In Silico Framework for Genomes Analysis

M. Saqib Nawaz[1] · **M. Zohaib Nawaz**[1,2] ·
Yongshun Gong[3] · **Philippe Fournier-Viger**[1,*] ·
Abdoulaye Baniré Diallo[4]

Date of receiving / date of acceptance

Abstract Genomes hold the complete genetic information of an organism. Examining and analyzing genomic data plays a critical role in properly understanding an organism, particularly the main characteristics, functionalities, and evolving nature of harmful viruses. However, the rapid increase in genomic data poses new challenges and demands for extracting meaningful and valuable insights from large and complex genomic datasets. In this paper, a novel Framework for Genome Data Analysis (F4GDA), is developed that offers various methods for the analysis of viral genomic data in various forms. The framework's methods can not only analyze the changes in genomes but also various genomes contents. As a case study, the genomes of five SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) VoC (variants of concern), which are divided into three types/groups on the basis of geographical locations, are analyzed using this framework to investigate (1) the nucleotides, amino acids and synonymous codon changes in the whole genomes of VoC as well as in the Spike protein, (2) whether different environments affect the rate of changes in genomes, (3) the variations in nucleotide bases, amino acids, and codon base compositions in VoC genomes, and (4) to compare VoC genomes with the reference genome sequence of SARS-CoV-2.

Keywords Genome Data · Nucleotide bases · Amino acid changes · SARS-CoV-2 · VoC (Variants of Concern)

[1] College of Computer Science and Software Engineerin, Shenzhen University, China E-mail: {*msaqibnawaz, philfv*}*@szu.edu.cn*
[2] Faculty of Computing and Information Technology, Department of Computer Science, University of Sargodha, Pakistan E-mail: *zohaib.nawaz@uos.edu.pk*
[3] School of Software, Shandong University. E-mail: ysgong@sdu.edu.cn
[3] Department of Computer Science, Université du Québec á Montréal, Canada E-mail: *diallo.abdoulaye@uqam.ca*
[*] Corresponding author

1 Introduction

Despite significant progress due to scientific advancements, infectious diseases remain prevalent around the world. Examining the genomes of viruses is critical for understanding their main characteristics, functionalities, evolving nature, and for developing effective vaccines or therapeutics that can provide long-term immunity. The genomes of a virus can now be sequenced rapidly from patients around the world and shared on public repositories such as GenBank [1] and GISAID [2]. However, the rapid growth in size and complexity of genomic data poses new challenges. It demands the use of high-performance computing (HPC) and the development of novel methods to extract meaningful and valuable insights from large and complex genomic datasets. Some key issues are the storage, management, and processing of massive biological data, as well as extracting useful information from the data. Thus, to analyze complex biological data, sophisticated and efficient computational approaches are now necessary. In particular, the recent COVID-19 pandemic [3, 4] impact showed to the world that efficient genomes analysis approaches, well-organized databases and search engines, preferably integrative ones capable of working across multiple data sources, are required to address pandemics. Furthermore, the third United Nations Sustainable Development Goal[1], which emphasizes the importance of "Good Health and Well-being," highlights the necessity of utilizing computational tools and computer-assisted techniques in healthcare research and development. These tools are crucial not only for combating emerging diseases but also for improving health outcomes. In this context, computational and computer-assisted tools can aid biomedical and biological scientists in investigating and uncovering important information within genomic data. Their use can not only expedite the process of discovering actionable insights for early intervention but can also contribute to an improved global response.

Computational biology and bioinformatics projects, such as Biopython [5], Bioconductor [6], and BioJava [7]), offer diverse tools for analyzing and comprehending genomic data. However, these projects utilize distinct algorithms for sequence alignment. Since the onset of COVID-19, numerous online databases (e.g. COG-UK [8], EpiCoV from GISAID, CARD [9]), search systems, and resources (e.g. CoV-GLUE [10], 2019nCoVR [11], CoV-Seq [12], VirusViz [13], CoVMT [14], coronaApp [15], MicroGMT [16], hcov19-variants) have been developed specifically for analyzing the genomes of SARS-CoV-2 and its variants. Nevertheless, the majority of these tools concentrate on specific accruing changes in genomes, operate solely on nucleotides or amino acids, and align and compare genomes with the reference genome sequence. Furthermore, these systems predominantly rely on other algorithms for sequence alignments and comparisons.

In this paper, a novel framework is introduced, called F4GDA (Framework for Genome Data Analysis), which comprises three primary methods to analyze genome data in (1) *nucleotide form*, (2) in *coding region form* and (3) in *protein form*. The first method identifies nucleotides changes in whole genomes. The second method is used to find nucleotides changes, including those arising from synonymous codons,

[1] undp.org/sustainable-development-goals/good-health

in whole genomes as well as in different genes within genomes. The third method detects amino acid changes in genomes and their constituent genes. Additionally, the framework offers the capability to analyze the overall composition of nucleotide bases (A%, C%, G% and T%) and amino acids, GC and AT/GC contents, and the codon base composition (A_3%, C_3%, G_3%, T_3%, GC_1%, GC_2%, GC_3%) in both whole genomes and different genes.

As a case study, the F4GDA framework is applied to analyze the genomes of five SARS-CoV-2 VoC (Alpha, Beta, Delta, Gamma and Omicron). This analysis encompasses genomes sourced from different countries (DC), the same country but different locations (SCDL) and the same country and location (SCSL). Compared to existing tools, the F4GDA framework does not depend on any sequence annotation and sequence alignment techniques, and it allows for the comparison and analysis of genomes in three distinct forms (nucleotide, coding region, and protein), among themselves and with the reference sequence. The framework can interpret and analyze genomes with a focus on both nucleotide and amino acid sequence variations. As a result, the F4GDA framework facilitates simple and rapid analysis of viral genomes, enabling the monitoring of evolutionary changes in both nucleotide and amino acid sequences across populations.

The rest of this paper is organized as follows: The related work is discussed in Section 2. The proposed F4GDA framework for the analysis of viruses genomic data in various forms is introduced in Section 3. Section 4 presents an evaluation of the framework's capabilities and performance. Finally, Section 5 concludes the paper with some remarks.

2 Related Word

Several research studies [17–26] have examined and identified frequent mutations types, single nucleotide polymorphisms (SNPs) and nucleotides/amino acid changes in SARS-CoV-2 genomes sourced from various countries and archived in online databases such as NCBI's GenBank and GISAID. For a deeper understanding of SNP data analysis, readers may refer to [27]. The study [28] developed a framework leveraging SNPs analysis for the early detection of Alzheimer's disease. The CoVsurver tool, integrated with GISAID, was employed in [29] to analyze genomes with patient-follow up status for mutations, revealing a direct correlation between mutations and clinical outcomes. . However, some of these studies lack details regarding the quality and nucleotide completeness of the collected genomes. Furthermore, genomes were analyzed solely in one form, relying on various alignment techniques to compare the collected genomes with the SARS-CoV-2 reference sequence (either NC_045122 or MN908947). In [23], the ORF1ab gene, which constitutes two thirds of the SARS-CoV-2 genome at the 5'UTR end and encodes PP1ab and PP1a polyproteins, was excluded from the analysis.

Shi et al. [30] proposed a development method for biological sequence analysis algorithms, integrating the formal partition-and-recur (PAR) method, component technology, generic programming and domain engineering. Some studies [31,32] employed pattern mining techniques on genome sequences to uncover interesting hidden patterns of nucleotides and amino acids, their relationships with each other, and their prediction(s) in genomes. Moreover, a mutation technique was used to find nucleotide and amino acid changes in genomes. However, these prediction models have limitations, as they can only predict the next nucleotide or amino acid in the genomes, and with low accuracy. The mutation analysis technique has the limitations that the length of the genomes must be the same and also the genomes should only contain nucleotides without any additional information. Pathen et al. [33] computed the mutations in nucleotides and codons in the whole genomes of COVID-19 for various countries. But each genome was only compared with the reference genome.

SARS-CoV-2 genomes were compared in [34] with other coronaviruses genomes for various parameters using bioinformatics tools. FLAT [35] and its improved version BPNIF [36] can be used for the detection of longest common consecutive subsequences (LCCS) in biological sequences.

Compared to previous works and recently developed tools, this study offers a unified framework to analyze genomic data in various forms. SARS-CoV-2 genomes for VoC in three forms, which are further categorized into three groups (DC, SCDL and SCSL), are analyzed using this framework. VoC genomes from the same group are compared not only with each other but also with the reference genome sequence to analyze them for various aspects. F4GDA is independent of any external tools or packages and is not only limited solely to SARS-CoV-2; it can be utilized to analyze genomes of other known viruses.

3 The F4GDA Framework

The proposed F4GDA framework for genome analysis (Figure 1) consists of: (1) Accepting the genomes in three forms (*nucleotide, coding region* and *protein*), and (2) Analyzing these genomes for (a) changes in nucleotides, codons (synonymous and non-synonymous) and amino acids, as well as (b) variations in various compositions such as the nucleotide bases, amino acids composition, nucleotides contents, and codons third base composition. This analysis can be conducted either on whole genomes or on specific regions of interest, such as the S protein. Further details are provided next.

F4GDA is utilized to analyze genomes of SARS-CoV-2 VoC sourced from GenBank [1]. The genomes are downloaded from GenBank in three forms. Consequently, there are three FASTA files for each VoC genomes: (1) Whole genomes in *nucleotide form*, (2) Genomes in *coding region form*, where each genome contains genes information and nucleotides encoding amino acids for each gene, and (3) Genomes in *protein form*, where each genome contains the genes information and the amino acid sequence for each gene.

Nucleotides other than the four basic nucleotide bases (A, C, G and T) are referred to here as ambiguous or *redundant nucleotides* (RN) due to their infrequent occurrence. Similarly, amino acids other than the 20 main amino acids are designated as ambiguous or *redundant amino acids* (RAA). When analyzing genomes, the primary focus is on the four bases and 20 amino acids, and it is assumed that RN and RAA encode the same base and amino acid, respectively, in both sequences. Note

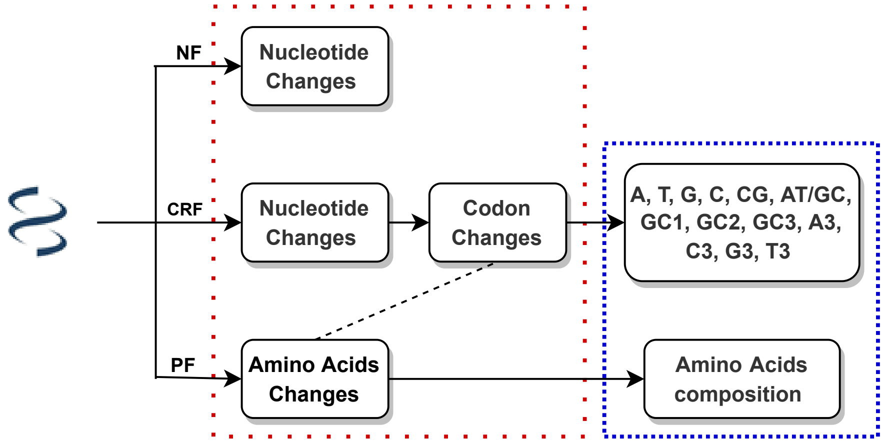

Fig. 1: Schematic of the F4GDA for genomes analysis. Dotted red box contains methods to analyze the genomes for nucleotides, amino acids and codons changes in three forms: *Nucleotide Form (NF)*, *Coding Region form (CRF)* and *Protein Form (PF)*. Dotted blue box contains methods to discover various contents inside genomes in *CRF* and *PF*.

that we investigated the RN and RAA in collected genomes and found that indeed they encode the same base and amino acid, respectively, in most cases. In the genetic code, there are 2 amino acids with a single codon choice, 9 amino acids with two codons, 1 amino acid with three codons, 5 amino acids with four codons, and 3 amino acids with six codons [37]. Thus, there are five synonymous codon families (SCF).

