

Abstract

Sequencing communities at scale: dealing with the data

Guy Cochrane (EBI, UK)

Sequence-based approaches to understanding environmental and host-associated communities of organisms are in growth, driven by rapid advance in, and decreasing cost of, sequencing technologies. This growth brings many technical and human challenges. On the technical side, increasing volume demands greater sophistication for data management and sharing; raw data have increasing breadth (e.g. from metabarcoding through metagenomics to metatranscriptomics); data processing follows different paths (e.g. assembly, gene catalogue construction). On the human side, a userbase of increasing size and diversity must necessarily be engaged in the process of sharing data, applying data standards and enabling the science supported by the sequencing methods. The European Nucleotide Archive (ENA), the database of record for public domain sequence data, has a long history in the environmental sequencing domain and has fostered data sharing amongst metagenomicists from the start. As the underlying data management system for EMBL-EBI Metagenomics and a core infrastructure within the emerging Metagenome Exchange, the ENA is an important resource for those involved at all scales in environmental sequencing.

In the talk, I will introduce ENA and outline its core services. I will discuss the technical and human challenges that arise from environmental sequencing and highlight a number of areas of work, including data standards, reporting systems and our new content discovery interface, in which we seek to rise to these challenges.