGTGTTGGTGGC AAACCTTGTATCAAAGTAGC CACTGTACAGTCTAAAATGT CAGATGTAAAGTGCACATCA 11900

11901 GTAGTCTTACTCTCAGTTTT GCAACAACTCAGAGTAGAAT CATTATCTAAATTGTGGGCT CAATGTGTCCAGTTACACAA TGACATTCTCTTAGCTAAAG 12000

12001 ATACTACTGAAGCCTTTGAA AAAATGGTTTCACTACTTTC TGTTTTGCTTTCCATGCAGG GTGCTGTAGACATAAACAAG CTTTGTGAAGAAATGCTGGA 12100

12101 CAACAGGGCAACCTTACAAG CTATAGCCTCAGAGTTTAGT TCCCTTCCATCATATGCAGC TTTTGCTACTGCTCAAGAAG CTTATGAGCAGGCTGTTGCT 12200

12201 AATGGTGATTCTGAAGTTGT TCTTAAAAAGTTGAAGAAGT CTTTGAATGTGGCTAAATCT GAATTTGACCGTGATGCAGC CATGCAACGTAAGTTGGAAA 12300

12301 AGATGGCTGATCAAGCTATG ACCCAAATGTATAAACAGGC TAGATCTGAGGACAAGAGGG CAAAAGTTACTAGTGCTATG CAGACAATGCTTTTCACTAT 12400

12401 GCTTAGAAAGTTGGATAATG ATGCACTCAACAACATTATC AACAATGCAAGAGATGGTTG TGTTCCCTTGAACATAATAC CTCTTACAACAGCAGCCAAA 12500

12501 CTAATGGTTGTCATACCAGA CTATAACACATATAAAAATA CGTGTGATGGTACAACATTT ACTTATGCATCAGCATTGTG GGAAATCCAACAGGTTGTAG 12600

12601 ATGCAGATAGTAAAATTGTT CAACTTAGTGAAATTAGTAT GGACAATTCACCTAATTTAG CATGGCCTCTTATTGTAACA GCTTTAAGGGCCAATTCTGC 12700

12701 TGTCAAATTACAGAATAATG AGCTTAGTCCTGTTGCACTA CGACAGATGTCTTGTGCTGC CGGTACTACACAAACTGCTT GCACTGATGACAATGCGTTA 12800

12801 GCTTACTACAACACAACAAA GGGAGGTAGGTTTGTACTTG CACTGTTATCCGATTTACAG GATTTGAAATGGGCTAGATT CCCTAAGAGTGATGGAACTG 12900

12901 GTACTATCTATACAGAACTG GAACCACCTTGTAGGTTTGT TACAGACACACCTAAAGGTC CTAAAGTGAAGTATTTATAC TTTATTAAAGGATTAAACAA 13000

13001 CCTAAATAGAGGTATGGTAC TTGGTAGTTTAGCTGCCACA GTACGTCTACAAGCTGGTAA TGCAACAGAAGTGCCTGCCA ATTCAACTGTATTATCTTTC 13100

13101 TGTGCTTTTGCTGTAGATGC TGCTAAAGCTTACAAAGATT ATCTAGCTAGTGGGGGACAA CCAATCACTAATTGTGTTAA GATGTTGTGTACACACACTG 13200

13201 GTACTGGTCAGGCAATAACA GTTACACCGGAAGCCAATAT GGATCAAGAATCCTTTGGTG GTGCATCGTGTTGTCTGTAC TGCCGTTGCCACATAGATCA 13300

13301 TCCAAATCCTAAAGGATTTT GTGACTTAAAAGGTAAGTAT GTACAAATACCTACAACTTG TGCTAATGACCCTGTGGTTT TTACACTTAAAAACACAGTC 13400

13401 TGTACCGTCTGCGGTATGTG GAAAGGTTATGGCTGTAGTT GTGATCAACTCCGCGAACCC ATGCTTCAGTCAGCTGATGC ACAATCGTTTTTAAACGGGT 13500

13501 TTGCGGTGTAAGTGCAGCCC GTCTTACACCGTGCGGCACA GGCACTAGTACTGATGTCGT ATACAGGGCTTTTGACATCT ACAATGATAAAGTAGCTGGT 13600

13601 TTTGCTAAATTCCTAAAAAC TAATTGTTGTCGCTTCCAAG AAAAGGACGAAGATGACAAT TTAATTGATTCTTACTTTGT AGTTAAGAGACACACTTTCT 13700

13701 CTAACTACCAACATGAAGAA ACAATTTATAATTTACTTAA GGATTGTCCAGCTGTTGCTA AACATGACTTCTTTAAGTTT AGAATAGACGGTGACATGGT 13800

13801 ACCACATATATCACGTCAAC GTCTTACTAAATACACAATG GCAGACCTCGTCTATGCTTT AAGGCATTTTGATGAAGGTA ATTGTGACACATTAAAAGAA 13900

13901 ATACTTGTCACATACAATTG TTGTGATGATGATTATTTCA ATAAAAAGGACTGGTATGAT TTTGTAGAAAACCCAGATAT ATTACGCGTATACGCCAACT 14000

14001 TAGGTGAACGTGTACGCCAA GCTTTGTTAAAAACAGTACA ATTCTGTGATGCCATGCGAA ATGCTGGTATTGTTGGTGTA CTGACATTAGATAATCAAGA 14100

14101 TCTCAATGGTAACTGGTATG ATTTCGGTGATTTCATACAA ACCACGCCAGGTAGTGGAGT TCCTGTTGTAGATTCTTATT ATTCATTGTTAATGCCTATA 14200

14201 TTAACCTTGACCAGGGCTTT AACTGCAGAGTCACATGTTG ACACTGACTTAACAAAGCCT TACATTAAGTGGGATTTGTT AAAATATGACTTCACGGAAG 14300

14301 AGAGGTTAAAACTCTTTGAC CGTTATTTTAAATATTGGGA TCAGACATACCACCCAAATT GTGTTAACTGTTTGGATGAC AGATGCATTCTGCATTGTGC 14400

14401 AAACTTTAATGTTTTATTCT CTACAGTGTTCCCACTTACA AGTTTTGGACCACTAGTGAG AAAAATATTTGTTGATGGTG TTCCATTTGTAGTTTCAACT 14500

14501 GGATACCACTTCAGAGAGCT AGGTGTTGTACATAATCAGG ATGTAAACTTACATAGCTCT AGACTTAGTTTTAAGGAATT ACTTGTGTATGCTGCTGACC 14600

14601 CTGCTATGCACGCTGCTTCT GGTAATCTATTACTAGATAA ACGCACTACGTGCTTTTCAG TAGCTGCACTTACTAACAAT GTTGCTTTTCAAACTGTCAA 14700

14701 ACCCGGTAATTTTAACAAAG ACTTCTATGACTTTGCTGTG TCTAAGGGTTTCTTTAAGGA AGGAAGTTCTGTTGAATTAA AACACTTCTTCTTTGCTCAG 14800

14801 GATGGTAATGCTGCTATCAG CGATTATGACTACTATCGTT ATAATCTACCAACAATGTGT GATATCAGACAACTACTATT TGTAGTTGAAGTTGTTGATA 14900

14901 AGTACTTTGATTGTTACGAT GGTGGCTGTATTAATGCTAA CCAAGTCATCGTCAACAACC TAGACAAATCAGCTGGTTTT CCATTTAATAAATGGGGTAA 15000

15001 GGCTAGACTTTATTATGATT CAATGAGTTATGAGGATCAA GATGCACTTTTCGCATATAC AAAACGTAATGTCATCCCTA CTATAACTCAAATGAATCAT 15100

15101 AAGTATGCCATTAGTGCAAA GAATAGAGCTCGCACCGTAG CTGGTGTCTCTATCTGTAGT ACTATGACCAATAGACAGTT TCATCAAAAATTATTGAAAT 15200

15201 CAATAGCCGCCACTAGAGGA GCTACTGTAGTAATTGGAAC AAGCAAATTCTATGGTGGTT GGCACAACATGTTAAAAACT GTTTATAGTGATGTAGAAAA 15300

15301 CCCTCACCTTATGGGTTGGG ATTATCCTAAATGTGATAGA GCCATGCCTTACATGCTTAG AATTATGGCCTCACTTGTTC TTGCTCGCAAACATACAACG 15400

15401 TGTTGTAGCTTGTCACACCG TTTCTATAGATTAGCTAATG AGTGTGCTCAAGTATTGAGT GAAATGGTCATGTGTGGCGG TTCACTATATGTTAAACCAG 15500

15501 GTGGAACCTCATCAGGAGAT GCCACAACTGCTTATGCTAA TAGTGTTTTTAACATTTGTC AAGCTGTCACGGCCAATGTT AATGCACTTTTATCTACTGA 15600

15601 TGGTAACAAAATTGCCGATA AGTATGTCCGCAATTTACAA CACAGACTTTATGAGTGTCT CTATAGAAATAGAGATGTTG ACACAGACTTTGTGAATGAG 15700

15701 TTTTACGCATATTTGCGTAA ACATTTCTCAATGATGATAC TCTCTGACGATGCTGTTGTG TGTTTCAATAGCACTTATGC ATCTCAAGGTCTAGTGGCTA 15800

15801 GCATAAAGAACTTTAAGTCA GTTCTTTATTATCAAAACAA TGTTTTTATGTCTGTAGCAA AATGTTGGACTGAGACTGAC CTTACTAAAGGACCTCATGA 15900