SARS-COV-2 genomes of VoC from different geographical locations are analyzed using the framework. Three cases are considered: (1) genomes from different countries (DC), (2) genomes from the same country but different locations (SCDL) and (3) genomes from the same country and the same location (SCSL). The earlier genome of a variant is compared with the new genome of same variant on the basis of collection dates. The main reason for selecting VoC genomes from DC, SCDL, and SCSL is to examine and investigate the changes in genomes from both similar and different environments. The following sections provide details of the methods developed within the F4GDA framework for analyzing genomes in various forms.

3.1 Nucleotide Form

Algorithm 1 presents the pseudocode for the analysis of genomes in *nucleotide form*. The algorithm takes two genomes (GS_1 and GS_2) as input and compares the nucleotides one by one. If nucleotides at a particular location do not match and they do not belong to RN, then the location and the changed nucleotides in both sequences are stored in $Diff$. The change rate (CR) is calculated by the following formula:

$$CR = \frac{M}{Y} \times 100 \tag{1}$$

where M is the total number of changed nucleotides and Y is the total number of nucleotides in the genomes.

Algorithm 1 Analysis of genomes in *nucleotide form*

Input: Genomes in *nucleotide form* (GS_1, GS_2)
Output: Locations, changed nucleotides, total changes and the change rate

```
$Diff \leftarrow \emptyset$;
RN $\leftarrow \{N, R, Y, K, M, S, W, B, D, H, V\}$
Y $\leftarrow$ min(len($GS_1$), len($GS_2$))
M $\leftarrow$ 0
for $i \leftarrow 1$ to Y do
    if $(GS_1(i) \neq GS_2(i)) \wedge (GS_1(i) \vee GS_2(i) \notin RN)$ then
        M $\leftarrow$ M +1;
        $+\!\!+$Diff $\leftarrow$ i, $GS_1(i)$, $GS_2(i)$
    end if
end for
Calculate $CR$ using Equation 1.
return Diff, M, CR
```

3.2 Coding Region Form

Algorithm 2 presents the pseudocode for analyzing genomes in *coding region form*. It takes genomes as input and compares the nucleotides in each gene within the genomes one by one. If the nucleotides at a specific location in the genes are different and they do not belong to NR, then the codon where the changed nucleotides are is found by invoking the *FindCodon* procedure. The changed codons go through the *CheckCodon* procedure to determine whether each changed codon is a SC or not. The locations, changed nucleotides, and changed codons are stored in $Diff$. Additionally, the gene in which the change occurred and whether the changed codon resulted in the same or a different amino acid are also recorded in $Diff$. The change rate is calculated using Equation 1.

The *FindCodon* procedure identifies the respective codon for the changed nucleotides. It takes as input the location (i) and the lines ($LGene_1, LGene_2$) within the genes where the change occurred, and returns the codon for the changed nucleotides. The *CheckCodon* procedure first determines the amino acids encoded by the codons. It then checks whether these amino acids belong to the SCF. If both amino acids belong to the SCF, the changed codon is considered a SC; otherwise, the changed codon produced results in a different amino acid.

Algorithm 2 Analysis of genomes in *coding region form*

Input: Genomes in *coding region form* (GS_1, GS_2)
Output: Locations, changed nucleotides and codons, SC or not, total changes and the overall change rate

```
$Diff \leftarrow \emptyset$;
RN $\leftarrow \{N, R, Y, K, M, S, W, B, D, H, V\}$
Y $\leftarrow$ min(len($GS_1$), len($GS_2$))
M $\leftarrow$ 0
for each $Gene_1 \in GS_1 \wedge Gene_2 \in GS_2$ do
    Z $\leftarrow$ min(len($Gene_1$), len($Gene_2$))
    for $i \leftarrow 1$ to Z do
        if $(Gene_1(i) \neq Gene_2(i)) \wedge (Gene_1(i) \vee Gene_2(i) \notin RN)$ then
            M $\leftarrow$ M +1;
            $CGene_1, CCene_2$ = FindCodon($i, LGene_1, LGene_2$);
            $CC_1, CC_2$ = CheckCodon($CGene_1, CGene_2$)
            $+\!\!+$Diff $\leftarrow i$, $Gene_1(i)$, $Gene_2(i)$, $CGene_1$, $CGene_2$, $CC_1$, $CC_2$
        end if
    end for
end for
Calculate CR using Equation 1
return $Diff$, M, CR
```

3.3 Protein Form

Algorithm 3 presents the pseudocode for analyzing amino acids in genome sequences in *protein form*. The algorithm takes genomes as input and compares the amino acids in each gene of the genomes, one by one. If amino acids at a specific location in the genes do not match and they do not belong to RAA, the location and changed amino acids are recorded in $Diff$ along with the genes information. The change rate (CR) is calculated using Equation 1, where M represents the total number of changed amino acids, and Y represents the total number of amino acids in the genomes.

Algorithm 3 is actually verifying that the amino acid changes in the genomes are due to non-synonymous codons that can be found by using Algorithm 2.

3.4 Nucleotide Bases, Amino Acids and Codon Base Compositions

The F4GDA framework is also utilized for analyzing genomes in *coding region form* and *protein form* to find:

- **Frequent Nucleotide Bases and Amino Acids**: The overall composition (occurrence) of four nucleotide bases and 20 amino acids. This can reveal the frequent nucleotides and amino acids present in the genomes.
- **GC and AT/GC Contents**: The occurrence of G+C (called GC content) and AT/GC ratio.
- **Codon Base Compositions**: The occurrence of four nucleotide bases and G+C at the third position of a codon $(A_3, C_3, T_3, G_3, GC_1, GC_2, GC_3,)$.

The procedure for calculating the GC, AT/GC contents, and four nucleotide bases compositions in genome(s) in *nucleotide form*, or in any specific gene of interest in

Algorithm 3 Analysis of genome sequences in *protein form*

Input: Genomes in *protein form* (GS_1, GS_2)
Output: Locations, changed amino acids, total changes and the overall change rate

```
$Diff \leftarrow \emptyset$;
RAA $\leftarrow \{B, J, X, Z\}$
Y $\leftarrow$ min(length($GS_1$), length($GS_2$))
M $\leftarrow 0$
for each $Gene_1 \in GS_1 \wedge Gene_2 \in GS_2$ do
    Z $\leftarrow$ min(len($Gene_1$), len($Gene_2$))
    for $i \leftarrow 1$ to Z do
        if $(Gene_1(i) \neq Gene_2(i)) \wedge (Gene_1(i) \vee Gene_2(i) \notin RAA)$ then
            M $\leftarrow$ M +1;
            Diff $\leftarrow$ i, $Gene_1(i), Gene_2(i)$
        end if
    end for
end for
Calculate $CR$ using Equation 1
return Diff, CR
```

genome(s) is not presented here. Similarly, the genomes in *protein form* are analyzed to find the composition of 20 amino acids.

Algorithm 4 outlines the steps for calculating the codon base, such as GC_1, GC_2, GC_3, A_3, C_3, G_3 and T_3 in genomes in *coding region form*. The procedure first finds the codon present in the genome(s). The list GC with 3 elements is used to store the C or G presence at the third position of a codon. Initially the three elements are set to 0. If C or G is the first (second) or (third) nucleotide in the codon, then GC[0] (GC[1]) or (GC[2]) is incremented by 1. The list $ACTG$ with 4 elements is used to store the total occurrence of A, C, G and T at the third position of a codon respectively.

F4GDA, a Python-based tool, is designed to accept genomes in both FASTA and TXT formats , without reliance on external library, packages, or tools. The code for the main methods will be made public following the peer review process.

4 Results and Discussion

SARS-CoV-2 genomes for five VoC are obtained from the NCBI's GenBank. Some records within it have limited research potential due to being smaller than the reference sequence (< 5,000 nucleotides) or containing a large number of ambiguous letters, such as RN and RAA. We filter and select only those genomes that are *complete*, with a length of at least 29,000 nucleotides, and are available in three forms. For each variant, their genomes are collected from DC, SCDL and SCSL. Thus each variant is divided intro three types/groups based on to geographical locations. Each genome is downloaded in three forms: *nucleotide form*, *coding region form* and *protein form*.

The VoC genomes are ordered in ascending order on the basis of their collection date. Only genomes with complete collection data (Year-Month-Day) are considered, excluding those with only year in the collection date. The earlier genome of a vari-

Algorithm 4 Codon Base Composition

Input: Genome(s) (GS)
Output: A3, C3, G3, T3, GC1, GC2, GC3

```
GC ← [0]× 3
ACTG ← [0]× 4
CC ← 0
for $i$ ← 1 to len(GS) do
    if (i % 3 == 0) then
        CC += 1
        if (GS[i-2] == C or GS[i-2] == G) then
            GC[1] += 1
        else if (GS[i-1] == C or GS[1-1] == G) then
            GC[2] += 1
        else if (GS[i] == C or GS[i] == G) then
            GC[3] += 1
        end if
        if (GS[i] == A  then
            ACTG[1] += 1
        else if (GS[i] == C then
            ACTG[2] += 1
        else if (GS[i] == G  then
            ACTG[3] += 1
        else if (GS[i] == T then
            ACTG[4] += 1
        end if
    end if
end for
GC1, GC2, GC3 = $\frac{GC[1]}{CC}, \frac{GC[2]}{CC}, \frac{GC[3]}{CC}$
A3, C, G3, T3 $\frac{ACGT[1]}{CC}, \frac{ACGT[2]}{CC}, \frac{ACGT[3]}{CC}, \frac{ACGT[4]}{CC}$
return $CG1 \times 100, CG2 \times 100, CG3 \times 100$
return $A3 \times 100, C3 \times 100, G3 \times 100, T3 \times 100$
```

ant is compared with the new genome of the same variant based on their collection dates. The information for some collected genome sequences for SARS-CoV-2 VoC are listed in the Appendix (Table A1). Only the accession numbers for genomes are provided in Table A1. All experiments were performed on a laptop equipped with an Intel Celeron processor and 16 GB of RAM.

4.1 Results for Genomes in Nucleotide Form

First, the changes that occurred in various genomes of VoC are presented in *nucleotide form* (Table 1). The format: $\frac{Alpha(Beta)}{Gamma(Delta)}Omicron$ is used to show the results for the five VoC. For example, consider the changes in the first two genomes (1 and 2) for DC with the following format: $\frac{22264(22107)}{10(21463)}(20980)$ in Table 1. This indicates that 22,264 changes occurred in the Alpha genome 2 when compared with the Alpha genome 1, 22,107 changes occurred in the Beta genome 2 when compared with the Beta genome 1, 10 changes occurred in the Gamma genome 2 when compared with the Gamma genome 1, 21,463 changes occurred in the Delta genome 2 when compared with the Delta genome 1 and 20,980 changes occurred in the Omicron genome 2 when compared with the Omicron genome 1.

