# Genomic variant analysis of COVID-19 Genomes by Variant Transforms

Haroon Zeb Khan**, Muhammad Munawar Iqbal**
** University of Engineering and Technology, Taxila, Pakistan
Contact: haroon.zeb@student.uettaxila.edu.pk

***Abstract*—Coronavirus strain SARS-CoV-2 is behaving like a cuttle fish, it adopts to any environment,the gloomy face of Coronavirus pandemic is keeping changing multifaceted by regenerating alike starfish as new variants came to surprise the scientists and researchers. In this work we employed Google Variant Transforms tool for processing of COVID–19 VCF files with inclusion of Google big query for genomic variant analysis enveloped on Google Cloud Platform (GCP). We have converted COVID–19 genomic sequences into VCF files by various bioinformatics tools. Google Variant Transforms preprocessor algorithm preprocess COVID–19 VCF files before subjected to process, generated a report about three scrutiny criteria about VCF files, the computation job for Google variant transforms and its preprocessor algorithm jobs managed by Google Dataflow with job graph. We attain a table of COVID–19 variants sites as a variant residue through google big query and displayed results with Google Data studio. Our main finding were storing VCF files into Big query for the analysis of COVID genomes. We aimed that this research will be fruitful for the combating COVID–19 variants.**

***Keywords— Big query, Coronavirus, Genomic Sequences, Variant Transforms, Variant Caller Format (VCF)***

## I. Introduction

Coronavirus Disease (COVID) emerged in the late December of year 2019. The main symptoms were patients complaining of shortness of breath, high-grade-fever, and the most probable differential diagnosis of pneumonia at first was the cause of illness emerged rapidly in the Wuhan province of China. It was believed from an infection of closed environment of seafood and wet markets, after the exposure of these infected markets with people, hospitals were flooded with manifold in the Hubei, Wuhan [1]. At this stage it was not differed as a merely pneumonia or season cold, but the infection of this contagious disease leads to inquisition of the very cause in January 2020, World health organization (WHO) announced that it is coronavirus named it first novel coronavirus-2019 [2]. Coronavirus-2019 is the successor to its predecessors, for example, severe acute respiratory syndrome (SARS-CoV-1) and Middle Eastern Respiratory Syndrome (MERS-CoV). Though the wet market contains some prohibited or banned animals such as cobras, pangolin, bats and the packed environment. later this was identified as infection from these animals to humans. Eventually, humans act as a carrier for the disease to spread dramatically [3]. People which have cardiovascular diseases, asthma, diabetic and aged people are at high risk than any individual and from children to old aged people the transmission is more and in turn children are passive carriers, and should not be overlooked [4], [5]. Afterwards, this SARS-CoV-2 spread at alarming rate throughout the world by the people travel recently amid Wuhan. Europe and first world countries exhibiting full-fledged health system were hit tremendously as compared to third-world countries which have inadequate health resources. Additionally, health workers and paramedical staff got infected rapidly and some of them had died [6]. For this purpose different sets of testing were exercised, For instance, Reverse Transcription Polymerase Chain Reaction (RT-PCR) test [7] and High Resolution Computer Tomography (HRCT) scanning [8] Specific High Sensitivity Enzymatic Reporter Unlocking (SHERLOCK) one-pot(SPOT) [9]. The (SHERLOCK) test takes just one hour and it gives result precisely.At some point, there were 235,434,191 confirmed cases , mortality rate was 4,811,825 and recovered cases were 212,243,461 on global basis [10]. But there are certain medications recommended for COVID–19 and the opposite is also true [11].

Our key Contribution in this research can be summarized as:

**1:** Conceptual research design of genomic idea, this includes SARS-CoV-2 genomes.

**2:** Extracting genomic variants of SARS-CoV-2 genomes, this led to 30 VCF files creation.

**3:** Setting Code for Variant Transforms

The overall structure of this research is enlisted as section I contains introduction. Section II contains related research. In Section III, we present the research methodology and experimental work and in same section we have explained the experimental setup and requirements.In section IV we provide the results and evaluations and in section V, we conclude the paper.

## II. Related Research

The motivation of Coronavirus genomics research is zenith which displays the prominent research that is done since the pandemic occurred. For example,

