

AEGIS Open Data Release

Version 1.0

Data Release Policy: Ancient Environmental Genomics Initiative for Sustainability project

The Ancient Environmental Genomics Initiative for Sustainability project (AEGIS) is a partnership led by the Globe Institute, University of Copenhagen Denmark with partners at the University of Cambridge UK, Wellcome Sanger Institute UK, Institut Pasteur France, the University of Berkeley USA, the University of Colorado at Boulder USA, the University of Zurich Switzerland, MARUM, Bremen, Germany, EMBL European Bioinformatics Institute, Wageningen University Netherlands, and Seoul National University Korea. AEGIS receives funding from the Novo Nordisk Foundation and the Wellcome Trust.

AEGIS intends that all the sequence data generated by the project will be openly available for reuse, and is therefore depositing raw and assembled data in the public sequence databases (European Nucleotide Archive [ENA], and from there in the other International Nucleotide Sequence Database Collaboration [INSDC] nodes - GenBank and the DNA Data Bank of Japan). Many physical specimens will also be deposited in relevant national museum collections.

AEGIS data are released freely for reuse for any purpose upon deposition in ENA, and the AEGIS partners encourage such community reuse, respecting the origins of the samples and any rights arising from those noted on the ENA accessions. Our intention is to rapidly publish all submitted full eukaryote genome assemblies as Genome Note data publications. Raw and assembled metagenomic data and ancient environmental DNA survey data will be published by deposition in ENA and INSDC, and through descriptive data notes. We expect that users of the data will give appropriate acknowledgement and citation, and encourage contact to establish collaborations on large scale data analysis.

AEGIS may also make available for download intermediate data and assemblies *via* a project website. These data and assemblies are provided as is as a service to the community, and we make no assurances as to their completeness or quality. Please note that intermediate data products will be altered before final submission to ENA and we cannot guarantee persistence or availability of intermediate files in the long term. We strongly recommend that published analyses are based on data and assemblies submitted to ENA/INSDC. The reference genome sequences submitted to ENA by AEGIS will be annotated and presented through the EBI Ensembl database, and the annotations presented through Ensembl should be regarded as the official AEGIS versions.

AEGIS will be publishing integrated analyses across the data we produce, both collectively, and as individual partners, and we encourage researchers planning large scale analyses to contact us so that we can collaborate effectively to mutual advantage.

About this document

Author: Mark Blaxter

Date: 06 Nov 2020

Version: 0.9 modified from DToL policy [MB]

Purpose: To deliver a data release statement for both pre-submission and post-submission AEGIS raw, metagenome and genome data that is clear and precise.

Status:

- 0.1 Released to AEGIS Leadership Team for comment and editing. 06 November 2024
- 1.0 Adopted by the AEGIS Steering Committee. 11/11/2024
 - With respect to authorship on Genome Notes, it was agreed that AEGIS authors would in general be covered by a single AEGIS consortium authorship, with only the sample provider and other appropriate non-AEGIS individual and consortium authors as necessary.
 - With respect to data submissions, it was agreed that the BioSample objects for primary aeDNA data sets would contain full location coordinates, typically to GPS resolution, but most likely not estimated ages for the sample, since these in general are not known or are subject to change up to publication. There would be a unique identifier that would cross-reference to other metadata at AEGIS. To be discussed further.