

## Global *Meningitis* Genome Library: The Power of Curation

Authors: Bratcher, H., Parfitt KM., Maiden, M., and Jolley, K.

Meningitis is a devastating disease and remains a major public health challenge in regions and countries around the world.

The goal of the 'Global Meningitis Genome Library' (GMGL) data collection within PubMLST, <http://pubmlst.org/gmgl>, is to coordinate and host publicly available, global genome libraries for curated sets of meningitis-causing bacterial pathogens. A curated genome library is defined as a coherent, preferably published, set of isolates that include isolate provenance records and annotated genome sequence data.

The database strives to include only non-personally identifiable provenance and epidemiological data linked to an isolate genome so that published analysis can be replicated. This data (year and country of isolation, sample source, age range, and patient sex) is used to stratify the information found in whole genome sequencing and offers a powerful opportunity for enhanced molecular surveillance.

The work focus is divided between four organisms: *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Streptococcus agalactiae*. The surveillance of serogroup B meningococcal (MenB) disease is of particular importance, given that it is the only remaining major disease-causing meningococcal serogroup for which a polysaccharide conjugate vaccine is unavailable.

The GMGL isolates make up 15% of the total global *Neisseria meningitidis* database. As of July 2023, there are 10,764 curated GMGL isolate records (10,741 genome assemblies). There are 262 publications, with currently still 1,445 isolates, still awaiting assignment to one or more publications. There were 426 and 319 new genomes added to the GMGL in 2022 and 2023, respectively, with data reuse of 1,214 genomes in 2022.

The inclusion of provenance data has been at the forefront of the GMGL curation, with all (100%) isolates now having country data. The inclusion of epidemiological year data has increased by 38.9% since February 2021, detailing this for 98.0% of isolates. Data for patient sex and isolation source are now available for 10.0% and 26.2% of isolates, respectively. Whilst we can celebrate the success of the successful increase of provenance data curation, we are still working hard on filling in the gaps of missing data. Our aim is to continue to gather information from publication authors for addition to the database.