15901 ATTTTGCTCTCAACATACAA TGCTAGTTAAACAGGGTGAT GATTATGTGTACCTTCCTTA CCCAGATCCATCAAGAATCC TAGGGGCCGGCTGTTTTGTA 16000

16001 GATGATATCGTAAAAACAGA TGGTACACTTATGATTGAAC GGTTCGTGTCTTTAGCTATA GATGCTTACCCACTTACTAA ACATCCTAATCAGGAGTATG 16100

16101 CTGATGTCTTTCATTTGTAC TTACAATACATAAGAAAGCT ACATGATGAGTTAACAGGAC ACATGTTAGACATGTATTCT GTTATGCTTACTAATGATAA 16200

16201 CACTTCAAGGTATTGGGAAC CTGAGTTTTATGAGGCTATG TACACACCGCATACAGTCTT ACAGGCTGTTGGGGCTTGTG TTCTTTGCAATTCACAGACT 16300

16301 TCATTAAGATGTGGTGCTTG CATACGTAGACCATTCTTAT GTTGTAAATGCTGTTACGAC CATGTCATATCAACATCACA TAAATTAGTCTTGTCTGTTA 16400

16401 ATCCGTATGTTTGCAATGCT CCAGGTTGTGATGTCACAGA TGTGACTCAACTTTACTTAG GAGGTATGAGCTATTATTGT AAATCACATAAACCACCCAT 16500

16501 TAGTTTTCCATTGTGTGCTA ATGGACAAGTTTTTGGTTTA TATAAAAATACATGTGTTGG TAGCGATAATGTTACTGACT TTAATGCAATTGCAACATGT 16600

16601 GACTGGACAAATGCTGGTGA TTACATTTTAGCTAACACCT GTACTGAAAGACTCAAGCTT TTTGCAGCAGAAACGCTCAA AGCTACTGAGGAGACATTTA 16700

16701 AACTGTCTTATGGTATTGCT ACTGTACGTGAAGTGCTGTC TGACAGAGAATTACATCTTT CATGGGAAGTTGGTAAACCT AGACCACCACTTAACCGAAA 16800

16801 TTATGTCTTTACTGGTTATC GTGTAACTAAAAACAGTAAA GTACAAATAGGAGAGTACAC CTTTGAAAAAGGTGACTATG GTGATGCTGTTGTTTACCGA 16900

16901 GGTACAACAACTTACAAATT AAATGTTGGTGATTATTTTG TGCTGACATCACATACAGTA ATGCCATTAAGTGCACCTAC ACTAGTGCCACAAGAGCACT 17000

17001 ATGTTAGAATTACTGGCTTA TACCCAACACTCAATATCTC AGATGAGTTTTCTAGCAATG TTGCAAATTATCAAAAGGTT GGTATGCAAAAGTATTCTAC 17100

17101 ACTCCAGGGACCACCTGGTA CTGGTAAGAGTCATTTTGCT ATTGGCCTAGCTCTCCACTA CCCTTCTGCTCGCATAGTGT ATACAGCTTGCTCTCATGCC 17200

17201 GCTGTTGATGCACTATGTGA GAAGGCATTAAAATATTTGC CTATAGATAAATGTAGTAGA ATTATACCTGCACGTGCTCG TGTAGAGTGTTTTGATAAAT 17300

17301 TCAAAGTGAATTCAACATTA GAACAGTATGTCTTTTGTAC TGTAAATGCATTGCCTGAGA CGACAGCAGATATAGTTGTC TTTGATGAAATTTCAATGGC 17400

17401 CACAAATTATGATTTGAGTG TTGTCAATGCCAGATTACGT GCTAAGCACTATGTGTACAT TGGCGACCCTGCTCAATTAC CTACACCACGCACATTGCTA 17500

17501 ACTAAGGGCACACTAGAACC AGAATATTTCAATTCAGTGT GTAGACTTATGAAAACTATA GGTCCAGACATGTTCCTCGG AACTTGTCTGCGTTGTCCTG 17600

17601 CTGAAATTGTTGACACTGTG AGTGCTTTGGTTTATGATAA TAAGCTTAAAGCACATAAAG ACAAATCAGCTCAATGCTTT AAAATGTTTTATAAGGGTGT 17700

17701 TATCACGCATGATGTTTCAT CTGCAATTAACAGGCCACAA ATAGGCGTGGTAAGAGAATT CCTTACACGTAACCCTGCTT GGAGAAAAGCTGTCTTTATT 17800

17801 TCACCTTATAATTCACAGAA TGCTGTAGCCTCAAAGATTT TGGGACTACCAACTCAAACT GTTGATTCATCACAGGGCTC AGAATATGACTATGTCATAT 17900

17901 TCACTCAAACCACTGAAACA GCTCACTCTTGTAATGTAAA CAGATTTAATGTTGCTATTA CCAGAGCAAAAGTAGGCATA CTTTGCATAATGTCTGATAG 18000

18001 AGACCTTTATGACAAGTTGC AATTTACAAGTCTTGAAATT CCACGTAGGAATGTGGCAAC TTTACAAGCTGAAAATGTAA CAGGACTCTTTAAAGATTGT 18100

18101 AGTAAGGTAATCACTGGGTT ACATCCTACACAGGCACCTA CACACCTCAGTGTTGACACT AAATTCAAAACTGAAGGTTT ATGTGTTGACATACCTGGCA 18200

18201 TACCTAAGGACATGACCTAT AGAAGACTCATCTCTATGAT GGGTTTTAAAATGAATTATC AAGTTAATGGTTACCCTAAC ATGTTTATCACCCGCGAAGA 18300

18301 AGCTATAAGACATGTACGTG CATGGATTGGCTTCGATGTC GAGGGGTGTCATGCTACTAG AGAAGCTGTTGGTACCAATT TACCTTTACAGCTAGGTTTT 18400

18401 TCTACAGGTGTTAACCTAGT TGCTGTACCTACAGGTTATG TTGATACACCTAATAATACA GATTTTTCCAGAGTTGGTGC TAAACCACCGCCTGGAGATC 18500

18501 AATTTAAACACCTCATACCA CTTATGTACAAAGGACTTCC TTGGAATGTAGTGCGTATAA AGATTGTACAAATGTTAAGT GACACACTTAAAAATCTCTC 18600

18601 TGACAGAGTCGTATTTGTCT TATGGGCACATGGCTTTGAG TTGACATCTATGAAGTATTT TGTGAAAATAGGACCTGAGC GCACCTGTTGTCTATGTGAT 18700

18701 AGACGTGCCACATGCTTTTC CACTGCTTCAGACACTTATG CCTGTTGGCATCATTCTATT GGATTTGATTACGTCTATAA TCCGTTTATGATTGATGTTC 18800

18801 AACAATGGGGTTTTACAGGT AACCTACAAAGCAACCATGA TCTGTATTGTCAAGTCCATG GTAATGCACATGTAGCTAGT TGTGATGCAATCATGACTAG 18900

18901 GTGTCTAGCTGTCCACGAGT GCTTTGTTAAGCGTGTTGAC TGGACTATTGAATATCCTAT AATTGGTGATGAACTGAAGA TTAATGCGGCTTGTAGAAAG 19000

19001 GTTCAACACATGGTTGTTAA AGCTGCATTATTAGCAGACA AATTCCCAGTTCTTCACGAC ATTGGTAACCCTAAAGCTAT TAAGTGTGTACCTCAAGCTG 19100

19101 ATGTAGAATGGAAGTTCTAT GATGCACAGCCTTGTAGTGA CAAAGCTTATAAAATAGAAG AATTATTCTATTCTTATGCC ACACATTCTGACAAATTCAC 19200

19201 AGATGGTGTATGCCTATTTT GGAATTGCAATGTCGATAGA TATCCTGCTAATTCCATTGT TTGTAGATTTGACACTAGAG TGCTATCTAACCTTAACTTG 19300

19301 CCTGGTTGTGATGGTGGCAG TTTGTATGTAAATAAACATG CATTCCACACACCAGCTTTT GATAAAAGTGCTTTTGTTAA TTTAAAACAATTACCATTTT 19400

19401 TCTATTACTCTGACAGTCCA TGTGAGTCTCATGGAAAACA AGTAGTGTCAGATATAGATT ATGTACCACTAAAGTCTGCT ACGTGTATAACACGTTGCAA 19500

19501 TTTAGGTGGTGCTGTCTGTA GACATCATGCTAATGAGTAC AGATTGTATCTCGATGCTTA TAACATGATGATCTCAGCTG GCTTTAGCTTGTGGGTTTAC 19600

19601 AAACAATTTGATACTTATAA CCTCTGGAACACTTTTACAA GACTTCAGAGTTTAGAAAAT GTGGCTTTTAATGTTGTAAA TAAGGGACACTTTGATGGAC 19700

19701 AACAGGGTGAAGTACCAGTT TCTATCATTAATAACACTGT TTACACAAAAGTTGATGGTG TTGATGTAGAATTGTTTGAA AATAAAACAACATTACCTGT 19800

19801 TAATGTAGCATTTGAGCTTT GGGCTAAGCGCAACATTAAA CCAGTACCAGAGGTGAAAAT ACTCAATAATTTGGGTGTGG ACATTGCTGCTAATACTGTG 19900

