PyPop: A Software Framework for Population Genomics: Analyzing Large-Scale Multi-Locus Genotype Data

A. Lancaster, M. Nelson, D. Meyer, R. Single, G. Thomson

Pacific Symposium on Biocomputing 8:514-525(2003)

PYPOP: A SOFTWARE FRAMEWORK FOR POPULATION GENOMICS: ANALYZING LARGE-SCALE MULTI-LOCUS GENOTYPE DATA

ALEX LANCASTER, MARK P. NELSON, DIOGO MEYER AND GLENYS THOMSON

Department of Integrative Biology, University of California, Berkeley,
3060 Valley Life Sciences, Berkeley, CA, 94720, USA
E-mail: alexl@socrates.berkeley.edu

RICHARD M. SINGLE

Department of Biometry, University of Vermont,
Hills Science Building, Burlington, VT, 05405, USA

Software to analyze multi-locus genotype data for entire populations is useful for estimating haplotype frequencies, deviation from Hardy-Weinberg equilibrium and patterns of linkage disequilibrium. These statistical results are important to both those interested in human genome variation and disease predisposition as well as evolutionary genetics. As part of the 13$^{\mathrm{th}}$ International Histocompatibility and Immunogenetics Working Group (IHWG), we have developed a software framework (PyPop). The primary novelty of this package is that it allows integration of statistics *across* large numbers of data-sets by heavily utilizing the XML file format and the R statistical package to view graphical output, while retaining the ability to inter-operate with existing software. Largely developed to address human population data, it can, however, be used for population based data for any organism. We tested our software on the data from the 13$^{\mathrm{th}}$ IHWG which involved data sets from at least 50 laboratories each of up to 1000 individuals with 9 MHC loci (both class I and class II) and found that it scales to large numbers of data sets well.

1 Introduction

Several major factors account for variation in the human genome: *mutation*, *random genetic drift*, *migration* (or *gene flow*) and *natural selection*. Understanding of the relative roles of these evolutionary forces is important for the study of both complex and Mendelian diseases, since they can affect our ability to identify and localize disease predisposing variants and our power to recognize underlying functional mechanisms through which predisposing genes can become relatively common in a population.

In particular, genetic systems which are highly polymorphic can implicate natural selection as an important factor in maintaining variation. Genetic systems such as the Human Leukocyte Antigen (HLA) region (the Major Histocompatibility Complex [MHC] for humans) are highly polymorphic. Six classical class I and II loci each contain up to 399 alleles [11].

Several basic population genetics statistics from multi-locus genotype data can help us understand these patterns of variation, and their implications for disease studies and evolutionary genetics. These statistics include, but are not limited to estimating haplotype frequencies, identifying deviation from Hardy-Weinberg equilibrium and locating patterns of linkage disequilibrium in a given population.

In population studies several implementations of programs and routines to calculate these basic statistics exist, but are mainly oriented towards analyzing statistics on a population-by-population basis. The ability to cross-correlate these statistical features across many population data-sets will enable the identification of features in the genetic data that can further our understanding of the functional and disease predisposing role of specific alleles, and conversely allow us to rule out others.

Currently, packages for analyzing population data already exist such as `Arlequin` [13], `PHYLIP` [1] and `Genepop` [3]. In general, however, they are not oriented towards large-scale cross-population data analyses. Analogous to the tools being developed for sequence analysis and search, we seek a framework in which basic statistical data from population genetic analyses can be housed, interrogated and visualized in such a way that important features of interest to both the biomedical investigator and the evolutionary biologist can be highlighted.

We also did not want to reinvent the wheel, so, where possible we can inter-operate with existing population genetic analysis packages (either as part of the framework or through file formats). In this way, our software, *PyPop* (*Py*thon for *Pop*ulation Genetics) can be viewed as an integrating framework which draws on the strengths of existing tools in the community.

1.1 IHWG: International Histocompatibility Working Group

The primary motivation for developing this project was our role as the 'Biostatistics Core', part of the International Histocompatibility Working Group (IHWG (`http://www.ihwg.org/`)). The IHWG collected population data on the HLA region and largely focused on the HLA classical class I (A, B, and C) (1.8 Mb) and class II (DR, DQ, and DP) (1.2 Mb) genes, which flank the class III region on chromosome 6 (ch. 6p21.31).

