COG-UK – Trusted Research Environment Data Dictionary v1.0 02/07/2021

The below table contains a summary of COG-UK SARS-CoV-2 sequencing data transferred to UK trusted research environments. COG-UK data will be updated weekly with full data refresh and versioning information.

The central_sample_id (COG-ID) will be removed from all files and replaced with a study specific identifier. **Attempts at reidentification of individuals or combining TRE held COG-UK data with public COG-UK data is expressly forbidden under the terms of the data sharing agreement.**

Table 1: COG-UK data summary

File	URL	Data fields	Brief summary of file	Methods used to generate	Updated / version
Metadata	https://cog-uk.s3.climb.ac.uk/phylogenetics/latest/cog_metadata.csv	sequence_name, country, adm1, is_pillar_2, sample_date, epi_week, lineage, lineages_version, lineage_conflict, lineage_ambiguity_score, scorpio_call, scorpio_support, scorpio_conflict, n439k, del_21765_6, t1001i, del_1605_3, p681h, q27stop, mutations, y453f, p323l, e484k, d614g, n501y, a222v	File contains basic public metadata, including sequence_name, location, date, pangolin lineage assignment, version and associated scores, scorpio VOC/VUI constellation call and associated scores, key spike protein mutations calls and a list of all nucleotide mutations found	Generated by datapipe (https://github.com/COG-UK/datapipe) version v1.0.2 Pangolin and scorpio always run with most up to date software/model files using auto-update, version provided in metadata	02/07/2021 - tagged datapipe version 07/07/2021- added metadata output and remove lines from metadata corresponding to duplicate central-sample-id 16/07/2021 - added civet3 outputs
All sequences	https://cog-uk.s3.climb.ac.uk/phylogenetics/latest/cog_all.fasta		Consensus sequences for all COG-UK samples after deduplication by `central_sample_id`	Generated by datapipe (https://github.com/COG-UK/datapipe) version v1.0.2	
Unmasked alignment	https://cog-uk.s3.climb.ac.uk/phylogenetics/latest/cog_unmasked_alignment.fasta		Full alignment of all above COG-UK sequences, discarding insertions relative to reference `MN908947.3`	Generated by datapipe (https://github.com/COG-UK/datapipe) version v1.0.2	
Alignment	https://cog-uk.s3.climb.ac.uk/phylogenetics/latest/cog_alignment.fasta		Trimmed and filtered alignment of above COG-UK sequences, removing sequences with >95% ambiguous bases, and 5' and 3' UTRs of each sequence masked by Ns. A subset of these are combined with global sequences to create the COG-UK phylogenetic tree (includes almost all sequences from last 35 days).	Generated by datapipe (https://github.com/COG-UK/datapipe) version v1.0.2	

Genomic_ variants_ table	/cephfs/covid/bham/results/variants/latest/naive_variant_table.csv [**Table structure to be confirmed – query whether additional data/time column can be included**]	"Position" INTEGER, "Reference_Base" TEXT, "Alternate_Base" TEXT, "Is_Indel" INTEGER);	Summary of mutations present in SARS-CoV-2 in comparison to xxx reference		

Publications making use of COG-UK data, should include the COG-UK consortium as an author – please contact: **coguk_pubs@medschl.cam.ac.uk** for details.