19901 ATCTGGGACTACAAAAGAGA TGCTCCAGCACATATATCTA CTATTGGTGTTTGTTCTATG ACTGACATAGCCAAGAAACC AACTGAAACGATTTGTGCAC 20000

20001 CACTCACTGTCTTTTTTGAT GGTAGAGTTGATGGTCAAGT AGACTTATTTAGAAATGCCC GTAATGGTGTTCTTATTACA GAAGGTAGTGTTAAAGGTTT 20100

20101 ACAACCATCTGTAGGTCCCA AACAAGCTAGTCTTAATGGA GTCACATTAATTGGAGAAGC CGTAAAAACACAGTTCAATT ATTATAAGAAAGTTGATGGT 20200

20201 GTTGTCCAACAATTACCTGA AACTTACTTTACTCAGAGTA GAAATTTACAAGAATTTAAA CCCAGGAGTCAAATGGAAAT TGATTTCTTAGAATTAGCTA 20300

20301 TGGATGAATTCATTGAACGG TATAAATTAGAAGGCTATGC CTTCGAACATATCGTTTATG GAGATTTTAGTCATAGTCAG TTAGGTGGTTTACATCTACT 20400

20401 GATTGGACTAGCTAAACGTT TTAAGGAATCACCTTTTGAA TTAGAAGATTTTATTCCTAT GGACAGTACAGTTAAAAACT ATTTCATAACAGATGCGCAA 20500

20501 ACAGGTTCATCTAAGTGTGT GTGTTCTGTTATTGATTTAT TACTTGATGATTTTGTTGAA ATAATAAAATCCCAAGATTT ATCTGTAGTTTCTAAGGTTG 20600

20601 TCAAAGTGACTATTGACTAT ACAGAAATTTCATTTATGCT TTGGTGTAAAGATGGCCATG TAGAAACATTTTACCCAAAA TTACAATCTAGTCAAGCGTG 20700

20701 GCAACCGGGTGTTGCTATGC CTAATCTTTACAAAATGCAA AGAATGCTATTAGAAAAGTG TGACCTTCAAAATTATGGTG ATAGTGCAACATTACCTAAA 20800

20801 GGCATAATGATGAATGTCGC AAAATATACTCAACTGTGTC AATATTTAAACACATTAACA TTAGCTGTACCCTATAATAT GAGAGTTATACATTTTGGTG 20900

20901 CTGGTTCTGATAAAGGAGTT GCACCAGGTACAGCTGTTTT AAGACAGTGGTTGCCTACGG GTACGCTGCTTGTCGATTCA GATCTTAATGACTTTGTCTC 21000

21001 TGATGCAGATTCAACTTTGA TTGGTGATTGTGCAACTGTA CATACAGCTAATAAATGGGA TCTCATTATTAGTGATATGT ACGACCCTAAGACTAAAAAT 21100

21101 GTTACAAAAGAAAATGACTC TAAAGAGGGTTTTTTCACTT ACATTTGTGGGTTTATACAA CAAAAGCTAGCTCTTGGAGG TTCCGTGGCTATAAAGATAA 21200

21201 CAGAACATTCTTGGAATGCT GATCTTTATAAGCTCATGGG ACACTTCGCATGGTGGACAG CCTTTGTTACTAATGTAAAT GCGTCATCATCTGAAGCATT 21300

21301 TTTAATTGGATGTAATTATC TTGGCAAACCACGCGAACAA ATAGATGGTTATGTCATGCA TGCAAATTACATATTTTGGA GGAATACAAATCCAATTCAG 21400

21401 TTGTCTTCCTATTCTTTATT TGACATGAGTAAATTTCCCC TTAAATTAAGGGGTACTGCT GTTATGTCTTTAAAAGAAGG TCAAATCAATGATATGATTT 21500

21501 TATCTCTTCTTAGTAAAGGT AGACTTATAATTAGAGAAAA CAACAGAGTTGTTATTTCTA GTGATGTTCTTGTTAACAAC TAAACGAACAATGTTTGTTT 21600

21601 TTCTTGTTTTATTGCCACTA GACTCTAGTCAGTGTGTTAA TCTTACAACCAGAACTCAAT TACCCCCTGCATACACTAAT TCTTTCACACGTGGTGTTTA 21700

21701 TTACCCTGACAAAGTTTTCA GATCCTCAGTTTTACATTCA ACTCAGGACTTGTTCTTACC TTTCTTTTCCAATGTTACTT GGTTCCATGCTATACATGTC 21800

21801 TCTGGGACCAATGGTACTAA GAGGTTTGATAACCCTGTCC TACCATTTAATGATGGTGTT TATTTTGCTTCCACTGAGAA GTCTAACATAATAAGAGGCT 21900

21901 GGATTTTTGGTACTACTTTA GATTCGAAGACCCAGTCCCT ACTTATTGTTAATAACGCTA CTAATGTTGTTATTAAAGTC TGTGAATTTCAATTTTGTAA 22000

22001 TGATCCATTTTTGGGTGTTT ATTACCACAAAAACAACAAA AGTTGGATGGAAAGTGAGTT CAGAGTTTATTCTAGTGCGA ATAATTGCACTTTTGAATAT 22100

22101 GTCTCTCAGCCTTTTCTTAT GGACCTTGAAGGAAAACAGG GTAATTTCAAAAATCTTAGG GAATTTGTGTTTAAGAATAT TGATGGTTATTTTAAAATAT 22200

22201 ATTCTAAGCACACGCCTATT AATTTAGTGCGTGATCTCCC TCAGGGTTTTTCGGCTTTAG AACCATTGGTAGATTTGCCA ATGGGTATTAACATCACTAG 22300

22301 GTTTCAAACTTTACTTGCTT TACATAGAAGTTATTTGACT CCTGGTGATTCTTCTTCAGG TTGGACAGCTGGTGCTGCAG CTTATTATGTGGGTTATCTT 22400

22401 CAACCTAGGACTTTTCTATT AAAATATAATGAAAATGGAA CCATTACAGATGCTGTAGAC TGTGCACTTGACCCTCTCTC AGAAACAAAGTGTACGTTGA 22500

22501 AATCCTTCACTGTAGAAAAA GGAATCTATCAAACTTCTAA CTTTAGAGTCCAACCAACAG AATCTATTGTTAGATTTCCT AATATTACAAACTTGTGCCC 22600

22601 TTTTGGTGAAGTTTTTAACG CCACCAGATTTGCATCTGTT TATGCTTGGAACAGGAAGAG AATCAGCAACTGTGTTGCTG ATTATTCTGTCCTATATAAT 22700

22701 TCCGCATCATTTTCCACTTT TAAGTGTTATGGAGTGTCTC CTACTAAATTAAATGATCTC TGCTTTACTAATGTCTATGC AGATTCATTTGTAATTAGAG 22800

22801 GTGATGAAGTCAGACAAATC GCTCCAGGGCAAACTGGAAA GATTGCTGATTATAATTATA AATTACCAGATGATTTTACA GGCTGCGTTATAGCTTGGAA 22900

22901 TTCTAACAATCTTGATTCTA AGGTTGGTGGTAATTATAAT TACCTGTATAGATTGTTTAG GAAGTCTAATCTCAAACCTT TTGAGAGAGATATTTCAACT 23000

23001 GAAATCTATCAGGCCGGTAG CACACCTTGTAATGGTGTTG AAGGTTTTAATTGTTACTTT CCTTTACAATCATATGGTTT CCAACCCACTAATGGTGTTG 23100

23101 GTTACCAACCATACAGAGTA GTAGTACTTTCTTTTGAACT TCTACATGCACCAGCAACTG TTTGTGGACCTAAAAAGTCT ACTAATTTGGTTAAAAACAA 23200

23201 ATGTGTCAATTTCAACTTCA ATGGTTTAACAGGCACAGGT GTTCTTACTGAGTCTAACAA AAAGTTTCTGCCTTTCCAAC AATTTGGCAGAGACATTGCT 23300

23301 GACACTACTGATGCTGTCCG TGATCCACAGACACTTGAGA TTCTTGACATTACACCATGT TCTTTTGGTGGTGTCAGTGT TATAACACCAGGAACAAATA 23400

23401 CTTCTAACCAGGTTGCTGTT CTTTATCAGGGTGTTAACTG CACAGAAGTCCCTGTTGCTA TTCATGCAGATCAACTTACT CCTACTTGGCGTGTTTATTC 23500

23501 TACAGGTTCTAATGTTTTTC AAACACGTGCAGGCTGTTTA ATAGGGGCTGAACATGTCAA CAACTCATATGAGTGTGACA TACCCATTGGTGCAGGTATA 23600

23601 TGCGCTAGTTATCAGACTCA GACTAATTCTCCTCGGCGGG CACGTAGTGTAGCTAGTCAA TCCATCATTGCCTACACTAT GTCACTTGGTGCAGAAAATT 23700

23701 CAGTTGCTTACTCTAATAAC TCTATTGCCATACCCACAAA TTTTACTATTAGTGTTACCA CAGAAATTCTACCAGTGTCT ATGACCAAGACATCAGTAGA 23800

23801 TTGTACAATGTACATTTGTG GTGATTCAACTGAATGCAGC AATCTTTTGTTGCAATATGG CAGTTTTTGTACACAATTAA ACCGTGCTTTAACTGGAATA 23900

