

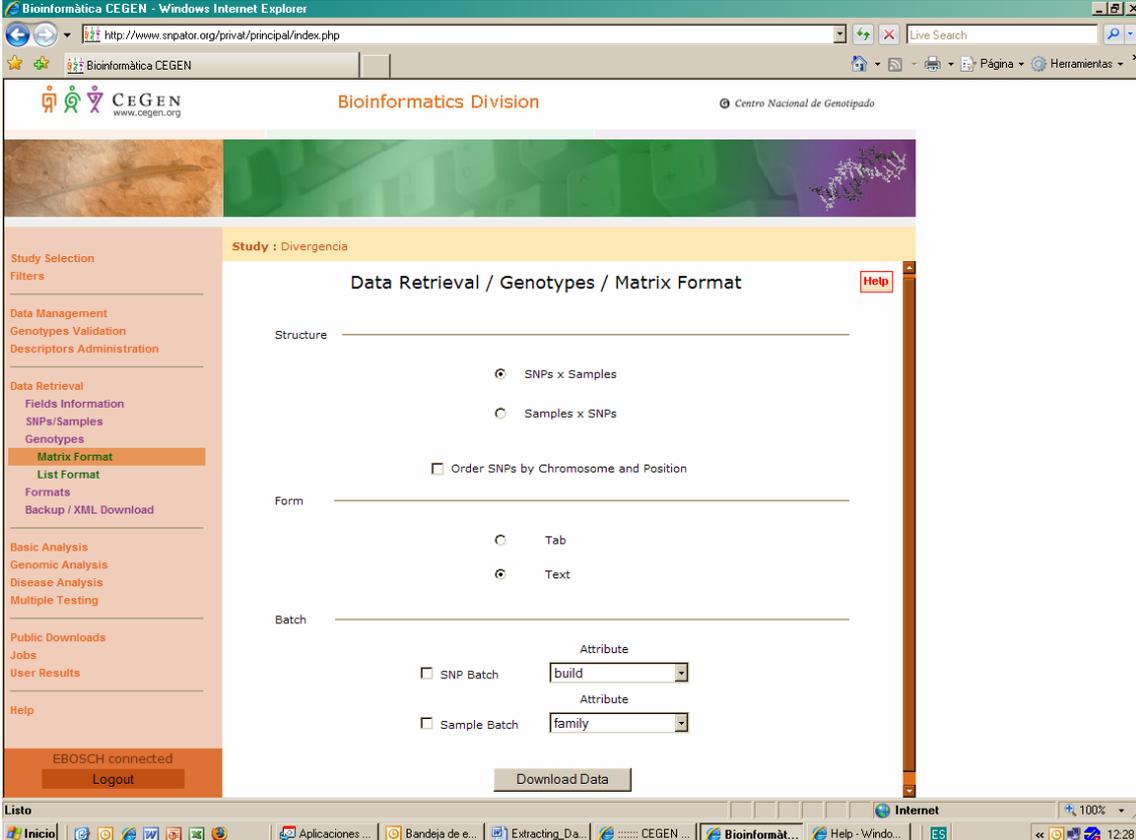
TUTORIAL

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".



The screenshot shows the SNPator web interface in Internet Explorer. The browser address bar shows the URL <http://www.snpator.org/privat/principal/index.php>. The page title is "Bioinformática CEGEN - Windows Internet Explorer". The main content area is titled "Data Retrieval / Genotypes / Matrix Format" and includes a "Help" button. The interface is for the study "Divergencia".

The "Structure" section has two radio buttons: "SNPs x Samples" (selected) and "Samples x SNPs".

The "Form" section has two radio buttons: "Tab" and "Text" (selected).

The "Batch" section has two checkboxes: "SNP Batch" and "Sample Batch".

Under "SNP Batch", there is a dropdown menu for "Attribute" with "build" selected.