As part of the Anthropology and Human Diversity component of the IHWG, data from populations from upwards of 50 laboratories was made available via a database housed at the Fred Hutchison Cancer Research Institute (the 'Database Core'). The IHWG provided standardized typing reagents to each lab involved in the component, which resulted in high resolution genotype data for each population.

The molecular characterization of these alleles thus allowed us to use both the allele frequency (the raw allele 'calls') and the underlying sequence information (the 'calls' can be converted into sequences if desired) as input to the analysis framework. `PyPop` does not distinguish between allele calls or sequence data and can transparently handle both.

1.2 Population genetic tests and statistics

The particular cross-population statistics we wished to address included: (1) conformity to Hardy-Weinberg expectations, (2) tests for balancing selection; (3) haplotype distributions and patterns of linkage disequilibrium among populations; and (4) other tests such as worldwide patterns of genetic differentiation.

Hardy-Weinberg Hardy-Weinberg equilibrium essentially states that unless there are counteracting forces, the frequencies of alleles will not change in a population and the expected genotype frequencies each generation are determined by the allele frequencies, and are termed Hardy-Weinberg proportions (HWP). In the context of multi-population analyses we can use deviation from HWP to determine whether this results from: (1) typing error (the first possibility investigated); (2) an 'admixed', or merged population; (3) operation of natural selection; or (4) inbreeding.

Ewens-Watterson test of neutrality This is a test with the null hypothesis of neutral evolution and determines the probability that the observed homozygosity under HWP, for a given sample size and observed number of alleles, is more extreme than the expected homozygosity under random neutral mutations and genetic drift (neutrality). This test can tell us whether selection, either directional (observed homozygosity > expected homozygosity) or balancing (observed homozygosity < expected homozygosity) is in operation on a particular locus across populations.

Haplotype estimation and linkage disequilibrium Linkage disequilibrium (LD) describes the non-random association of alleles at different genetic loci. Through estimating haplotype frequencies, it is possible to estimate LD in a population, the presence of significant LD can be due to history for very closely linked genes and can also indicate the operation of selection.

Other statistics Other individual-level population genetic statistics can also be calculated, such as F_{st} , which describes the apportionment of genetic diversity within subpopulations, relative to a larger population, allowing us the estimate the amount of admixture in a population.

1.3 Requirements

The data sets we wished to analyze were highly heterogeneous. Datasets varied considerably in the number of loci typed (from two to nine), number of individuals sampled (from less than 50 to 1000), and number of alleles at each locus (from 5 to 179). Given that we had such highly diverse data sets, we nevertheless wished to generate analyses that could be integrated in a systematic way, this led to a set of requirements for the software framework:

- **modular** each analysis (e.g. Hardy Weinberg) can be run stand-alone or as part of a battery of tests
- **configurable** analyses can be switched on or off as required by the user in a simple configuration file
- **a filter** all output from the analyses should be available as input to other programs
- **standardized output** output should be generated in the open standard format XML
- **integrating** platform should allow simple integration of modules written in other languages (e.g., C) and/or third party software

1.4 Integrating with existing software

Where possible, we would like to integrate with existing software. `Arlequin` [13] and `PHYLIP` [1] are packages that deal with some aspect of population genetic statistics. In addition, we had some existing in-house software, such as `emhaplofreq`, a program for haplotype estimation for the highly polymorphic HLA loci, and `gthwe`, a program for implementing Guo and Thompson's Hardy-Weinberg exact test [4].

In using these programs, however, we realized a need for a tool that could integrate features of existing software, and where needed, implement missing features that would realize our goals of doing large-scale population genetic data analysis. Specific limitations that we wished to address were:

1. **Modularity** Programs should allow different options to be set for analysis and produce content that is easily parsable, not a monolithic output set of statistics (such as unstructured plain ASCII text or HTML).
2. **Batch-ability** It should be simple to set up an entire job in an unattended 'batch mode', involving the creation of (or modification of) a configuration file with a text editor, followed by the invocation of a script.

Software that relies on a ‘captive user interface’, for a population to be analyzed and requires user interaction such as mouse clicks and menus, makes it difficult to analyze hundreds of data sets.

3. **Scalability** It should be straightforward, for example, to gather a single statistic for several populations and display it in a table without a laborious manual search across many files. This relates partly to the previous point: often existing software was oriented towards smaller and less heterogeneous datasets. More typical in evolutionary genetics studies is integrating results across several populations at one or two loci, rather than many populations with a large number of loci and high variability in the number of individuals.

