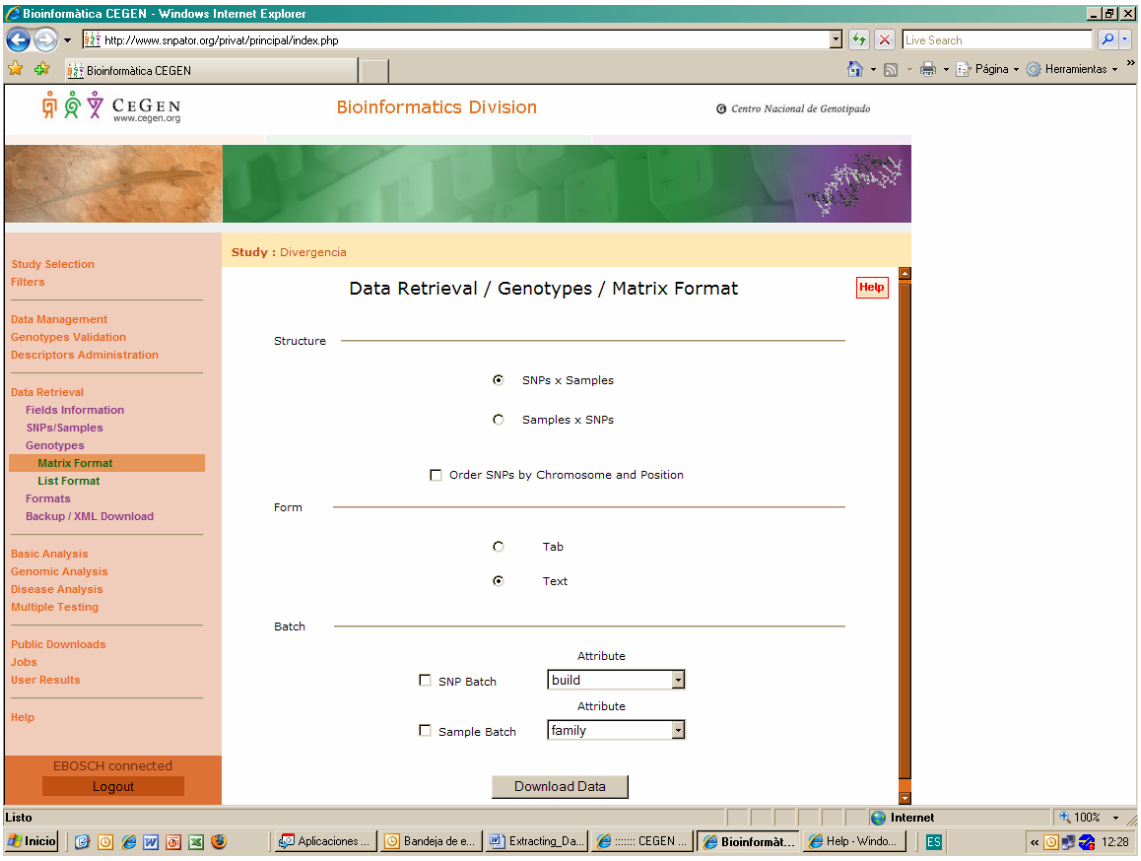
TUTORIAL
Extracting Data with SNPator

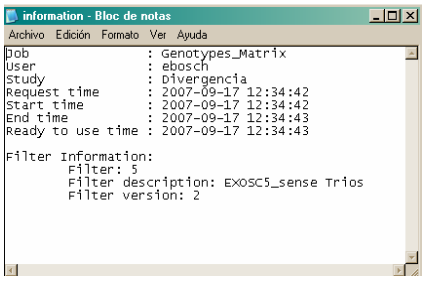
Genotype Data from a given study in SNPator can be retrieved in several different formats. You have the option of obtaining just the genotypes corresponding to a given number of Samples and SNPs in a list or a matrix format (see **Matrix Format** and **List Format**); you can retrieve your genotypes in different ready-to-use files, which you can use directly as an input files for a number of different software applications such as **Arlequin**, **Phase**, **Haploview**, **Association Cluster Detector** and **Multifactor Dimensionality Reduction** (see examples for Arlequin, Phase and Haploview); or you can also retrieve the genotypes from your study together with different Sample Attributes in a **Generic Format** that is suitable to be imported by calculus or statistical packages so you can therefore keep, transform and work with your data as you like.