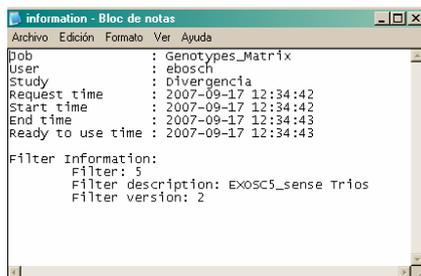
Under "Sample Batch", there is a dropdown menu for "Attribute" with "family" selected.

A "Download Data" button is located at the bottom of the form.

The left sidebar contains a navigation menu with categories: "Study Selection", "Data Management", "Data Retrieval", "Basic Analysis", "Public Downloads", and "Help". The "Data Retrieval" section is expanded, showing "Fields Information", "SNPs/Samples", "Genotypes", "Matrix Format" (highlighted), "List Format", "Formats", and "Backup / XML Download".

The bottom of the browser window shows the Windows taskbar with several open applications, including "Aplicaciones...", "Bandeja de e...", "Extracting_Da...", "CEGEN...", "Bioinformát...", and "Help - Windo...". The system clock shows 12:28.

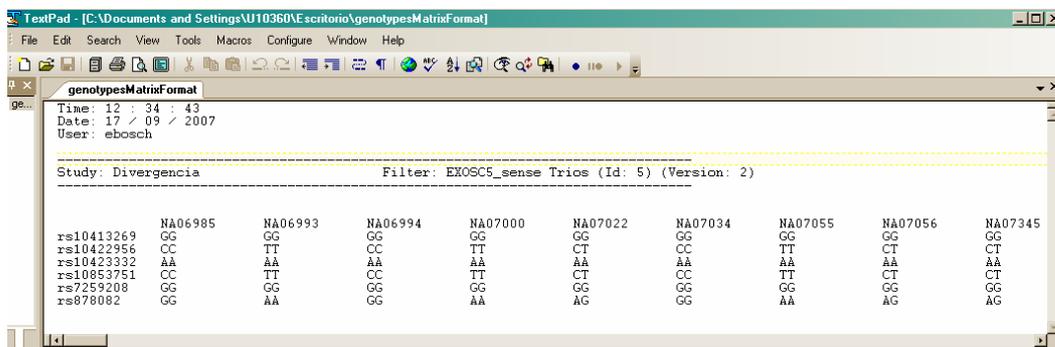
A **Report** file separated by tabs or spaces will be generated depending on the settings of the second option. The **Matrix Format Report** will be sent to the **User Results** section. Download to your computer the **ZIP Matrix Format Report** file or open the zip file directly. This ZIP contains two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



```

information - Bloc de notas
Archivo Edición Formato Ver Ayuda
Job : Genotypes_Matrix
User : ebosch
Study : Divergencia
Request time : 2007-09-17 12:34:42
Start time : 2007-09-17 12:34:42
End time : 2007-09-17 12:34:43
Ready to use time : 2007-09-17 12:34:43

Filter Information:
Filter: 5
Filter description: EXOSC5_sense Trios
Filter version: 2
  