Table 1: Results for VoC genomes in *nucleotide form*

Location	Genomes	Nucleotides changes	CR
DC	1 → 2	$\frac{22264(22107)}{10(21463)}(20980)$	$\frac{74.8042(74.1895)}{0.0335(72.0670)}(70.677)$
DC	2 → 3	$\frac{22201(21948)}{22282(22309)}(21214)$	$\frac{74.5926(73.5374)}{74.6049(74.9076)}(71.466)$
DC	3 → 4	$\frac{22199(21286)}{21926(22074)}(21986)$	$\frac{74.5858(71.4679)}{73.3949(74.0340)}(71.014)$
DC	4 → 5	$\frac{21215(22093)}{22221(21974)}(22147)$	$\frac{71.3708(74.1774)}{74.5446(73.9093)}(74.451)$
DC	5 → 6	$\frac{21099(21450)}{22284(21866)}(21214)$	$\frac{70.8900(71.9726)}{74.9268(73.5461)}(71.214)$
DC	6 → 7	$\frac{21760(22232)}{22264(21483)}(22268)$	$\frac{73.8737(74.7068)}{74.8596(71.8927)}(74.857)$
DC	7 → 8	$\frac{21106(22029)}{21875(21352)}(22159)$	$\frac{71.3546(74.0246)}{73.9195(71.5501)}(74.491)$
DC	8 → 9	$\frac{22147(22062)}{21839(21311)}(21587)$	$\frac{74.4111(74.2328)}{73.7978(71.4127)}(72.4809)$
SCDL	1 → 2	$\frac{22131(21385)}{21283(22283)}(22056)$	$\frac{74.3999(71.7617)}{71.3978(74.6774)}(71.1178)$
SCDL	2 → 3	$\frac{22137(21456)}{22282(22092)}(21399)$	$\frac{74.4200(71.8697)}{74.7492(74.1491)}(71.910)$
SCDL	3 → 4	$\frac{22150(21415)}{21789(21373)}(22063)$	$\frac{74.3937(71.7660)}{73.0684(71.7359)}(74.168)$
SCDL	4 → 5	$\frac{21370(21465)}{21179(6)}(22289)$	$\frac{71.8053(71.9336)}{71.5676(0.0200)}(74.928)$
SCDL	5 → 6	$\frac{21362(21399)}{21759(22026)}(22018)$	$\frac{71.7785(71.7629)}{73.5275(73.7716)}(73.962)$
SCDL	6 → 7	$\frac{16(21145)}{15(22118)}(22179)$	$\frac{0.5354(71.0503)}{0.0502(74.2837)}(74.403)$
SCDL	7 → 8	$\frac{23(21387)}{21501(22272)}(21307)$	$\frac{0.0770(72.0198)}{72.0760(74.8135)}(71.514)$
SCDL	8 → 9	$\frac{21312(21237)}{21394(22273)}(21328)$	$\frac{71.6682(71.5146)}{71.8473(74.816)}(71.5848)$
SCSL	1 → 2	$\frac{21905(22023)}{21186(21946)}(22237)$	$\frac{74.0559(73.9085)}{71.1488(73.6600)}(74.741)$
SCSL	2 → 3	$\frac{21240(21248)}{21315(21360)}(21036)$	$\frac{71.1248(71.5526)}{71.4707(71.7741)}(70.704)$
SCSL	3 → 4	$\frac{20501(21391)}{21271(22040)}(21752)$	$\frac{68.8808(72.0332)}{71.3443(74.0591)}(72.995)$
SCSL	4 → 5	$\frac{20854(21440)}{21302(21777)}(21652)$	$\frac{70.0702(71.9970)}{71.5384(72.9825)}(72.711)$
SCSL	5 → 6	$\frac{21936(22179)}{21340(22100)}(1)$	$\frac{74.1607(74.3272)}{71.3831(74.0848)}(0.0033)$
SCSL	6 → 7	$\frac{21933(21953)}{21348(21484)}(21357)$	$\frac{74.1505(73.6365)}{71.9394(72.0248)}(71.788)$
SCSL	7 → 8	$\frac{22183(21675)}{21350(21635)}(21381)$	$\frac{74.3971(72.7065)}{71.9460(72.5310)}(71.801)$
SCSL	8 → 9	$\frac{33(22043)}{21532(21241)}(22089)$	$\frac{0.1105(74.0651)}{71.1799(71.1791)}(74.178)$

In the *nucleotide form*, we found that a very large number of nucleotides ($>$ 20,000) were changed in most of the VoC genomes, resulting in very high change rates. For some genomes, the changes were less ($\leq$ 33 nucleotides). The average changes in VoC were: Alpha (18,961.54), Beta (21,668.7), Gamma (19,856.16), Delta (20,910.75) and Omicron (20,820.70). The slightly lower changes in Alpha/Gamma were due to three/two occasions where fewer nucleotides (in the range 10-33) were changed. The average changes in genomes from DC were 21,282.2, in genomes from SCDL were 19,541.8, and in genomes from SCSL were 20,556.7. VoC genomes from DC have more changes than those from from SCDL and SCSL. In contrast, SCSL genomes have slightly more changes than SCDL. In fact, this difference is significant, according to the p-value (0.5723 $>$ 0.05) obtained after using the Wilcoxon test [38] on nucleotides changes in genomes from SCDL and SCSL. High changes in VoC genomes from DC suggest that different environments have an effect on nucleotides changes.

4.2 Results for Genomes in Coding Form

Next, SAR-CoV-2 VoC genomes in *coding region form* are analyzed for nucleotide changes and CR in various genes. In total 12 genes were considered, which are: *ORF1ab, ORF1a, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, ORF10* and *N*. Some nucleotide changes (NC) produce the synonymous codon (SC) that encodes the same amino acid. The complete results for total NC (TNC), SC, NC in 12 genes and CR in VoC genomes from DC, SCDL and SCSL are listed in the Appendix (Table A2). *ORF10* is not included in Table A2 as no changes occurred. Table 2 lists the changes in the S protein, TNC, SC and CR with the format: $\frac{Alpha(Beta)}{Gamma(Delta)}(Omicron)$. For example, the entry $\frac{1(4)}{1(2401)}(0)$ in Table 2 for DC indicates that 1 change occurred in the S protein of Alpha genome 2 when compared with the S protein of Alpha genome 1, 4 changes occurred in the S protein of Beta genome 2 when compared with the S protein of Beta genome 1, 1 change occurred in the S protein of Gamma genome 2 when compared with the S protein of of Gamma genome 1, and so on.

The average total nucleotide changes (TNC) in VoC are: Alpha (221.29), Beta (934), Gamma (212.29), Delta (2633.58), and Omicron (1413.33). Delta genomes exhibit more nucleotides changes compared to genomes from other variants, followed by Omicron and Beta. As discussed in Section 2.1, the S protein binds to the human host cell membrane by interacting with the host ACE2 receptor. Thus, blocking or preventing the binding of S proteins with ACE2 receptors is considered the primary and most important approach to block the cell entry of SARS-CoV-2. Consequently, the S protein is the most significant target for COVID-19 vaccine, drugs, and therapeutic research.

The average nucleotide changes in the S protein in VoC are: Alpha (203.41), Beta (182.66), Gamma (196.04), Delta (643.20) and Omicron (568.5). This means that in Delta genomes, 25% of nucleotide changes were in the S protein. For Omicron genomes, 40% of nucleotide changes were in the S protein. Whereas for Alpha, Beta, and Gamma, the percentage of nucleotide changes occurring in the S protein were: 92%, 19% and 92%, respectively. The average nucleotide changes in the S protein in VoC genomes from DC were 649.05, in VoC genomes from SCDL were 296.97, and in VoC genomes from SCSL were 130.25. Thus, VoC genomes from DC have more nucleotides changes, followed by SCDL and SCSL. The average nucleotide changes in VoC genomes from DC were 1,149.82, in VoC genomes from SCDL were 763.55 and in VoC genomes from SCSL were 1,335.55. The high average for SCSL is due to the fact that one genome sequence in Gamma generated changes ($>$ 26,000) with its successor and predecessor. Otherwise, DC has more nucleotide changes than SCDL.

The average nucleotide changes that produce the SC in VoC are: Alpha (9.58), Beta (27.91), Gamma (11.37), Delta (58.6) and Omicron (35.12). In Alpha genomes, there are approximately 4.49% nucleotide changes due to SC resulted in the silent mutation. For other variants this percentage is Beta (2,98%), Gamma (5.35%), Delta

Table 2: Results for VoC genomes in *coding region form*

Loc	Genomes	S	TNC	SC	CR
DC	1 → 2	$\frac{1(4)}{1(2401)}(0)$	$\frac{15(18)}{14(2427)}(0)$	$\frac{7(7)}{7(52)}(0)$	$\frac{0.035(0.042)}{0.4949(5.71)}(0)$
	2 → 3	$\frac{2(3)}{2330(6)}(2264)$	$\frac{12(14)}{2333(33)}(11939)$	$\frac{3(8)}{46(15)}(278)$	$\frac{0.028(0.032)}{5.497(0.127)}(28.13)$
	3 → 4	$\frac{4(2)}{2330(2399)}(2264)$	$\frac{16(13)}{2337(2440)}(11943)$	$\frac{6(9)}{47(58)}(278)$	$\frac{0.037(0.030)}{5.504(5.74)}(28.14)$
	4 → 5	$\frac{2(2)}{0(1)}(1)$	$\frac{22(19)}{8(47)}(9)$	$\frac{9(15)}{5(16)}(2)$	$\frac{0.052(0.044)}{0.212(0.110)}(0.0212)$
	5 → 6	$\frac{2421(2)}{0(2400)}(1)$	$\frac{2447(19)}{13(2426)}(10)$	$\frac{53(12)}{5(54)}(2)$	$\frac{5.80(0.044)}{0.056(5.71)}(0.0235)$
	6 → 7	$\frac{2422(2)}{0(2)}(2)$	$\frac{2439(17)}{25(51)}(11)$	$\frac{51(10)}{13(14)}(3)$	$\frac{5.78(0.040)}{0.084(0.120)}(0.0259)$
	7 → 8	$\frac{7(3)}{0(2400)}(2)$	$\frac{34(18)}{31(2415)}(12)$	$\frac{6(11)}{14(54)}(7)$	$\frac{0.080(0.042)}{0.120(5.68)}(0.028)$
	8 → 9	$\frac{3(2)}{1(1)}(2274)$	$\frac{39(19)}{21(11)}(2276)$	$\frac{7(6)}{6(1)}(54)$	$\frac{0.092(0.044)}{0.494(0.025)}(5.63)$
SCDL	1 → 2	$\frac{2(2141)}{1(482)}(0)$	$\frac{6(10992)}{9(190)}(813)$	$\frac{2(239)}{1(6)}(30)$	$\frac{0.0142(25.89)}{0.056(0.447)}(2.42)$
	2 → 3	$\frac{3(2207)}{1(184)}(0)$	$\frac{11(11062)}{10(188)}(0)$	$\frac{3(244)}{4(5)}(0)$	$\frac{0.0260(34.08)}{0.056(0.445)}(0)$
	3 → 4	$\frac{0(2)}{2(2)}(1)$	$\frac{5(20)}{13(9)}(7)$	$\frac{3(9)}{10(5)}(2)$	$\frac{0.011(0.047)}{0.030(0.028)}(0.0164)$
	4 → 5	$\frac{0(0)}{2(1)}(2275)$	$\frac{14(17)}{15(7)}(2277)$	$\frac{7(7)}{9(4)}(54)$	$\frac{0.0379(0.40)}{0.035(0.021)}(5.36)$
	5 → 6	$\frac{1(1)}{3(0)}(2274)$	$\frac{31(22)}{18(4)}(2277)$	$\frac{11(10)}{8(4)}(54)$	$\frac{0.0735(0.051)}{0.042(0.009)}(5.36)$
	6 → 7	$\frac{0(1)}{4(0)}(0)$	$\frac{20(20)}{18(7)}(1)$	$\frac{8(6)}{11(5)}(0)$	$\frac{0.0474(0.047)}{0.042(0.014)}(0.007)$
	7 → 8	$\frac{0(3)}{3(0)}(0)$	$\frac{26(23)}{23(5)}(1)$	$\frac{11(14)}{10(3)}(0)$	$\frac{0.061(0.054)}{0.054(0.014)}(0.007)$
	8 → 9	$\frac{2(2)}{3(2)}(0)$	$\frac{32(18)}{36(9)}(2275)$	$\frac{13(9)}{12(3)}(52)$	$\frac{0.075(0.042)}{0.084(0.042)}(5.36)$
SCSL	1 → 2	$\frac{0(0)}{1(3)}(0)$	$\frac{6(17)}{9(8)}(5)$	$\frac{0(8)}{4(1)}(3)$	$\frac{0.014(0.040)}{0.035(0.049)}(0.011)$
	2 → 3	$\frac{1(0)}{1(2574)}(0)$	$\frac{9(11)}{13(26385)}(3)$	$\frac{2(7)}{6(543)}(0)$	$\frac{0.021(0.025)}{0.306(62.15)}(0.212)$
	3 → 4	$\frac{1(0)}{3(2574)}(5)$	$\frac{16(2)}{20(26482)}(19)$	$\frac{3(2)}{9(536)}(5)$	$\frac{0.037(0.004)}{0.047(62.37)}(0.044)$
	4 → 5	$\frac{0(0)}{3(1)}(5)$	$\frac{12(4)}{31(11)}(15)$	$\frac{4(2)}{11(5)}(5)$	$\frac{0.028(0.01)}{0.073(0.025)}(0.035)$
	5 → 6	$\frac{0(2)}{3(2)}(1)$	$\frac{15(6)}{30(16)}(1)$	$\frac{5(2)}{13(10)}(0)$	$\frac{0.035(0.02)}{0.070(0.042)}(0.007)$
	6 → 7	$\frac{0(2)}{3(0)}(1)$	$\frac{8(17)}{24(2)}(9)$	$\frac{2(7)}{12(0)}(5)$	$\frac{0.018(0.04)}{0,0565(0.014)}(0.282)$
	7 → 8	$\frac{5(1)}{3(2)}(0)$	$\frac{33(21)}{21(16)}(13)$	$\frac{10(11)}{6(7)}(7)$	$\frac{0.078(0.049)}{0.0495(0.042)}(0.030)$
	8 → 9	$\frac{5(2)}{7(0)}(0)$	$\frac{41(27)}{23(17)}(4)$	$\frac{13(15)}{6(7)}(2)$	$\frac{0.097(0.063)}{0.054(0.070)}(0.014)$