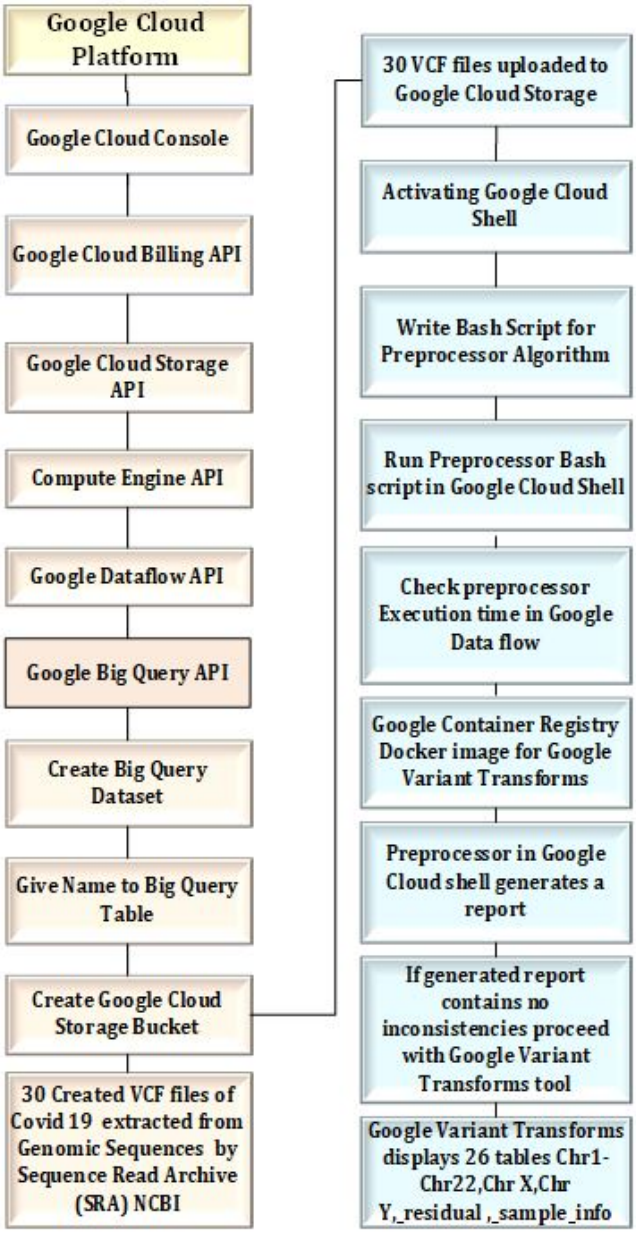



Fig. 1. Variant Transforms Methodology step by step for COVID-19 VCF files

Qian Guo et.al devised a platform Virus Host Prediction(VHP) which can anticipate the prospective virus host and the pattern of infection by deep learning technique. VHP platform was evidently efficient predictive in terms of virus host and the risk of 2019-nCoV is comparatively high. But, there were mistakes in the sequencing and the assembly of the two corona virus [12].

Refat khan et.al, predicted COVID–19 through analysis of nucleotide mutations of the genomes by proposing Long Short Term Memory (LSTM) with recurrent neural network (RNN) time series based shows the outcome of forecasting 400th patient with respect to the frequency of the mutation, but insertion, deletion was not applied [13]. Zhi-Jian et.al, studied the viral genome by an effective tool, BioAider for tracking SARS-CoV-2 and its implementation involving the transmission among humans. This implies to the genomic variations and annotations to guide researchers towards treatment, but with limited genomic data of SARS-CoV-2 [14].

Bo Ram Beck et.al, proposed a deep-learning based forecasting antiviral medications that can be useful to counteract the COVID–19, SARS-CoV-2, which is targeted drug approach ,Molecule transformer-drug target interaction(MTDTI) that suggest a list of antiviral drugs which can be used as for the treatment of SARS-CoV-2. However, no antiviral drug is evidentiary effective [15].

Sara Cleemput et.al, developed an internet-based typing tool acts as a genome spy which can accurately identify and classify coronavirus (SARS-cov-2) by analysis in seconds. However, this method has limitations if genomic sequences exceeds 2000. there is a possibility of incorrect results [16].

Jasper woo chan et.al, did analysis of novel coronavirus by extracting genome from a patient suffering from pneumonia. The results reveal that it is closely related with bats after compared with other types of coronavirus genomes. However, no another host infectivity other than bats, civets or camels [17].

In the literature we have identified the research gap which is to compute genomic variance is the complement between a genetic sample and a reference genome. The pavement of diagnosis of patient disease and to devise innovative treatment methods is the key for the attainment of therapeutic objective. Every variant corresponding sample are retained as particular file format which is variant file format (VCF). In contrast the files are not regarded for the data science and machine learning on such genomic data. Data particularly variants for example, mutations in coronavirus can be identified and retracted to extract meaningful insights to analyses variants and to unravel the genesis of the disease. Secondly, for genomic variance computation and processing demands a state-of-the -art technology such as Google cloud platform (GCP) because it has a specific suite for genomic analysis for instance, Cloud life sciences suite. Additionally, to the best of our knowledge for the genomic variants there was no processing platform of high standard that can handle gigantic variants data and processed in no time. For example, Variant Transforms has flourished Mayo Clinic genomic data platform and sparked many VCF files by Google's Variant Transforms.

