Antao *BMC Bioinformatics* 2010, **11**(Suppl 12):S10
http://www.biomedcentral.com/1471-2105/11/S12/S10

BMC Bioinformatics

PROCEEDINGS Open Access

interPopula: a Python API to access the HapMap Project dataset

Tiago Antao

From The 11th Annual Bioinformatics Open Source Conference (BOSC) 2010
Boston, MA, USA. 9-10 July 2010

Abstract

Background: The HapMap project is a publicly available catalogue of common genetic variants that occur in humans, currently including several million SNPs across 1115 individuals spanning 11 different populations. This important database does not provide any programmatic access to the dataset, furthermore no standard relational database interface is provided.

Results: interPopula is a Python API to access the HapMap dataset. interPopula provides integration facilities with both the Python ecology of software (e.g. Biopython and matplotlib) and other relevant human population datasets (e.g. Ensembl gene annotation and UCSC Known Genes). A set of guidelines and code examples to address possible inconsistencies across heterogeneous data sources is also provided.

Conclusions: interPopula is a straightforward and flexible Python API that facilitates the construction of scripts and applications that require access to the HapMap dataset.

Background

The HapMap project [1] (http://hapmap.ncbi.nlm.nih.gov/) is an effort to identify and catalogue genetic similarities and differences in humans. The project makes information available on single nucleotide polymorphisms (SNPs), and it more recently added information on copy number variation (CNV). HapMap phase 3 includes data on 1115 individuals (around 1.5 million SNPs per individual) spanning 11 populations while phase 2 included only 4 populations (270 individuals) but more than 3.5 million SNPs per individual. This dataset can be useful in a multitude of situations from finding genes that affect human health to evolutionary research about the human species or for genome-wide association studies. All of the information generated is released into the public domain and can be downloaded with minimal constraints. The HapMap project provides access to the data in bulk form (via FTP download), a web interface [2] which includes a genome browser [3] and the data mining application HapMart based on Biomart [4]. Programmatic and relational database interfaces are not offered though some API support is implemented by external parties such as a generic Perl API for variation datasets in Ensembl [5], BioPerl's Bio::PopGen module [6] or the GGtools package [7] for R/Bioconductor. Most existing libraries support only a subset of features (e.g. parsing of HapMap file formats or creating a local database) making the construction of scripts and applications more complex as basic data manipulation functionality must be built as least partially. Furthermore, there is no known Python library supporting HapMap data.

Implementation

interPopula provides a Python API to access the HapMap dataset. Interfaces to all HapMap phases are supported including phase 2 data with fewer populations but more SNPs genotyped per individual and phase 3 covering more populations. interPopula provides access to frequency, genotype, linkage disequilibrium and phasing datasets. The recent CNV dataset is also supported along with family relationships for the 5 populations

Correspondence: tra@popgen.eu
Liverpool School of Tropical Medicine, L3 5QA, Liverpool, UK

where sampling was performed for family trios (mother, father and one offspring).

Support for annotation information that is commonly needed to process HapMap data is also provided through an API to both the UCSC Known Genes dataset [8] from the UCSC genome browser database [9] and the Ensembl gene annotation database [10].

The API was constructed according to the following design guidelines:

1. The API is straightforward and self-contained. The core API requires only a Python interpreter, has no extra dependencies and minimal administrative overhead.

2. Downloaded data is stored on an SQL database for faster access. All data is stored using sqlite [11] which is natively supported in Python thus lowering the maintenance costs of the system. interPopula can also be connected to enterprise-grade databases which support multiple users, concurrent usage and large datasets for which the standard sqlite backend might not be enough (a PostgreSQL example is provided).

3. Data management (i.e. downloading from the HapMap site and local database construction) is fully automated: the required data subset is downloaded on demand only once and stored locally, reducing the load on both the client and server.

4. While SQL interfaces are made available from both the UCSC and Ensembl projects for their annotation databases, interPopula uses the same implementation strategy for the HapMap dataset: files are intelligently downloaded and locally stored. This provides a consistent interface to these two datasets which provide important annotation information frequently used to process HapMap data.

5. The framework is extensible and designed to be easily integrated with other Python tools and external databases. The web site provides several examples of integration with standard tools used in Python for bioinformatics such as Biopython [12], NumPy [13] and matplotlib [14].

6. Integration with Biopython allows for access to the Entrez SNP database and the population genetics tools supported by Biopython such as Genepop [15] allowing automated analysis of datasets.