(2.22%), and Omicron (2.48%). Delta has the lowest rate for silent mutations among all the variants. It is important to point out here that it is possible for two or three consecutive nucleotide changes in a codon to only change one amino acid; these are not included in the SC.

4.3 Results for Genomes in Protein Form

SAR-CoV-2 VoC genomes in *protein from* for amino acids and CR in 12 genes is also analyzed. The complete results for total amino acids (TAA), amino acid changes in 12 genes (excluding ORF10) and CR in VoC genomes from DC, SCDL and SCSL are listed in the Table A3 of Appendix. Table 3 lists the amino acid changes in the S protein, TAA and CR with the format: $\frac{Alpha(Beta)}{Gamma(Delta)}(Omicron)$.

The average TAA changes in VoC are: Alpha (101.95), Beta (407.66), Gamma (95.16), Delta (1,155.54) and Omicron (615.16). Delta genomes have more amino acid changes compared to genomes from other variants, followed by Omicron and Beta. The average amino acid changes in the S protein in VoC were: Alpha (90.54),

Table 3: Results for VoC strains in *protein form*

Loc	Genomes	S	TAA	MR
DC	$1 \rightarrow 2$	$\frac{0(3)}{3(1050)}(0)$	$\frac{8(11)}{7(1069)}(0)$	$\frac{0.056(0.077)}{0.049(7.55)}(0)$
DC	$2 \rightarrow 3$	$\frac{1(2)}{1017(5)}(985)$	$\frac{9(6)}{1017(18)}(5190)$	$\frac{0.064(0.042)}{7.19(0.127)}(36.72)$
DC	$3 \rightarrow 4$	$\frac{2(2)}{1018(1048)}(985)$	$\frac{10(4)}{1020(1075)}(5194)$	$\frac{0.071(0.028)}{7.21(7.59)}(36.75)$
DC	$4 \rightarrow 5$	$\frac{2(2)}{0(1)}(1)$	$\frac{13(4)}{3(31)}(7)$	$\frac{0.092(0.028)}{0.021(0.21)}(0.049)$
DC	$5 \rightarrow 6$	$\frac{1071(1)}{0(1048)}(1)$	$\frac{1085(7)}{8(1065)}(8)$	$\frac{7.72(0.049)}{0.056(7.52)}(0.056)$
DC	$6 \rightarrow 7$	$\frac{1072(1)}{0(2)}(1)$	$\frac{1080(7)}{12(37)}(8)$	$\frac{7.68(0.049)}{0.084(0.26)}(0.056)$
DC	$7 \rightarrow 8$	$\frac{7(1)}{3(1048)}(1)$	$\frac{24(7)}{13(1054)}(5)$	$\frac{0.170(0.049)}{0.120(7.45)}(.035)$
DC	$8 \rightarrow 9$	$\frac{3(1)}{1(1)}(990)$	$\frac{28(13)}{15(10)}(990)$	$\frac{0.199(0.09)}{0.106(0.706)}(7.00)$
SCDL	$1 \rightarrow 2$	$\frac{1(930)}{0(87)}(0)$	$\frac{4(4790)}{8(91)}(341)$	$\frac{0.028(33.87)}{0.056(0.643)}(2.42)$
SCDL	$2 \rightarrow 3$	$\frac{2(959)}{0(88)}(0)$	$\frac{8(4819)}{6(90)}(0)$	$\frac{0.0569(34.08)}{0.042(0.636)}(0)$
SCDL	$3 \rightarrow 4$	$\frac{0(1)}{1(1)}(1)$	$\frac{2(11)}{3(4)}(5)$	$\frac{0.014(0.077)}{0.021(0.028)}(0.035)$
SCDL	$4 \rightarrow 5$	$\frac{0(0)}{1(0)}(990)$	$\frac{9(10)}{6(3)}(990)$	$\frac{0.064(0.070)}{0.042(0.021)}(7.00)$
SCDL	$5 \rightarrow 6$	$\frac{1(0)}{1(0)}(990)$	$\frac{20(12)}{10(0)}(991)$	$\frac{0.124(0.084)}{0.070(0)}(7.01)$
SCDL	$6 \rightarrow 7$	$\frac{0(0)}{1(0)}(0)$	$\frac{12(14)}{7(2)}(1)$	$\frac{0.085(0.099)}{0.049(0.014)}(0.007)$
SCDL	$7 \rightarrow 8$	$\frac{0(2)}{1(0)}(0)$	$\frac{15(9)}{13(2)}(1)$	$\frac{0.106(0.054)}{0.091(0.014)}(0.007)$
SCDL	$8 \rightarrow 9$	$\frac{1(2)}{3(2)}(990)$	$\frac{19(9)}{24(6)}(991)$	$\frac{0.135(0.063)}{0.169(0.042)}(7.01)$
SCSL	$1 \rightarrow 2$	$\frac{0(0)}{0(1)}(0)$	$\frac{6(9)}{5(7)}(2)$	$\frac{0.042(0.063)}{0.035(0.049)}(0.014)$
SCSL	$2 \rightarrow 3$	$\frac{1(0)}{1(1132)}(0)$	$\frac{7(4)}{7(11545)}(3)$	$\frac{0.049(0.028)}{0.4949(81.65)}(0.021)$
SCSL	$3 \rightarrow 4$	$\frac{1(0)}{2(1132)}(5)$	$\frac{13(0)}{11(11591)}(14)$	$\frac{0.092(0)}{0.077(81.97)}(0.099)$
SCSL	$4 \rightarrow 5$	$\frac{0(0)}{2(1)}(5)$	$\frac{8(2)}{20(6)}(10)$	$\frac{0.056(0.014)}{0.141(0.042)}(0.070)$
SCSL	$5 \rightarrow 6$	$\frac{0(2)}{1(1)}(1)$	$\frac{10(4)}{17(9)}(1)$	$\frac{0.071(0.028)}{0.120(0.063)}(0/007)$
SCSL	$6 \rightarrow 7$	$\frac{0(2)}{1(0)}(1)$	$\frac{6(10)}{12(2)}(4)$	$\frac{0.042(0.070)}{0.084(0.014)}(0.0282)$
SCSL	$7 \rightarrow 8$	$\frac{4(1)}{2(1)}(0)$	$\frac{23(10)}{15(6)}(6)$	$\frac{0.163(0.070)}{0.106(0.042)}(0.042)$
SCSL	$8 \rightarrow 9$	$\frac{4(1)}{4(0)}(0)$	$\frac{28(12)}{17(10)}(2)$	$\frac{0.019(0.084)}{0.120(0.070)}(0.014)$

Beta (79.70), Gamma (43.25), Delta (277.12) and Omicron (267.79). This means that in Delta strains, 24% of amino acid changes were in the S protein. For Omicron, 40.28% of amino acid changes were in the S protein. Whereas for Alpha, Beta and Gamma, the percentage of amino acid changes occurred in the S protein are: 89%, 19% and 46% respectively. Is is interesting to see that nucleotide and amino acid changes in the S protein share the same ratio for the variants excluding Gamma. The average amino acid changes in the S protein in VoC genomes from DC were 258.97, in VoC genomes from SCDL were 126.4 and in VoC genomes from SCSL were 57.67. The average TAA changes in VoC genomes from DC were 504.4, in VoC genomes from SCDL were 333.95 and in VoC genomes from SCSL were 586.95. Thus, VoC strains from SCSL have more amino changes, followed by VoC strains from DC and SCDL. The reason for high changes in SCSL is discussed next.

Amino acids change rates for VoC genomes are shown in Figure 2. The changes in Delta is the highest followed by Omicron and Beta. It can be seen that changes in two Delta genomes from SCSL were very high (approximately 81% CR). This has played a significant part in increasing the changes rate in SCSL. For the same genomes, 56% of nucleotides and amino acid changes were found in the ORF1ab.