23901 GCTGTTGAACAAGACAAAAA CACCCAAGAAGTTTTTGCAC AAGTCAAACAAATTTACAAA ACACCACCAATTAAAGATTT TGGTGGTTTTAATTTTTCAC 24000

24001 AAATATTACCAGATCCATCA AAACCAAGCAAGAGGTCATT TATTGAAGATCTACTTTTCA ACAAAGTGACACTTGCAGAT GCTGGCTTCATCAAACAATA 24100

24101 TGGTGATTGCCTTGGTGATA TTGCTGCTAGAGACCTCATT TGTGCACAAAAGTTTAACGG CCTTACTGTTTTGCCACCTT TGCTCACAGATGAAATGATT 24200

24201 GCTCAATACACTTCTGCACT GTTAGCGGGTACAATCACTT CTGGTTGGACCTTTGGTGCA GGTGCTGCATTACAAATACC ATTTGCTATGCAAATGGCTT 24300

24301 ATAGGTTTAATGGTATTGGA GTTACACAGAATGTTCTCTA TGAGAACCAAAAATTGATTG CCAACCAATTTAATAGTGCT ATTGGCAAAATTCAAGACTC 24400

24401 ACTTTCTTCCACAGCAAGTG CACTTGGAAAACTTCAAGAT GTGGTCAACCAAAATGCACA AGCTTTAAACACGCTTGTTA AACAACTTAGCTCCAATTTT 24500

24501 GGTGCAATTTCAAGTGTTTT AAATGATATCCTTTCACGTC TTGACAAAGTTGAGGCTGAA GTGCAAATTGATAGGTTGAT CACAGGCAGACTTCAAAGTT 24600

24601 TGCAGACATATGTGACTCAA CAATTAATTAGAGCTGCAGA AATCAGAGCTTCTGCTAATC TTGCTGCTACTAAAATGTCA GAGTGTGTACTTGGACAATC 24700

24701 AAAAAGAGTTGATTTTTGTG GAAAGGGCTATCATCTTATG TCCTTCCCTCAGTCAGCACC TCATGGTGTAGTCTTCTTGC ATGTGACTTATGTCCCTGCA 24800

24801 CAAGAAAAGAACTTCACAAC TGCTCCTGCCATTTGTCATG ATGGAAAAGCACACTTTCCT CGTGAAGGTGTCTTTGTTTC AAATGGCACACACTGGTTTG 24900

24901 TAACACAAAGGAATTTTTAT GAACCACAAATCATTACTAC AGACAACACATTTGTGTCTG GTAACTGTGATGTTGTAATA GGAATTGTCAACAACACAGT 25000

25001 TTATGATCCTTTGCAACCTG AATTAGACTCATTCAAGGAG GAGTTAGATAAATATTTTAA GAATCATACATCACCAGATG TTGATTTAGGTGACATCTCT 25100

25101 GGCATTAATGCTTCAGTTGT AAACATTCAAAAAGAAATTG ACCGCCTCAATGAGGTTGCC AAGAATTTAAATGAATCTCT CATCGATCTCCAAGAACTTG 25200

25201 GAAAGTATGAGCAGTATATA AAATGGCCATGGTACATTTG GCTAGGTTTTATAGCTGGCT TGATTGCCATAGTAATGGTG ACAATTATGCTTTGCTGTAT 25300

25301 GACCAGTTGCTGTAGTTGTC TCAAGGGCTGTTGTTCTTGT GGATCCTGCTGCAAATTTGA TGAAGACGACTCTGAGCCAG TGCTCAAAGGAGTCAAATTA 25400

25401 CATTACACATAAACGAACTT ATGGATTTGTTTATGAGAAT CTTCACAATTGGAACTGTAA CTTTGAAGCAAGGTGAAATC AAGGATGCTACTCCTTCAGA 25500

25501 TTTTGTTCGCGCTACTGCAA CGATACCGATACAAGCCTCA CTCCCTTTCGGATGGCTTAT TGTTGGCGTTGCACTTCTTG CTGTTTTTCATAGCGCTTCC 25600

25601 AAAATCATAACCCTCAAAAA GAGATGGCAACTAGCACTCT CCAAGGGTGTTCACTTTGTT TGCAACTTGCTGTTGTTGTT TGTAACAGTTTACTCACACC 25700

25701 TTTTGCTCGTTGCTGCTGGC CTTGAAGCCCCTTTTCTCTA TCTTTATGCTTTAGTCTACT TCTTGCAGAGTATAAACTTT GTAAGAATAATAATGAGGCT 25800

25801 TTGGCTTTGCTGGAAATGCC GTTCCAAAAACCCATTACTT TATGATGCCAACTATTTTCT TTGCTGGCATACTAATTGTT ACGACTATTGTATACCTTAC 25900

25901 AATAGTGTAACTTCTTCAAT TGTCATTACTTCAGGTGATG GCACAACAAGTCCTATTTCT GAACATGACTACCAGATTGG TGGTTATACTGAAAAATGGG 26000

26001 AATCTGGAGTAAAAGACTGT GTTGTATTACACAGTTACTT CACTTCAGACTATTACCAGC TGTACTCAACTCAATTGAGT ACAGACACTGGTGTTGAACA 26100

26101 TGTTACCTTCTTCATCTACA ATAAAATTGTTGATGAGCCT GAAGAACATGTCCAAATTCA CACAATCGACGGTTCATCCG GAGTTGTTAATCCAGTAATG 26200

26201 GAACCAATTTATGATGAACC GACGACGACTACTAGCGTGC CTTTGTAAGCACAAGCTGAT GAGTACGAACTTATGTACTC ATTCGTTTCGGAAGAGACAG 26300

26301 GTACGTTAATAGTTAATAGC GTACTTCTTTTTCTTGCTTT CGTGGTATTCTTGCTAGTTA CACTAGCCATCCTTACTGCG CTTCGATTGTGTGCGTACTG 26400

26401 CTGCAATATTGTTAACGTGA GTCTTGTAAAACCTTCTTTT TACGTTTACTCTCGTGTTAA AAATCTGAATTCTTCTAGAG TTCCTGATCTTCTGGTCTAA 26500

26501 ACGAACTAAATATTATATTA GTTTTTCTGTTTGGAACTTT AATTTTAGCCATGGCAGATT CCAACGGTACTATTACCGTT GAAGAGCTTAAAAAGCTCCT 26600

26601 TGAACAATGGAACCTAGTAA TAGGTTTCCTATTCCTTACA TGGATTTGTCTTCTACAATT TGCCTATGCCAACAGGAATA GGTTTTTGTATATAATTAAG 26700

26701 TTAATTTTCCTCTGGCTGTT ATGGCCAGTAACTTTAGCTT GTTTTGTGCTTGCTGCTGTT TACAGAATAAATTGGATCAC CGGTGGAATTGCTATCGCAA 26800

26801 TGGCTTGTCTTGTAGGCTTG ATGTGGCTCAGCTACTTCAT TGCTTCTTTCAGACTGTTTG CGCGTACGCGTTCCATGTGG TCATTCAATCCAGAAACTAA 26900

26901 CATTCTTCTCAACGTGCCAC TCCATGGCACTATTCTGACC AGACCGCTTCTAGAAAGTGA ACTCGTAATCGGAGCTGTGA TCCTTCGTGGACATCTTCGT 27000

27001 ATTGCTGGACACCATCTAGG ACGCTGTGACATCAAGGACC TGCCTAAAGAAATCACTGTT GCTACATCACGAACGCTTTC TTATTACAAATTGGGAGCTT 27100

27101 CGCAGCGTGTAGCAGGTGAC TCAGGTTTTGCTGCATACAG TCGCTACAGGATTGGCAACT ATAAATTAAACACAGACCAT TCCAGTAGCAGTGACAATAT 27200

27201 TGCTTTGCTTGTACAGTAAG TGACAACAGATGTTTCATCT CGTTGACTTTCAGGTTACTA TAGCAGAGATATTACTAATT ATTATGAGGACTTTTAAAGT 27300

27301 TTCCATTTGGAATCTTGATT ACATCATAAACCTCATAATT AAAAATTTATCTAAGTCACT AACTGAGAATAAATATTCTC AATTAGATGAAGAGCAACCA 27400

27401 ATGGAGATTGATTAAACGAA CATGAAAATTATTCTTTTCT TGGCACTGATAACACTCGCT ACTTGTGAGCTTTATCACTA CCAAGAGTGTGTTAGAGGTA 27500

27501 CAACAGTACTTTTAAAAGAA CCTTGCTCTTCTGGAACATA CGAGGGCAATTCACCATTTC ATCCTCTAGCTGATAACAAA TTTGCACTGACTTGCTTTAG 27600

27601 CACTCAATTTGCTTTTGCTT GTCCTGACGGCGTAAAACAC GTCTATCAGTTACGTGCCAG ATCAGTTTCACCTAAACTGT TCATCAGACAAGAGGAAGTT 27700

27701 CAAGAACTTTACTCTCCAAT TTTTCTTATTGTTGCGGCAA TAGTGTTTATAACACTTTGC TTCACACTCAAAAGAAAAAC AGAATGATTGAACTTTCATT 27800

27801 AATTGACTTCTATTTGTGCT TTTTAGCCTTTCTGCTATTC CTTGTTTTAATTATGCTTAT TATCTTTTGGTTCTCACTTG AACTGCAAGATCATAATGAA 27900

