Aya Ahmed AbduElmenaem Mohammed

Third Year [Medical Informatics Department]

CoV-Seq, a New Tool for SARS-CoV-2 Genome Analysis and Visualization: Development and Usability Study

Abstract

COVID-19 was quickly declared a global pandemic after its discovery in late 2019. SARS-CoV-2 genomes are being sequenced and shared on public repositories at a rapid rate. Scientists must periodically refresh data sets in order to keep up with these changes, which is a time-consuming and labor-intensive process. Furthermore, scientists with little bioinformatics or programming experience may struggle. To fix these issues, we created CoV-Seq, an interactive web server that allows researchers to analyze SARS-CoV-2 genomes quickly and easily. In Python and JavaScript, CoV-Seq is implemented. This article includes the URLs for the web server and source code. We created CoV-Seq, a web-based tool for analyzing custom SARS-CoV-2 sequences quickly and easily. The web server includes an interactive module for personalized sequence analysis and a weekly modified index of genetic variations for all publicly available SARS-CoV-2 sequences. We think CoV-Seq would aid in our knowledge of COVID-19's genetic underpinnings.

Introduction

The novel coronavirus SARS-CoV-2 has sparked an epidemic of viral pneumonia and has turned into a global pandemic since its discovery in late 2019. While attempts to curb its spread, SARS-CoV-2 had infected nearly 33 million patients and killed nearly 1 million people worldwide as of late September 2020 . Scientists sequenced the SARS-CoV-2 genome to learn more about its evolution and genetics. To fix these issues, we created the CoV-Seq framework. A data processing system that takes FASTA sequences and produces variant callers in variant call format (VCF) and open reading frame (ORF) predictions is part of CoV-Seq. The pipeline detects and annotates gene mutations while filtering reduced sequences, removing duplicates, aligning sequences, and identifying and filtering low-quality sequences. We have a web server that allows non-programmers to quickly analyze custom sequences. An integrated genome visualizer and tabulated views of genetic variations and ORF projections are included in the web interface Both of the findings are available for download for further review. We also have a present predominantly for increased processing in settings. We compiled SARS-CoV-2 molecules from the Global Initiative on Exchanging Bird Flu Sample, the Bio technology Information, the European Nucleic acid Database, and China National GeneBank to make data sharing easier.

Related Works

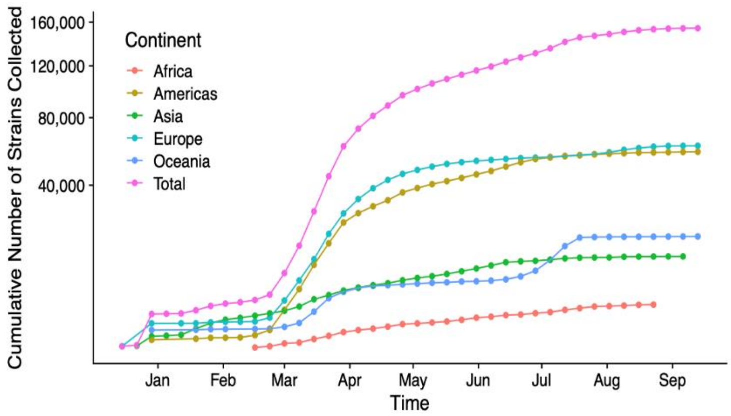
https://publichealth.jmir.org/2020/4/e23542?utm_source=TrendMD&utm_medium=cpc&utm_campaign=JMIR_TrendMD_0
Peter Forster et al., J Med Internet Res, 2020.
https://bioinform.jmir.org/2021/1/e25995/citations?utm_source=TrendMD&utm_medium=cpc&utm_campaign=JMIR_Bioinformatics_and_Biotechnology_TrendMD_0
Emilio Mastriani et al., JMIR Bioinformatics and Biotechnology, 2021.
https://www.genomeweb.com/informatics/center-genomic-regulations-covid-19-viral-beacon-digs-thousands-sars-cov-2-genomes?utm_source=TrendMD&utm_medium=TrendMD&utm_campaign=1&trendmd-shared=1#.X2tf0WhKhPY
Neil Versel, Genome Web.
https://gh.bmj.com/content/6/1/e004408?utm_campaign=bmjgh&utm_content=consumer&utm_medium=cpc&utm_source=trendmd&utm_term=usage-042019
Lu Lu et al., Global Health, 2021.
-Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534 [http://europepmc.org/abstract/MED/32087114][https://dx.doi.org/10.1016/S1473-3099(20)30120-1] [https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=32087114&dopt=Abstract]