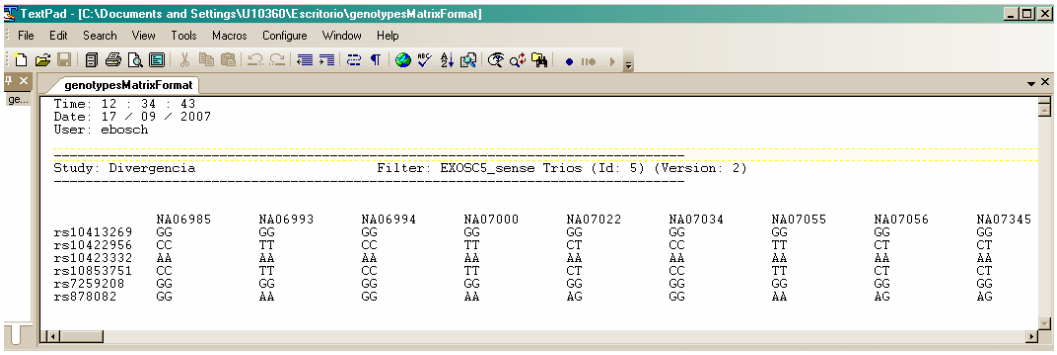
1. Matrix Format:

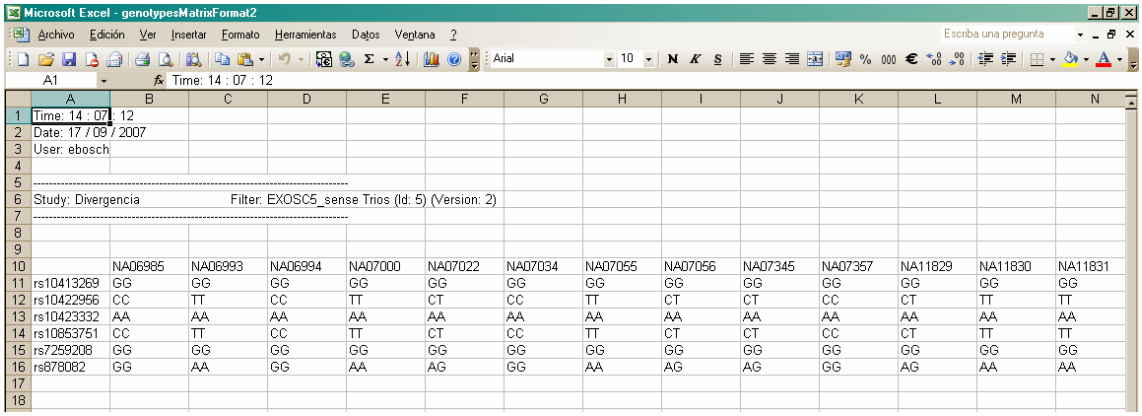
Under the **Matrix Format** option in the **Data Retrieval - Genotypes** menu, genotypes can be downloaded in a matrix that can be easily imported in most data programs as Excel, SPSS, etc. You can choose whether to put **SNPs** in rows (getting **SNPs** x **Samples**) or in columns (**Samples** x **SNPs**). **SNPs** will be ordered in alphabetical order unless you select the option "**Order SNPs by Chromosome and Position**".



A **Report** file separated by tabs or spaces will be generated depending on the settings of the second option. The **Matrix Format Report** will be sent to the **User Results** section. Download to your computer the **ZIP Matrix Format Report** file or open the zip file directly. This ZIP contains two files (the **information** file and the **genotypeMatrixformat** file), which can be opened as a text or with Excel. They look as follows:





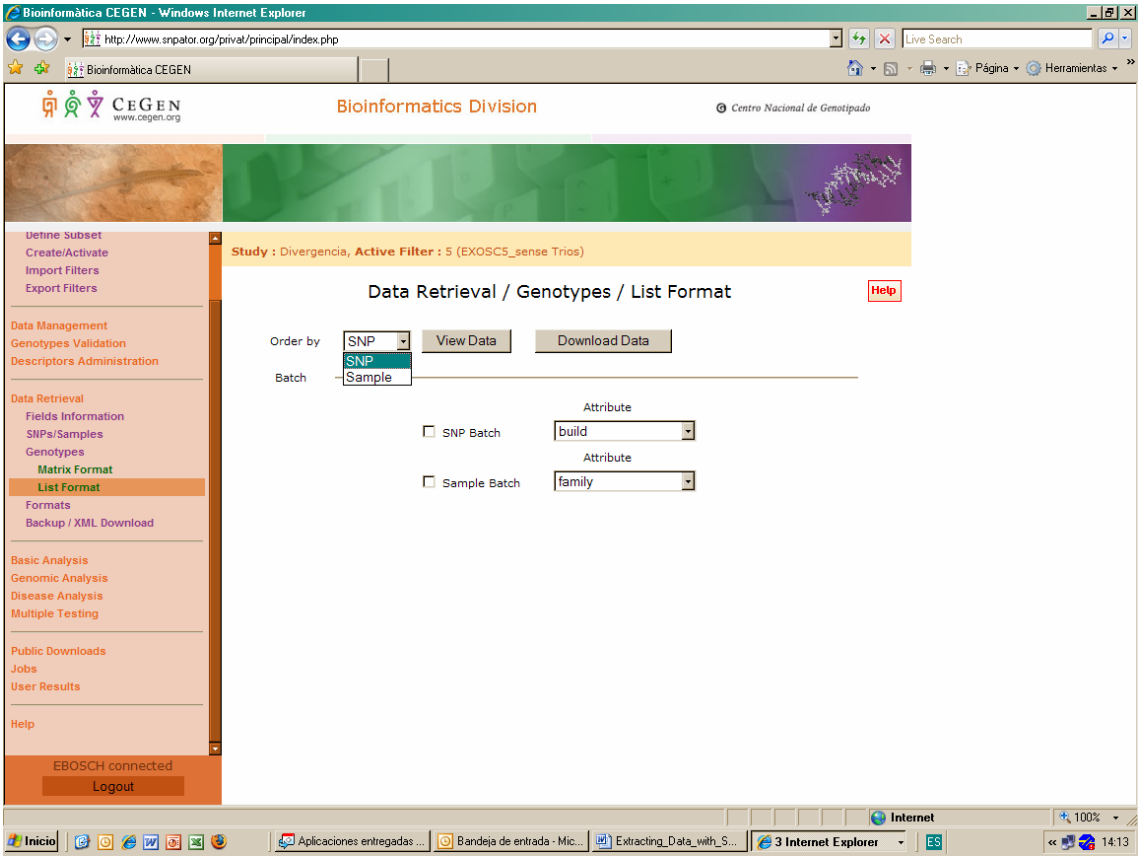


You can use **Batch Mode** option by selecting one of the fields in the **Samples** or **SNP** table in order obtain as many Genotype Matrices as different values are in those fields, using each time only those **samples** or **SNPs** that have each of the values. For instance, if you have defined your SNPs in a *"gene"* field as "BCA1" or BCA2", selecting "gene" as the attribute of the **SNP batch mode** will result in having two runs of the **Matrix Format** genotypes retrieval, taking separately SNPs in BCA1 and BCA2. If you select at the same time **Sample** and **SNP batch** fields, you are going to obtain as many runs of the process as

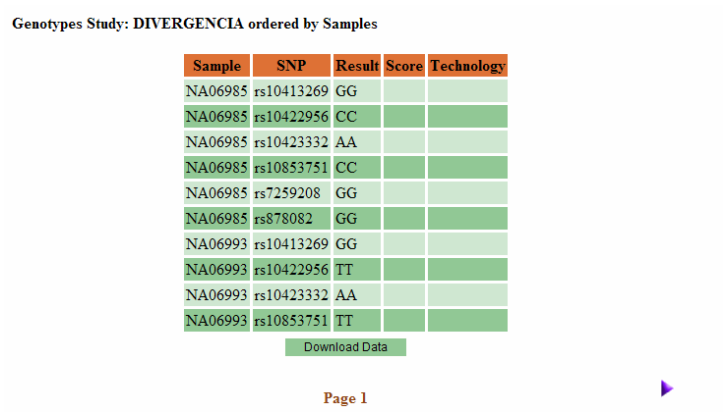
all possible combinations of the values of **samples** (for example, Cases and Controls) and **SNPs** in the fields you selected.

2. List Format:

An ordered list of all **Genotypes** can be obtained using the **Genotypes - List Format** option in the **Data Retrieval** menu. **Results** can be ordered by **SNP** or **Sample** and as in other applications in SNPator, a **Batch Mode** option is available.



The **View Data** option allows for a quick browsing of the data without generating a **Report.**



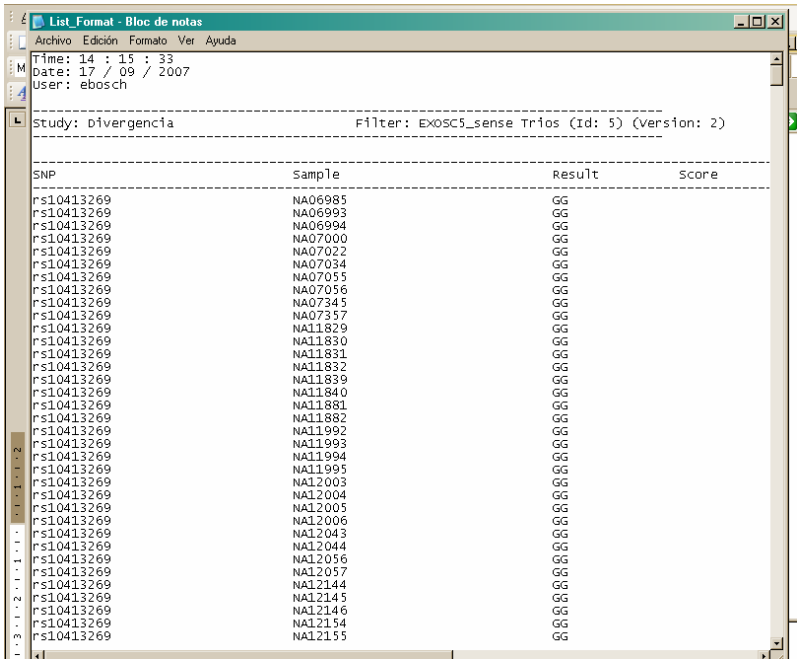
Genotypes Study: DIVERGENCIA ordered by Samples