4. **Open-source** We plan to release our software under an open source license [9]. Software that is open-source allows others to extend and re-use components, allows interoperation via an open and published interfaces, and can reduce duplication of effort within the community. Some existing software that was not open-source required reverse engineering of their file formats and run-time behaviour in order to be able to communicate and write interfaces to them.

5. **Cross-platform** Software should be cross platform, and not be tied to proprietary features. In particular it should be be available to run high-performance UNIX platforms such as GNU/Linux or Solaris as well as Windows and Macintosh platforms.

2 Method

2.1 Overall design

To integrate the data analyses for multiple populations, the analysis pipeline for `PyPop` has been designed in two major parts, or phases, the first is the basic population genetic analyses, and the second integrates the results of these analyses across multiple populations. The overall data and work flow is shown in Figure 1.

2.2 Implementation

We decided to implement the project in the object-oriented scripting language Python [10]. It is an interpreted language allowing for rapid prototyping of modules and has a convenient standard library of functions. Through use of the

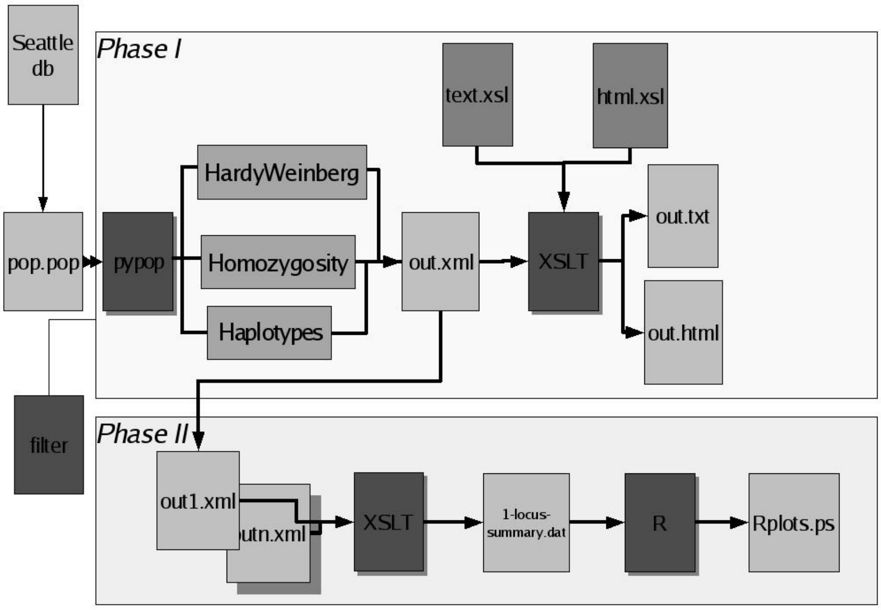


Figure 1. Work flow

the Simple Wrapper Interface Generator (SWIG [12]), it is also straightforward to 'wrap' existing code in C, C++ or Java and call it from Python.

In line with our philosophy of integrating with 'best of breed' open-source components, we leveraged the work of the Numeric Python [8] project, which provides efficient data structures for holding large arrays of data, and `R`, the open-source implementation [6] of the programming language `S`. As our XSLT parser, we chose the fast C-based `libxslt` [7] from the GNOME project [2].

2.3 Data flow

In phase I, before analysis begins, the multi-locus genotype data for each population (stored in a text file with the `.pop` extension), is passed through a filter module for data cleaning. Next, the basic population genetic analyses are run for each population file from the database. The filter module is a set of rules that traps any allele name which does not have a close match in the database of HLA alleles [5] maintained by Steve Marsh of the Anthony Nolan Trust in the UK. Although at present this is HLA-specific, it has been designed in such a way that it is simple to write a filter for other types of population data, and if desired (if the data is already 'clean'), it can be simply switched off. Next the analysis begins. Depending on the users configuration, this can include basic allele count data; the Hardy-Weinberg statistics; haplotype estimates; and the Ewens-Watterson test of neutrality (represented by the green modules in Figure 1). The results of the analyses are stored in an XML file format as shown in Figure 2.

```
<dataanalysis date="2002-07-09-05-01-18">
   <filename>ukimid_nireland.pop</filename>
   <pypop-version>DEVEL_VERSION</pypop-version>
   <populationdata>
     <labcode>UKIMID</labcode>
     <method>SSOP</method>
     <ethnic>Irish</ethnic>
     <contin>Europe</contin>
     <collect>Northern Ireland</collect>
     <latit>54 degrees 40 minutes north</latit>
     <longit>6 degrees 45 minutes west</longit>
     <complex>3</complex>
     <popname>NIreland</popname>
     <totals>
       <indivcount>1000</indivcount>
       <allelecount>2000</allelecount>
       <locuscount>9</locuscount>
       <lociWithDataCount>4</lociWithDataCount>
     </totals>
   </populationdata>
   <locus name="A">
     <allelecounts>
       <untypedindividuals>0</untypedindividuals>
       <indivcount>1000</indivcount>
       <allelecount>2000</allelecount>
       <distinctalleles>26</distinctalleles>
       <allele name="0101">
         <frequency>0.20200 </frequency><count>404</count>
       </allele>
...
```