27901 ACTTGTCACGCCTAAACGAA CATGAAATTTCTTGTTTTCT TAGGAATCATCACAACTGTA GCTGCATTTCACCAAGAATG TAGTTTACAGTCATGTACTC 28000

28001 AACATCAACCATATGTAGTT GATGACCCGTGTCCTATTCA CTTCTATTCTAAATGGTATA TTAGAGTAGGAGCTAGAAAA TCAGCACCTTTAATTGAATT 28100

28101 GTGCGTGGATGAGGTTGGTT CTAAATCACCCATTCAGTAC ATCGATATCGGTAATTATAC AGTTTCCTGTTTACCTTTTA CAATTAATTGCCAGGAACCT 28200

28201 AAATTGGGTAGTCTTGTAGT GCGTTGTTCGTTCTATGAAG ACTTTTTAGAGTATCATGAC GTTCGTGTTGTTTTAGATTT CATCTAAACGAACAAACTAA 28300

28301 AATGTCTGATAATGGACCCC AAAATCAGCGAAATGCACCC CGCATTACGTTTGGTGGACC CTCAGATTCAACTGGCAGTA ACCAGAATGGAGAACGCAGT 28400

28401 GGGGCGCGATCAAAACAACG TCGGCCCCAAGGTTTACCCA ATAATACTGCGTCTTGGTTC ACCGCTCTCACTCAACATGG CAAGGAAGACCTTAAATTCC 28500

28501 CTCGAGGACAAGGCGTTCCA ATTAACACCAATAGCAGTCC AGATGACCAAATTGGCTACT ACCGAAGAGCTACCAGACGA ATTCGTGGTGGTGACGGTAA 28600

28601 AATGAAAGATCTCAGTCCAA GATGGTATTTCTACTACCTA GGAACTGGGCCAGAAGCTGG ACTTCCCTATGGTGCTAACA AAGACGGCATCATATGGGTT 28700

28701 GCAACTGAGGGAGCCTTGAA TACACCAAAAGATCACATTG GCACCCGCAATCCTGCTAAC AATGCTGCAATCGTGCTACA ACTTCCTCAAGGAACAACAT 28800

28801 TGCCAAAAGGCTTCTACGCA GAAGGGAGCAGAGGCGGCAG TCAAGCCTCTGCTCGTTCCT CATCACGTAGTCGCAACAGT TCAAGAAATTCAACTCCAGG
28900

28901 CAGCAGTAGGGGAACTTCTC CTGCTAGAATGGCTGGCAAT GGCGGTGATGCTGCTCTTGC TTTGCTGCTGCTTGACAGAT TGAACCAGCTTGAGAGCAAA
29000

29001 ATGTCTGGTAAAGGCCAACA ACAACAAGGCCAAACTGTCA CTAAGAAATCTGCTGCTGAG GCTTCTAAGAAGCCTCGGCA AAAACGTACTGCCACTAAAG 29100

29101 CATACAATGTAACACAAGCT TTCGGCAGACGTGGTCCAGA ACAAACCCAAGGAAATTTTG GGGACCAGGAACTAATCAGA CAAGGAACTGATTACAAACA 29200

29201 TTGGCCGCAAATTGCACAAT TTGCCCCCAGCGCTTCAGCG TTCTTCGGAATGTCGCGCAT TGGCATGGAAGTCACACCTT CGGGAACGTGGTTGACCTAC

29300

29301 ACAGGTGCCATCAAATTGGA TGACAAAGATCCAAATTTCA AAGATCAAGTCATTTTGCTG AATAAGCATATTGACGCATA CAAAACATTCCCACCAACAG 29400

29401 AGCCTAAAAAAGGACAAAAAG AAGAAGGCTGATGAAACTCA AGCCTTACCGCAGAGACAGA AGAAACAGCAAACTGTGACT CTTCTTCCTGCTGCAGATTT

29500

29501 GGATGATTTCTCCAAACAAT TGCAACAATCCATGAGCAGT GCTGACTCAACTCAGGCCTA AACTCATGCAGACCACACAA GGCAGATGGGCTATATAAAC 29600

29601 GTTTTCGCTTTTCCGTTTAC GATATATAGTCTACTCTTGT GCAGAATGAATTCTCGTAAC TACATAGCACAAGTAGATGT AGTTAACTTTAATCTCACAT 29700

29701 AGCAATCTTTAATCAGTGTG TAACATTAGGGAGGACTTGA AAGAGCCACCACATTTTCAC CGAGGCCACGCGGAGTACGA TCGAGTGTACAGTGAACAAT 29800

29801 GCTAGGGAGAGCTGCCTATA TGGAAGAGCCCTAATGTGTA AAATTAATTTTAGTAGTGCT ATCCCCATGTGATTTTAATA GCTTCTTAGGAGAATGACAA 29900

29901 AAAAAAAATCACATGGGGAT AGCAC 29925

```
EMBOSS_001         1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHS     50
                     |||||||||.||||||||||||||||||||||||||||||||||||||||
EMBOSS_001         1 MFVFLVLLSLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHS     50

EMBOSS_001        51 TQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNI    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001        51 TQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNI    100

EMBOSS_001       101 IRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNK    150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       101 IRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNK    150

EMBOSS_001       151 SWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGY    200
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       151 SWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGY    200

EMBOSS_001       201 FKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLT    250
                     ||||||||||||||||||||||||||||||||||||||.|||||||||||||
EMBOSS_001       201 FKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRLQTLLALHRSYLT    250

EMBOSS_001       251 PGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETK    300
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       251 PGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETK    300

EMBOSS_001       301 CTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASV    350
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       301 CTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASV    350

EMBOSS_001       351 YAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF    400
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       351 YAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF    400

EMBOSS_001       401 VIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN    450
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       401 VIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN    450

EMBOSS_001       451 YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPT    500
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       451 YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPT    500

EMBOSS_001       501 NGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTG    550