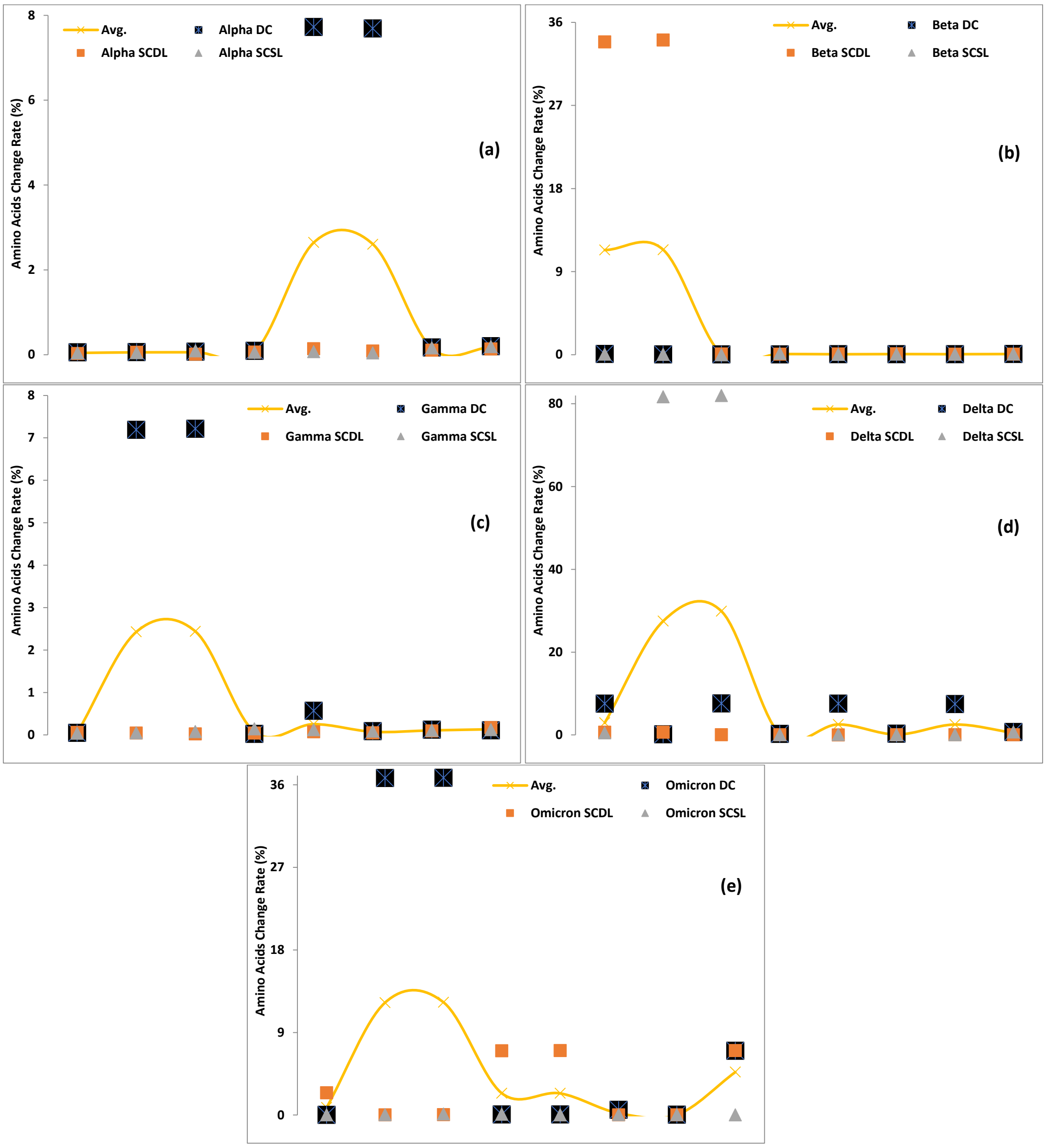

Fig. 2: Amino acids CR for VoC genomes

Such huge changes can be caused by sequencing error as ORF1ab is prone to such errors because it is present at the one end (5‘UTR) of genomes. Such high changes in nucleotides and amino acids in ORF1ab were also found in some sequences from Beta and Omicron.

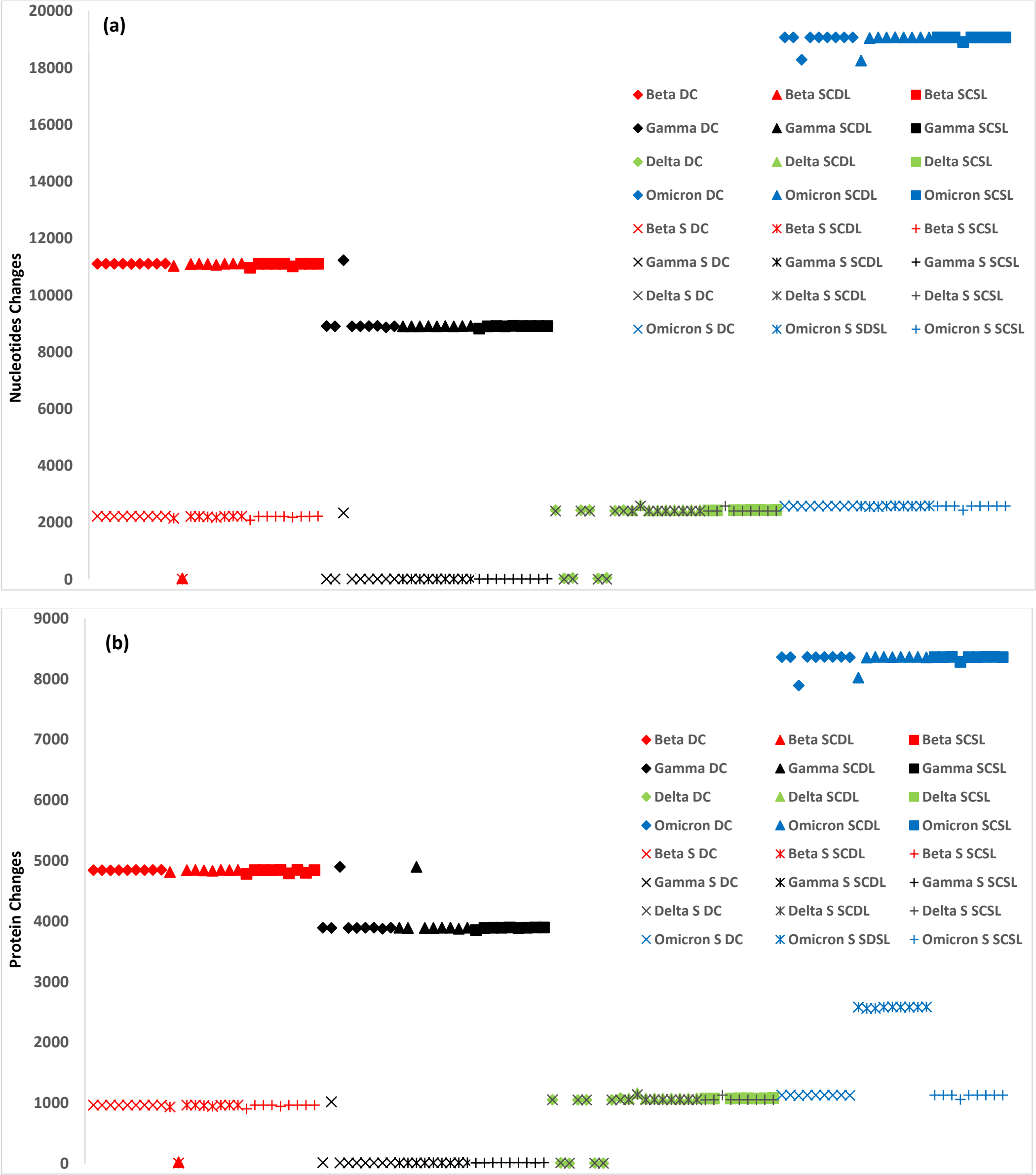

Fig. 3: Nucleotides and amino acid changes in VoC genomes when compared with the *RefSeq*. (a) Nucleotides changes in the whole genomes and in the S protein, and (b) amino acid changes in the whole genomes and in the S protein

For SARS-CoV-2, the reference sequence (*RefSeq*) *NC_045512* in NCBI GenBank with Pango Lineage B is the first genome sequence. This sequence was released in January 2020 by the Public Health Clinical Center and School of Public Health in Shanghai, China [3]. The nucleotides and amino acid changes in VoC genomes when compared with the *RefSeq* are also analyzed (Figure 3). Omicron variant genomes have most nucleotide and amino acid changes compared to other three variants. It is interesting to see that nucleotides and amino acid changes decrease from Beta to Gamma to Delta (15% and 65% decrease from Beta to Gamma and Gamma to Delta respectively). However the increase in changes from Delta to Omicron is huge, with approximately 6.38 times more changes. However, most of the changes in Omicron genomes were not in the S protein. Omicron genomes have approximately 1.2 times more changes in the S protein compared to Delta. Gamma genomes have the least number of changes in the S protein when compared to other three variants. Each of the variant genomes when analyzed with the *RefSeq* have more nucleotides and amino acid changes (8 times more for whole genomes and 4 times for the S gene) compared to nucleotide changes and amino acid changes in each variant genomes when analyzed with each other.

4.4 Nucleotides, Amino Acids and Codon Base Compositions

The composition (occurrence frequency) were computed for the following : (1) nucleotide bases, (2) amino acids, (3) GC and AT/GC contents, and (4) the nucleotide bases at the third position of a codon (A_3, C_3, G_3, T_3) and GC (G or C) at the first, second and third codon position (GC_1, GC_2, GC_3). These composition were computed in all genomes (Table 4) and the S protein (Table 5) of variants respectively.

In five VoC genomes, A and T make up for approximately 62% and the remaining 38% belongs to nucleotides C and G. Similarly, for amino acids, Leucine (L) is the most frequent amino acid with composition of 9.6%, followed by Valine (V) (8.21%) and Thr (T) (7.12%). These results are consistent with the results obtained in [31, 34, 39]. In [31], the Apriori algorithm was used to identify frequent nucleotide bases and it was found that A (29.88%) and T (32.12%) contribute 62%, and C(18.34%) and G (19.64%) contribute the remaining 38%, on average. Similarly, in [34], [39], MEGA X [40], Alfree [41] were used to find frequent nucleotides in SARS-CoV-2 genomes. It was found that A and T contribute 62%, 62.1%, and C and G contribute 38%, 37.9%, approximately.

The average GC and AT ratios were approximately 37.83% and 62.17% respectively, which indicates that all the SARS-CoV-2 genomes are AT rich. The AT/GC ratio for all variants was the same. In all the genomes, GC_1 (46.84%) is the most frequent, followed by GC_2 (38.45%) and GC_3 (28.07%). This means that AT_3 (71.93%) is most frequent, followed by AT_2 (61.46%) and AT_1 (53.16%). Thus, the third position in a codon is also AT rich across the genomes. At the third position in the codon, it is observed that on average T (43.76%) was most frequent, followed by A (28.04%), C (15.15%) and G (12.92%).

In the S protein, GC_1, GC_3, A_3 and G_3 were low. Whereas, GC, GC_2, C_3 and T_3 were high. We find some interesting information related to compositions percentage in variants. For example, GC_1 and A_3 in Delta and Omicron were high compared to Beta and Gamma. Similarly GC_2 was high in Delta (40.72%) and A_3 was high in

Table 4: GC contents and Nucleotides/AAs frequencies

Contents	Alpha	Beta	Gamma	Delta	Omicron
A	29.909	29.865	29.863	29.872	29.897
C	18.084	18.077	18.084	18.070	18.074
G	19.750	19.754	19.746	19.756	19.751
T	32.256	32.303	32.305	32.301	32.276
Ala A	6.846	6.826	6.837	6.837	6.818
Arg R	3.403	3.417	3.405	3.414	3.421
Asn N	5.426	5.424	5.407	5.427	5.376
Asp D	5.088	5.085	5.096	5.066	5.097
Cys C	3.056	3.070	3.068	3.069	3.069
Gln Q	3.646	3.639	3.662	3.667	3.623
Glu E	4.813	4.808	4.794	4.806	4.823
Gly G	5.935	5.939	5.923	5.951	5.909
His H	1.865	1.874	1.859	1.853	1.889
Ile I	5.137	5.172	5.163	5.145	5.181
Leu L	9.685	9.664	9.665	9.683	9.632
Lys K	5.940	5.897	5.919	5.917	5.964
Met M	2.216	2.206	2.206	2.213	2.207
Phe F	5.001	4.996	5.009	4.988	5.028
Pro P	3.905	3.929	3.928	3.889	3.922
Ser S	6.722	6.729	6.731	6.759	6.710
Thr T	7.541	7.494	7.501	7.500	7.497
Trp W	1.108	1.110	1.109	1.109	1.110
Tyr Y	4.529	4.552	4.568	4.553	4.550
Val V	8.126	8.157	8.139	8.136	8.163
GC	37.835	37.831	37.830	37.826	37.825
AT/GC	1.643	1.643	1.643	1.643	1.643
GC1	46.685	46.880	46.877	46.878	46.886
GC2	38.434	38.465	38.460	38.493	38.406
GC3	28.041	28.065	28.090	28.070	28.099
A3	28.089	28.066	28.064	28.089	28.115
C3	15.120	15.162	15.147	15.138	15.157
G3	12.920	12.903	12.943	12.931	12.942
T3	43.717	43.795	43.786	43.809	43.711

Table 5: CG Contents in Spike Protein

Contents	Alpha	Beta	Gamma	Delta	Omicron
GC	38.460	38.619	38.543	38.618	38.545
AT/GC	1.600	1.589	1.594	1.589	1.594
GC1	45.288	44.965	44.813	45.248	45.110
GC2	39.859	39.930	39.996	40.727	39.536
GC3	26.700	26.576	26.587	26.582	26.579
A3	26.974	26.861	26.997	27.033	27.252
C3	15.933	15.917	15.914	15.883	15.691
G3	10.766	10.658	10.672	10.698	10.887
T3	46.267	46.156	46.415	46.384	45.920

Delta and Omicron, compared to other variants. On the other hand, C_3 was low in Delta and Omicron (15.88% and 15.69% respectively). These differences needs further study to investigate their possible association or role in the change rates increase or decrease among variants.