Figure 2. Extract from sample XML output file

XML was chosen as the output format because: 1) it can be read as input by other programs and; 2) it is readily transformable into human-readable form (e.g., a text file) or web-form (e.g., HTML) via XSLT (eXtensible Stylesheet Language for Transformations) [14].

In phase II, the results of the data analyses of the individual data files are integrated, and the benefits of using XML as the storage and exchange format are realized. There are two major benefits: one relevant for displaying results of each individual run, the second, far more powerful benefit, for aggregating the data for cross-population meta-analyses and transforming it as input for third party packages.

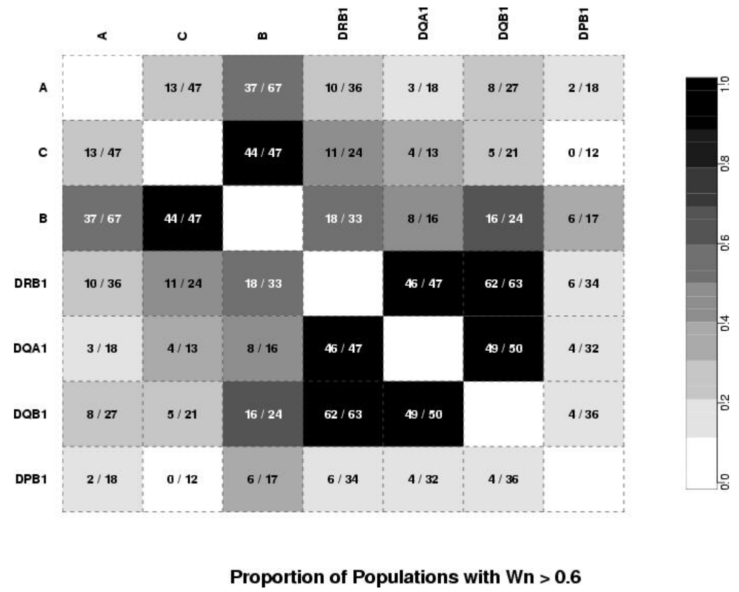
First, since many of the analyses (notably estimating the significance of all pairwise linkage disequilibrium) can take a considerable amount of time, especially if the population consists of many individuals and is highly polymorphic, the *analysis generation* stage is decoupled from the *analysis presentation* stage. This enables tweaking of the ‘human-readable’ text output of the individual files of the presentation without completely re-running the analyses. A sample text output is shown in Figure 3. The same XML content can also be used to generate an HTML version or web version of the same data set.

```
Performed on the ’ukimid_nireland.pop’ file at: 2002-07-09-05-01-18

Population Summary
==================

       Lab code: UKIMID
  Typing method: SSOP
      Ethnicity: Irish
      Continent: Europe
Collection site: Northern Ireland
       Latitude: 54 degrees 40 minutes north
      Longitude: 6 degrees 45 minutes west
Population Name: NIreland
[...]

1.1. Allele Counts [A]
----------------------

Untyped individuals: 0
Sample Size (n): 1000
Allele Count (2n): 2000
Distinct alleles (k): 26

Counts ordered by frequency   | Counts ordered by name
Name      Frequency (Count)   | Name      Frequency (Count)
0201      0.27400   548       | 0101      0.20200   404
0101      0.20200   404       | 0201      0.27400   548
```

Figure 3. Extract from sample plain text output generated from XML data

Second, the results of Phase I are used to investigate patterns of variation across populations and within populations. Examples include comparisons of evidence of selection, using the homozygosity test of neutrality for a given locus for all populations (e.g., DPA1 for all populations), or comparing the sets of loci in significant linkage disequilibrium across populations. This is implemented using XSLT to transform all the desired sets of output XML files into tables of data that are read by the statistical package `R`.

