Lifebit SARS-CoV-2 Dataset

To accelerate progress and help researchers around the world save time in combating the Covid-19 virus, Lifebit has made freely available, on a global basis, the dataset for SARS-CoV-2 from NCBI.

The dataset consists of viral sequence data for SARS-CoV-2 and includes all next generation sequencing runs for SARS-CoV-2 from SRA, associated metadata, and the virus reference genome. The dataset contains more than 25GB of raw sequence data from 13 different projects and over 300 FASTQ files.

DETAILS

- Data is hosted in AWS S3 to facilitate collaboration in the cloud
- Previously, the data was hosted on AWS in the US region only; now it is also available on AWS in the EU, eu-west-1
- Because Lifebit has stored the raw FASTQ files, researchers no longer need to use the SRA toolkit to download and convert files from .sra format, saving time
- Lifebit provides an independent backup of the data

TECHNICAL

The AWS S3 bucket is:
s3://lifebit-sars-cov-2.

This contains a folder for the reference genome and the raw sequence data/reads. The reads folder contains the metadata for all reads. The reads are grouped together/organised by Project ID. The data can be downloaded using AWS (e.g. `aws s3 sync`) or by downloading the public links (e.g. with `wget`).

For a complete list of available files please see:

https://github.com/lifebit-ai/SARS-CoV-2/blob/master/assets/ucsc/aws_https_links.txt

hello@lifebit.ai

lifebit