Note that the SARS-CoV-2 genomes in *coding region* and *protein* forms contain two genes ORF1ab (that combines ORF1a and ORF1b) and ORF1a. As ORF1a contents is present in ORF1ab, thus the changes found in ORF1a will also be present in ORF1ab. This duplication increases the nucleotides and amino acid changes. We did not exclude the ORF1a gene as we considered all the genomic data present in the genomes. The methods for genomes analysis in genomes in *coding region* and *protein* forms have one requirement. The additional genes information should be present at the same location (line). In the future, we plan to make this framework more generic that can analyze nucleotides and amino acids without considering the requirement of same locations for genes information as well as adding codon usage bias (CUB) measures in the framework.

5 Conclusion

A framework, called F4GDA, was introduced, which offers various methods to analyze genomic data in various forms for (1) nucleotides and amino acid changes, (2) silent changes due to synonymous codons, and (3) the compositions of nucleotides, amino acids, GC, AT/GC contents, and codon bases.

To demonstrate the usefulness of F4GDA, a case study was presented with genomes of five VoC (divided into three geographical locations) for SARS-CoV-2. The obtained results reveal that genomes for VoC from DC have more changes than SCDL and SCSL. Delta variant genomes have more nucleotides and amino acid changes, followed by the Omicron and Beta variants. One quarter (approximately 24%) of nucleotides and amino acid changes were found in the S protein in Delta. Delta (Gamma) genomes have the least (most) silent changes due to synonymous codons, followed by Omicron, Beta and Alpha. VoC genomes analysis with the SARS-CoV-2 reference sequence revealed that nucleotides and amino acid changes decreased from Beta to Gamma to Delta. However, the changes were increased from Delta to Omicron (6.38 times more changes). The GC_1 and A_3 ratios in the S protein of Delta and Omicron were high compared to Beta and Gamma. Similarly the GC_2 was high in Delta compared to other variants.

F4GDA is not limited to analyzing SARS-CoV-2 and can be used for the comparison and analysis of DNA viruses, metagenomic data, and even human DNA. The framework implementation does not depend on any external libraries/packages, tools, and the genomes can be provided in both FASTA and TXT formats. We believe that F4GDA will be particularly helpful to scientists, biomedical experts, and members of the biocybernetics community who have limited or no programming knowledge. F4GDA offers many interesting future work opportunities, such as:

1. Adding and analyzing more genomic data, not only from NCBI GenBank but also from other databases, such as GISAID, for SARS-COV-2 variants.
2. Enhancing the framework's generality by: (a) Analyzing genomes with varying positions for genes information in *coding region* and *protein* forms, (b) Incorporating the codon usage bias (CUB) or synonymous codon usage bias (SCUB) [42] measures in the framework, (c) Analyzing the difference in various nucleotides at codon third position for their possible association with the nucleotides change rates, (d) Examining and categorizing the nucleotide changes into various types, such as missense, silent and nonsense changes, and (e) Finding the locations of

frequent nucleotides and amino acid changes. This willfacilitate the examination and identification of single nucleotide polymorphisms (SNPs) in variants and their various genes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

Authors did not receive funding for this work.

References

1. E. W. Sayers, M. Cavanaugh, K. Clark, J. Ostell, K. D. Pruitt, and I. Karsch-Mizrachi, “Genbank,” *Nucleic Acids Research*, vol. 48, pp. D84–D86, 2019.
2. K. Kali, G. Saberwal, and G. Sharma, “The lag in SARS-CoV-2 genome submissions to GISAID,” *Nature Biotechnology*, vol. 39, pp. 1058–1060, 2021.
3. F. Wu et al, “A new coronavirus associated with human respiratory disease in China,” *Nature*, vol. 579, pp. 265–529, 2020.
4. H. Zhang, K. Saravanan, Y. Yang, M. Hossain, J. Li, X. Ren, Y. Pan, and Y. Wei, “Deep learning based drug screening for novel coronavirus 2019-ncov,” *Interdisciplinary Sciences: Computational Life Sciences*, vol. 12, no. 3, pp. 368–376, 2020.
5. P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. de Hoon, “Biopython: Freely available Python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.
6. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Y. Yang, and J. Zhang, “Bioconductor: Open software development for computational biology and bioinformatics,” *Genome Biology*, vol. 5, 2004.
7. A. Prlìc, A. Yates, S. E. Bliven, P. W. Rose, J. Jacobsen, P. V. Troshin, M. Chapman, P. E. Bourne, and S. Willis, “BioJava: An open-source framework for bioinformatics in 2012,” *Bioinformatics*, vol. 28, pp. 2693–2695, 08 2012.
8. C.-. COG-UK Consortium, “An integrated national scale SARS-CoV-2 genomic surveillance network,” *Lancet Microbe*, vol. 1, pp. e99–e100, 2020.
9. L. Shen, D. Maglinte, D. Ostrow, U. Pandey, M. Bootwalla, A. Ryutov, A. Govindarajan, , A. R. Judkins, and X. Gai, “Children’s hospital los angeles covid-19 analysis research database (card) - a resource for rapid sars-cov-2 genome identification using interactive online phylogenetic tools,” *biorXiv*, 2020.
10. J. Singer, R. Gifford, M. Cotten, and D. Robertson, “CoV-GLUE: A web application for tracking SARS-CoV-2 genomic variation,” *Preprints*, 2020.
11. Z. Gong et al, “An online coronavirus analysis platform from the national genomics data center,” *Zoological Research*, vol. 41, no. 6, pp. 705–708, 2020.
12. B. Liu, K. Liu, H. Zhang, L. Zhang, Y. Bian, and L. Huang, “Cov-seq, a new tool for sars-cov-2 genome analysis and visualization: Development and usability study,” *Journal of Medical Internet Research*, vol. 22, no. 10, p. e22299, 2020.
13. A. Bernasconi et al, “VirusViz: Comparative analysis and effective visualization of viral nucleotide and amino acid variants,” *Nucleic Acids Research*, vol. 49, no. 15, pp. e90–e90, 2021.
14. I. Alam, A. Radovanovic, R. Incitti, A. A. Kamau, M. Alarawi, E. I. Azhar, and T. Gojobori, “CovMT: an interactive SARS-CoV-2 mutation tracker, with a focus on critical variants,” *Lancet Infectious Diseases*, vol. 21, p. P602, 2021.

15. D. Mercatelli, L. Triboli, E. Fornasari, F. Ray, and F. M. Giorgi, “Coronapp: A web application to annotate and monitor SARS-CoV-2 mutations,” *Journal of Medical Virology*, vol. 93, no. 5, pp. 3238–3245, 2021.
16. Y. Xing, X. Li, X. Gao, and Q. Dong, “MicroGMT: A mutation tracker for SARS-CoV-2 and other microbial genome sequences,” *Frontiers in Microbiology*, vol. 11, 2020.
17. S. Weber, C. Ramirez, and W. Doerfler, “Signal hotspot mutations in SARS-CoV-2 genomes evolve as the virus spreads and actively replicates in different parts of the world,” *Virus Research*, vol. 289, p. 198170., 2020.
18. F. Yuan, L. Wang, Y. Fang, and L. Wang, “Global SNP analysis of 11,183 SARS-CoV-2 strains reveals high genetic diversity,” *Transboundary and Emerging Diseases*, vol. 68, pp. 3288–3304, 2021.
19. N. Ghosh, I. Saha, S. Nandi, and N. Sharma, “Characterisation of SARS-CoV-2 clades based on signature snps unveils continuous evolution,” *Methods*, vol. 203, pp. 282–296, 2021.
20. B. Z. Sia, W. X. Boon, Y. Y. Yap, S. Kumar, and C. H. Ng, “Prediction of the effects of nonsynonymous variants on SARS-CoV-2 proteins,” *F1000Research*, vol. 11, 2022.
21. M. Pachetti, B. Marini, F. Benedetti, F. Giudici, E. Mauro, P. Storici, C. Masciovecchio, S. Angeletti, M. Ciccozzi, R. C. Gallo, D. Zella, and R. I. 11, “Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant,” *Journal of Translational Medicine*, vol. 18, 2020.
22. I. Saha, N. Ghosh, N. Sharma, and S. Nandi, “Hotspot Mutations in SARS-CoV-2,” *Frontiers in Genetics*, vol. 12, 2021.
23. M. Zelenova, A. Ivanova, S. Semyonov, and Y. Gankin, “Analysis of 329,942 SARS-CoV-2 records retrieved from GISAID database,” *Computers in Biology and Medicine*, vol. 138, p. 10948, 2021.
24. N. Ghosh, S. Nandi, and I. Saha, “A review on evolution of emerging SARS-CoV-2 variants based on spike glycoprotein,” *International Immunopharmacology*, vol. 105, p. 108565, 2022.
25. E. C. Rouchka, J. H. Chariker, and D. Chung, “Variant analysis of 1040 SARS-CoV-2 genomes,” *PLoS ONE*, vol. 15, 2020.
26. T. Koyama, D. Platt, and P. L, “Variant analysis of SARS-CoV-2 genomes,” *Bulletin of the World Health Organization*, vol. 98, 2020.
27. X. Ding and X. Guo, “A survey of snp data analysis,” *Big Data Mining and Analytics*, vol. 1, no. 3, pp. 173–190, 2018.
28. H. Ahmed, H. Soliman, and M. Elmogy, “Early detection of alzheimer’s disease using single nucleotide polymorphisms analysis based on gradient boosting tree,” *Computers in Biology and Medicine*, vol. 146, p. 105622, 2022.
29. A. Nagy, S. Pongor, and B. Gyõrffy, “Different mutations in SARS-CoV-2 associate with severe and mild outcome,” *International Journal of Antimicrobial Agents*, vol. 75, 2021.
30. H. Shi, H. Chen, Q. Yang, J. Wang, and H. Shi, “A method for bio-sequence analysis algorithm development based on the par platform,” *Big Data Mining and Analytics*, vol. 6, no. 1, pp. 11–20, 2023.
31. M. S. Nawaz, P. Fournier-Viger, A. Shojaee, and H. Fujita, “Using artificial intelligence techniques for COVID-19 genome analysis,” *Applied Intelligence*, vol. 51, no. 3, pp. 3086–3103, 2021.
32. M. S. Nawaz, P. Fournier-Viger, M. Aslam, W. Li, Y. He, and X. Niu, “Using alignment-free and pattern mining methods for sars-cov-2 genome analysis,” *Appl. Intell.*, vol. 53, no. 19, pp. 21920–21943, 2023.
33. R. K. Pathan, M. Biswas, and M. U. Khandaker, “Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model,” *Chaos, Solitions and Fractals*, vol. 138, p. 110018, 2020.
34. A. Chandra1 and S. B. Chandra, “A bioinformatic analysis of the spike glycoprotein & evolution of COVID-19,” *Medicine Science*, vol. 11, pp. 171–175, 2022.
35. M. Issa, A. E. Hassanien, D. Oliva, A. Helmi, I. Ziedan, and A. Alzohairy, “ASCA-PSO: Adaptive sine cosine optimization algorithm integrated with particle swarm for pairwise local sequence alignment,” *Expert Systems with Applications*, vol. 99, pp. 56–70, 2018.
36. M. Issa, A. M. Helmi, A. H. Elsheikh, and M. A. Elaziz, “A biological sub-sequences detection using integrated BA-PSO based on infection propagation mechanism: Case study COVID-19,” *Expert Systems with Applications*, vol. 189, p. 116063, 2022.
37. F. Wright, “The ’effective number of codons’ used in a gene,” *Gene*, vol. 87, pp. 23–29, 1990.
38. D. Rey and M. Neuhäuser, *Wilcoxon-Signed-Rank Test*, pp. 1658–1659. Springer, 2011.
39. M. S. Nawaz et al, “COVID-19 genome analysis using alignment-free methods,” in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA AIE), 316-328*, 2021.

