November 2011

User Guide

DIVERGENOME: a bioinformatics platform to assist the analysis of genetic variation

Manual developed by Wagner C. S. Magalhaes, Maira R. Rodrigues and Eduardo Tarazona-Santos

- ✓ DIVERGENOME is available at: **http://www.pggenetica.icb.ufmg.br/divergenome/**
- ✓ Supported Operating Systems: Windows, Linux32bits and Linux64bits and MAC-OS.
- ✓ Privacy policy:

This platform system and its documentation are freely available for academics purposes.

Suggested citation:

Magalhães, WCS; Rodrigues, M; Silva, D et al*:* **DIVERGENOME: a bioinformatics platform to assist the analysis of genetic variation.** Submitted.

Address for correspondence:

E-mail: **divergenome@gmail.com**

Table of Contents

What is the DIVERGENOME?

DIVERGENOME is a web accessible open-source platform (http://www.pggenetica.icb.ufmg.br/divergenome) to assist the analysis of genetic and epidemiologic datasets. It was developed to help investigators in data storage and analysis for population genetics and genetic epidemiology studies. The platform contains two components. The first component, DIVERGENOMEdb, is a relational database developed using MySQL. The second component, DIVERGENOMEtools, is a dynamic pipeline composed of a set of scripts, developed using the programming language Perl, and a graph-based coordination algorithm, that enables the conversion of both queries submitted to the database and independent files to many popular file formats required by well known software in population genetics and genetic epidemiology.

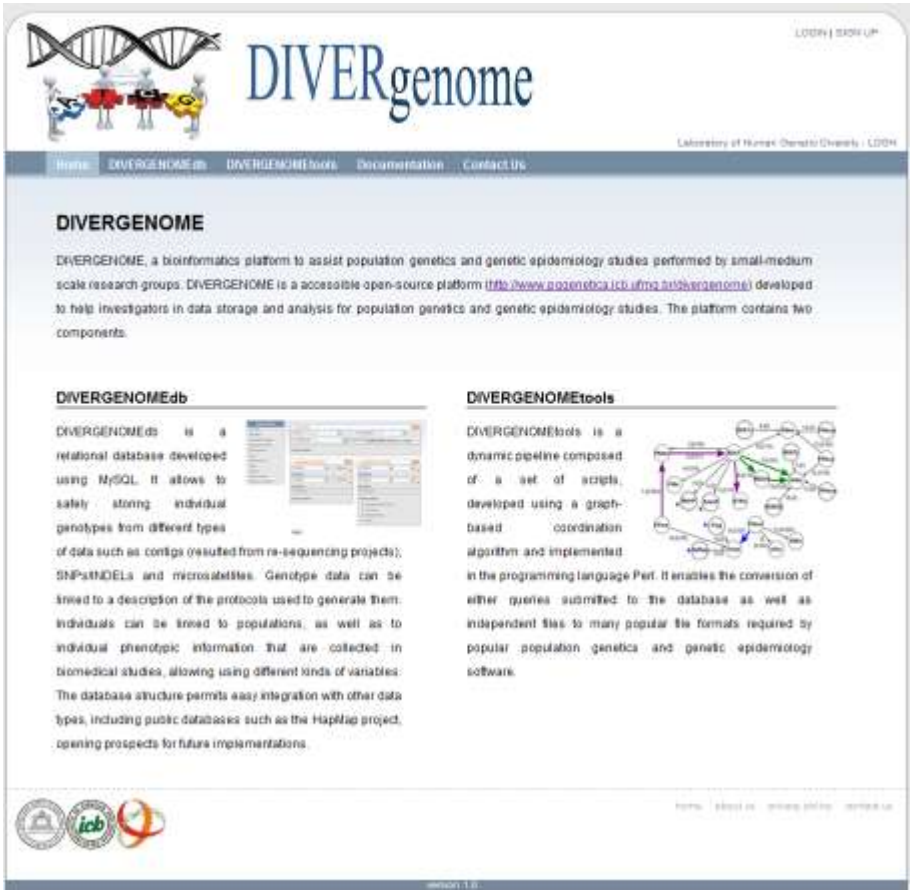
DIVERGENOMEdb is helpful to safely store individual genotypes from three different types of data: contigs (resulted from re-sequencing projects), SNPs/INDELs, and microsatellites. Genotype data can be linked to a description of protocols used to generate them. Individuals can be linked to populations, as well as to individual phenotypic information that are collected in genetic epidemiology studies using different kinds of variables.

DIVERGENOMEtools is useful for re-sequencing studies as well other types of studies including SNPs and other kinds of polymorphisms of haploid and diploid data in humans. Its main functionalities are the following:

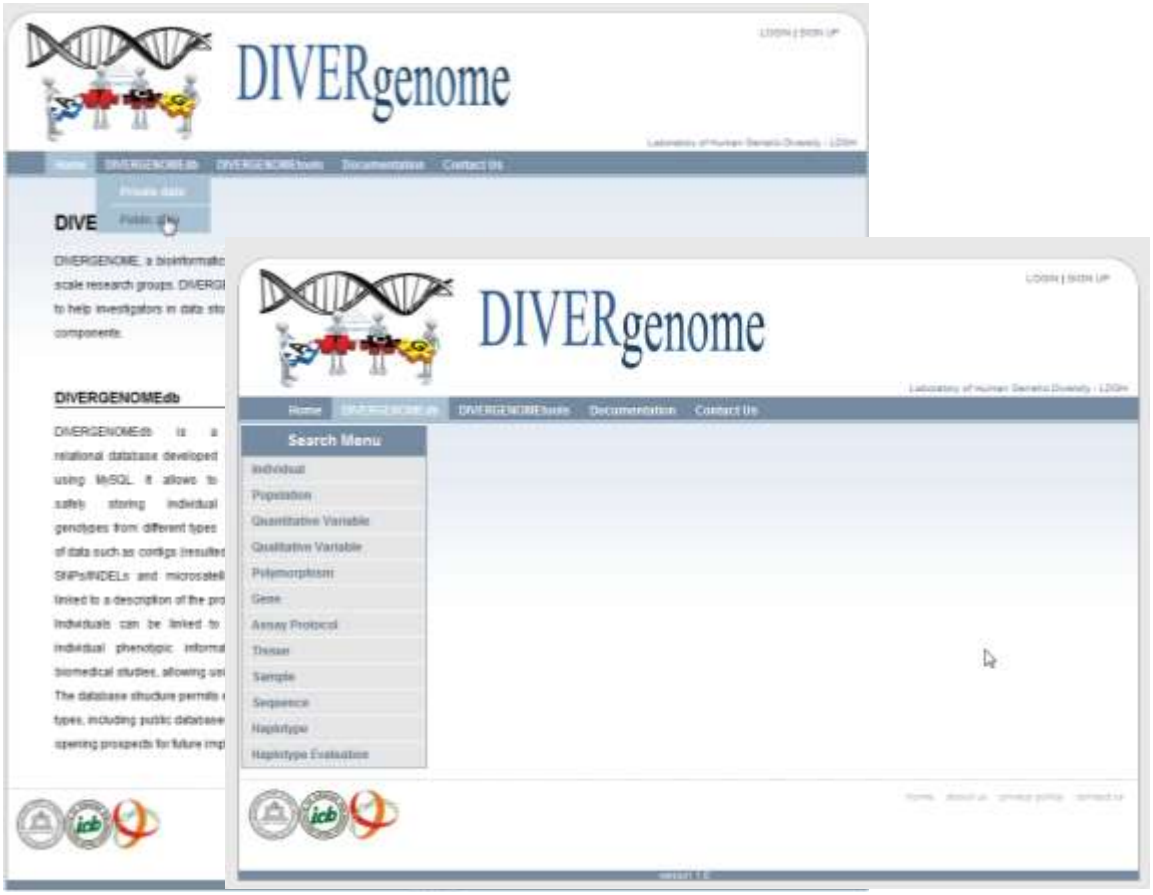
- Re-format genotypes called by PolyPhred into a matrix of genotypes with individuals as rows and segregating-sites as columns (SDAT format);
- Prepare input files for haplotype inferences using the popular software PHASE and fastPHASE;
- Prepare input files for the software Haploview;
- Prepare input files for the software Structure;
- Re-format SDAT format to Nexus format;
- Re-format SDAT format to Sweep format;
- Prepare input files for packages of the R platform;

- Handle PHASE output file that contain only polymorphic sites to reconstruct the inferred haplotypes including polymorphic and monomorphic sites in FASTA format, as required by population genetics software for re-sequencing data such as DNAsp.

How to use DIVERGENOMEdb



DIVERGENOME stores and link information on genotypes, polymorphisms, individuals, populations, and individual phenotypes and even more important, to organize all these data in the format of Projects. These can be defined by their coordinator (e.g., Principal Investigators) as public (FIGURE 2), when the managed data is intended to be visualized by unregistered users (e.g., for published data), or as private, when data should be accessed only by users who have been granted permission by the project coordinator.



Users must first register before have access to private projects (data).

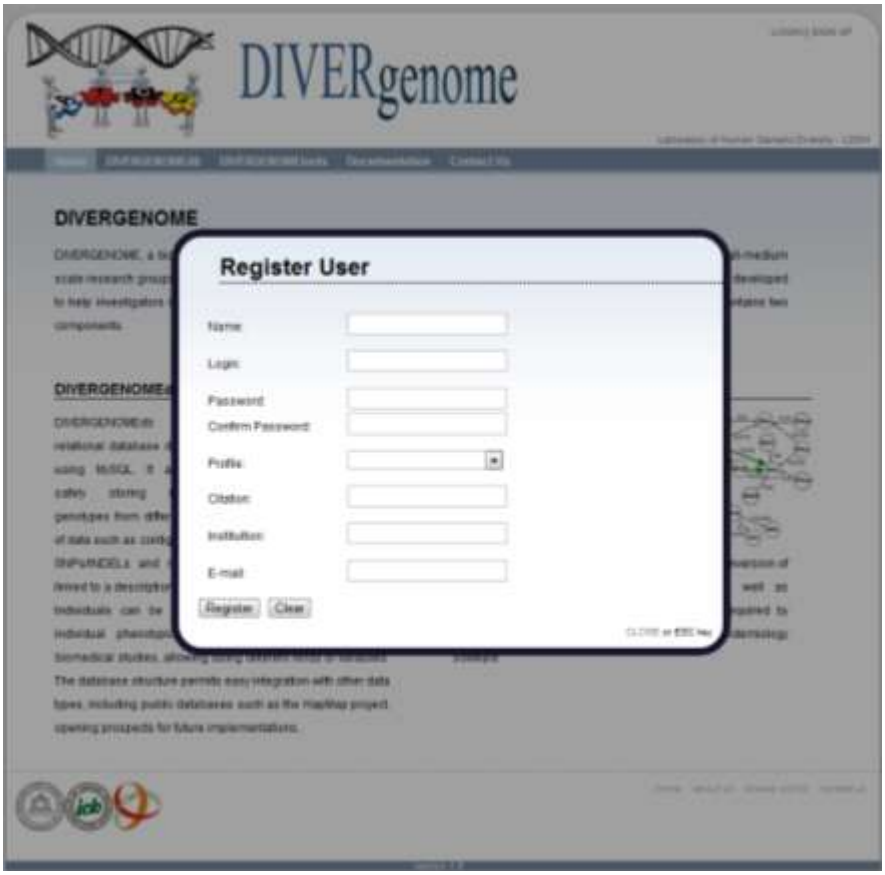
1. Logging into the system

Using any web browser, connect to the following URL:

http://www.pggenetica.icb.ufmg.br/divergenome/

Inside the DIVERGENOME home page, a login window will appear in the right corner. You should now type the **Username** and **Password** and press the **"Login"** Button. Take care of entering upper and lower case correctly.