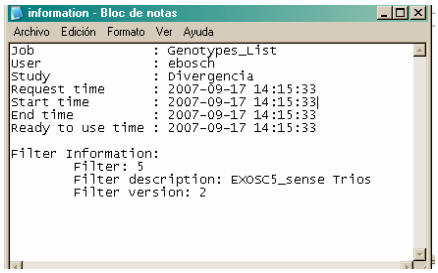
Sample	SNP	Result	Score	Technology
NA06985	rs10413269	GG		
NA06985	rs10422956	CC		
NA06985	rs10423332	AA		
NA06985	rs10853751	CC		
NA06985	rs7259208	GG		
NA06985	rs878082	GG		
NA06993	rs10413269	GG		
NA06993	rs10422956	TT		
NA06993	rs10423332	AA		
NA06993	rs10853751	TT		

Download Data

Page 1

The **Download Data** option creates a **List Format Report** which will be sent to the **User Results** section. This ZIP contains two files (the **information** file and the **List_Format** file), which can be opened as a text or with Excel. They look as follows:





3. Data Retrieval - Format Phase:

A properly formatted and **ready-to-use** file that you can use as an input file for the **Phase** software is generated in the **Phase** option under the **Data Retrieval – Formats menu**.

What you need to do is as follows:

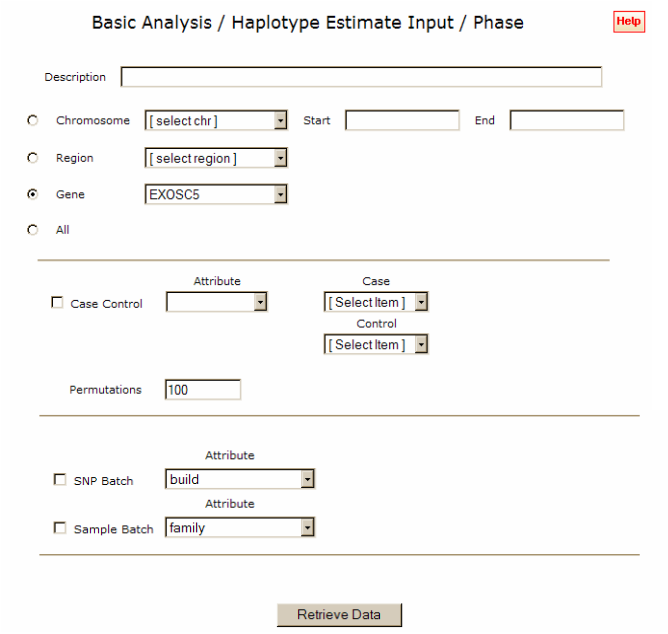
1. Enter a Description to identify the process in the top box.

2. Select which of the **SNPs** in your **Study** are going to be used. You can do it in several ways:

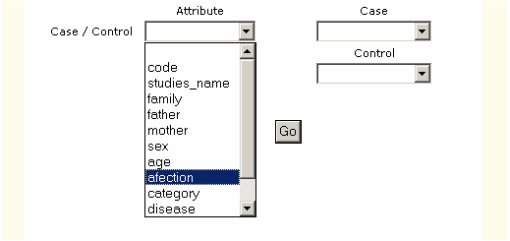
- By position: You can specify a chromosome and the positions which delimit the region you are interested in.

- By Region: Depending on the content of the field "**Region**" in your **SNP Table**.

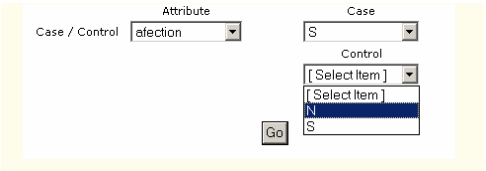
- By Gene: Depending on the content of the field "**Gene**" in your **SNP Table**.



3. If the "**Case Control**" box is ticked, the **Phase** input file generated by **SNPator** will be ready to be run with the Case/Control option on. In this case, the field that will be used to distinguish cases from controls must be selected:



Then, you have to select a value for cases and controls:

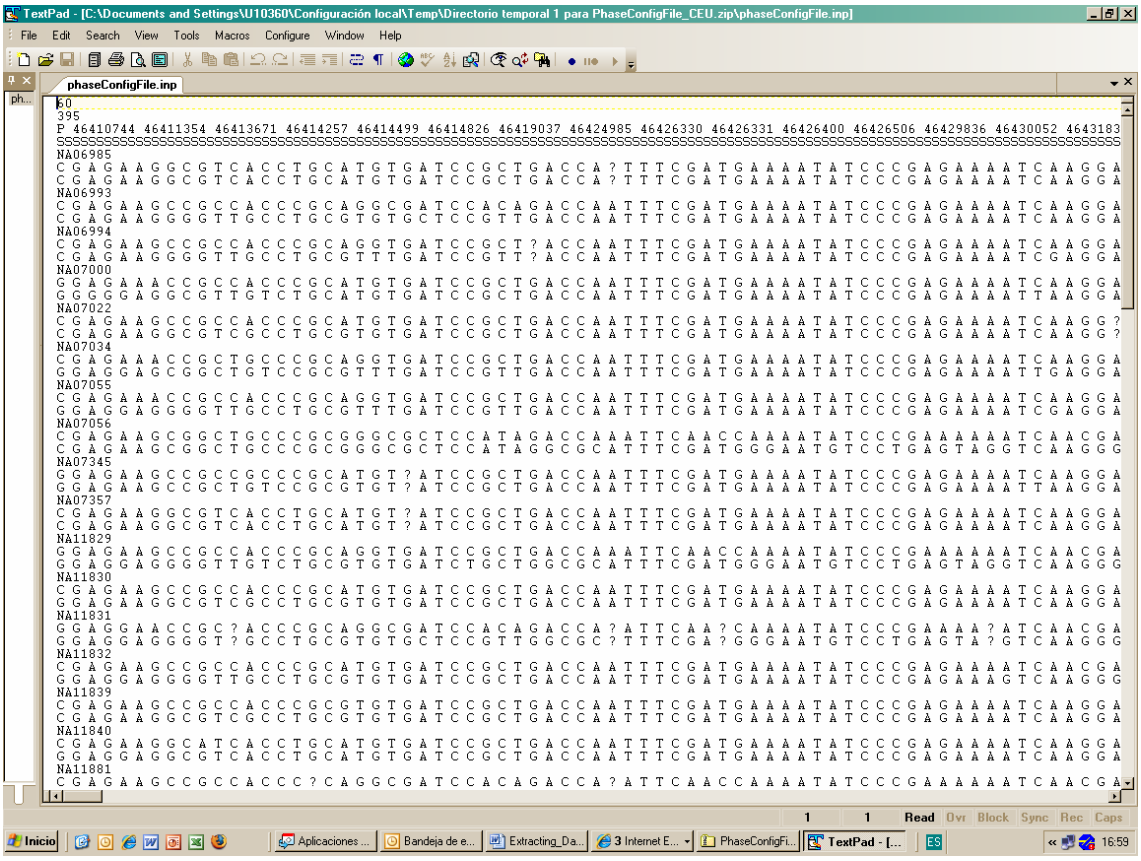


Samples with values other than the ones declared here will be discarded and not used in the process. If a blank " " is selected, that means that there are **Samples** with no value in the field and that you can select those as a case or control.

4. Use the **Batch Mode** option to run the Phase process as many times as different values in the Samples and/ore SNPs fields have been selected.

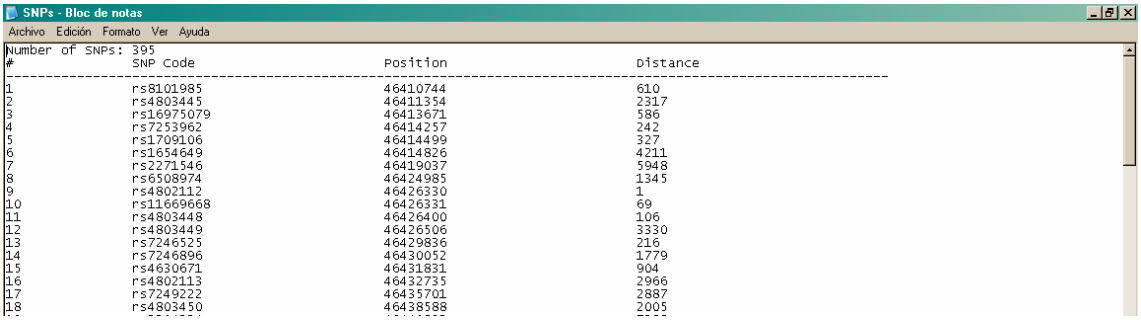
5. Finally, after pressing go the resulting Phase input file is sent to the **User Results** section. It will be a *.zip file containing several files:

- phaseConfigFile.inp
This is the Phase input file that has to be given to Phase in order to perform haplotype estimations.



- SNPs.txt
A text file containing the list of **SNPs** that have been used to build the phase input file. It provides info on:

- SNP code
- Position
- Distance to next SNP



- SNPs.snp
Another **SNP** info text with a format suitable for other programs such as Sweep.

- information.txt
SNPator information to identify the job: date, time, user, study, and filter.

- status.txt
Report of possible errors in the process.

4. Data Retrieval - Arlequin Format:

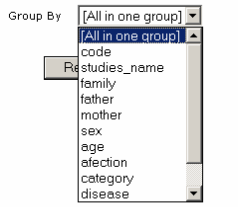
A properly formatted and **ready-to-use** file that you can use as an input file for the **Arlequin** software is generated in the **Arlequin** option under the **Data Retrieval – Formats menu**. The resulting file is formatted to perform linkage disequilibrium tests in Arlequin. Once this option is executed, a *.zip file appears in the **User results** section containing:

- information.txt
SNPator related information to identify the job: date, time, user, study and filter.