```



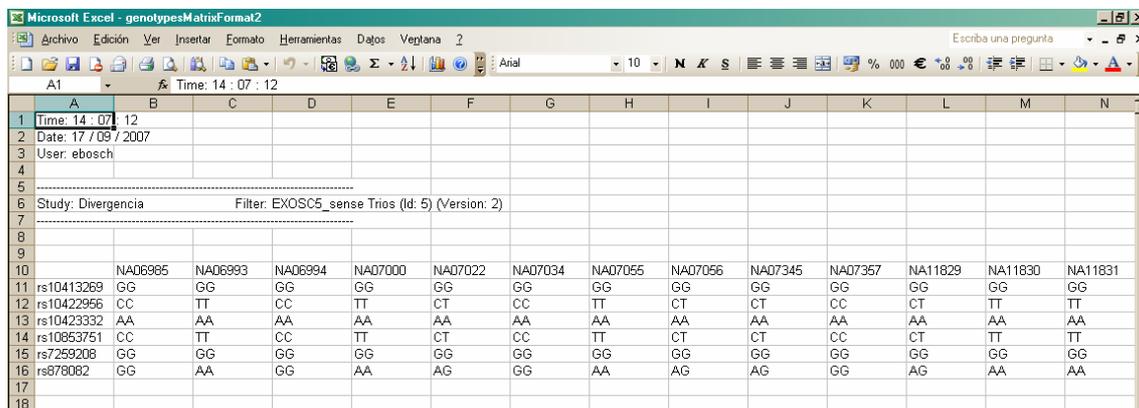
```

TextPad - [C:\Documents and Settings\U10360\Escritorio\genotypesMatrixFormat]
File Edit Search View Tools Macros Configure Window Help

genotypesMatrixFormat
Time: 12 : 34 : 43
Date: 17 / 09 / 2007
User: ebosch

-----
Study: Divergencia Filter: EXOSC5_sense Trios (Id: 5) (Version: 2)
-----

rs10413269 NA06985 NA06993 NA06994 NA07000 NA07022 NA07034 NA07055 NA07056 NA07345
rs10422956 GG GG GG GG GG GG GG GG GG
rs10422956 CC TT CC TT CT CT CC TT CT
rs10423332 AA AA AA AA AA AA AA AA AA
rs10853751 CC TT CC TT CT CT CC TT CT
rs7259208 GG GG GG GG GG GG GG GG GG
rs878082 GG AA GG AA AG GG AA AG GG AG AA AG
  
```



	NA06985	NA06993	NA06994	NA07000	NA07022	NA07034	NA07055	NA07056	NA07345	NA07357	NA11829	NA11830	NA11831
1	rs10413269	GG											
2	rs10422956	CC	TT	CC	TT	CT	CT	CT	CC	CT	CT	TT	TT
3	rs10423332	AA											
4	rs10853751	CC	TT	CC	TT	CT	CT	CT	CC	CT	CT	TT	TT
5	rs7259208	GG											
6	rs878082	GG	AA	GG	AA	AG	GG	AA	AG	AG	GG	AG	AA

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 screenshot shows the SNPator web application interface. The browser title is 'Bioinformática CEGEN - Windows Internet Explorer'. The URL is 'http://www.snpator.org/privat/principal/index.php'. The page header includes 'Bioinformática CEGEN', 'Bioinformatics Division', and 'Centro Nacional de Genotipado'. The main content area is titled 'Data Retrieval / Genotypes / List Format'. It shows a 'Study : Divergencia, Active Filter : 5 (EXOSCS_sense Trios)'. Below this, there are options to 'Order by' (SNP, Sample) and 'Batch' (SNP Batch, Sample Batch). There are also 'View Data' and 'Download Data' buttons. The interface is in Spanish.

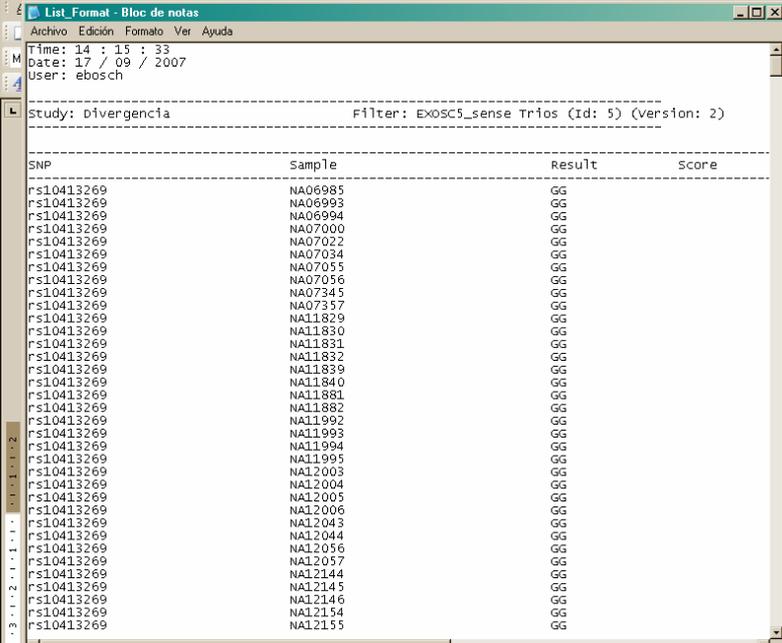
The **View Data** option allows for a quick browsing of the data without generating a **Report**.