2. Registering new Users



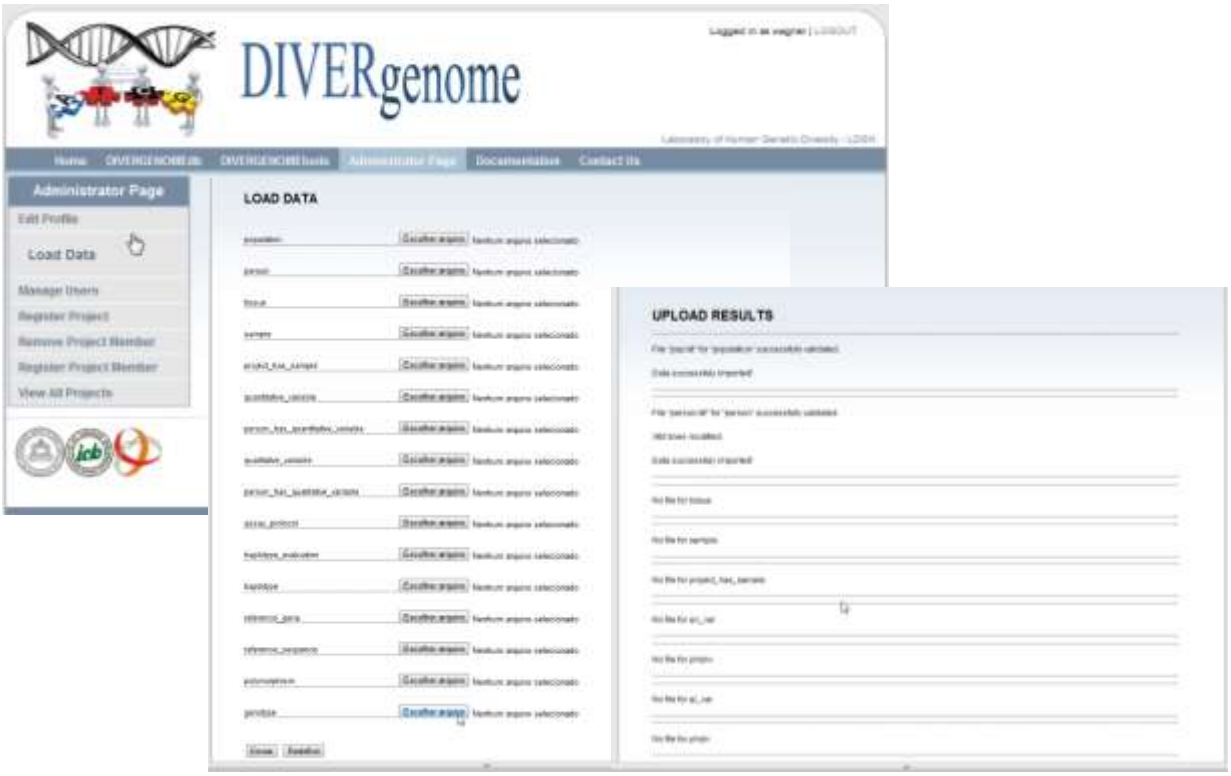
After complete this form your account will be registered.

After verifying your account information, your account will receive the status "waiting" and you should wait administrator approval. (FIGURE 3)


DIVERgenome
Logged in as wagner | LOGOUT
Home
DIVERGENOME.db
DIVERGENOME.tools
Contact Us
Waiting List
Show All
A B C D E F G H I J K L M
N O P Q R S T U W X Y Z
User Name: Eduardo M. Tarazona Santos
Login: eduardo
Administrator
Researcher
User
Waiting
Ready
Declined
Update Values
Delete User
Wagner C. Santos Magalhaes

3. Loading data

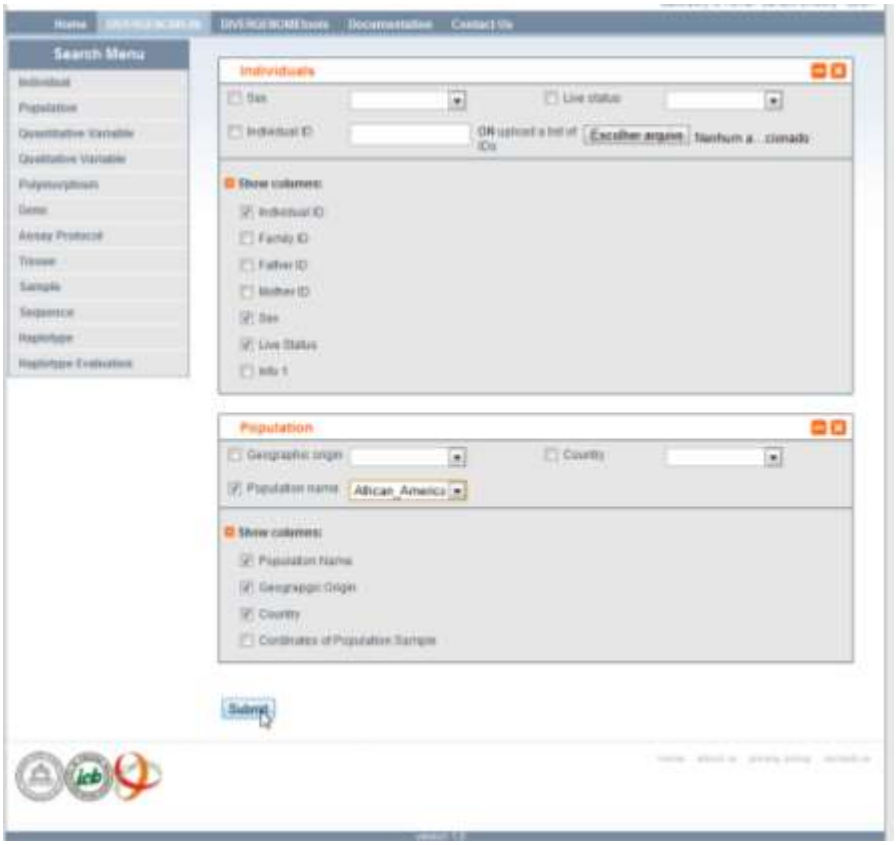
Data entry is carried out only by Administrators and Project Coordinators, as stated before, using the Web interface (described in following sections). At the moment, it is possible to upload files in CSV (comma separated value) and tab-delimited formats. Users can upload their own data as well as complementary data from different public data sources. Filling the example form (Excel file), saving it and uploading after.