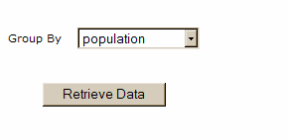
- arlequinConfigFile.inp
The input file for Arlequin

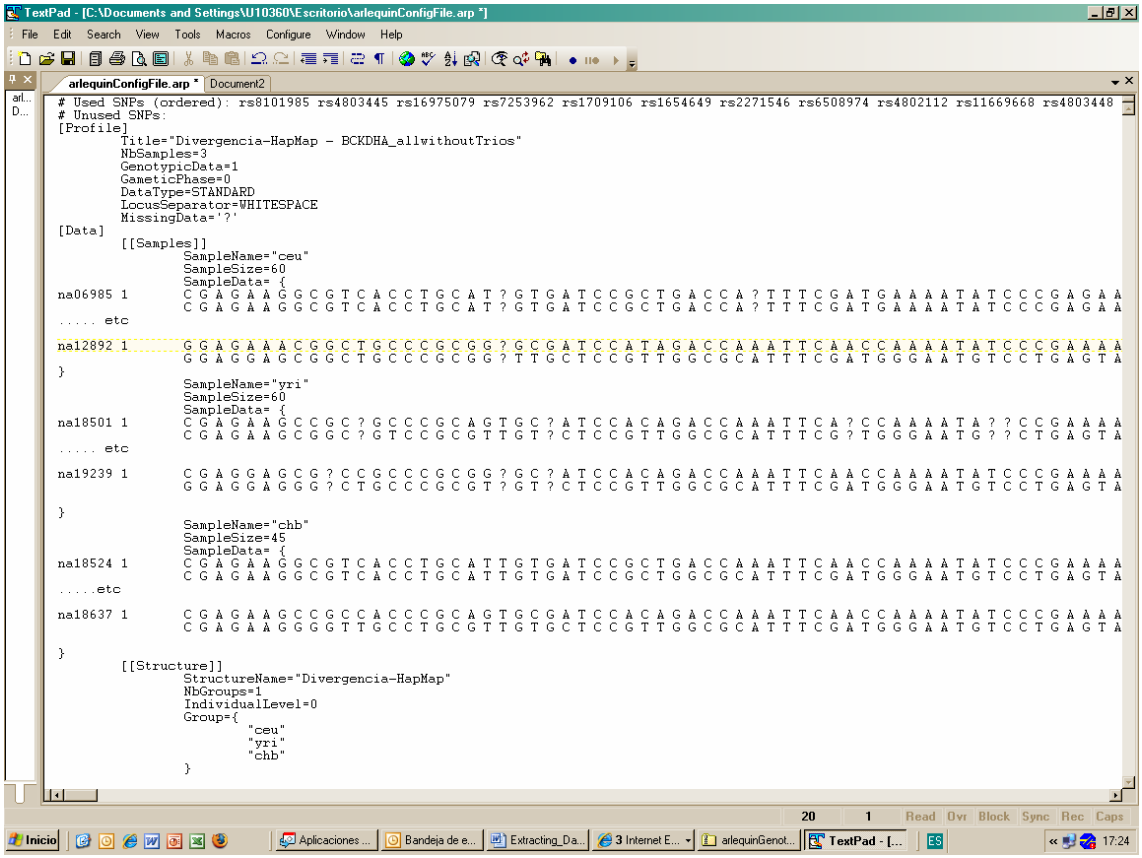
This file will contain two comment lines at the top indicating which **SNPs** have been used and which **SNPs** have been discarded in the creation of the file.

Moreover, the data will be put together in only one sample or divided into different samples depending on the value of the box Group By:



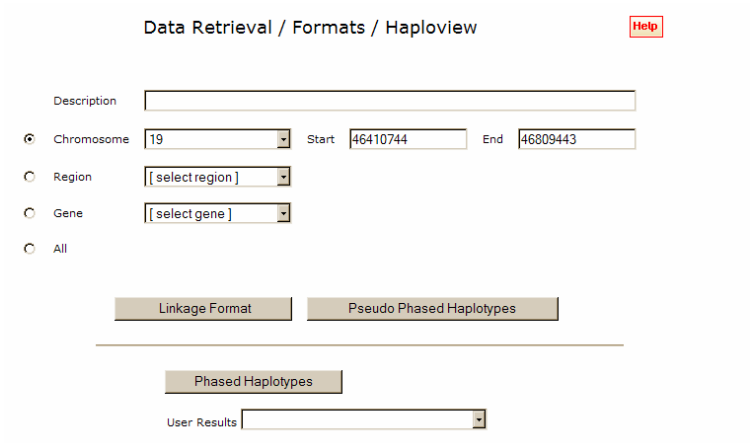
If you select a field here, the values of that field will be used as categories to create different samples in the file. For example, in a study with samples from three populations, if we group by population data, will appear as follows:





4. Data Retrieval - Haploview Format:

A properly formatted and **ready-to-use** file that you can use as an input file for the **Haploview** software is generated in the **Haploview** option under the **Data Retrieval – Formats menu**.