Methods

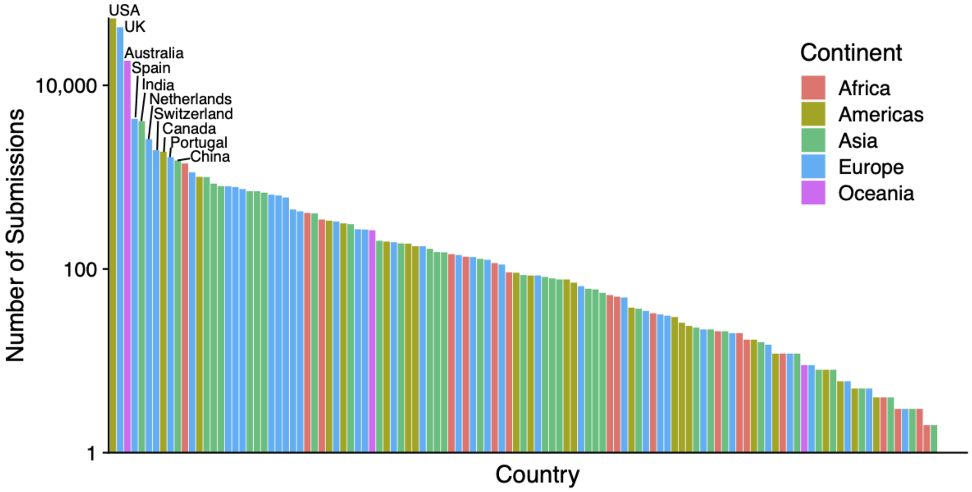
The entirety of currently accessible SARS-CoV-2 sequenced genomes are stored in the GISAID, NCBI, ENA, and CNGB databases. The ability to download information in batches is available in all databases. To handle the firmware upgrade procedure, we use Firefox. GISAID, NCBI, ENA, and CNGB sequences for SARS-CoV-2 were combined. Several samples constituted unfinished chromosomes, with just a specific gene in such cases. We used a liberal threshold of 25,000 nucleotides to extract these genomes, which excluded noticeably fragmented genomes while maintaining genomic information. Because NCBI and ENA are both members of the International Nucleotide Sequence Database Collaboration (INSDC), we found duplicate submissions that we were able to eliminate by comparing accession IDs. Dual submissions can also show in GISAID and INSDC with separate accession IDs. If two reports contained identical genomic sequences, we judged them to be questionable duplications. For variant calling, we employed a bespoke Python script that took into account single nucleotide polymorphisms (SNPs), insertions, and deletions. We used buff tools to left-normalize each variant and deleted sample with too many variations, which might indicate a sequencing mistake. Because we excluded samples with extraordinarily large numbers of variations while preserving the majority of samples, we adopted a liberal threshold of 350 variations.

Results

Over time, the number of SARS-CoV-2 genomes reported to GISAID, NCBI, ENA, and CNGB has risen. The total number of segments by collecting date is shown in (Figure 1). Korea's patterns have been consistently increasing until January, while combinations from other continents expanded gradually at first but then accelerated dramatically in March. It's worth noting that, while the number of entries corresponds with the total number of instances, the number of comments does not necessarily represent the real number of instances (Figure 2). Five European nations (United Kingdom, Spain, Portugal, the Netherlands, and Switzerland), two Asian nations (India and China), two North American nations, and one Ocean country make up the top 10 nations with the most entries.



(Figuer1)



(Figure2)