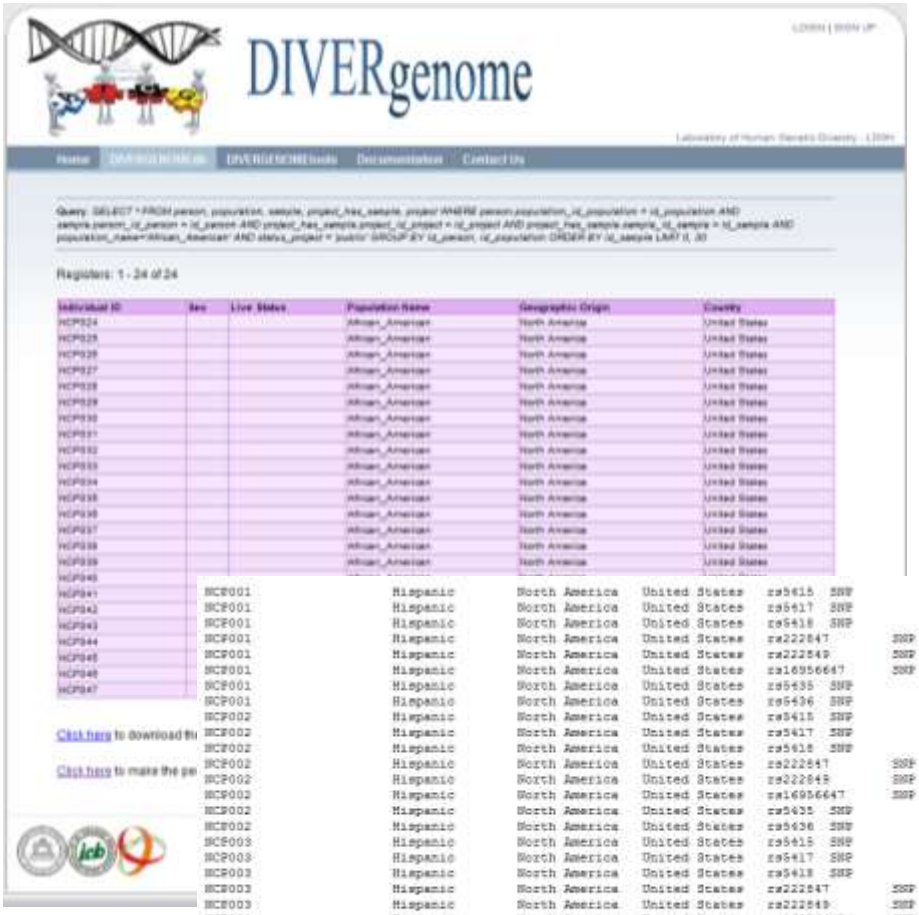


4. Searching

Almost all data retrieval options come with an example “value” that will guide you.

Combining Different tables





HCP001	Hispanic	North America	United States	rs5615	SNP		CC		
HCP001	Hispanic	North America	United States	rs5617	SNP		CC		
HCP001	Hispanic	North America	United States	rs5618	SNP			GG	
HCP001	Hispanic	North America	United States	rs222847		SNP			TT
HCP001	Hispanic	North America	United States	rs222849		SNP			TC
HCP001	Hispanic	North America	United States	rs16956647		SNP			CC
HCP001	Hispanic	North America	United States	rs5435	SNP		CT		
HCP001	Hispanic	North America	United States	rs5436	SNP		CC		
HCP002	Hispanic	North America	United States	rs5615	SNP		CT		
HCP002	Hispanic	North America	United States	rs5617	SNP		CA		
HCP002	Hispanic	North America	United States	rs5618	SNP			GA	
HCP002	Hispanic	North America	United States	rs222847		SNP			TT
HCP002	Hispanic	North America	United States	rs222849		SNP			TC
HCP002	Hispanic	North America	United States	rs16956647		SNP			CC
HCP002	Hispanic	North America	United States	rs5435	SNP		CC		
HCP002	Hispanic	North America	United States	rs5436	SNP		CC		
HCP003	Hispanic	North America	United States	rs5615	SNP		CC		
HCP003	Hispanic	North America	United States	rs5617	SNP		CA		
HCP003	Hispanic	North America	United States	rs5618	SNP			GA	
HCP003	Hispanic	North America	United States	rs222847		SNP			TT
HCP003	Hispanic	North America	United States	rs222849		SNP			CC
HCP003	Hispanic	North America	United States	rs16956647		SNP			CC
HCP003	Hispanic	North America	United States	rs5435	SNP		CC		
HCP003	Hispanic	North America	United States	rs5436	SNP		CC		
HCP004	Hispanic	North America	United States	rs5615	SNP		CC		
HCP004	Hispanic	North America	United States	rs5617	SNP		AA		
HCP004	Hispanic	North America	United States	rs5618	SNP			AA	
HCP004	Hispanic	North America	United States	rs222847		SNP			TT
HCP004	Hispanic	North America	United States	rs222849		SNP			CC
HCP004	Hispanic	North America	United States	rs16956647		SNP			CC
HCP004	Hispanic	North America	United States	rs5435	SNP		CC		
HCP004	Hispanic	North America	United States	rs5436	SNP		CC		
HCP005	Hispanic	North America	United States	rs5615	SNP		CC		
HCP005	Hispanic	North America	United States	rs5617	SNP		AA		
HCP005	Hispanic	North America	United States	rs5618	SNP			AA	
HCP005	Hispanic	North America	United States	rs222847		SNP			TT
HCP005	Hispanic	North America	United States	rs222849		SNP			CC
HCP005	Hispanic	North America	United States	rs16956647		SNP			CC
HCP005	Hispanic	North America	United States	rs5435	SNP		CC		
HCP005	Hispanic	North America	United States	rs5436	SNP		CC		
HCP006	Hispanic	North America	United States	rs5615	SNP		TT		
HCP006	Hispanic	North America	United States	rs5617	SNP		CC		
HCP006	Hispanic	North America	United States	rs5618	SNP			GG	
HCP006	Hispanic	North America	United States	rs222847		SNP			TT
HCP006	Hispanic	North America	United States	rs222849		SNP			TT
HCP006	Hispanic	North America	United States	rs16956647		SNP			CC
HCP006	Hispanic	North America	United States	rs5435	SNP		TT		
HCP006	Hispanic	North America	United States	rs5436	SNP		CC		
HCP007	Hispanic	North America	United States	rs5615	SNP		TT		
HCP007	Hispanic	North America	United States	rs5617	SNP		CC		
HCP007	Hispanic	North America	United States	rs5618	SNP			GG	
HCP007	Hispanic	North America	United States	rs222847		SNP			TT
HCP007	Hispanic	North America	United States	rs222849		SNP			TT
HCP007	Hispanic	North America	United States	rs16956647		SNP			CC
HCP007	Hispanic	North America	United States	rs5435	SNP		TT		
HCP007	Hispanic	North America	United States	rs5436	SNP		CC		
HCP008	Hispanic	North America	United States	rs5615	SNP		CC		