A sample output of this process is shown in Figure 4. This is an output plot from `R` and depicts for each of the loci analyzed as part of the IHWG workshop, the proportion of populations in which the linkage disequilibrium, as measured by the W_n statistic, exceed 0.6.



	A	C	B	DRB1	DQA1	DQB1	DPB1
A		13 / 47	37 / 67	10 / 36	3 / 18	8 / 27	2 / 18
C	13 / 47		44 / 47	11 / 24	4 / 13	5 / 21	0 / 12
B	37 / 67	44 / 47		18 / 33	8 / 16	16 / 24	6 / 17
DRB1	10 / 36	11 / 24	18 / 33		46 / 47	62 / 63	6 / 34
DQA1	3 / 18	4 / 13	8 / 16	46 / 47		49 / 50	4 / 32
DQB1	8 / 27	5 / 21	16 / 24	62 / 63	49 / 50		4 / 36
DPB1	2 / 18	0 / 12	6 / 17	6 / 34	4 / 32	4 / 36	

Figure 4. Sample graphical output from meta-analysis: proportions of populations with linkage disequilibrium measure Wn greater than 0.6

The XML output (both the individual population-level data files, and the aggregated multi-population data) can be transformed (via XSLT) into formats suitable for input to other programs. Currently we have a prototype module for generating input for `PHYLIP` and a prototype module to generate `Arlequin .arp` files.

3 Results and discussion

3.1 Analyzing IHWG workshop data

The `PyPop` framework was used to analyze the full set of IHWG data from both the 12th and 13th workshops. With 119 separate data-sets (some of which individually required up to 9 days to complete the basic per-population statistics), it was straightforward to set up a batch program to generate the individual output analyses. For each file, an individual XML file was generated. From these individual files, using XSLT stylesheets as described above in phase II of PyPop, it was straightforward to generate the input data files for the statistical package, R. The R code was set up to generate overall graphs for many population genetic parameters of interest. In particular, the analysis allows 'slicing' of the data along a number of axes. As an example, we can view Hardy Weinberg deviation, for all geographic regions at a given locus, or view all loci for a given region. As another example, we can view the number of populations that have data for a given region (Figure 5), or number of populations for which data was provided for a given locus (Figure 6).

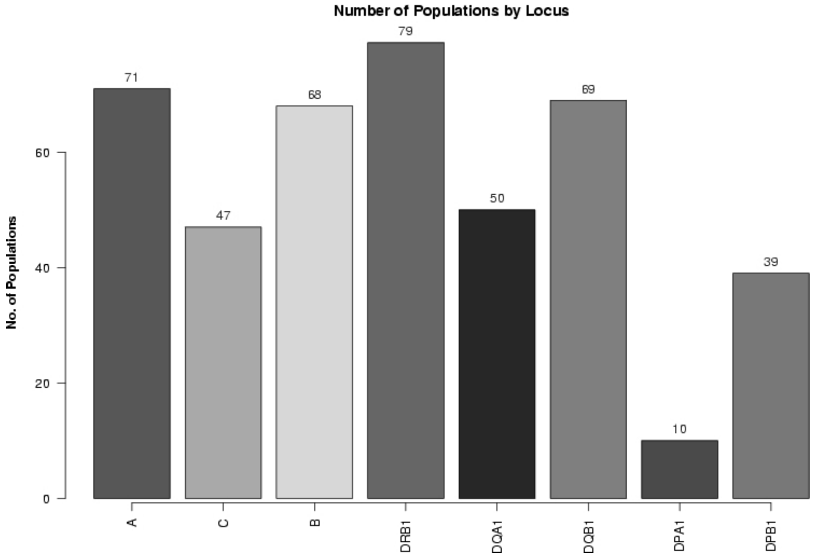


Figure 5. Sample metadata output: viewing number of populations by locus

This 'meta-analysis' code is modular and can cope with non-HLA data, and thus can potentially be useful for other large-scale population genetic analyses. The significance is that the complete pipeline, from analyzing the individual data files, to the generation of the overall 'meta' output can be completely automated from the command-line. Further, the flexibility of the XML format, allows us to easily extract and output data for future data analyses without requiring regeneration of the basic population genetics statistics.

