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: a bioinformatics platform to assist the analysis of genetic variation

✓ 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:

Address for correspondence:
E-mail: divergenome@gmail.com
Table of Contents

- What is the DIVERGENOME? ................................................................. 4
- How to use DIVERGENOMEdb ............................................................ 5
  1. Logging into the system ................................................................. 6
  2. Registering new Users ................................................................. 7
  3. Loading data .................................................................................. 9
  4. Searching ...................................................................................... 10
      Combining Different tables ............................................................ 11
- How to Use DIVERGENOMEtools .................................................... 13
- PolyPhred Output format ................................................................. 14
- Prettybase format ........................................................................... 16
- SDAT format .................................................................................... 17
- PHASE input format ........................................................................ 18
- PHASE Output format ..................................................................... 19
- DNAsp Input format ......................................................................... 21
- Structure .......................................................................................... 22
- NEXUS ............................................................................................. 22
- R PACKAGES ................................................................................... 22
- HAPLOVIEW .................................................................................... 23
- SWEEP .............................................................................................. 23
- References ...................................................................................... 23
- Acknowledgements ......................................................................... 24
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: a bioinformatics platform to assist the analysis of genetic variation
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**

![Image of combining different tables in DIVERGENOME](image_url)
**DIVERGENOME: a bioinformatics platform to assist the analysis of genetic variation**

![DIVERGENOME Platform Screenshot]

**Table Example**

<table>
<thead>
<tr>
<th>Individual ID</th>
<th>Sex</th>
<th>State</th>
<th>Population Name</th>
<th>Geographical Origin</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCP001</td>
<td>M</td>
<td>NY</td>
<td>North America</td>
<td>United States</td>
<td>NY</td>
</tr>
<tr>
<td>HCP002</td>
<td>F</td>
<td>CA</td>
<td>North America</td>
<td>United States</td>
<td>CA</td>
</tr>
<tr>
<td>HCP003</td>
<td>M</td>
<td>TX</td>
<td>North America</td>
<td>United States</td>
<td>TX</td>
</tr>
<tr>
<td>HCP004</td>
<td>F</td>
<td>GA</td>
<td>North America</td>
<td>United States</td>
<td>GA</td>
</tr>
<tr>
<td>HCP005</td>
<td>M</td>
<td>TX</td>
<td>North America</td>
<td>United States</td>
<td>TX</td>
</tr>
<tr>
<td>HCP006</td>
<td>F</td>
<td>CA</td>
<td>North America</td>
<td>United States</td>
<td>CA</td>
</tr>
<tr>
<td>HCP007</td>
<td>M</td>
<td>CA</td>
<td>North America</td>
<td>United States</td>
<td>CA</td>
</tr>
<tr>
<td>HCP008</td>
<td>F</td>
<td>TX</td>
<td>North America</td>
<td>United States</td>
<td>TX</td>
</tr>
</tbody>
</table>

**Legend:**
- **NH:** North America
- **CA:** California
- **TX:** Texas
- **NY:** New York
- **GA:** Georgia
- **NY:** New York
- **CA:** California
- **TX:** Texas
- **GA:** Georgia

**Note:** The platform allows for the analysis and visualization of genetic variation data across different populations and geographical regions.
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 \texttt{BEGIN\_GENOTYPE} and \texttt{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.
DIVERGENOME: a bioinformatics platform to assist the analysis of genetic variation

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:

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

Sample lines of a Prettybase genotype file are illustrated bellow:

```
1369  NCP001  TC
1369  NCP002  TT
1369  NCP003  TT
1369  NCP004  TT
1369  NCP005  TT
1369  NCP006  TT
1369  NCP007  TT
1369  NCP008  TT
1369  NCP009  TT
1369  NCP010  CC
1369  NCP011  TT
1369  NCP012  TC
1369  NCP013  TC
1369  NCP014  TT
1369  NCP015  TT
1369  NCP016  TT
1369  NCP017  TT
1369  NCP018  TT
1369  NCP019  TT
1369  NCP020  TT
1369  NCP021  TT
1369  NCP022  CC
1369  NCP023  TT
1369  NCP024  TT
1369  NCP025  CC
1369  NCP026  TC
1369  NCP027  TC
1369  NCP028  CC
1369  NCP029  CC
1369  NCP030  TT
1369  NCP031  TT
1369  NCP032  TT
```

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

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

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.
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:

```
... 
Number of Loci: 3
Positions of loci: 15 35 55
END INPUT_SUMMARY

BEGIN LIST_SUMMARY
  1 CCC 1
  2 CCT 1
  3 CTT 1
END LIST_SUMMARY

BEGIN BESTPAIRS_SUMMARY
  NCP038 : (1, 2)
  NCP039 : (2, 3)
  NCP039 : (1, 3)
END BESTPAIRS_SUMMARY

... 
```

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 and end positions of each fragment respect to the reference sequence separated by a comma.

   Example:
   
   ![Row 1: number of fragments](image1)
   
   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:

```
>Seq_ref  
GCGGAAGGAGAGGTTCATTGAGAAACCATGAAAGGAGCTTGTGTACGAGATTCTCTGGGAGAGGGAGA 
>NCP001a  
NNNNNNNNNNNGCGGAAGGAGGAGATGAGGAGTT--CAGAAGCGGAATCA--TGGGTGGAGGAGCTTTCA 
>NCP001b  
NNNNNNNNNNNGCGGAAGGAGGAGATGAGGAGTT--CAGAAGCGGAATCA--TGGGTGGAGGAGCTTTCA 
>NCP002a  
NNNNNNNNNNNGCGGAAGGAGGAGATGAGGAGTT--CAGAAGCGGAATCA--TGGGTGGAGGAGCTTTCA 
>NCP002b  
NNNNNNNNNNNGCGGAAGGAGGAGATGAGGAGTT--CAGAAGCGGAATCA--TGGGTGGAGGAGCTTTCA 
>NCP003a  
NNNNNNNNNNNGCGGAAGGAGGAGATGAGGAGTT--CAGAAGCGGAATCA--TGGGTGGAGGAGCTTTCA 
>NCP003b  
NNNNNNNNNNNGCGGAAGGAGGAGATGAGGAGTT--CAGAAGCGGAATCA--TGGGTGGAGGAGCTTTCA 
>NCP004a  
NNNNNNNNNNNGCGGAAGGAGGAGATGAGGAGTT--CAGAAGCGGAATCA--TGGGTGGAGGAGCTTTCA 
>NCP004b  
NNNNNNNNNNNGCGGAAGGAGGAGATGAGGAGTT--CAGAAGCGGAATCA--TGGGTGGAGGAGCTTTCA 
```

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**


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 Research Agency (FAPEMIG).