PyPop - a software pipeline for large-scale multilocus population genomics

Alex K. Lancaster1,3*, Richard M. Single2, Owen D. Solberg1, Mark P. Nelson1 and Glenys Thomson1

April 29, 2010

1 Department of Integrative Biology, University of California, Berkeley, 3060 Valley Life Sciences, Berkeley CA 94720-3140 USA
2 Department of Mathematics & Statistics, University of Vermont, 306 Mansfield House, Burlington, VT 05405-1455 USA
3 Present address: Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142-1479 USA
*Correspondence: alexl@wi.mit.edu

PyPop (Python for Population Genomics) is an open-source framework for performing large-scale population genetic analyses on multilocus genotype and allele frequency data. It computes tests and measures of Hardy-Weinberg equilibrium (locus-level and individual genotype-level), linkage disequilibrium, and selection, and estimates multi-locus haplotypes. PyPop supplements and extends existing population genetic software incorporating them as modules, modified to accommodate highly polymorphic data, rather than reimplimenting them from scratch. It facilitates evolutionary analyses by integrating population genetic statistics within and across populations.

Originally developed to analyze the highly polymorphic genetic data of the human leukocyte antigen region of the human genome, PyPop has applicability to any kind of multilocus genetic data. It was the primary platform for evolutionary analysis of data collected for a major NIH-funded collaborative grant that included over 30 laboratories and 200 populations (Lancaster et al., 2007a,b). PyPop has also been successfully used in studies by our group, with collaborators, and in publications by many independent research teams in over 70 peer reviewed papers.

PyPop deploys a standard Extensible Markup Language (XML) output format and integrates the results of multiple analyses on various populations that were performed at different times into a common output format that can be read into a spreadsheet. The XML output format allows PyPop to be embedded as part of larger analysis pipelines. It also features an Application Programming Interface (API) allowing functionality to be incorporated into other programs. This talk will focus on recent features of PyPop which include the prefILTERing of the input genotype data and the ability to translate arbitrary allele names into full amino acid or nucleotide sequences.

All code is made available under the terms of the GNU General Public License (GNU GPL):

- Homepage: http://www.pypop.org/
- Latest source tarball: http://www.pypop.org/pypop-0.7.0.tar.gz

References
