NaDa Samy mahmoud

Paper Name: *COVID SEQUENCE, a New Tool for SARS-CoV-2 Genome Analysis and Visualization: Development and Usability Study.*

Abstract:

The genomes of SARS **COVID-19** rapidly sequenced and to keep up with updates and evolution scientists want to refresh and re-clean data sets but scientists have limited with Bioinformatics tools and programming to analyse the sequences so to handle these problems they developed **COVID -19** SEQ by using "web server "which facilitate analysing the sequence implemented in python and JavaScript using web server the results when we have a new sequence **COVID -19** SEQ predicts Gene Boundaries ,the locations of genes ,identifies genetic variations ,identifying elements on the genome ,a process called gene prediction and attaching biological information to these elements by using **A command- line interface** is available for high throughput processing so the conclusions scientists have developed the SARS-COV-2 sequences and they handled web service for fast and easy analysis so the web server provides an interactive module for analysis and annotations the genome they thought that understanding the genome will help to know a whole sequences an predict any variations in the sequences .

Introduction:

Coronavirus are a large family of viruses that are known to cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory syndrome and severe acute respiratory syndrome "SARS", coronavirus can transmitted from person to person and to understand its evolution and genetics scientists have sequenced SARS COVID-19 from patients and make statistics from group of people with different ages and genders so now we have a huge data to take a genomic sequences to keep up with the latest updates developments scientists need to frequently download and clean a new data sets , **COVID -19** SEQ consists of different components : a data analysis (**FASTA** sequences and generates variant call sets in **VCF** *"variant call format* "also known as **processed files** : includes a genomic data and **ORF** *"open reading frames"* is the part of reading frame that has the ability to be translated , the pipeline automatically filters low –quality sequences and remove duplicate sequences , performs sequence alignment and identifies ,annotates genetic variant , we use a web server to enable the rapid analysis of sequences without

using programming web interface includes an interactive genome visualizer and tabulated displays of genetic variants and ORF predictions further we use command-line interface to facilitate data sharing .

Related work:

Comparison With Prior Works the existing software packages VAPID **[1],** and viral Genome ORF reader(VIGOR) **[2]**, focus on **gene annotations** to our knowledge a software package that identifies, annotates, and visualizes genetic variants of SARS-CoV-2 has not previously been created [1][2], Researchers developed the COV-Seq-2web server for fast and easy analysis of **SARS-COV-2** sequences, they aim that **COV-SEQ** will help with improving our understanding of the genetics of **COVID-19**,in the future ,they plan to expand the scope of COV-SEQ to include other viruses but now **CoV-Seq-2 is currently limited to SARS-CoV-2 sequences** ,The web server does not allow custom reference sequences other than **COVID-19** We chose to focus on this virus because it has constituted the majority of processing requests during the **COVID-19 pandemic**. They plan to provide additional functionality to accept custom reference sequences in a future release, **COVID -19** SEQ consists of different components : a data analysis that takes **FASTA** sequences and generates variant call sets in **VCF** ***"variant call format*** "also known as **processed files** : includes a **genomic data** and **ORF** ***"open reading frames"*** is the part of reading frame that has the ability to be translated, they use a web server to enable the rapid analysis of sequences without using programming web interface.

Materials and Methods :

1-Data collection: the most sequences for SARS-COV-19 genomic are collected in (GISAID,NCBI,ENA,CNGB)databases provide the option of downloading data in groups using **selenium python tool "** https://selenium-python.readthedocs.io/" .

2- Preprocessing: They aggregated COVID-19 sequences from databases and filtered these genomes using lenient cut off of 25,000 nucleotides so doing removed incomplete genomes while retaining complete genomes (NCBI, ENA)are part of international nucleotide sequence database collaboration "**INSDC**" they make a comparing by the 2 ways and removed duplicate sequences and if they have identical genomic sequences these suspect duplications were marked in the data but not removed cause can infect multiple patients.

3-pairwise sequence Alignment: they performed pairwise alignment against the reference sequence NCBI by using multiple Alignment using "**FAST FOURIER TRANSFORM**"

4-varient calling: they used a custom Python script for variant calling, in which they considered single nucleotide polymorphisms (SNPs), insertions, and deletions.

5-ORF Boundry Detection: they used a method similar to viral Annotation pipeline and identification "**VAPID**" and translated the coordinates of ORF boundaries from the

reference genome to the query genome by using their pairwise alignment for multi-segments ORF they annonated each segment independently and combined them afterward.

6-interactive visualization: the COVID-19 sequence web server is hosted on "**AWS**" amzon web services ,after the user submits data throw either a text box for a single sequence or file upload for an arbitrary number of sequences the back end program will perform pairwise sequence alignment ,variant calling and **ORF** boundry detection to generate results in **VCF** and **JSON** formats and the front end templates will then render genome sequences on a new page using **JSON** data the results page highlights mutations on submitted sequences against the reference sequence and shows details upon cursor hover Users can also zoom in to check the details of the genomic sequence. Selecting a specific ORF will expand the display to show the mutation table, the ORF table, and the gene sequence for the selected ORF.

Results:

In this paper they have described COVID-19 sequences consists of several components by using a web server and command- line interface CLI or "**terminal**" processing many sequences at once and enables rapid analysis ,interactive visualization module accepts custom sequences as input and display **the genetic variant and ORF boundary detection** on an interactive genome browser , to downstream analysis with publicly available data, CoV-Seq-2 provides downloadable analysis results with sequence metadata and genetic mutations, to facilitate downstream analysis they aggregated sars-cov-2 from (GISAID,NCBI,ENA,CNGB) and annontetion genetic variant ,location and collection data for each sequence, all aggregate information can be download from the web server, we provide statistics on the geographical and chronological distributions of sequence submissions and encourage further analysis by other scientists.

Acknowledgment:

We acknowledge Dr.Sara for her highly appreciated efforts with us and hope this paper be a good one to other students who want to know about COVID-19 sequences and tools used in this field.

References:

1. Shean RC, Makhsous N, Stoddard GD, Lin MJ, Greninger AL. VAPiD: a lightweight cross-platform viral annotation pipeline and identification tool to facilitate virus genome submissions to NCBI GenBank. BMC Bioinformatics 2019 Jan 23;20(1):48
2. Wang S, Sundaram JP, Spiro D. VIGOR, an annotation program for small viral genomes. BMC Bioinformatics 2010 Sep 07;11:451 .