interPopula: Database and tool integration for population genetics using Python

Tiago Antao

Liverpool School of Tropical Medicine, University of Liverpool, UK. E-mail: tiagoantao@gmail.com

URL: http://popgen.eu/soft/interPop – Code URL: https://launchpad.net/interpopula – License: GPL v3

Population genetics suffers from a structural lack of Application Programming Interfaces (APIs) to interact with existing public databases. InterPopula address that problem by providing a Python library to interact with several of such databases. The initial version of the library and supporting scripts is mainly concerned with human population genetics and genomics by providing support to the HapMap project and also the UCSC Know Genes database (a part of the UCSC Genome Browser). HapMap – http://hapmap.ncbi.nlm.nih.gov/ – is a freely available dataset of human DNA sequence variation (SNP based) currently covering 11 different human populations around the globe. UCSC Know Genes is a database constructed from Swiss-Prot/TrEMBL (UniProt) and the associated mRNA data from Genbank using a fully automated process to predict genes along a genome.

InterPopula has three distinct objectives:

1. Provide an API to access the HapMap dataset and the UCSC Know Genes database (and, in the future, other UCSC Genome Browser tables). Currently, and as far as we know, there is no alternative API to interPopula, in any language, to access HapMap data. The Known Gene database (for which no API is also known, though it is possible to interact with the database with an SQL interface) supports more species, therefore making interPopula of (limited) use also for non-human studies. The API also allows for the export of HapMap data in Genepop file format. This format, being the de facto standard in non-sequence based population genetics, permits the analysis of HapMap data using a vast array of software applications which are able to import Genepop files.

2. Make available a set of scripts – based on the above library – that can serve not only as useful utilities, but also as examples of database and external tool integration. Currently we provide examples of integration with Entrez databases (nucleotide and SNP), the Genepop population genetics suite and charting libraries. Integration with Entrez databases and Genepop is achieved through Biopython.

3. A set of guidelines and scripts was developed in order to facilitate a consistent view across heterogeneous databases. HapMap, UCSC Known Gene and the Entrez databases might not be fully consistent among themselves and, if care is not taken, database integration efforts might lead to erroneous results.

InterPopula includes a core library, a set of scripts and documentation. The code includes unit testing support in order to maintain code quality. The development infrastructure is based on distributed version control (Bazaar) over the Launchpad hosting platform.

Future development plans include the support of more databases from the UCSC Genome Browser. As this repository supports multiple species, this will make a interPopula less centered on humans and more a general purpose (multi-species) population genetics suite. Long term efforts might include supporting other databases, depending on user feedback.