7. Facilities to export HapMap data to Genepop format are provided enabling (non-automated) analysis of the HapMap dataset with the plethora of population genetics software which support this format. Data export can also be use to initialize population genetics simulators like the Python-based simuPOP [16] allowing computational simulations to be initialised with real datasets.

8. A large set of scripts is included, serving both as utilities to analyse the data, as well 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.

9. A set of guidelines and scripts was developed in order to facilitate a consistent view across heterogeneous databases. HapMap, Ensembl, 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. The main pitfall is the usage of different NCBI reference builds across different databases, most notably HapMap is still based on build 36 whereas other databases either support multiple builds or only the most recent build 37.

10. A robust open-source software development process is put in place: a full public web based platform (hosted on Launchpad) is used to maintain the code infrastructure and unit tests approach 100% coverage.

Results

interPopula can be used to create a wide range of applications and scripts based on the HapMap dataset. The most commonly expected usage pattern will be for genome wide association studies, though the example presented here will be of a different nature.

As an example of usage, we present a population comparison of all the genotyped SNPs for a gene. We will plot the F_{st} statistic for all Lactase SNPs between two HapMap populations: Utah residents with Northern and Western European ancestry (CEU) and Yoruban in Ibadan, Nigeria (YRI). These populations are known to differ in their tolerance to lactose [17]. This example uses genotype information from HapMap and also demonstrates the integration facilities with UCSC Known Genes (to retrieve gene position and exon data), matplotlib (used for plotting), Biopython and Genepop (used to calculate F_{st}).

This example, which is quite complex in terms of integration between several databases and tools can be broken down into the following steps:

1. Load the Known Genes database. The version pertaining build 36 should be loaded to assure consistency with HapMap.

2. Determine relevant information about Lactase from Known Genes. The following information is needed: The chromosome on which it is located, the start and end positions in the chromosome and all exon positions.

3. Load HapMap genotype information for the CEU and YRI populations for the relevant chromosome.

4. Retrieve all the HapMap SNP ids between the start and end positions in the chromosome.

5. Export a Genepop formatted file with two populations including all HapMap SNPs for Lactase.

6. Call the Genepop application via Biopython to calculate the F_{st} for all markers.

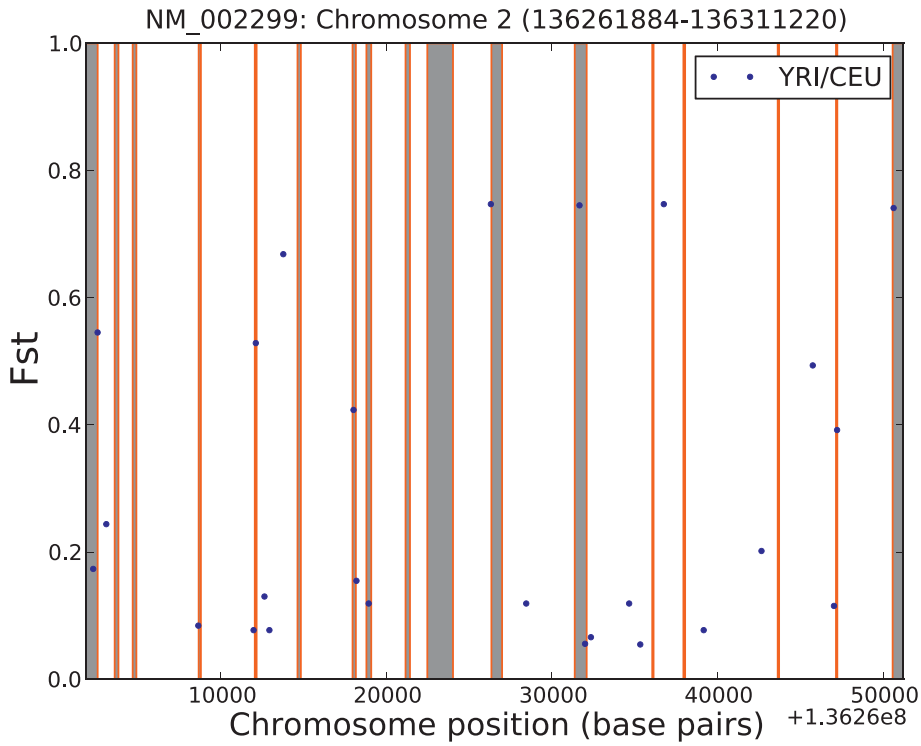

Figure 1 F_{st} for Lactase between 2 HapMap populations. F_{st} between CEU and YRI populations for all Lactase SNPs on the HapMap database. The X-axis reports the position on chromosome 2 (the value on the lower right is the absolute offset from the beginning of the chromosome), the Y-axis the F_{st} value. The dots represent the F_{st} values for existing SNPs on the HapMap database. The red boxes represent exon positions. To construct this chart HapMap frequency data and the UCSC Known Genes database were consulted. Biopython and Genepop were used to compute the F_{st} statistic.