How to Use DIVERGENOMEtools

DIVERGENOMEtools' web page is shown in Figure 1. To start using it, the user needs to follow a couple of simple steps, described below and illustrated in Figure 8.

1. Choose your input file format to be converted by selecting one of the data formats listed in the left column. A brief description of each data format is shown when you place the mouse over a data format's name.

2. Choose the desired output format into which your input file will be converted. A list of possible output formats is listed in the right column. Output formats that cannot be generated by converting your selected input format are disabled. A brief description of each data format available for conversion is shown when you place the mouse over a data format's name.

3. Press "Submit" to proceed.

4. Upload boxes will appear according to your choice of input and output formats. Some conversions require more than one input file. You should upload all the required files and press "Submit".

5. The converted file is shown as a web link (in blue). Some conversions result in more than one output file, depending on the selected output format. You should download all resulting files. To download, right click the web link and choose "Save link as". Alternatively, click on the web link and the resulting file will be displayed in a separate page; right click on this separate page and choose "Save as".

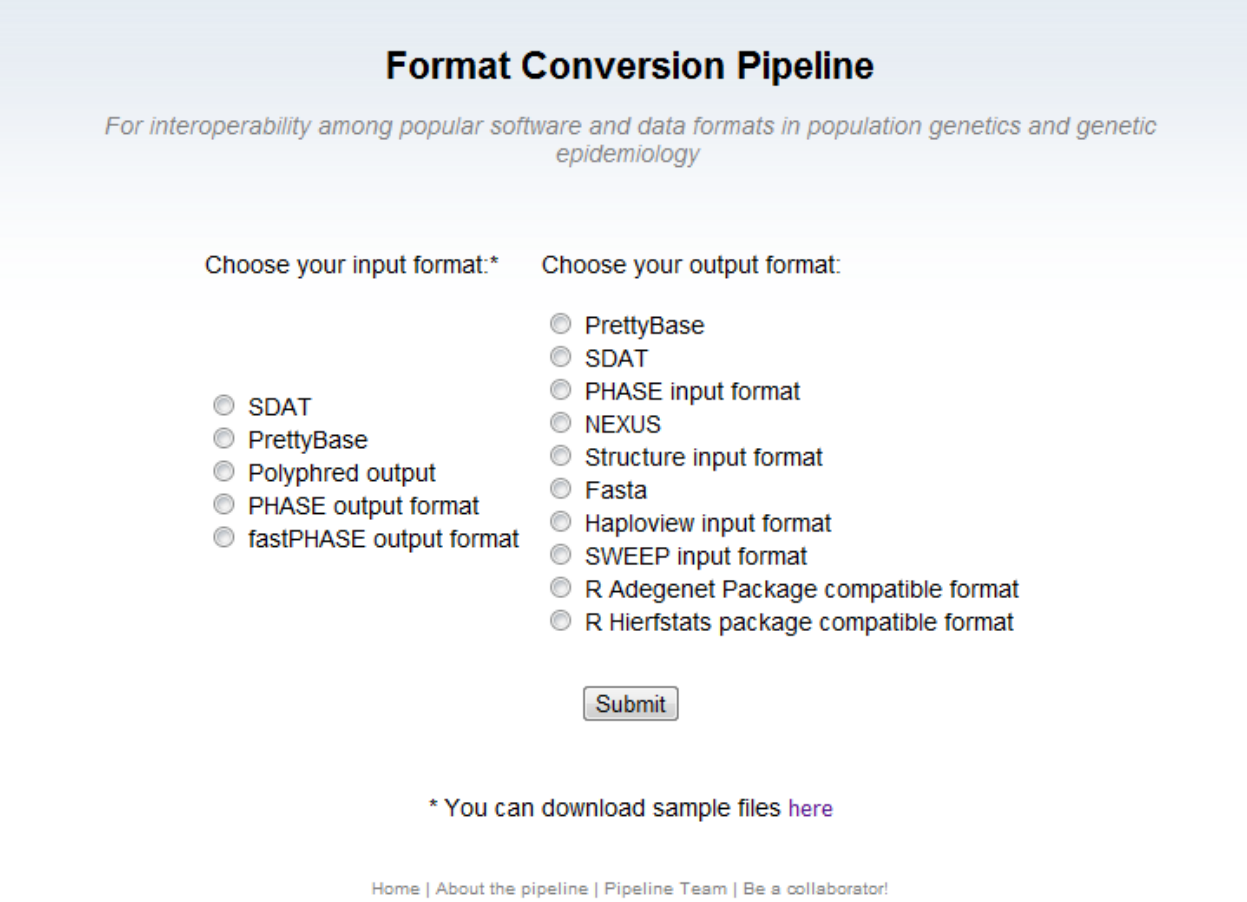
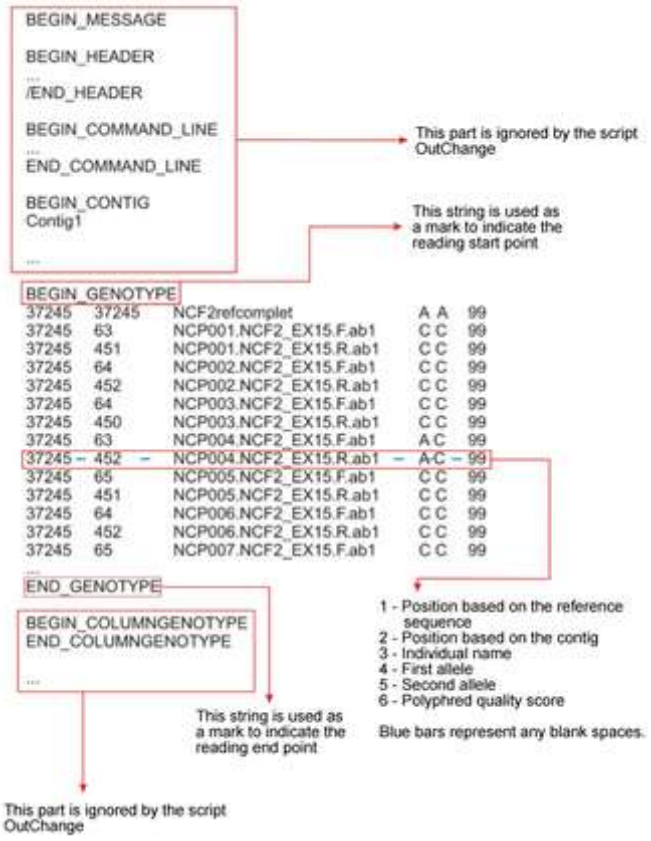