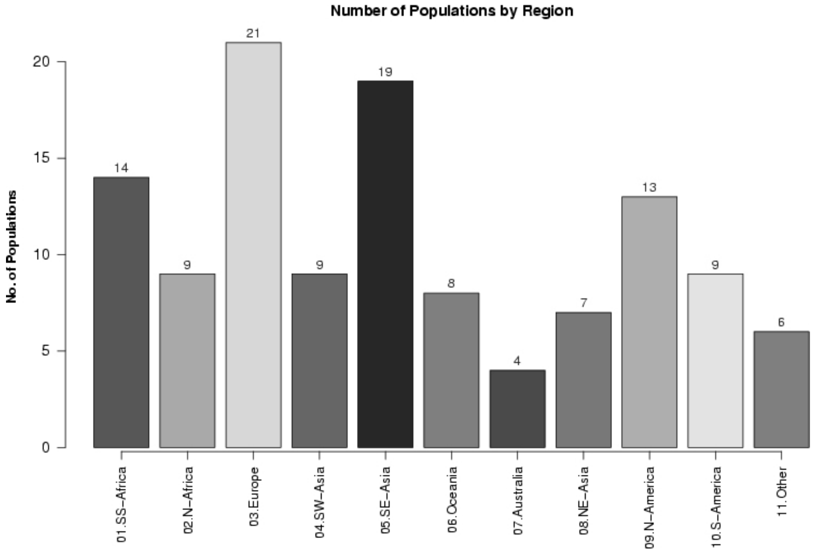


Figure 6. Sample metadata output, a second way to view data: number of populations within each region

3.2 Role and significance

It's important to note that PyPop is not attempting to supplant or replace other, more established, population genetics analysis packages. The basic population genetic statistics computed by PyPop are not new in and of themselves, nor is the ability to perform analyses on multiple populations (Arlequin can currently do this). Indeed, PyPop can use existing packages to calculate them. However the approach of integrating information in highly heterogeneous datasets on a large scale is new, and not currently available in the evolutionary genetics community. In addition to tests unavailable in existing projects, the uniqueness of the present project is that it is intended to be a high-throughput system that enables population genetics to join the realm of genomics.

3.3 Future directions

In the future we plan to continue development on the modules that can inter-operate with PHYLIP and Arlequin. We also plan to take the Ewens-Watterson test of neutrality down to the amino-acid level (by considering each amino acid site as a genetic locus), necessitating the translation of the allele 'calls' into sequence data (where possible). This will result in a useful module for those wishing to analyze allele data at the sequence level. Longer-term possibilities include developing a graphical front-end, possibly web-based, integration with data mining tools such as clustering analysis, and currently under discussion is integration with the NCBI's new `dbMHC` database [15]. We also plan to release `PyPop` under an open source license. Details will be made available at the Thomson lab website: `http://allele5.biol.berkeley.edu/`

Acknowledgements

This work has benefited from the support of NIH grant AI49213 (13th IHW). Thanks to Steve Mack.

References

1. J. Felsenstein, 'PHYLIP Phylogeny Inference Package (Version 3.2) `http://evolution.genetics.washington.edu/phylip.html`' *Cladistics* **5**:164-166 (1989).
2. GNOME Project `http://www.gnome.org/`.
3. Genepop `ftp://ftp.cefe.cnrs-mop.fr/genepop`.
4. SW Guo and EA Thompson, 'Performing the exact test of Hardy-Weinberg proportion for multiple alleles' *Biometrics* **48**:361-72 (1992).
5. IMGT/HLA Database `http://www.ebi.ac.uk/imgt/hla/`.
6. Ross Ihaka and Robert Gentleman, 'R: A Language for Data Analysis and Graphics `http://www.r-project.org/`' *Journal of Computational and Graphical Statistics* **5**(3):299-314 (1996).
7. libxslt, XSLT C library for GNOME `http://xmlsoft.org/`.
8. Numeric Python `http://numpy.sf.net/`.
9. Open Source Initiative `http://www.opensource.org/`.
10. Python `http://www.python.org/`.
11. D A Rhodes and J Trowsdale, 'Genetics and molecular genetics of the MHC' *Rev. Immunogenetics* **1**(1):21-31 (1999).
12. Simple Wrapper Interface Generator `http://www.swig.org/`.
13. S. Schneider, D. Roessli, and L. Excoffier, 'Arlequin: A software for population genetics data analysis. `http://lgb.unige.ch/arlequin/`', Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva (2000).
14. eXtensible Stylesheet Language `http://www.w3.org/Style/XSL/`.
15. dbMHC `http://www.ncbi.nlm.nih.gov/IEB/Research/GVWG/MHC/`.