## III. Research Methodology and experimental work

This section includes the methodology of the research that contains various steps. First, we downloaded Coronavirus (SARS-CoV-2) COVID–19 reference genome from the National Center for Biotechnology Information (NCBI) website in the

FASTA format [18]. This reference genome was indexed by Burrows Wheeler Alignment (BWA) producing resulting indexed reference genome. Secondly, we downloaded SARS-CoV-2 genome sequences through NCBI Sequence Read Archive (SRA) study SRP266465 involves COVID–19 patients of Massachusetts General Hospital data given by Broad Institute of Harvard and MIT to NCBI. After that we used some of the accessions [19]. There were a total of 30 SRA accessions which can be listed with their respective SRA accession ID, sequence type ,number of RNA nitrogenous bases. For example, adenine, cytosine, uracil guanine bases, lastly the size of the genomic sequence as shown in Table 1. We squashed each SRA accession using fasterq-dump [20] to split each accession in two parts in the fasterq format these parts were run through BWA with eight threads as threshold. This multi-threading results in one Sequence Alignment Map (SAM) file using paired read strategy. Fourthly, we converted SAM file into Binary Alignment Map (BAM) file using SAM tools. Further the BAM file was sorted. Fifthly, we indexed the reference genome using SAM tools. For the generation of variant likelihoods, the sorted BAM file produced raw Binary variant calling format (BCF) file which was then processed to generated both variant BCF files and finally it generates Variant Caller Format (VCF) files respectively using BCF tools. For 30 SRA accessions we generate 30 VCF files. we opted Google Cloud Platform GCP Cloud Life sciences suite for our work. These 30 VCF files were uploaded to Google Cloud Storage Bucket created genomevt-v-1, in the multi-region as location type, we had implemented Google Variant Transforms for the processing of 30 VCF files. The step-by-step approach for the Google Variant Transforms as shown in figure 1

Comprises of first uploading VCF files into the cloud storage and then run bash script using wild card approach for all the 30 VCF files at once through Variant Transforms Preprocessor in Google Cloud Shell which could check for any inconsistency in the subjected VCF files. Google Dataflow compute the Variant transforms job VCF-to-big query-preprocess lasts for approximately 10 minutes. The preprocess prints a report about the VCF files in tab separated values (TSV) format as VCF-VCFreport-report.tsv and also resolved headers of all 30 VCF files in the VCF format, as resolved-headers.VCF. For our scope of research, the VCF report in the tsv fortunately displayed three scrutiny criteria for the processed VCF files a) No Headers Conflicts Found, b) No Inferred Headers Found, c) No Malformed Records Found. It is worth to mention that before subjecting VCF files to Google variant Transforms. First run through Variant transforms preprocessor algorithm because it could check for any inconsistency and the reverse of this if we subject our VCF files without running preprocessor

TABLE I
30 Genomic Covid-19 RNA Sequences

| S.No | SRA Accession ID | Assay Type | Cases | Bytes |
|---|---|---|---|---|
| 1 | SRR11953670 | RNA-Seq | 155.19 M | 50.14 Mb |
| 2 | SRR11953671 | RNA-Seq | 329.73 M | 112.15 Mb |
| 3 | SRR11953672 | RNA-Seq | 287.81M | 98.84 Mb |
| 4 | SRR11953673 | RNA-Seq | 1.83G | 602.02 Mb |
| 5 | SRR11953674 | RNA-Seq | 210.79 M | 65.32 Mb |
| 6 | SRR11953675 | RNA-Seq | 375.25 M | 126.16 Mb |
| 7 | SRR11953676 | RNA-Seq | 34.65 M | 17.00 Mb |
| 8 | SRR11953677 | RNA-Seq | 286.86 M | 91.17 Mb |
| 9 | SRR11953678 | RNA-Seq | 426.54 M | 141.15 Mb |
| 10 | SRR11953679 | RNA-Seq | 4.73 G | 1.49 Gb |
| 11 | SRR11953680 | RNA-Seq | 122.49 M | 38.41 Mb |
| 12 | SRR11953681 | RNA-Seq | 172.60 M | 60.65 Mb |
| 13 | SRR11953683 | RNA-Seq | 86.34 M | 26.99 Mb |
| 14 | SRR11953684 | RNA-Seq | 88.02 M | 28.05 Mb |
| 15 | SRR11953685 | RNA-Seq | 249.69 M | 82.19 Mb |
| 16 | SRR11953686 | RNA-Seq | 227.45 M | 71.69 Mb |
| 17 | SRR11953687 | RNA-Seq | 28.38 M | 14.17 Mb |
| 18 | SRR11953688 | RNA-Seq | 7.39 M | 2.30 Mb |
| 19 | SRR11953689 | RNA-Seq | 368.23 M | 119.66 Mb |
| 20 | SRR11953690 | RNA-Seq | 126.67 M | 39.56 Mb |
| 21 | SRR11953691 | RNA-Seq | 168.78 M | 53.01 Mb |
| 22 | SRR11953692 | RNA-Seq | 175.44 M | 55.67 Mb |
| 23 | SRR11953693 | RNA-Seq | 255.16 M | 82.28 Mb |
| 24 | SRR11953694 | RNA-Seq | 261.46 M | 83.45 Mb |
| 25 | SRR11953695 | RNA-Seq | 94.99 M | 34.19 Mb |
| 26 | SRR11953696 | RNA-Seq | 179.42 M | 54.99 Mb |
| 27 | SRR11953697 | RNA-Seq | 330.67 M | 109.22 Mb |
| 28 | SRR11953698 | RNA-Seq | 262.83 Mb | 91.46 Mb |
| 29 | SRR11953699 | RNA-Seq | 1.95 G | 667.67 Mb |
| 30 | SRR11953700 | RNA-Seq | 27.79 M | 9.82 Mb |

then if any of the VCF file has inconsistency will abort Variant transforms for running further. In case of inconsistency among VCF files, the preprocessor will run accordingly and it will display any deformity in the above mentioned three scrutiny's criteria.so it is caution for the researcher not to run these malformed VCF files to variant transforms.