Figure 1: Web interface of DIVERGENOMEtools.

PolyPhred Output format

The Polyphred output format is the output file generated by the software PolyPhred during the analysis of a contig of sequences. Polyphred automatically makes genotype calls for each site identified as variable within a contig. If a reference sequence is used (included using the SudoPhred option of PolyPhred), as assumed by the pipeline, each polymorphism is identified by its position in the reference sequence. The pipeline parses the PolyPhred output file to get the relevant information, which is (the individual genotypes) in between the strings **BEGIN_GENOTYPE** and **END_GENOTYPE.** In this section, each row represents a genotype

call for a read and has five fields (columns): (1) the polymorphism identified by its position in the reference sequence, (2) the position of the polymorphism in the considered read, (3) the name of the read, as in the file exported from the automated sequencer, (4) the called genotype, (5) the quality score of the genotype call. It is important that the individual names in item (3) have equal IDs because otherwise they will be treated as different individuals (see the topic PolyPhred Output Start Point: Checking Sample IDs above). At last, the information showed before and after **BEGIN_GENOTYPE** and **END_GENOTYPE,** respectively**,** is ignored by the pipeline.



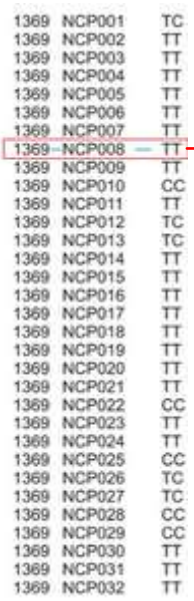
The PolyPhred file can be used to obtain 3 other files: Prettybase format, SDAT format and PHASE Input. These are described in the next sections.

Prettybase format

For the purpose of this pipeline, Prettybase is a tab delimited format which describes the data found by the PolyPhred software. The following table describes the accepted values for each column of the Prettybase format:

Column	Description
Site position	An integer uniquely identifying the locus
Individual ID	A string of characters uniquely identifying the individual
First allele	One character chosen from the set (A,G,C,T,?) , with '?' for unknown
Second allele	One character chosen from the set (A,G,C,T,?) , with '?' for unknown

Sample lines of a Prettybase genotype file are illustrated bellow:



- Column 1: Coordinates of variable sites based on the reference sequence

- Column 2: A string of characters uniquely identifying the individual

- Column 3: Genotypes

SDAT format

SDAT is a tab-delimited ASCII file containing a matrix of genotypes where each row represents a sample, each column represents a locus and the element (sample i, locus j) of the matrix if the genotype for the sample i for the locus j. More specifically:

Row 1: '\tab' in the first column, each locus name in subsequent columns.

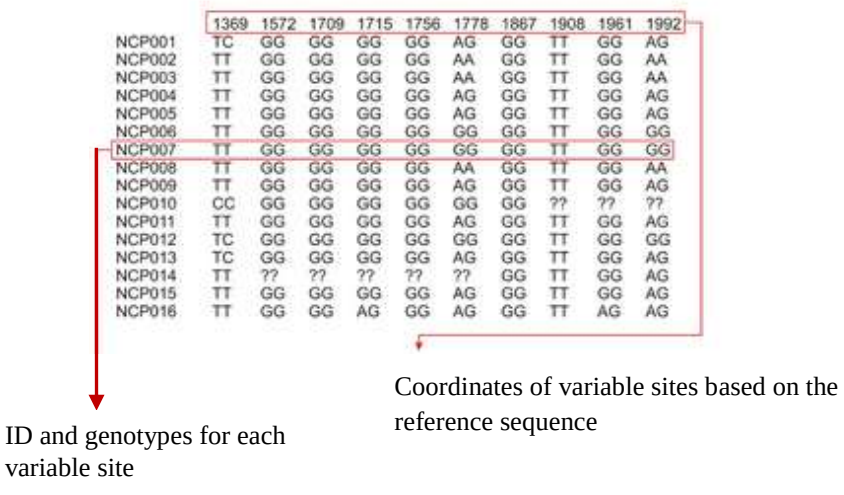
Row 2 and subsequent: sample name in the first column, genotypes for that sample for each locus in the subsequent columns.

Column 1: '\tab' in the first row, each sample name in subsequent rows

Column 2: locus name in the first row, genotypes for that locus for each sample in subsequent rows.

Genotypes are coded as a two character string with alleles 'ACGT' or the character '?' for missing alleles.