Genotypes Study: DIVERGENCIA ordered by Samples

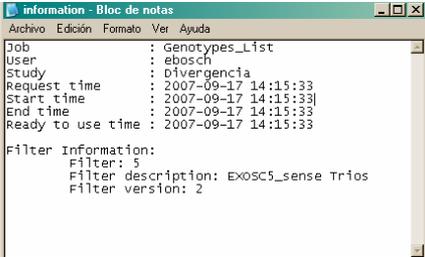
Sample	SNP	Result	Score	Technology
NA06985	rs10413269	GG		
NA06985	rs10422956	CC		
NA06985	rs10423332	AA		
NA06985	rs10853751	CC		
NA06985	rs7259208	GG		
NA06985	rs878082	GG		
NA06993	rs10413269	GG		
NA06993	rs10422956	TT		
NA06993	rs10423332	AA		
NA06993	rs10853751	TT		

Download Data

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:



SNP	Sample	Result	Score
rs10413269	NA06985	GG	
rs10413269	NA06993	GG	
rs10413269	NA06994	GG	
rs10413269	NA07000	GG	
rs10413269	NA07022	GG	
rs10413269	NA07034	GG	
rs10413269	NA07055	GG	
rs10413269	NA07056	GG	
rs10413269	NA07345	GG	
rs10413269	NA07357	GG	
rs10413269	NA11829	GG	
rs10413269	NA11830	GG	
rs10413269	NA11831	GG	
rs10413269	NA11832	GG	
rs10413269	NA11839	GG	
rs10413269	NA11840	GG	
rs10413269	NA11881	GG	
rs10413269	NA11882	GG	
rs10413269	NA11992	GG	
rs10413269	NA11993	GG	
rs10413269	NA11994	GG	
rs10413269	NA11995	GG	
rs10413269	NA12003	GG	
rs10413269	NA12004	GG	
rs10413269	NA12005	GG	
rs10413269	NA12006	GG	
rs10413269	NA12043	GG	
rs10413269	NA12044	GG	
rs10413269	NA12056	GG	
rs10413269	NA12057	GG	
rs10413269	NA12144	GG	
rs10413269	NA12145	GG	
rs10413269	NA12146	GG	
rs10413269	NA12154	GG	
rs10413269	NA12155	GG	



```

Job      : Genotypes_List
User     : ebosch
Study    : Divergencia
Request time : 2007-09-17 14:15:33
Start time  : 2007-09-17 14:15:33
End time    : 2007-09-17 14:15:33
Ready to use time : 2007-09-17 14:15:33

Filter Information:
Filter: 5
Filter description: EXOSC5_sense Trios
Filter version: 2

```

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**.

Basic Analysis / Haplotype Estimate Input / Phase Help

Description

Chromosome Start End

Region

Gene

All

Case Control

Permutations

SNP Batch

Sample Batch

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:

Case / Control

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

Case / Control

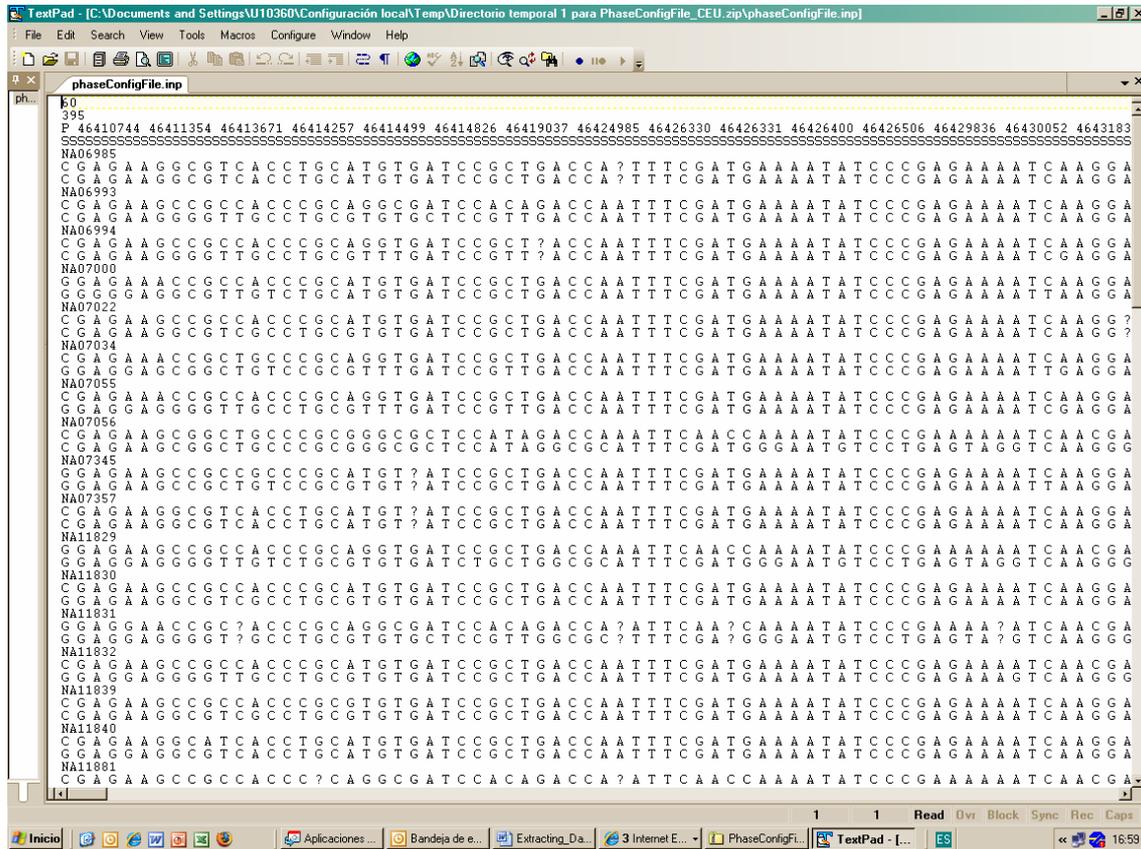
Samples with values other than the ones declared here will be discarded and not used in the process. If a blank " " is selected, that means that there are **Samples** with no value in the field and that you can select those as a case or control.

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

The screenshot shows a Notepad window titled 'SNPs - Bloc de notas'. The window contains a table with the following data:

#	SNP Code	Position	Distance
1	rs8101985	46410744	610
2	rs4803445	46411354	2317
3	rs16975079	46413671	586
4	rs7253952	46414257	242
5	rs1708106	46414499	327
6	rs1654649	46414826	4211
7	rs2271546	46419037	5948
8	rs6508974	46424985	1345
9	rs4802112	46426330	1
10	rs11669668	46426331	69
11	rs4803448	46426400	106
12	rs4803449	46426506	3330
13	rs7246525	46429836	216
14	rs7246896	46430052	1779
15	rs4630671	46431831	904
16	rs4802113	46432735	2966
17	rs7249222	46435701	2887
18	rs4803450	46438588	2005

- 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

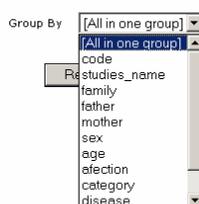
SNPator related information to identify the job: date, time, user, study and filter.