### *A. experimental setup and requirements*

This section complements implementation of the research including hardware required and software required for the presented research. We have implemented this work by having hp Pavilion laptop with CPU Intel Core -i5 7th Generation, Hard drive 1TB, RAM 6GB, GPU NVidia GEFORCE 940MX with CUDA 10.00, Tensor RT 6.0.1.5 Operating Systems Windows 10 64 bit and for creating COVID -19 VCF files we used Xubuntu, 20.04 LTS.in a virtualized environment. In addition, we used Google Cloud platform integrated with both mentioned operating systems respectively.

## IV. results and evaluation

This section describes the results obtained in this work as in the methodology section we used wild card approach for the 30 VCF files present in the Google Cloud storage Bucket named as genomevt-v-1. In addition, we implemented Google big query. First, we created Big Query dataset named gencovt in the United States (US) data as geographic location. We named our table for our VCF files result as cvariant-covid19, we run a variant transforms bash script in Google Cloud Shell, Google Dataflow

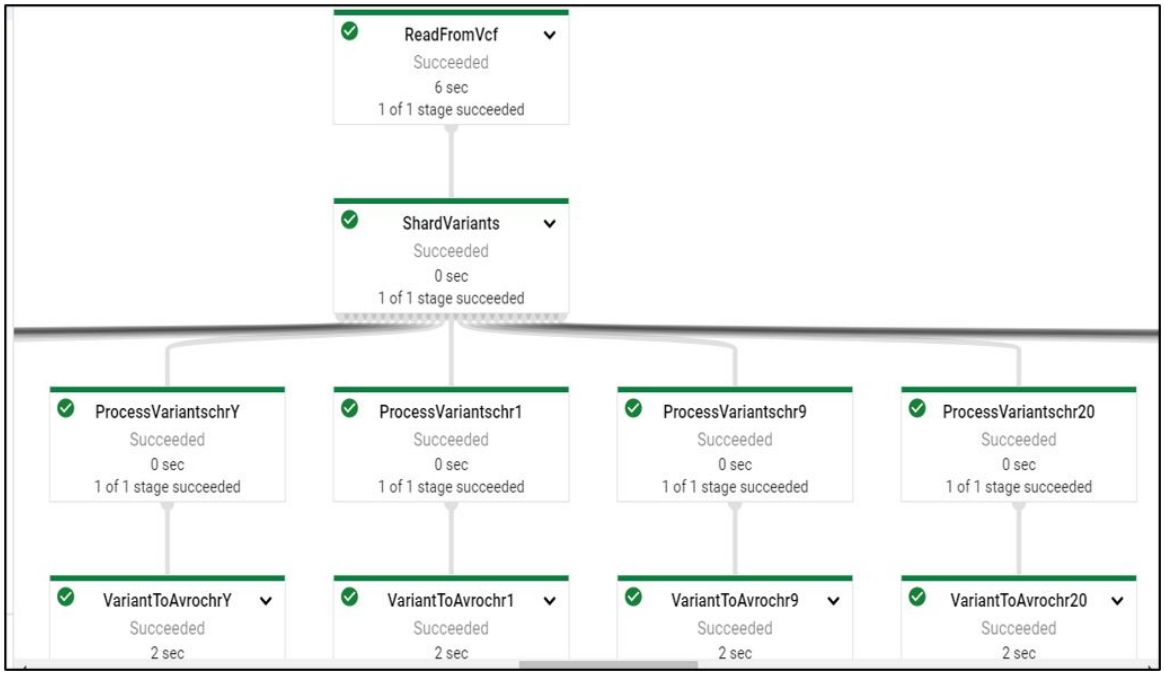



Fig. 2. Variant Transform Algorithm

computed the job VCF-to-big query, it lasted for approximately 15 minutes. After completion of the variant transforms job in the big query it generated a total of 26 tables for variants of SARS-CoV-2, which, can be listed here (from Chr1-Chr22) are autosomal chromosomes, whereas, cvariant-covid19-ChrX, cvaraint-covid19-ChrY, are (pair of Sex Chromosomes), cvaraint-covid19-residual, (SARS-CoV-2 Variants table) cvariant-covid19-sample-info (listed 30 VCF files as Sample Table) as shown in table II. respectively.