Example:



	1369	1572	1709	1715	1756	1778	1867	1908	1961	1992
NCP001	TC	GG	GG	GG	GG	AG	GG	TT	GG	AG
NCP002	TT	GG	GG	GG	GG	AA	GG	TT	GG	AA
NCP003	TT	GG	GG	GG	GG	AA	GG	TT	GG	AA
NCP004	TT	GG	GG	GG	GG	AG	GG	TT	GG	AG
NCP005	TT	GG	GG	GG	GG	AG	GG	TT	GG	AG
NCP006	TT	GG	GG	GG	GG	GG	GG	TT	GG	GG
NCP007	TT	GG	GG	GG	GG	GG	GG	TT	GG	GG
NCP008	TT	GG	GG	GG	GG	AA	GG	TT	GG	AA
NCP009	TT	GG	GG	GG	GG	AG	GG	TT	GG	AG
NCP010	CC	GG	GG	GG	GG	GG	GG	??	??	??
NCP011	TT	GG	GG	GG	GG	AG	GG	TT	GG	AG
NCP012	TC	GG	GG	GG	GG	GG	GG	TT	GG	GG
NCP013	TC	GG	GG	GG	GG	AG	GG	TT	GG	AG
NCP014	TT	??	??	??	??	??	GG	TT	GG	AG
NCP015	TT	GG	GG	GG	GG	AG	GG	TT	GG	AG
NCP016	TT	GG	GG	AG	GG	AG	GG	TT	AG	AG

Coordinates of variable sites based on the reference sequence

ID and genotypes for each variable site

PHASE input format

The input file accepted by the software PHASE specifies the following information in a tab delimited style:

Row 1: Number of individuals to be analyzed,

Row 2: Number of genotyped loci/sites,

Row3: The accepted type of loci: SNP (S) or microsatellites (M),

Successive rows: the genotypes for each individual.

The default structure for the input file can be represented as follows:

Number of Individuals

Number of Loci

P Position(1) Position(2) Position (Number of Loci)

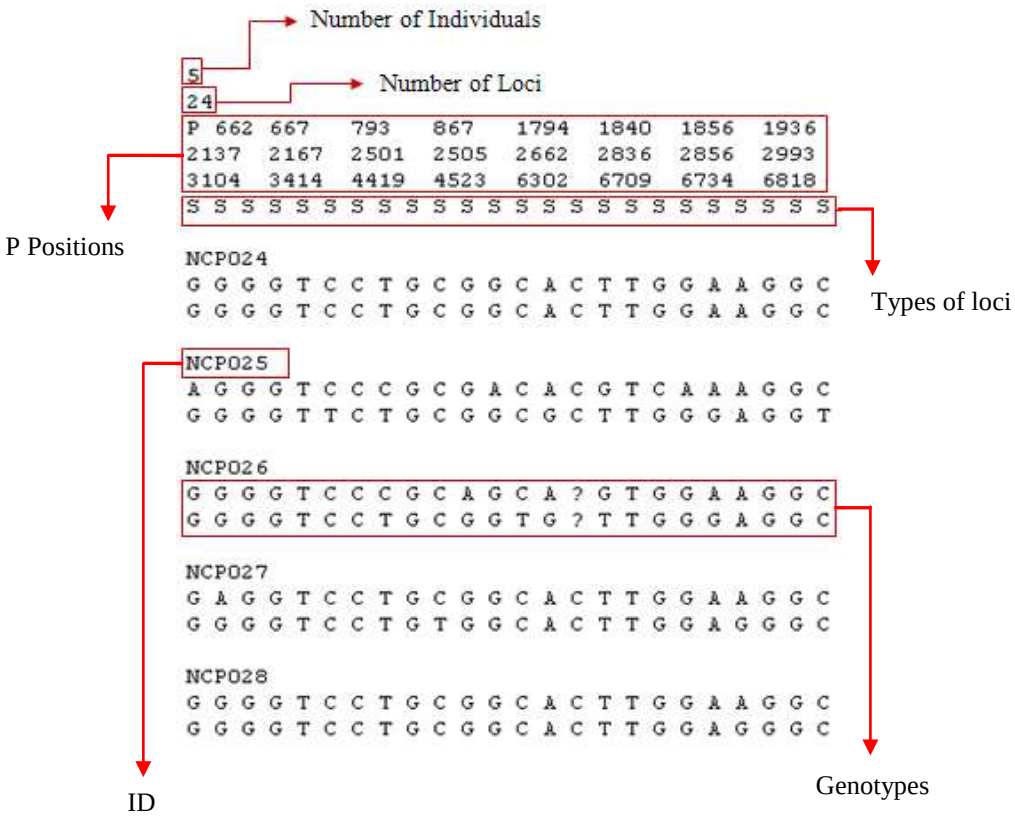
Locus Type (1) Locus Type (2) ... Locus Type (Number of Loci)

ID(1)

Genotype(1)

ID(2)

Example:



Additional information and different options and specific input flags may be found on the PHASE's documentation page at: http://stephenslab.uchicago.edu/instruct2.1.pdf.

PHASE Output format

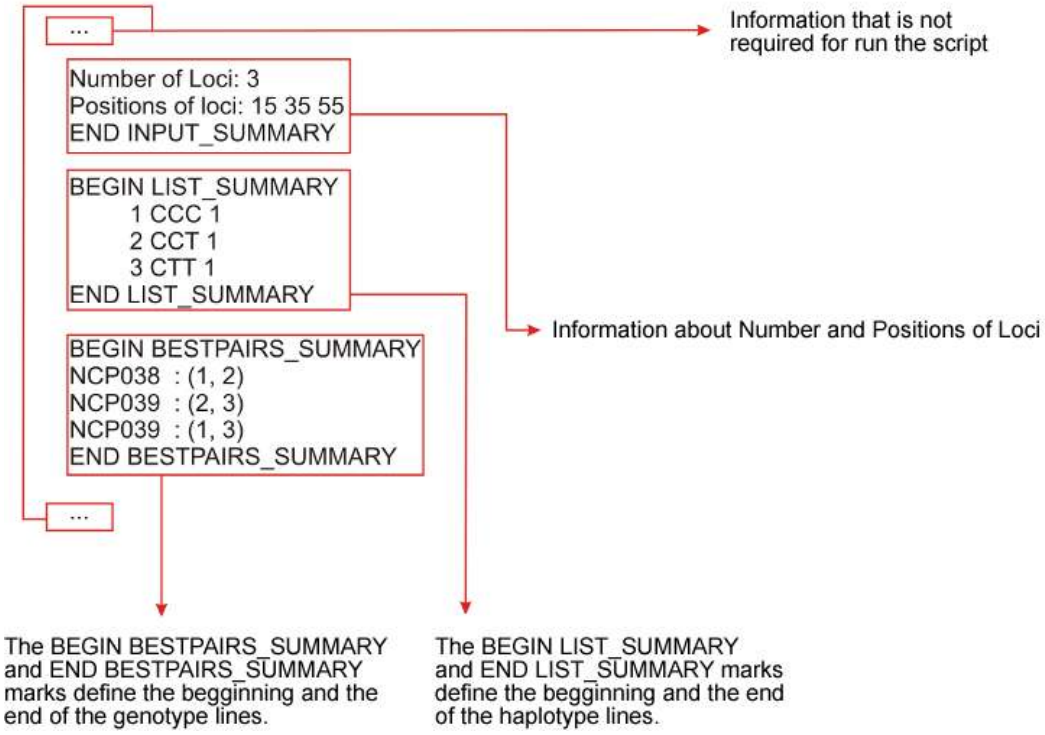
Among the different files produced by PHASE, the pipeline works on the summary output file, usually characterized by:

- A header containing the version number of the software and credits.
- A copy of the command line used to run the program.
- A list of haplotypes in the \best" reconstruction, with a summary of the frequency with which each haplotype occurred in this \best" reconstruction (note that these are not

supposed to be population frequency estimates; frequency estimates are given in the freqs _file).

- A list of the best haplotype guess for each individual, with parentheses () at positions where the phase was difficult to infer, and square brackets [ ] around alleles that were difficult to infer. Specifically, the bracketed positions indicate those positions where phase certainty (respectively genotype certainty) was < p (respectively < q), where the thresholds p and q can be set by the user at runtime with the -p and -q options (e.g. use -p0.8 to set the phase threshold to 80%). The default thresholds are p = q = 90%.

Example:



Additional information may be found on the PHASE documentation page at http://stephenslab.uchicago.edu/instruct2.1.pdf

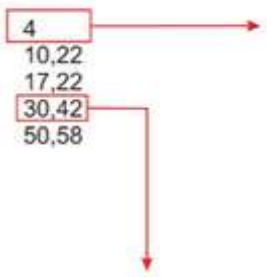
DNAsp Input format

With the pipeline it is also possible to create a FASTA file to be used as input to run the software DNAsp. To do this, the user needs to provide three different files:

1. The PHASE output file (described above);
2. A Fragments file;

   This file informs the number of sequenced fragments and the start e end positions of each fragment respect to the reference sequence separated by a comma.

   Example:



Row 1: number of fragments

Start and end positions of each fragments separated by comma.

3. A Reference sequence in FASTA format (described below);

   This file is a FASTA file used as a reference for the positions of each polymorphism identified by Polyphred (it must be the same file used to run Polyphred).

The FASTA file generated by the pipeline with the three files described above can then be used as input for the software DNAsp. This FASTA file format begins with the symbol '>' in the first line of the file; the sequence name is the first word after that symbol. Sequence names

can be up to 20 characters, blank spaces and tabs are not allowed. Additional characters in this line are considered to be comments. The sequence data starts in the second line. Nucleotide data can be written in one or more lines.

Example of FASTA format:

Structure

The input file accepted by the software Structure (Pritchard et al., 2000).

NEXUS

File composed of a number of blocks, such as TAXA, CHARACTERS, and TREES blocks.

R PACKAGES

Two R package formats are available for conversion: Adegenet (Jombart and Ahmed, 2011) and Hierfstat (de Meeus and Goudet, 2007). Both formats are tab delimited files containing a matrix of genotypes where each row represents a sample, the first column represents population

information, the following columns represent loci and the elements of the matrix are the genotypes.

HAPLOVIEW

The input file accepted by the software Haploview (Barrett et al., 2005).

SWEEP

The input file accepted by the software SWEEP (Sabeti et al., 2002).

References

Barrett JC, Fry B, Maller J, Daly MJ. 2005. **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics* 21:263-265.

de Meeus T, Goudet J. 2007. **A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels**. *Infect Genet Evol* 7:731-5.

Jombart T, Ahmed I. 2011**. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data**. *Bioinformatics*.

Montgomery KT IO, Li L, Loomis S, Obourn V, Kucherlapati R.: **PolyPhred analysis software for mutation detection from fluorescence-based sequence data**. *Current Protocol in Human Genetics* 2008, **Oct**(Oct):Chapter 7:Unit 7.16.

Pritchard JK, Stephens M, Donnelly P. 2000. **Inference of population structure using multilocus genotype data**. *Genetics* 155:945-959.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R: **DnaSP, DNA polymorphism analyses by the coalescent and other methods**. *Bioinformatics* 2003, **19**(18):2496-2497.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ and others. 2002. **Detecting recent positive selection in the human genome from haplotype structure**. *Nature* 419:832-837.

Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data**. *American Journal of Human Genetics* 2001, **68**(4):978-989.

Acknowledgements

We are grateful to Douglas Santos and Eduardo Galvão for his informatics technical assistance, Dr. Sérgio D Pena, Dr. Alexandre Pereira, Dr. Emanuel Dias Neto, Dr. Mirella Moro, and members of the Laboratory of Translational Genomics and the Core Genotyping Facility from the National Cancer Institute for their suggestions and criticisms. Members of the Laboratory of Human Genetic Diversity collaborated testing DIVERGENOME and with suggestions. We are also grateful to Dr. Peter E.M. Taschner (LOVD), Dr. Mike Feolo (dbGAP) and Dr. Carlos Morcillo (SNPator) for clarifying aspects of their bioinformatics platforms. Fogarty International Center and National Cancer Institute (5R01TW007894) funded this study. The study and its participants also received funding and fellowships from the following Brazilian agencies: Brazilian National Research Council (CNPq), Ministry of Education (CAPES), Ministry of Health (PNPD-Saúde Program) and the Minas Gerais State Reaearch Agency (FAPEMIG).