40. S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura, “MEGA X: Molecular evolutionary genetics analysis across computing platforms,” *Molecular Biology and Evolution*, vol. 35, no. 6, pp. 1547–1549, 2018.
41. A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski, “Alignment-free sequence comparison: Benefits, applications, and tools,” *Genome Biology*, vol. 18, pp. 1–17, 2017.
42. S. K. Behura and D. W. Severson, “Codon usage bias: causative factors, quantification methods and genome-wide patterns: With emphasis on insect genomes,” *Biological Reviews*, vol. 88, no. 1, pp. 49–61, 2013.

Appendix

Table A1: Accession number of genome sequences collected from NCBI GenBank

No	Loc	Alpha	Beta	Gamma	Delta	Omicron
1	DC	MZ292138	OK091660	MZ611956	MW989805	OM212472
2		BS001082	OL980871	MW642248	MZ401459	OL672836
3		MZ914594	MZ944846	MZ277386	MZ397171	OL869974
4		OL548855	MZ317890	MW938104	OM180022	OL988214
5		OL548845	MZ068154	MZ427312	OL779162	OM011974
6		OK433390	MZ413998	MZ477746	OK067236	OM131552
7		OK550252	OL691513	OK433609	OK104589	OM635094
8		MZ914469	OK511530	OK550231	OM180371	BS002408
9		OL517745	OL779106	OM148349	OL336682	OM621556
1	SCDL	MW59728	MW580574	MW520923	MW989805	OL717062
2		OL615972	MW763126	MZ779649	MZ185411	OM322711
3		OL514263	MW808712	MW994525	MZ434516	OM444886
4		OL368928	MW938318	OL538280	OL522772	OM635624
5		OK511092	MZ195719	OK547875	OL524558	OM570097
6		OL522043	MZ927137	OL522250	OL535295	OM615201
7		OL532082	OK233737	OL532393	OL580287	OM622129
8		OL525541	OK183005	OK410835	OL764554	OM618869
9		OK367254	OK171528	OK262610	OL675618	OM621556
1	SCSL	OK547686	MW721421	MW963205	MZ283541	OL764360
2		MW519728	MZ779858	MW869051	MZ434516	OM038110
3		MZ635623	MW909347	MZ286697	MZ491480	OM159296
4		OL468875	MW939890	MZ356676	OL532097	OM197959
5		OL368929	MW938318	MZ386214	OK552549	OM228212
6		OK547770	MZ166476	OL522593	OL685808	OM360659
7		MZ450166	MZ228902	OK121613	OL427108	OM372298
8		OL524451	MZ315302	OL525287	OL419184	OM500947
9		OL535573	MZ307356	OK630532	OM258868	OM619789

Table A2: Results for VoC strains in *coding region form*

Loc	Genomes	ORF1ab	ORF1a	S	ORF3a	E	M	ORF6	ORF7a	ORF7b	ORF8	N	TNC	SC	CR
DC	$1 \rightarrow 2$	$\frac{8(6)}{6(15)}(0)$	$\frac{5(4)}{5(9)}(0)$	$\frac{1(4)}{1(2401)}(0)$	$\frac{1(1)}{1(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(2)}{1(0)}(0)$	$\frac{15(18)}{14(2427)}(0)$	$\frac{7(7)}{7(52)}(0)$	$\frac{0.035(0.042)}{0.4949(5.71)}(0)$
	$2 \rightarrow 3$	$\frac{6(6)}{1(12)}(7319)$	$\frac{4(3)}{0(10)}(1546)$	$\frac{2(3)}{2330(6)}(2264)$	$\frac{0(0)}{1(1)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(3)}(0)$	$\frac{0(1)}{1(0)}(810)$	$\frac{12(14)}{2333(33)}(11939)$	$\frac{3(8)}{46(15)}(278)$	$\frac{0.024(0.032)}{5.497(0.127)}(28.13)$
	$3 \rightarrow 4$	$\frac{7(7)}{4(20)}(7321)$	$\frac{4(2)}{1(15)}(1548)$	$\frac{4(2)}{2330(2399)}(2264)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(4)}(0)$	$\frac{1(1)}{2(0)}(810)$	$\frac{16(13)}{2337(2440)}(11943)$	$\frac{6(9)}{47(58)}(278)$	$\frac{0.037(0.030)}{5.504(5.74)}(28.14)$
	$4 \rightarrow 5$	$\frac{10(9)}{4(26)}(4)$	$\frac{9(4)}{2(18)}(4)$	$\frac{2(2)}{0(1)}(1)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(2)}{2(0)}(0)$	$\frac{22(19)}{8(47)}(9)$	$\frac{9(15)}{5(16)}(2)$	$\frac{0.052(0.044)}{0.212(0.110)}(0.0212)$
	$5 \rightarrow 6$	$\frac{12(9)}{7(14)}(4)$	$\frac{9(4)}{6(9)}(4)$	$\frac{2421(2)}{0(2400)}(1)$	$\frac{1(0)}{0(0)}(1)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(2)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{2(2)}{0(0)}(10)$	$\frac{2447(19)}{13(2426)}(10)$	$\frac{53(12)}{5(54)}(2)$	$\frac{5.80(0.044)}{0.056(5.71)}(0.0235)$
	$6 \rightarrow 7$	$\frac{8(7)}{13(24)}(4)$	$\frac{4(4)}{12(17)}(4)$	$\frac{2422(2)}{0(2)}(2)$	$\frac{1(0)}{0(0)}(1)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(4)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{2(2)}{0(3)}(0)$	$\frac{2439(17)}{25(51)}(11)$	$\frac{51(10)}{13(14)}(3)$	$\frac{5.78(0.040)}{0.084(0.120)}(0.0259)$
	$7 \rightarrow 8$	$\frac{13(7)}{15(7)}(4)$	$\frac{11(6)}{13(3)}(4)$	$\frac{7(3)}{0(2400)}(2)$	$\frac{1(0)}{0(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{2(2)}{3(2)}(1)$	$\frac{34(18)}{31(2415)}(12)$	$\frac{6(11)}{14(54)}(7)$	$\frac{0.080(0.042)}{0.120(5.68)}(0.028)$
	$8 \rightarrow 9$	$\frac{18(8)}{9(4)}(1)$	$\frac{13(7)}{9(3)}(1)$	$\frac{3(2)}{1(1)}(2274)$	$\frac{2(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{2(2)}{4(1)}(0)$	$\frac{39(19)}{21(11)}(2276)$	$\frac{7(6)}{6(1)}(54)$	$\frac{0.092(0.044)}{0.494(0.025)}(5.63)$
SCDL	$1 \rightarrow 2$	$\frac{3(7311)}{6(4)}(0)$	$\frac{0(1539)}{3(3)}(0)$	$\frac{2(2141)}{1(482)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(813)$	$\frac{6(10992)}{9(190)}(813)$	$\frac{2(239)}{1(6)}(30)$	$\frac{0.0142(25.89)}{0.056(0.447)}(2.42)$
	$2 \rightarrow 3$	$\frac{6(7313)}{6(1)}(0)$	$\frac{1(1540)}{3(1)}(0)$	$\frac{3(2207)}{1(184)}(0)$	$\frac{0(2)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{1(0)}{0(1)}(0)$	$\frac{11(11062)}{10(188)}(0)$	$\frac{3(244)}{4(5)}(0)$	$\frac{0.0260(25.05)}{0.056(0.445)}(0)$
	$3 \rightarrow 4$	$\frac{2(9)}{6(4)}(2)$	$\frac{2(5)}{5(2)}(2)$	$\frac{0(2)}{2(2)}(1)$	$\frac{0(2)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(1)}(0)$	$\frac{5(20)}{13(9)}(7)$	$\frac{3(9)}{10(5)}(2)$	$\frac{0.011(0.047)}{0.030(0.028)}(0.0164)$
	$4 \rightarrow 5$	$\frac{7(7)}{6(3)}(1)$	$\frac{7(7)}{5(2)}(1)$	$\frac{0(0)}{2(1)}(2275)$	$\frac{0(1)}{1(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{2(0)}{1(1)}(0)$	$\frac{16(17)}{15(7)}(2277)$	$\frac{5(7)}{9(4)}(54)$	$\frac{0.0379(0.40)}{0.035(0.021)}(5.36)$
	$5 \rightarrow 6$	$\frac{16(11)}{6(2)}(2)$	$\frac{11(8)}{5(0)}(1)$	$\frac{1(1)}{3(0)}(2274)$	$\frac{1(1)}{0(0)}(0)$	$\frac{0(9)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(9)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{1(9)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{4(1)}(0)$	$\frac{31(22)}{18(4)}(2277)$	$\frac{11(10)}{8(4)}(54)$	$\frac{0.0735(0.051)}{0.042(0.009)}(5.36)$
	$6 \rightarrow 7$	$\frac{9(9)}{6(3)}(1)$	$\frac{6(8)}{5(3)}(0)$	$\frac{0(1)}{4(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{2(0)}{(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(2)}{3(0)}(0)$	$\frac{20(20)}{18(7)}(1)$	$\frac{8(6)}{11(5)}(0)$	$\frac{0.0474(0.047)}{0.042(0.014)}(0.007)$
	$7 \rightarrow 8$	$\frac{12(10)}{11(3)}(1)$	$\frac{6(6)}{6(2)}(0)$	$\frac{0(3)}{3(0)}(0)$	$\frac{0(1)}{2(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{2(2)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{4(1)}{1(0)}(0)$	$\frac{26(23)}{23(5)}(1)$	$\frac{11(14)}{10(3)}(0)$	$\frac{0.061(0.054)}{0.054(0.014)}(0.007)$
	$8 \rightarrow 9$	$\frac{15(8)}{19(4)}(2274)$	$\frac{9(5)}{11(3)}(0)$	$\frac{2(2)}{3(2)}(0)$	$\frac{0(1)}{2(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{2(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{3(1)}{1(0)}(0)$	$\frac{32(18)}{36(9)}(2275)$	$\frac{13(9)}{12(3)}(52)$	$\frac{0.075(0.042)}{0.084(0.042)}(5.36)$
SCSL	$1 \rightarrow 2$	$\frac{3(10)}{5(3)}(1)$	$\frac{2(6)}{3(1)}(1)$	$\frac{0(0)}{1(3)}(0)$	$\frac{0(1)}{0(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(2)$	$\frac{6(17)}{9(8)}(5)$	$\frac{0(8)}{4(1)}(3)$	$\frac{0.014(0.040)}{0.035(0.049)}(0.011)$
	$2 \rightarrow 3$	$\frac{4(4)}{6(14777)}(1)$	$\frac{2(3)}{5(9033)}(1)$	$\frac{1(0)}{1(2574)}(0)$	$\frac{0(1)}{1(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(2)}{0(0)}(1)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{9(11)}{13(26385)}(3)$	$\frac{2(7)}{6(543)}(0)$	$\frac{0.021(0.025)}{0.306(62.15)}(0.212)$
	$3 \rightarrow 4$	$\frac{9(1)}{8(14842)}(5)$	$\frac{4(0)}{7(8866)}(4)$	$\frac{1(0)}{3(2574)}(5)$	$\frac{1(0)}{2(0)}(1)$	$\frac{0(0)}{0(0)}(1)$	$\frac{0(1)}{0(0)}(1)$	$\frac{1(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(2)$	$\frac{16(2)}{20(26482)}(19)$	$\frac{3(2)}{9(536)}(5)$	$\frac{0.037(0.004)}{0.047(62.37)}(0.044)$
	$4 \rightarrow 5$	$\frac{5(2)}{14(4)}(4)$	$\frac{3(2)}{9(4)}(3)$	$\frac{0(0)}{3(1)}(5)$	$\frac{1(0)}{2(0)}(0)$	$\frac{0(0)}{0(0)}(1)$	$\frac{1(0)}{0(1)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{1(0)}{2(1)}(1)$	$\frac{12(4)}{31(11)}(15)$	$\frac{4(2)}{11(5)}(5)$	$\frac{0.028(0.01)}{0.073(0.025)}(0.035)$
	$5 \rightarrow 6$	$\frac{6(2)}{14(7)}(0)$	$\frac{4(2)}{8(6)}(0)$	$\frac{0(2)}{3(2)}(1)$	$\frac{1(0)}{1(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{2(0)}{2(1)}(0)$	$\frac{15(6)}{30(16)}(1)$	$\frac{5(2)}{13(10)}(0)$	$\frac{0.035(0.0141)}{0.070(0.042)}(0.007)$
	$6 \rightarrow 7$	$\frac{3(6)}{12(1)}(2)$	$\frac{2(5)}{8(1)}(2)$	$\frac{0(2)}{3(0)}(1)$	$\frac{1(1)}{0(0)}(2)$	$\frac{0(0)}{1(0)}(0)$	$\frac{0(2)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{2(0)}{0(0)}(2)$	$\frac{8(17)}{24(2)}(9)$	$\frac{2(7)}{12(0)}(5)$	$\frac{0.018(0.04)}{0.0565(0.014)}(0.282)$
	$7 \rightarrow 8$	$\frac{15(9)}{10(7)}(4)$	$\frac{10(7)}{4(6)}(4)$	$\frac{5(1)}{3(2)}(0)$	$\frac{0(2)}{1(0)}(2)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{0(1)}{1(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{2(1)}{1(1)}(3)$	$\frac{33(21)}{21(16)}(13)$	$\frac{10(11)}{6(7)}(7)$	$\frac{0.078(0.049)}{0.0495(0.042)}(0.030)$
	$8 \rightarrow 9$	$\frac{20(14)}{6(7)}(2)$	$\frac{13(10)}{5(6)}(2)$	$\frac{5(2)}{7(0)}(0)$	$\frac{1(0)}{2(0)}(0)$	$\frac{0(0)}{0()}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{1(0)}{0()}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{2(1)}{1(3)}(0)$	$\frac{41(27)}{23(17)}(4)$	$\frac{13(15)}{6(7)}(2)$	$\frac{0.097(0.063)}{0.054(0.070)}(0.014)$