The procedure that one needs to follow depends of the type of **Haploview** file that you want to generate: Linkage format, Pseudo Phased Haplotypes or Phased Haplotypes.

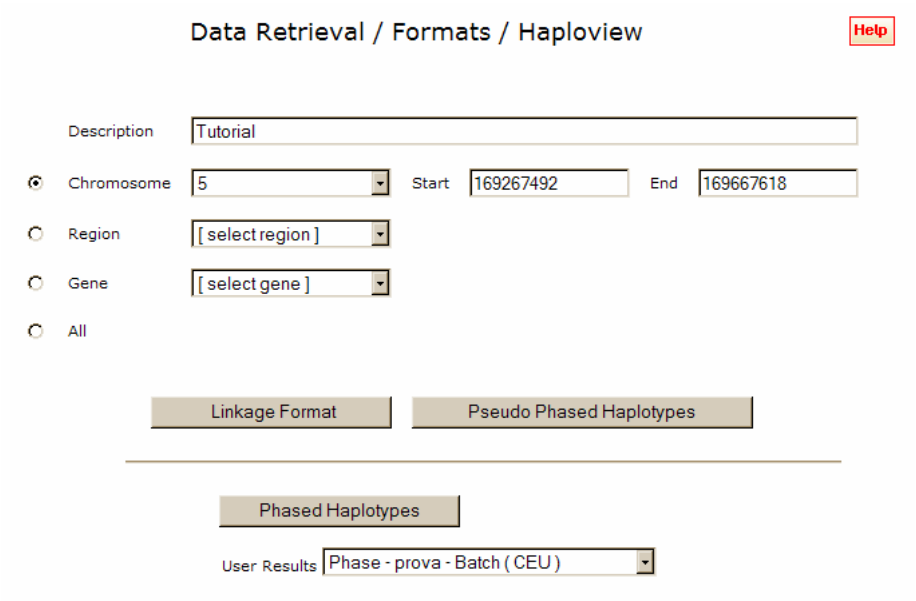
For example **Phased Haplotype** data for Haploview's input can be obtained as follows:

The **Haploview** file will be generated from a previously run **PHASE** haplotype estimate and, consequently, with known genetic phase.

1. Enter a Description to identify the process.

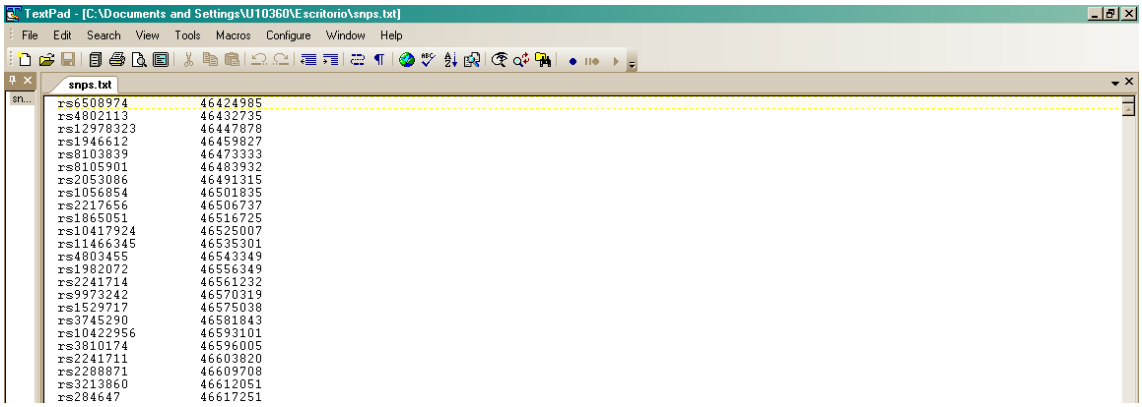
2. Select a previously run **PHASE** result in the "User results" combobox