TABLE II
VARAINT TRANSFORM GENERATED BIGQUERY TABLE

| S.No | Big query data set and table name | Big query Nested table names | Chromosome type |
|---|---|---|---|
| 1 | gencovt, cvariant_covid19 | cvariant_covid19__Chr1 | Autosmal |
| 2 | gencovt, cvariant_covid19 | cvariant_covid19__Chr2 | Autosmal |
| 3 | gencovt, cvariant_covid19 | cvariant_covid19__Chr3 | Autosmal |
| 4 | gencovt, cvariant_covid19 | cvariant_covid19__Chr4 | Autosmal |
| 5 | gencovt, cvariant_covid19 | cvariant_covid19__Chr5 | Autosmal |
| 6 | gencovt, cvariant_covid19 | cvariant_covid19__Chr6 | Autosmal |
| 7 | gencovt, cvariant_covid19 | cvariant_covid19__Chr7 | Autosmal |
| 8 | gencovt, cvariant_covid19 | cvariant_covid19__Chr8 | Autosmal |
| 9 | gencovt, cvariant_covid19 | cvariant_covid19__Chr9 | Autosmal |
| 10 | gencovt, cvariant_covid19 | cvariant_covid19__Chr10 | Autosmal |
| 11 | gencovt, cvariant_covid19 | cvariant_covid19__Chr11 | Autosmal |
| 12 | gencovt, cvariant_covid19 | cvariant_covid19__Chr12 | Autosmal |
| 13 | gencovt, cvariant_covid19 | cvariant_covid19__Chr13 | Autosmal |
| 14 | gencovt, cvariant_covid19 | cvariant_covid19__Chr14 | Autosmal |
| 15 | gencovt, cvariant_covid19 | cvariant_covid19__Chr15 | Autosmal |
| 16 | gencovt, cvariant_covid19 | cvariant_covid19__Chr16 | Autosmal |
| 17 | gencovt, cvariant_covid19 | cvariant_covid19__Chr17 | Autosmal |
| 18 | gencovt, cvariant_covid19 | cvariant_covid19__Chr18 | Autosmal |
| 19 | gencovt, cvariant_covid19 | cvariant_covid19__Chr19 | Autosmal |
| 20 | gencovt, cvariant_covid19 | cvariant_covid19__Chr20 | Autosmal |
| 21 | gencovt, cvariant_covid19 | cvariant_covid19__Chr21 | Autosmal |
| 22 | gencovt, cvariant_covid19 | cvariant_covid19__Chr22 | Autosmal |
| 23 | gencovt, cvariant_covid19 | cvariant_covid19__ChrX | sex |
| 24 | gencovt, cvariant_covid19 | cvariant_covid19__ChrY | sex |
| 25 | gencovt, cvariant_covid19 | cvariant_covid19__residual (Variants Table) | —— |
| 26 | gencovt, cvariant_covid19 | cvariant_covid19__sample_info (Sample Table) | —— |

In these 26 tables created by Variant Transforms tool in the big query only two tables cvariant-covid19-residual and cvariant-covid19-sample-info was non empty while the rest of tables were empty, which describes that the dataset was based on viral genome of coronavirus, as viruses does not possess chromosomes, however, the living organisms retains chromosomes. The sample-info table contains information about the sample-id, sample-name, file-path, and ingestion-time, while, the residual table comprises of all relevant information describing the variants. Through big query depending on the extraction of information big query uses SQL query for the display of results in the tabular form. We presented the google dataflow computation job graph for variant transform algorithm as shown in figure 2. we shown Google variant transforms google data flow job graph which is different from the preprocessor graph as shown in figure 3.The algorithm of this job graph by google data flow starts from reading VCF files and then it shards all variants from all subjected VCF files and afterwards it process variants first autosomal chromosomes and sex chromosomes finally it processed variants transformed to Avro chromosomes, where Avro is file format for storing processed variants data. The Google big query generated two non-empty tables one describes the 30 VCF files sample details as shown in table III and the other table consists of variants information as shown in table IV. However, Table III is fixed whereas table IV is presented here as one of data science scenario which is obtained from Master table of big query containing 2924 rows correspondingly 2924 records. We created table III which has reference-name, reference-bases, alternate bases. alt, start-position ,end-position, Mann-Whitney U Test values which comprises of Read position bias (RPB), Base

