

# INSDC Statement on SARS-CoV-2 sequence data sharing during COVID-19

Guy Cochrane (EMBL-EBI), Ilene Karsch-Mizrachi (NCBI-NLM-NIH) and Masanori Arita (DDBJ) on behalf of the International Nucleotide Sequence Database Collaboration

The databases of the International Nucleotide Sequence Database Collaboration (INSDC; <http://www.insdc.org/>) capture, organise, preserve and present nucleotide sequence data as part of the open scientific record. INSDC member institutions – the EMBL European Bioinformatics Institute (EMBL-EBI)<sup>1</sup>, the NIG DNA Data Bank of Japan (NIG-DDBJ)<sup>2</sup> and the National Library of Medicine's National Center for Biotechnology Information at NIH (NCBI)<sup>3</sup> – are committed to the continued delivery of this critical element of scientific infrastructure.

The global COVID-19 crisis has brought an urgent need for the rapid open sharing of data relating to the outbreak. Most importantly, access to sequence data from the SARS-CoV-2 viral genome is essential for our understanding of the biology and spread of COVID-19. To aid in that effort, all three INSDC members have prioritized processing of SARS-CoV-2 sequence data and have streamlined the submission process.

Availability of data through INSDC databases provides:

- Rapid open access – INSDC quickly makes submitted data freely available to everyone, without restrictions on reuse
- Linkage of raw sequence read data to genome assemblies, providing researchers with the ability to validate the integrity of assemblies and investigate asserted mutations and changes in genome sequences
- Integration of SARS-CoV-2 sequences with entirety of INSDC data, including related coronaviruses genome sequences, enabling comparison across species
- Linkage of sequences to the published literature, enhancing the discovery process
- Integrated data analysis tools, such as BLAST, to further understanding of the virus

In support of the global response to the COVID-19 crisis, the INSDC calls upon the research community to:

- Submit raw SARS-CoV-2 data to the databases of the INSDC
- Submit consensus/assembled SARS-CoV-2 data to the databases of the INSDC
- Provide information relating to the sequenced isolate or sample as part of the sequence submission; minimally the time and place of isolation/sampling and an isolate/sample identifier should be provided to maximise the value of the sequences.
- In cases where scientists have already established submissions to other databases, these submissions should continue in parallel to the INSDC submission

The integration of INSDC databases with the global bioinformatics data infrastructure, including tools, secondary databases, compute capacity and curation processes, assures the rapid dissemination of data and drives its maximal impact.

---

<sup>1</sup> <https://www.ebi.ac.uk/>

<sup>2</sup> <https://www.ddbj.nig.ac.jp/>

<sup>3</sup> <https://www.ncbi.nlm.nih.gov/>

In addition to these fundamental roles of INSDC member institutions in the sharing of viral sequence data, each institution has rapidly established COVID-19-specific programmes and resources: the [European COVID-19 Data Platform](#) from EMBL-EBI, the DDBJ's [Research Data Resources on New Coronavirus](#) and the [NCBI SARS-CoV-2 Resources](#). These resources both demonstrate the connectedness of INSDC databases to broader bioinformatics initiatives and serve to add immediate value to COVID-19 research.