Table A3: Results for VoC strains in *protein form*

Loc	Genomes	ORF1ab	ORF1a	S	ORF3a	E	M	ORF6	ORF7a	ORF7b	ORF8	N	PC	MR
DC	$1 \rightarrow 2$	$\frac{4(3)}{3(19)}(0)$	$\frac{3(1)}{1(7)}(0)$	$\frac{0(3)}{3(1050)}(0)$	$\frac{1(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(2)}{0(0)}(0)$	$\frac{8(11)}{7(1069)}(0)$	$\frac{0.056(0.077)}{0.049(7.55)}(0)$
	$2 \rightarrow 3$	$\frac{5(2)}{0(5)}(3204)$	$\frac{3(1)}{0(5)}(662)$	$\frac{1(2)}{1017(5)}(985)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(1)}(339)$	$\frac{9(6)}{1017(18)}(5190)$	$\frac{0.064(0.042)}{7.19(0.127)}(36.72)$
	$3 \rightarrow 4$	$\frac{4(1)}{2(14)}(3206)$	$\frac{3(0)}{0(11)}(662)$	$\frac{2(2)}{1018(1048)}(985)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(1)}{0(1)}(339)$	$\frac{10(4)}{1020(1075)}(5194)$	$\frac{0.071(0.028)}{7.21(7.59)}(36.75)$
	$4 \rightarrow 5$	$\frac{5(1)}{2(17)}(3)$	$\frac{5(0)}{1(12)}(3)$	$\frac{2(2)}{0(1)}(1)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(1)}{0(0)}(0)$	$\frac{13(4)}{3(31)}(7)$	$\frac{0.092(0.028)}{0.021(0.21)}(0.049)$
	$5 \rightarrow 6$	$\frac{5(3)}{4(9)}(3)$	$\frac{5(1)}{4(5)}(3)$	$\frac{1071(1)}{0(1048)}(1)$	$\frac{1(0)}{0(1)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(2)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{2(1)}{0(0)}(0)$	$\frac{1085(7)}{8(1065)}(8)$	$\frac{7.72(0.049)}{0.056(7.52)}(0.056)$
	$6 \rightarrow 7$	$\frac{2(2)}{6(16)}(3)$	$\frac{2(1)}{6(11)}(3)$	$\frac{1072(1)}{0(2)}(1)$	$\frac{1(0)}{0(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(1)}(0)$	$\frac{0(1)}{0(1)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{2(2)}{0(3)}(0)$	$\frac{1080(7)}{12(37)}(8)$	$\frac{7.68(0.049)}{0.084(0.26)}(0.056)$
	$7 \rightarrow 8$	$\frac{10(2)}{7(2)}(1)$	$\frac{6(2)}{7(0)}(1)$	$\frac{7(1)}{3(1048)}(1)$	$\frac{0(0)}{0(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{1(2)}{0(2)}(1)$	$\frac{24(7)}{13(1054)}(5)$	$\frac{0.170(0.049)}{0.120(7.45)}(.035)$
	$8 \rightarrow 9$	$\frac{12(5)}{5(3)}(0)$	$\frac{10(5)}{5(3)}(0)$	$\frac{3(1)}{1(1)}(990)$	$\frac{1(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{1(2)}{4(1)}(0)$	$\frac{28(13)}{15(10)}(990)$	$\frac{0.199(0.09)}{0.106(0.706)}(7.00)$
SCDL	$1 \rightarrow 2$	$\frac{2(3200}{5(2)}(0)$	$\frac{0(659)}{3(1)}(0)$	$\frac{1(930)}{0(87)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(341)$	$\frac{4(4790)}{8(91)}(341)$	$\frac{0.028(33.87)}{0.056(0.643)}(2.42)$
	$2 \rightarrow 3$	$\frac{4(3200)}{4(1)}(0)$	$\frac{1(659)}{2(0)}(0)$	$\frac{2(959)}{0(88)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(1)}(0)$	$\frac{8(4819)}{6(90)}(0)$	$\frac{0.0569(34.08)}{0.042(0.636)}(0)$
	$3 \rightarrow 4$	$\frac{1(5)}{1(2)}(1)$	$\frac{1(3)}{1(1)}(1)$	$\frac{0(1)}{1(1)}(1)$	$\frac{0(2)}{0(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(1)$	$\frac{2(11)}{3(4)}(5)$	$\frac{0.014(0.077)}{0.021(0.028)}(0.035)$
	$4 \rightarrow 5$	$\frac{4(4)}{2(2)}(0)$	$\frac{4(4)}{1(1)}(0)$	$\frac{0(0)}{1(0)}(990)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{1(0)}(0)$	$\frac{9(10)}{6(3)}(990)$	$\frac{0.064(0.070)}{0.042(0.021)}(7.00)$
	$5 \rightarrow 6$	$\frac{10(5)}{3(0)}(1)$	$\frac{6(5)}{3(0)}(0)$	$\frac{1(0)}{1(0)}(990)$	$\frac{1(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{3(0)}(0)$	$\frac{20(12)}{10(0)}(991)$	$\frac{0.124(0.084)}{0.070(0)}(7.01)$
	$6 \rightarrow 7$	$\frac{5(6)}{2(1)}(1)$	$\frac{3(6)}{2(1)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{2(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{2(0)}(2)$	$\frac{12(14)}{7(2)}(1)$	$\frac{0.085(0.099)}{0.049(0.014)}(0.007)$
	$7 \rightarrow 8$	$\frac{6(4)}{6(1)}(0)$	$\frac{3(2)}{3(1)}(0)$	$\frac{0(2)}{1(0)}(0)$	$\frac{0(0)}{2(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{2(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{3(1)}{1(0)}(0)$	$\frac{15(9)}{13(2)}(1)$	$\frac{0.106(0.054)}{0.091(0.014)}(0.007)$
	$8 \rightarrow 9$	$\frac{8(4)}{12(2)}(0)$	$\frac{5(1)}{6(2)}(0)$	$\frac{1(2)}{3(2)}(990)$	$\frac{0(0)}{2(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{2(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{3(1)}{1(0)}(0)$	$\frac{19(9)}{24(6)}(991)$	$\frac{0.135(0.063)}{0.169(0.042)}(7.01)$
SCSL	$1 \rightarrow 2$	$\frac{3(5)}{3(3)}(0)$	$\frac{2(3)}{2(3)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(1)}{0(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(1)$	$\frac{6(9)}{5(7)}(2)$	$\frac{0.042(0.063)}{0.035(0.049)}(0.014)$
	$2 \rightarrow 3$	$\frac{3(1)}{3(6470)}(1)$	$\frac{2(1)}{2(3943)}(1)$	$\frac{1(0)}{1(1132)}(0)$	$\frac{0(1)}{1(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{7(4)}{7(11545)}(3)$	$\frac{0.049(0.028)}{0.4949(81.65)}(0.021)$
	$3 \rightarrow 4$	$\frac{7(0)}{4(6500)}(3)$	$\frac{4(0)}{4(3958)}(1)$	$\frac{1(0)}{2(1132)}(5)$	$\frac{1(0)}{1(0)}(0)$	$\frac{0(0)}{0(0)}(1)$	$\frac{0(0)}{0(1)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(1)$	$\frac{13(0)}{11(11591)}(14)$	$\frac{0.092(0)}{0.077(81.97)}(0.099)$
	$4 \rightarrow 5$	$\frac{3(1)}{9(2)}(2)$	$\frac{2(1)}{7(2)}(1)$	$\frac{0(0)}{2(1)}(5)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{1(0)}{0(0)}(1)$	$\frac{0(0)}{0(1)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{1(0)}{2(0)}(1)$	$\frac{8(2)}{20(6)}(10)$	$\frac{0.056(0.014)}{0.141(0.042)}(0.070)$
	$5 \rightarrow 6$	$\frac{4(1)}{8(4)}(0)$	$\frac{2(1)}{5(4)}(0)$	$\frac{0(2)}{1(1)}(1)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{2(0)}{1(0)}(0)$	$\frac{10(4)}{17(9)}(1)$	$\frac{0.071(0.028)}{0.120(0.063)}(0/007)$
	$6 \rightarrow 7$	$\frac{2(4)}{6(1)}(0)$	$\frac{1(3)}{4(1)}(0)$	$\frac{0(2)}{1(0)}(1)$	$\frac{1(0)}{0(0)}(1)$	$\frac{0(0)}{1(0)}(0)$	$\frac{0(1)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{2(0)}{0(0)}(2)$	$\frac{6(10)}{12(2)}(4)$	$\frac{0.042(0.070)}{0.084(0.014)}(0.0282)$
	$7 \rightarrow 8$	$\frac{10(4)}{5(2)}(1)$	$\frac{6(3)}{4(2)}(1)$	$\frac{4(1)}{2(1)}(0)$	$\frac{0(1)}{1(0)}(1)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{2(1)}{1(1)}(3)$	$\frac{23(10)}{15(6)}(6)$	$\frac{0.163(0.070)}{0.106(0.042)}(0.042)$
	$8 \rightarrow 9$	$\frac{13(5)}{4(3)}(1)$	$\frac{8(5)}{4(3)}(1)$	$\frac{4(1)}{4(0)}(0)$	$\frac{1(0)}{2(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{0(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{0(0)}{1(0)}(0)$	$\frac{1(0)}{0(0)}(0)$	$\frac{0(0)}{0(1)}(0)$	$\frac{1(1)}{1(3)}(0)$	$\frac{28(12)}{17(10)}(2)$	$\frac{0.019(0.084)}{0.120(0.070)}(0.014)$