TABLE III
RESIDUE TABLE FOR COVID-19 VCF FILES

| S.No | Start Position | End Position | Reference Name | Reference Bases | MQB | MQSB | BQB | PV4 | RPB | SGB | VBD | Record count |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 18484 | 19872 | NC_045512.2 | C | 1 | 1 | 0.909091 | 1 | 0.818182 | -0.453602 | 0.6 | 4 |
| 2 | 28360 | 28360 | NC_045512.2 | T | 0.999953 | 0.995664 | 0.975468 | 1 | 0.233929 | -0.693147 | 0.130402 | 3 |
| 3 | 2416 | 2416 | NC_045512.2 | C | 1 | 0.999987 | 1 | 1 | 1 | -0.693147 | 0.997474 | 3 |
| 4 | 28337 | 28337 | NC_045512.2 | G | 1 | 0.999138 | 1 | 1 | 1 | -0.693147 | 0.991373 | 3 |
| 5 | 28883 | 28883 | NC_045512.2 | G | 0.990479 | 0.868433 | 0.0578885 | 1 | 0.231979 | -0.693147 | 8.62216e-7 | 3 |
| 6 | 16875 | 16883 | NC_045512.2 | TACAAC AAC | null | 0.95494 | null | 1 | null | -0.453602 | 0.505913 | 3 |
| 7 | 25843 | 25843 | NC_045512.2 | A | 1 | 1 | 0.825177 | 1 | 0.00006248 | -0.651104 | 0.000019721 | 3 |
| 8 | 25843 | 25843 | NC_045512.2 | A | 1 | 1 | 0.887746 | 1 | 0.109217 | -0.616816 | 0.000481013 | 3 |
| 9 | 26233 | 26233 | NC_045512.2 | G | 1 | 1 | 1 | 1 | 1 | -0.693147 | 0.000041484 | 3 |
| 10 | 23403 | 23403 | NC_045512.2 | A | 1 | 0.994879 | 1 | 1 | 1 | -0.693147 | 0.143395 | 3 |
| 11 | 14408 | 14408 | NC_045512.2 | C | 0.999265 | 0.027975 | 0.0822216 | 1 | 0.0979562 | -0.693147 | 0.0350496 | 3 |
| 12 | 3037 | 3037 | NC_045512.2 | C | 1 | 0.984749 | 1 | 1 | 1 | -0.693147 | 0.00192569 | 3 |
| 13 | 25563 | 25563 | NC_045512.2 | G | 1 | 0.999986 | 1 | 1 | 1 | -0.693147 | 1.66242e-7 | 3 |
| 14 | 18848 | 18848 | NC_045512.2 | C | 1 | 1 | 0.948064 | 1 | 0.0265162 | -0.556411 | 0.0162466 | 3 |
| 15 | 19872 | 19872 | NC_045512.2 | G | 1 | 1 | 0.366897 | 1 | 0.960687 | -0.590765 | 0.00187095 | 3 |

TABLE IV
COVID-19 30 VCF FILES SAMPLE$_i nfoTable$

| S.No | Sample id | Sample name | File path | Ingestion date | Record count |
|---|---|---|---|---|---|
| 1. | 3216550734874377043 | SRR11953670.sort.bam | gs://genomevt_v-1/vcf/SRR11953670.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 2. | 352489629759615578 | SRR11953671.sort.bam | gs://genomevt_v-1/vcf/SRR11953671.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 3. | 8227663251833541077 | SRR11953672.sort.bam | gs://genomevt_v-1/vcf/SRR11953672.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 4. | 8357062031852223960 | SRR11953673.sort.bam | gs://genomevt_v-1/vcf/SRR11953673.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 5. | 2806340148173514704 | SRR11953674.sort.bam | gs://genomevt_v-1/vcf/SRR11953674.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 6. | 2582873836293763020 | SRR11953675.sort.bam | gs://genomevt_v-1/vcf/SRR11953675.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 7. | 979035968381476808 | SRR11953676sort.bam | gs://genomevt_v-1/vcf/SRR11953676.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 8. | 5277621863161793105 | SRR11953677.sort.bam | gs://genomevt_v-1/vcf/SRR11953677.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 9. | 566504859183568350 | SRR11953678sort.bam | gs://genomevt_v-1/vcf/SRR11953678var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 10. | 3833680178983260965 | SRR11953679.sort.bam | gs://genomevt_v-1/vcf/SRR11953679.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 11. | 8533412563562316116 | SRR11953680sort.bam | gs://genomevt_v-1/vcf/SRR11953680.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 12. | 3519872865606448254 | SRR11953681.sort.bam | gs://genomevt_v-1/vcf/SRR11953681.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 13. | 3662038835489796241 | SRR11953683.sort.bam | gs://genomevt_v-1/vcf/SRR11953683.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 14. | 5890621020233963195 | SRR11953684.sort.bam | gs://genomevt_v-1/vcf/SRR11953684.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 15. | 5316490262417333523 | SRR11953685.sort.bam | gs://genomevt_v-1/vcf/SRR11953685.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 16. | 150785628643285227 | SRR11953686.sort.bam | gs://genomevt_v-1/vcf/SRR11953686.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 17. | 7413376885039050630 | SRR11953687.sort.bam | gs://genomevt_v-1/vcf/SRR11953687.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 18. | 668647478787251403 | SRR11953688.sort.bam | gs://genomevt_v-1/vcf/SRR11953688.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 19. | 2040752609674665819 | SRR11953689.sort.bam | gs://genomevt_v-1/vcf/SRR11953689.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 20. | 3189181725568227223 | SRR11953690.sort.bam | gs://genomevt_v-1/vcf/SRR11953690.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 21. | 7689594917302935664 | SRR11953691sort.bam | gs://genomevt_v-1/vcf/SRR11953691.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 22. | 3349372544095480697 | SRR11953692.sort.bam | gs://genomevt_v-1/vcf/SRR11953692.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 23. | 7351685286731755226 | SRR11953693.sort.bam | gs://genomevt_v-1/vcf/SRR11953693.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 24. | 8127585703603632648 | SRR11953694.sort.bam | gs://genomevt_v-1/vcf/SRR11953694.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 25. | 2543651269845840959 | SRR11953695.sort.bam | gs://genomevt_v-1/vcf/SRR11953695.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 26. | 1311637592895431099 | SRR11953696.sort.bam | gs://genomevt_v-1/vcf/SRR11953696var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 27. | 1542741616589133359 | SRR11953697.sort.bam | gs://genomevt_v-1/vcf/SRR11953697.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 28. | 5909100411738514128 | SRR11953698.sort.bam | gs://genomevt_v-1/vcf/SRR11953698.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 29. | 3344475419929969819 | SRR11953699.sort.bam | gs://genomevt_v-1/vcf/SRR11953699.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |
| 30. | 5297186131858853415 | SRR11953700.sort.bam | gs://genomevt_v-1/vcf/SRR11953700.var-final.vcf | 2021-03-09 13:04:00 UTC | 1 |