3. Press the "**Phased Haplotypes**" button. The resulting file can be found in the **User Results** section.

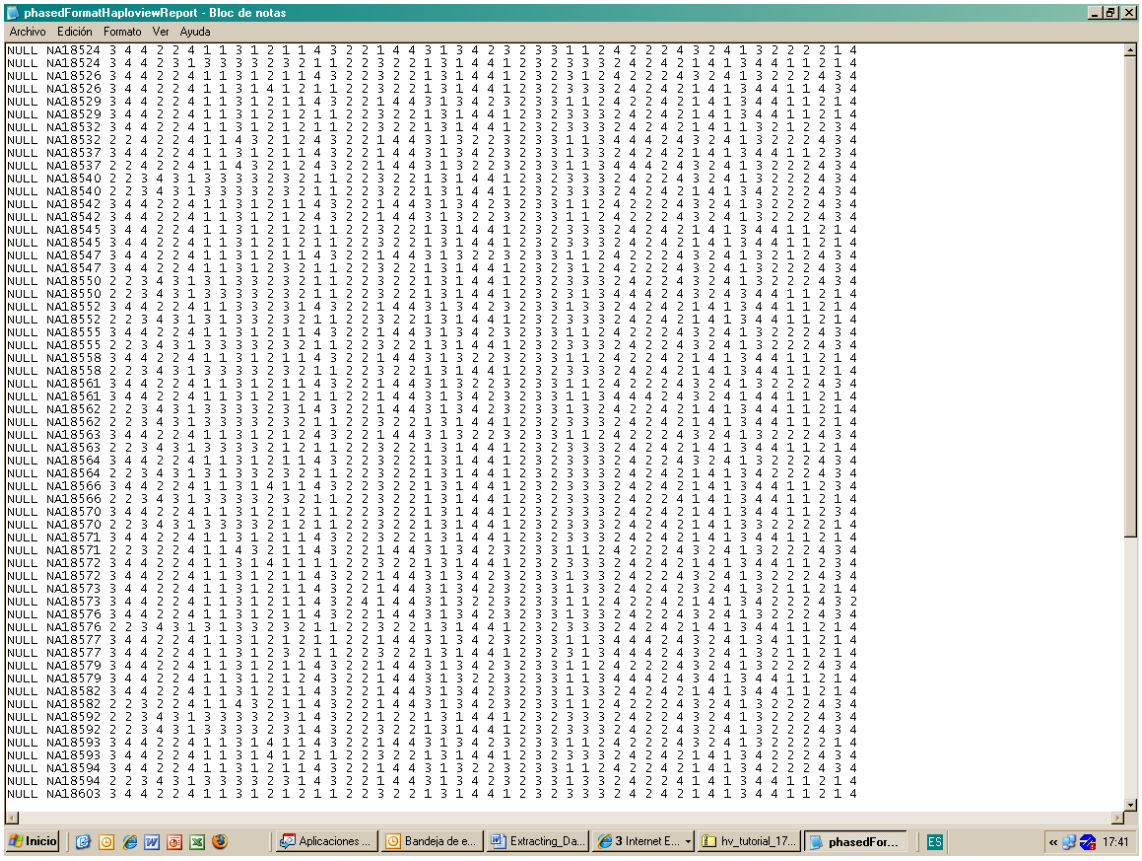


There should be a *.zip file containing:

- snps.txt
List of **SNPs** with their positions in the format required by Haploview

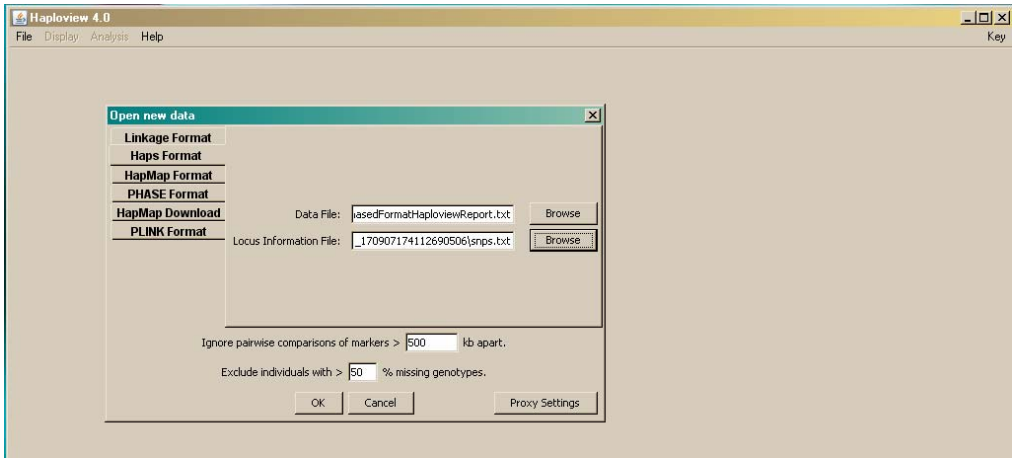


- phasedFormatHaploviewReport.txt
The input file for Haploview with the genotype information.



- information.txt
SNPator information to identify the job: date, time, user, study and filter.

4. Upload the **phasedFormatHaploviewReport.txt** as Data file and the **snps.txt as Locus Information File** in the Haps Format option for opening your data in Haploview


Haploview 4.0
File Display Analysis Help
Key
Open new data
Linkage Format
Haps Format
HapMap Format
PHASE Format
HapMap Download
PLINK Format
Data File:
asedFormatHaploviewReport.txt
Browse
Locus Information File:
_17090717411269O506\snps.txt
Browse
Ignore pairwise comparisons of markers > 500 kb apart.
Exclude individuals with > 50 % missing genotypes.
OK
Cancel
Proxy Settings