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       501 NGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTG    550

EMBOSS_001       551 VLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP    600
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       551 VLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP    600
```

Figure 7.1 Difference of discovered strain 1 spike protein sequence and original strain spike protein sequence

```
EMBOSS_001       601 GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCL    650
                     ..:||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       601 EVDTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCL    650

EMBOSS_001       651 IGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLG    700
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       651 IGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLG    700

EMBOSS_001       701 AENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECS    750
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       701 AENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECS    750

EMBOSS_001       751 NLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGF    800
                     ||||||||||||||.||||||||||||||||||||||||||||||||||||
EMBOSS_001       751 NLLLQYGSFCTQLNRGLTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGF    800

EMBOSS_001       801 NFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLI    850
                     ||||||||||||||||||||||||||||||||||||||||||||||||||.|
EMBOSS_001       801 NFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDHI    850

EMBOSS_001       851 CAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM    900
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       851 CAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM    900

EMBOSS_001       901 QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQD    950
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       901 QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQD    950

EMBOSS_001       951 VVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGR   1000
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       951 VVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGR   1000

EMBOSS_001      1001 LQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLM   1050
                     ||||||||||||||||||||||||||||||..:||||||||||||||||||
EMBOSS_001      1001 LQSLQTYVTQQLIRAAEIRASANLAATKMSLRLLGQSKRVDFCGKGYHLM   1050

EMBOSS_001      1051 SFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGT   1100
                     ||||||||||||||||||||||||||||||...||||||||||||||||||
EMBOSS_001      1051 SFPQSAPHGVVFLHVTYVPAQEKNFTTAPAKSDDGKAHFPREGVFVSNGT   1100

EMBOSS_001      1101 HWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKE   1150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      1101 HWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKE   1150

EMBOSS_001      1151 ELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL   1200
                     ||||||||||||.||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      1151 ELDKYFKNHTSADVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL   1200

EMBOSS_001      1201 QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSC   1250
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      1201 QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSC   1250

EMBOSS_001      1251 GSCCKFDEDDSEPVLKGVKLHYT   1273
                     |||||||||||||||||||||||
EMBOSS_001      1251 GSCCKFDEDDSEPVLKGVKLHYT   1273
```

Figure 7.1 Difference of discovered strain 1 spike protein sequence and original strain spike protein sequence

```
EMBOSS_001         1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHS     50
                     ||||||||||||||||||||||||||.:.||||||||||||||||||||||
EMBOSS_001         1 MFVFLVLLPLVSSQCVNLTTRTQLPKSDTNSFTRGVYYPDKVFRSSVLHS     50

EMBOSS_001        51 TQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNI    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001        51 TQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNI    100

EMBOSS_001       101 IRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNK    150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       101 IRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNK    150

EMBOSS_001       151 SWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGY    200
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       151 SWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGY    200

EMBOSS_001       201 FKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLT    250
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       201 FKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLT    250