7. Plot the calculated $F_{st}s$ along with the exon positions.

The result of this example is shown in Figure 1. The X-axis reports the position along chromosome 2, F_{st}in on the Y-axis, the dots represent the F_{st} values for existing SNPs on the HapMap database and the red boxes are the exon positions (17 in the case of Lactase). Interpreting the results of this specific application of interPopula is beyond the scope of this manuscript but at least two different interpretations are possible: (i) SNPs where F_{st} is above approximately 0.45 are candidates for positive selection (as around 95% of F_{st}values for humans are below 0.45 [18]) or (ii) the F_{st} statistic is noisy when applied to a single marker [19] . The above example was constructed using the UCSC Known Genes database but the programmer can alternatively use the Ensembl gene annotation database instead.

This example (script IGFstGene.py in the distribution), along with more than 20 others including data export, connection to enterprise-grade databases, analysis of the distribution of the number of exons per gene, the distribution of genes per chromosome are made available with interPopula.

In order to illustrate interPopula's basic API, Figure 2 shows a commented script which provides useful functionality. In this example the HapMap frequency database is consulted to report the frequency of both alleles for each SNP within a certain chromosome interval. The code example is less than one page in length and there are only 4 API calls to achieve the complete functionality. This is one case illustrating the ease of use of the API. All scripts provided with interPopula are documented to the level of the example presented and automated documentation covering the full API is extracted from the source using epydoc (http://epydoc.sourceforge.net/).

The part of the API devoted to both UCSC Known Genes and the Ensembl gene annotation database can be used stand-alone to access both databases, i.e., it can be used for application and scripts that have no relationship with the HapMap data. interPopula's UCSC and Ensembl APIs can be used to access also non-human data as genome annotations are available for other species. This is especially useful with the Ensembl dataset as it makes available gene annotation information for many other species. Users should note that the quality of the datasets for other species varies as more effort is

```
from interPopula import Config
from interPopula.HapMap.Frequency import Frequency

#configuration directory
Config.dataDir = "."

#Han Chinese Lactase information
pop = "CHB"
chr = 2
startChr = 136261855
endChr = 136311220

#Lets get the Frequency information
freqDB = Frequency("2010-05_phaseIII")

# We require a chromosome and population
freqDB.requireChrPop(chr, pop)

# We need to get the RSIds for the interval
RSs = freqDB.getRSsForInterval(chr, startChr, endChr)

print "rsid allele1 freqAllele1 allele2 freqAllele2"
for rs in RSs:
    #We get frequency information
    freqSNP = freqDB.getPopSNPs(pop, rs)
    #a1 retrives allele 1 (A,C,T,G), a2 does the same for 2
    a1   = freqSNP[5]
    a2   = freqSNP[6]
    #frequency of a1 homozygotes
    a1a1 = freqSNP[7]
    #frequency of a2 homozygotes
    a2a2 = freqSNP[8]
    #frequency of heterozygotes
    a1a2 = freqSNP[9]
    #gets the frequency of allele 1
    fa1 = (2.0*a1a1+a1a2)/(2*a1a1 + 2*a2a2 + 2*a1a2)
    print rs, a1, fa1, a2, 1 - fa1

#example output
#rsid allele1 freqAllele1 allele2 freqAllele2)
#730005 C 0.727941176471 T 0.272058823529
#872151 C 0.658088235294 T 0.341911764706
#1042712 C 0.36496350365 G 0.63503649635
#[...]
```

Figure 2 Example code to print the frequency of HapMap SNPs. This example describes how to consult the HapMap frequency database to retrieve the allele frequencies for a set of SNPs in a section of a chromosome.

put in the curation of human data (e.g. while for humans the chromosome information is normally the chromosome number, for cats - *Felis catus* - it is mostly scaffold data). Stand-alone example script examples are provided for both datasets.

Future development efforts for interPopula will focus on supporting large datasets. As sequencing costs continue to decrease and the sequencing of complete genomes becomes commonplace it is clear that the backend infrastructure will have to be redesigned to support the large amounts of data generated by such efforts. In this context, supporting the 1000 genomes project [20] is a natural extension for interPopula as many of the samples used in this project come from the HapMap dataset. While the API for UCSC and Ensembl extensions provides access to other species data, the main focus of interPopula will remain providing robust and well-maintained APIs for publicly available human genomic datasets which lack a standardized Python API or relational interface.