quality Bias (BQB) , Mapping quality bias (MQB),and Mapping quality vs Strand bias(MQSB) and the final field Variant Distance Bias (VDB), which checks for random collection of variant bases in the region where mapped reads are present.

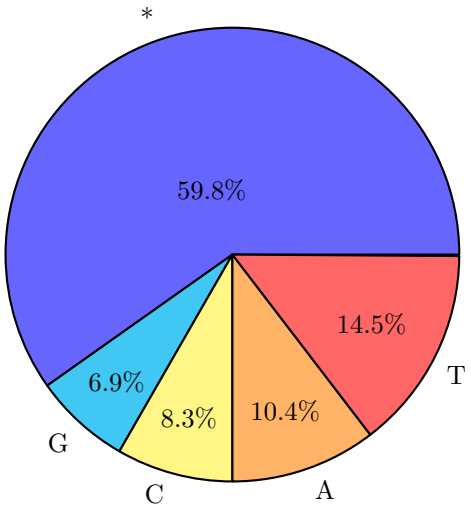
The tables created by google big query and produced results by google data studio and visualization represented by stacked bar graph indicates Mann-Whitney U test for VCF files we created, describes maximum value of 1 and symbolized as (*) and all the remaining refrence bases close to value of 1 as shown in figure 3. The pie chart talks about alternate bases.alt approximately 59.8 % describes alternate bases symbolically represented as (*) ,14.5% presents Thymine (T),10.4 tells about Adenine (A), 8.3 % describes cytosine (C) participation and finally 6.9 % presents Guanine(G) with respect to alternate bases.alt as shown in figure 3.

The retrospective literature that we presented did not have this research that we exploited using Coronavirus genomic variant analyses on the enveloped Google Cloud platform: Google Variant Transforms, Big query, Dataflow so this research is incomparable.

## V. CONCLUSIONS

There is no doubt that world is facing challenging situation of Coronavirus pandemic. It is been now three years on. Globally, uptil now 236,726,445 coronavirus cases in which 4,834,210 mortality rate and approximately 213,850,684 were recovered. First world countries are effected badly as compared with Third world countries despite of full-fledged healthcare systems and the reverse is also true. In the present time amid covid 19 now it has transformed as covid 20 because of a variety of variants of covid For example UK variant B.1.17 just appeared in the late December 2020 which was found to be 80% transmissible than

Fig. 3. Pie chart displays alternate bases.alt




Wuhan variant. Similarly the other variant which is called South African variant lineage of B1.351 variant. we have used Google Cloud Platform Cloud Life sciences Suite. So in the essence of our work summaries there is going to be more variants as time progresses.The importance of variants will be paramount leading to drug or vaccine discovery for the Coronavirus research. In this work we created 30 variant caller format files from Coronavirus genome sequences using its reference genome, then we first run Variant Transforms tool Preprocessor algorithm for the 30 VCF files, then after successful execution it generated report which found no inconsistency which was bottleneck for further analysis of variants using variant transforms running bash script. It is good practice to first run VCF files through variant transforms preprocessor tool, the reason behind this it alarms researcher for inconsistency in the VCF files and caution for further analysis of the variants. Moreover, we created 26 tables, out of 26 tables, 2 tables were reserved for residual and sample information. The processing time calculated by the Google data flow for the preprocess phase it accounted for 10 minutes and for non-Preprocess phase it accounted for nearly 15 minutes.

## References


[1] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, "The proximal origin of sars-cov-2," *Nature medicine*, vol. 26, no. 4, pp. 450–452, 2020.

[2] J. Riou and C. L. Althaus, "Pattern of early human-to-human transmission of wuhan 2019 novel coronavirus (2019-ncov), december 2019 to january 2020," *Eurosurveillance*, vol. 25, no. 4, p. 2000058, 2020.