EMBOSS_001       251 PGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETK    300
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       251 PGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETK    300

EMBOSS_001       301 CTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASV    350
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       301 CTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASV    350

EMBOSS_001       351 YAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF    400
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       351 YAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF    400

EMBOSS_001       401 VIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN    450
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       401 VIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN    450

EMBOSS_001       451 YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPT    500
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       451 YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPT    500

EMBOSS_001       501 NGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTG    550
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       501 NGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTG    550

EMBOSS_001       551 VLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP    600
                     ||||||||||||||||||||||.||||||||||||||||||||||||||||
EMBOSS_001       551 VLTESNKKFLPFQQFGRDIADTIDAVRDPQTLEILDITPCSFGGVSVITP    600

EMBOSS_001       601 GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCL    650
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       601 GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCL    650

EMBOSS_001       651 IGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLG    700
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       651 IGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLG    700
```

Figure 7.2: Difference of discovered strain 2 spike protein sequence and original strain spike protein sequence

```
EMBOSS_001       701 AENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECS    750
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       701 AENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECS    750

EMBOSS_001       751 NLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGF    800
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       751 NLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGF    800

EMBOSS_001       801 NFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLI    850
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       801 NFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLI    850

EMBOSS_001       851 CAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM    900
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       851 CAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM    900

EMBOSS_001       901 QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQD    950
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       901 QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQD    950

EMBOSS_001       951 VVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGR   1000
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       951 VVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGR   1000

EMBOSS_001      1001 LQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLM   1050
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      1001 LQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLM   1050

EMBOSS_001      1051 SFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGT   1100
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      1051 SFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGT   1100

EMBOSS_001      1101 HWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKE   1150
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      1101 HWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKE   1150

EMBOSS_001      1151 ELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL   1200
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      1151 ELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL   1200

EMBOSS_001      1201 QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSC   1250
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      1201 QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSC   1250

EMBOSS_001      1251 GSCCKFDEDDSEPVLKGVKLHYT   1273
                     |||||||||||||||||||||||
EMBOSS_001      1251 GSCCKFDEDDSEPVLKGVKLHYT   1273
```

Figure 7.2: Difference of discovered strain 2 spike protein sequence and original strain spike protein sequence