Conclusions

interPopula is a flexible, straightforward Python API to the HapMap project. It strives to integrate with both common Python bioinformatics and scientific libraries and other genomic databases that are commonly used in conjunction with the HapMap dataset. interPopula makes HapMap data processing possible inside Python, thus opening the possibility for the development of a plethora of interesting applications and scripts that make use of this important resource for human population genomics studies.

Availability and requirements

Project name interPopula

Project home pagehttp://popgen.eu/soft/interPop/. Development site: https://launchpad.net/interpopula

Operating systems Platform independent

Programming language Python

Other requirements Optionally NumPy, Biopython, Genepop and matplotlib

License GNU GPL version 3

Any restrictions to use by non-academics None

List of abbreviations used

API - Application Programming Interface; CEU - HapMap population sample comprising Utah residents with Northern and Western European ancestry; CNV - Copy Number Variation; SNP - Single Nucleotide Polymorphism; SQL - Structured Query Language; UCSC - University of California, Santa Cruz; YRI -HapMap population sample comprising Yoruban in Ibadan, Nigeria

Acknowledgments

TA was supported by research grants SFRH/BD/30834/2006 and PTDC/BIA-BDE/65625/2006 from Fundação para a Ciência e Tecnologia (FCT), Portugal.

I would like to thank the two anonymous reviewers for their comments and for pointing out important bugs related to data retrieval and support for multiple dataset versions in the HapMap core component of interPopula. This article has been published as part of *BMC Bioinformatics* Volume 11 Supplement 12, 2010: Proceedings of the 11th Annual Bioinformatics Open Source Conference (BOSC) 2010. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/11?issue=S12.

Competing interests

The author declares that he has no competing interests.

Published: 21 December 2010

References

1. Consortium IH, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JTF, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449(7164)**:851-861.
2. Thorisson GA, Smith AV, Krishnan L, Stein LD: **The International HapMap Project Web site.** *Genome Res* 2005, **15(11)**:1592-1593.
3. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12(10)**:1599-1610.
4. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart—biological queries made easy.** *BMC Genomics* 2009, **10**:22.
5. Rios D, McLaren WM, Chen Y, Birney E, Stabenau A, Flicek P, Cunningham F: **A database and API for variation, dense genotyping and resequencing data.** *BMC Bioinformatics* 2010, **11**:238.
6. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12(10)**:1611-1618.

7. Carey VJ, Morgan M, Falcon S, Lazarus R, Gentleman R: **GGtools: analysis of genetics of gene expression in bioconductor.** *Bioinformatics* 2007, **23(4)**:522-523.
8. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The UCSC Known Genes.** *Bioinformatics* 2006, **22(9)**:1036-1046.
9. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2010.** *Nucleic Acids Res* 2010, **38(Database issue)**:D613-D619.
10. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M: **The Ensembl automatic gene annotation system.** *Genome Res* 2004, **14(5)**:942-950.
11. **The SQLite database engine.** [http://www.sqlite.org/].
12. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL: **Biopython: freely available Python tools for computational molecular biology and bioinformatics.** *Bioinformatics* 2009.
13. Oliphant TE: *Guide to NumPy* Provo, UT; 2006 [http://www.tramy.us/].
14. Hunter JD: **Matplotlib: A 2D Graphics Environment.** *Computing in Science and Engg* 2007, **9(3)**:90-95.
15. Rousset F: **genepop'007: a complete re-implementation of the genepop software for Windows and Linux.** *Molecular Ecology Resources* 2008, **8**:103-106.
16. Peng B, Kimmel M: **simuPOP: a forward-time population genetics simulation environment.** *Bioinformatics* 2005, **21(18)**:3686-3687.
17. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P: **Convergent adaptation of human lactase persistence in Africa and Europe.** *Nat Genet* 2007, **39**:31-40.
18. Akey J, Zhang G, Zhang K, Jin L, Shriver M: **Interrogating a High-Density SNP Map for Signatures of Natural Selection.** *Genome Research* 2002, **12**:1805-1814.
19. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM: **Genomic signatures of positive selection in humans and the limits of outlier approaches.** *Genome Res* 2006, **16(8)**:980-989.
20. **The 1000 genomes project.** [http://1000genomes.org].

doi:10.1186/1471-2105-11-S12-S10
Cite this article as: Antao: interPopula: a Python API to access the HapMap Project dataset. *BMC Bioinformatics* 2010 **11**(Suppl 12):S10.