[3] P.-I. Lee and P.-R. Hsueh, "Emerging threats from zoonotic coronaviruses-from sars and mers to 2019-ncov," *Journal of*

[4] L. Mousavizadeh and S. Ghasemi, "Genotype and phenotype of covid-19: Their roles in pathogenesis," *Journal of Microbiology, Immunology and Infection*, vol. 54, no. 2, pp. 159–163, 2021.

*microbiology, immunology, and infection*, vol. 53, no. 3, p. 365, 2020.

[5] A. L. Mueller, M. S. McNamara, and D. A. Sinclair, "Why does covid-19 disproportionately affect older people?" *Aging (albany NY)*, vol. 12, no. 10, p. 9959, 2020.

[6] M. Mehdi, M. Waseem, M. H. Rehm, N. Aziz, S. Anjum, and M. A. Javid, "Depression and anxiety in health care workers during covid-19." *Biomedica*, vol. 36, 2020.

[7] A. Tahamtan and A. Ardebili, "Real-time rt-pcr in covid-19 detection: issues affecting the results," *Expert review of molecular diagnostics*, vol. 20, no. 5, pp. 453–454, 2020.

[8] M. Lins, J. Vandevenne, M. Thillai, B. R. Lavon, M. Lanclus, S. Bonte, R. Godon, I. Kendall, J. De Backer, and W. De Backer, "Assessment of small pulmonary blood vessels in covid-19 patients using hrct," *Academic radiology*, vol. 27, no. 10, pp. 1449–1455, 2020.

[9] M. Patchsung, K. Jantarug, A. Pattama, K. Aphicho, S. Suraritdechachai, P. Meesawat, K. Sappakhaw, N. Leelahakorn, T. Ruenkam, T. Wongsatit *et al.*, "Clinical validation of a cas13-based assay for the detection of sars-cov-2 rna," *Nature biomedical engineering*, vol. 4, no. 12, pp. 1140–1149, 2020.

[10] D. Worldometer, "Covid-19 coronavirus pandemic," *World Health Organization, www. worldometers. info*, 2020.

[11] C. Wu, Y. Liu, Y. Yang, P. Zhang, W. Zhong, Y. Wang, Q. Wang, Y. Xu, M. Li, X. Li *et al.*, "Analysis of therapeutic targets for sars-cov-2 and discovery of potential drugs by computational methods," *Acta Pharmaceutica Sinica B*, vol. 10, no. 5, pp. 766–788, 2020.

[12] Q. Guo, M. Li, C. Wang, P. Wang, Z. Fang, S. Wu, Y. Xiao, H. Zhu *et al.*, "Host and infectivity prediction of wuhan 2019 novel coronavirus using deep learning algorithm," *BioRxiv*, 2020.

[13] 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, Solitons & Fractals*, vol. 138, p. 110018, 2020.

[14] Z.-J. Zhou, Y. Qiu, Y. Pu, X. Huang, and X.-Y. Ge, "Bioaider: An efficient tool for viral genome analysis and its application in tracing sars-cov-2 transmission," *Sustainable cities and society*, vol. 63, p. 102466, 2020.

[15] B. R. Beck, B. Shin, Y. Choi, S. Park, and K. Kang, "Predicting commercially available antiviral drugs that may act on the novel coronavirus (sars-cov-2) through a drug-target interaction deep learning model," *Computational and structural biotechnology journal*, vol. 18, pp. 784–790, 2020.

[16] S. Cleemput, W. Dumon, V. Fonseca, W. Abdool Karim, M. Giovanetti, L. C. Alcantara, K. Deforche, and T. De Oliveira, "Genome detective coronavirus typing tool for rapid identification and characterization of novel coronavirus genomes," *Bioinformatics*, vol. 36, no. 11, pp. 3552–3555, 2020.

[17] J. F.-W. Chan, K.-H. Kok, Z. Zhu, H. Chu, K. K.-W. To, S. Yuan, and K.-Y. Yuen, "Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting wuhan," *Emerging microbes & infections*, vol. 9, no. 1, pp. 221–236, 2020.

[18] X. Xia, "Dating the common ancestor from an ncbi tree of 83688 high-quality and full-length sars-cov-2 genomes," *Viruses*, vol. 13, no. 9, p. 1790, 2021.

[19] National Center for Biotechnology Information, "Run Selector :: NCBI," 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP266465{\&}o=acc{_}s{\%}3Aa{\&}s=SRR11953670,SRR11953671,SRR11953672,SRR11953673,SRR11953674,SRR11953675,SRR11953676,SRR11953677,SRR11953678,SRR11953679,SRR11953680,SRR11953681,SRR11953683,SRR11953684,SRR11953685,SRR1

[20] R. Li, K. Hu, H. Liu, M. R. Green, and L. J. Zhu, "Onestoprnaseq: a web application for comprehensive and efficient analyses of rna-seq data," *Genes*, vol. 11, no. 10, p. 1165, 2020.