Figure 7.3: Difference of MR discovered strain 1 spike protein sequence and the MR assembled contig spike protein sequence

```
EMBOSS_001      732 TTEILPVSMTK---------TSVDCT---------MYICGDSTECSNLLL    763
                    ..:||.:.:.:         ..|.||         :|.|..:....|.:|
EMBOSS_001      718 YPQILLLVLPQKFYQCLPRHQIVQCTFGMIQLNAAIYCCNMAVFVHNTVL    767

EMBOSS_001      764 QYGSFCTQLNRALTGIAVEQDKNTQEVFAQVK----------QIYKTPP-    802
                         ...||:..|........|......|:|          ||::...
EMBOSS_001      768 -----LELLNKTKTPKKFVHKSNKFTKHHQLKILVVLIFHKFQIHQNQAR    812

EMBOSS_001      803 ---IKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC    849
                       :|.:.....|:|     .|..:.:..|:   :.|.:..|:.:
EMBOSS_001      813 GHLLKIYFSTKHLQLL-----ASSNNMVIALV---ILLLEPSFVHK----    850

EMBOSS_001      850 LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS------    893
                        :...|:.....:.:.:|..||...::     :||.|.:..
EMBOSS_001      851 ----SLTALLFCHLCSQIRLLNTLLHCRVL-----SLLVGPLVQVLHYKY    891

EMBOSS_001      894 ----GWTFGAGA-ALQIPFAMQ----------MAYRFNGIGVTQNVLYEN    928
                        .|..|... .|...|:|:          :..:|....:.|....||
EMBOSS_001      892 HVLCKWLIGLMVLELHRMFSMRTKNLPTNLIVLLAKFKTHFLPQQANLEN    941

EMBOSS_001      929 QKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSN    978
                    .|:    :::.:.|:...|::.|..|.:.|
EMBOSS_001      942 FKM----WSTKMHKLTRLLNNLAPILVQFQ--------------------    967

EMBOSS_001      979 FGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTY-VTQQLIRAAEI   1027
                       :..:...:|::|          ||....:.|...: |.:.::...|:
EMBOSS_001      968 ---VFMISFHVLTKL----------RLKCKLIGSQADFNVCRHMLNNLEL   1004

EMBOSS_001      1028 RASANLAATKMSEC---VLGQSKRVDFCGKGYHLMSFPQSAPHGVVF---   1071
                     :.|..|....:..|   .|...|.:.|..:.  ::..|..:.|.:||
EMBOSS_001      1005 QKSELLLILLLLNCQSVYLDNQKELIFVERA--IILCPSLSQHLMVFSCM   1052

EMBOSS_001      1072 ---LH--------VTYVPAQEKNFTTAPAI---CHDGKAHFPREGVFVSN   1107
                        ||        :.:| ..|..|....::   .|.|.    .:|:|:::
EMBOSS_001      1053 LMSLHKKRTSQLLLPFV-MMEHTFLVKVSLFQMAHTGL----HKGIFMNH   1097

EMBOSS_001      1108 -----GTHWF-------VTQRNFYEPQIITTDNTFVSGNCDVVIGIVNN-   1144
                          .||..       |:...|   .|:...||....:.: ::.|:::
EMBOSS_001      1098 KSLLQTTHLCLVTVMLEVSTTQF---MILCNLNTHSRRSIN-ILRIIHHQ   1143

EMBOSS_001      1145 -TVYDPLQPELD-SFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEI   1192
                      .:...|...|. ::|::|                      ..:::.:|
EMBOSS_001      1144 MLIVTSLALMLQLTYKKKL----------------------TASMRLQI   1170

EMBOSS_001      1193 DRLNEVAKNLN--ESLIDLQELGK-----------------YEQYIKWPW   1223
                     ..|:.::|||.  .|::.:..|.:                 |:|.:....
EMBOSS_001      1171 MNLSSISKNLESMSSIMAMVHLARFYSWLDCHSNGDNYALLYDQLLLSQG   1220

EMBOSS_001      1224 YIWLGFIAGLIAIV-----MVTIMLCCMTSCCSCLKGCCSCGSCCKFDED   1268
                     .::||.:..:..:.     .:|:.:
EMBOSS_001      1221 LLFLGILLQIRRLASAQRSQITLHI------------------------   1245

EMBOSS_001      1269 DSEPVLKGVKLHYT   1282

EMBOSS_001      1246 --------------   1245
```

Figure 7.3: Difference of MR discovered strain 1 spike protein sequence and MR assembled contig spike protein sequence

```
EMBOSS_001         1 MFLLTTKRTMFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDK     50
                     ||||||||||||||||||.||||||||||||||||||||||||||||||
EMBOSS_001         1 MFLLTTKRTMFVFLVLLPLDSSQCVNLTTRTQLPPAYTNSFTRGVYYPDK     50

EMBOSS_001        51 VFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVY    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001        51 VFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVY    100

EMBOSS_001       101 FASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFL    150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       101 FASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFL    150

EMBOSS_001       151 GVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLRE    200
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       151 GVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLRE    200

EMBOSS_001       201 FVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTL    250
                     ||||||||||||||||||||||||||||||||||||||||:||||||||||
EMBOSS_001       201 FVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPMGINITRFQTL    250

EMBOSS_001       251 LALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDC    300
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       251 LALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDC    300

EMBOSS_001       301 ALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEV    350
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       301 ALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEV    350

EMBOSS_001       351 FNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLC    400
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       351 FNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLC    400

EMBOSS_001       401 FTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNL    450
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       401 FTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNL    450

EMBOSS_001       451 DSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFP    500
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       451 DSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFP    500

EMBOSS_001       501 LQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNF    550
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       501 LQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNF    550

EMBOSS_001       551 NFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCS    600
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       551 NFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCS    600

EMBOSS_001       601 FGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSN    650
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       601 FGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSN    650

EMBOSS_001       651 VFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQS    700
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       651 VFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQS    700

EMBOSS_001       701 IIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMY    750
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       701 IIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMY    750
```

Figure 7.4: Difference of MR discovered strain 2 spike protein sequence and MR assembled contig spike protein sequence

```
EMBOSS_001      751 ICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKT    800
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      751 ICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKT    800

EMBOSS_001      801 PPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCL    850
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      801 PPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCL    850

EMBOSS_001      851 GDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAG    900
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      851 GDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAG    900

EMBOSS_001      901 AALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSST    950
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      901 AALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSST    950

EMBOSS_001      951 ASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEV   1000
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001      951 ASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEV   1000

EMBOSS_001     1001 QIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVD   1050
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001     1001 QIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVD   1050

EMBOSS_001     1051 FCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPR   1100
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001     1051 FCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPR   1100

EMBOSS_001     1101 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPL   1150
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001     1101 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPL   1150

EMBOSS_001     1151 QPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAK   1200
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001     1151 QPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAK   1200

EMBOSS_001     1201 NLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCC   1250
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001     1201 NLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCC   1250

EMBOSS_001     1251 SCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT   1282
                    ||||||||||||||||||||||||||||||||
EMBOSS_001     1251 SCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT   1282
```

Figure 7.4: Difference of MR discovered strain 2 spike protein sequence and MR assembled contig spike protein sequence