How good are Triple IDs in ENA? (Topic 2)

Aim/problem/goal

Large sets of records in the European Nucleotide Archive (ENA) reference specimens (e.g. as specimen_voucher, bio_material, or culture_collection) through the use of GBIF triple IDs, that is a concatenation of institution code, collection code and catalog number. However, it is unclear whether these links correctly reference specimen records in GBIF as, for example, the collection code is sometimes omitted. The goal is to investigate how reliable the triple IDs are and to develop methods for improving them by inspecting additional data items (e.g. gathering date and country). Reliable links between ENA sequences and GBIF specimens would (a) allow users to follow links between both infrastructures, (b) feed metadata on voucher specimens from GBIF to ENA (which are often poorer on ENA), and (c) add publication references on ENA sequences to the corresponding GBIF specimens.

Method

The planned method for this task was to

1. Download ENA sequence records based on vouchers
   (referenced by specimen_voucher or bio_material or culture_collection)
2. Download/access potential specimens from GBIF
   (identified by inst/coll code and/or catalog number)
3. Develop matching algorithm
   - Triple ID based, if it exists
   - If not by examining metadata items (taxonomy, gathering date/country)
4. Result: List of (potential) matches (maybe with flag)

However, it soon became clear that these steps had already been done by GBIF. Shortly before the hackathon, GBIF had imported the sequences of interest from ENA. As all GBIF records, the clustering algorithm processed these records, already grouping 720k ENA sequences in clusters with their corresponding specimens. This team therefore joined the team that was working on enhancing the GBIF clustering algorithms (Topic 3) and worked on increasing this number by improving the algorithm.

Results and Conclusion

See summary topic 3 (Suppl. material 3)