- 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:



```

alequinConfigFile.apr * Document2
# Used SNPs (ordered): rs8101985 rs4803445 rs16975079 rs7253962 rs1709106 rs1654649 rs2271546 rs6508974 rs4802112 rs11669668 rs4803448
# Unused SNPs:
[[Profile]]
Title="Divergencia-HapMap - BCKDHA_allwithoutTrios"
NbSamples=3
GenotypicData=1
GeneticPhase=0
DataType=STANDARD
LocusSeparator=WHITESPACE
MissingData=?
[[Data]]
[[Samples]]
SampleName="ceu"
SampleSize=60
SampleData={
na06985 1 C G A G A A G G C G T C A C C T G C A T ? G T G A T C C G C T G A C C A ? T T T C G A T G A A A A T A T C C C G A G A A
..... etc C G A G A A G G C G T C A C C T G C A T ? G T G A T C C G C T G A C C A ? T T T C G A T G A A A A T A T C C C G A G A A
}
na12892 1 G G A G A A A C G G C T G C C C G C G G ? G C G A T C C A T A G A C C A A A T T C A A C C A A A A T A T C C C G A A A A
G G A G G A G C G G C T G C C C G C G G ? T T G C T C C G T T G G C G C A T T T C G A T G G G A A T G T C T G A G T A
}
SampleName="yri"
SampleSize=60
SampleData={
na18501 1 C G A G A A G C C G C ? G C C C G C A G T G C ? A T C C A C A G A C C A A A T T C A ? C C A A A A T A ? ? C C G A A A A
..... etc C G A G A A G C G G C ? G T C C G C G T T G T ? C T C C G T T G G C G C A T T T C G ? T G G G A A T G ? ? C T G A G T A
}
na19239 1 C G A G G A G C G ? C C G C C C G C G G ? G C ? A T C C A C A G A C C A A A T T C A A C C A A A A T A T C C C G A A A A
G G A G G A G G G ? C T G C C C G C G T ? G T ? C T C C G T T G G C G C A T T T C G A T G G G A A T G T C T G A G T A
}
}
SampleName="chb"
SampleSize=45
SampleData={
na18524 1 C G A G A A G G C G T C A C C T G C A T T G T G A T C C G C T G A C C A A A T T C A A C C A A A A T A T C C C G A A A A
.....etc C G A G A A G G C G T C A C C T G C A T T G T G A T C C G C T G G C G C A T T T C G A T G G G A A T G T C T G A G T A
}
na18637 1 C G A G A A G C C G C C A C C C G C A G T G C G A T C C A C A G A C C A A A T T C A A C C A A A A T A T C C C G A A A A
C G A G A A G G G G T T G C G T T G T G C T C C G T T G G C G C A T T T C G A T G G G A A T G T C T G A G T A
}
}
[[Structure]]
StructureName="Divergencia-HapMap"
NbGroups=1
IndividualLevel=0
Group={
"ceu"
"yri"
"chb"
}
    
```

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.

Data Retrieval / Formats / Haploview Help

Description

Chromosome Start End

Region

Gene

All

User Results

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.

Data Retrieval / Formats / Haploview Help

Description

Chromosome Start End

Region

Gene

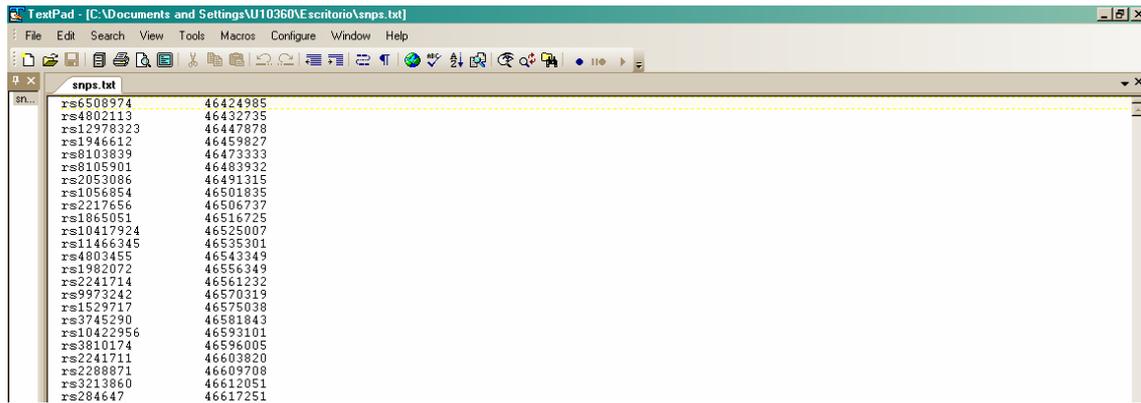
All

User Results

There should be a *.zip file containing:

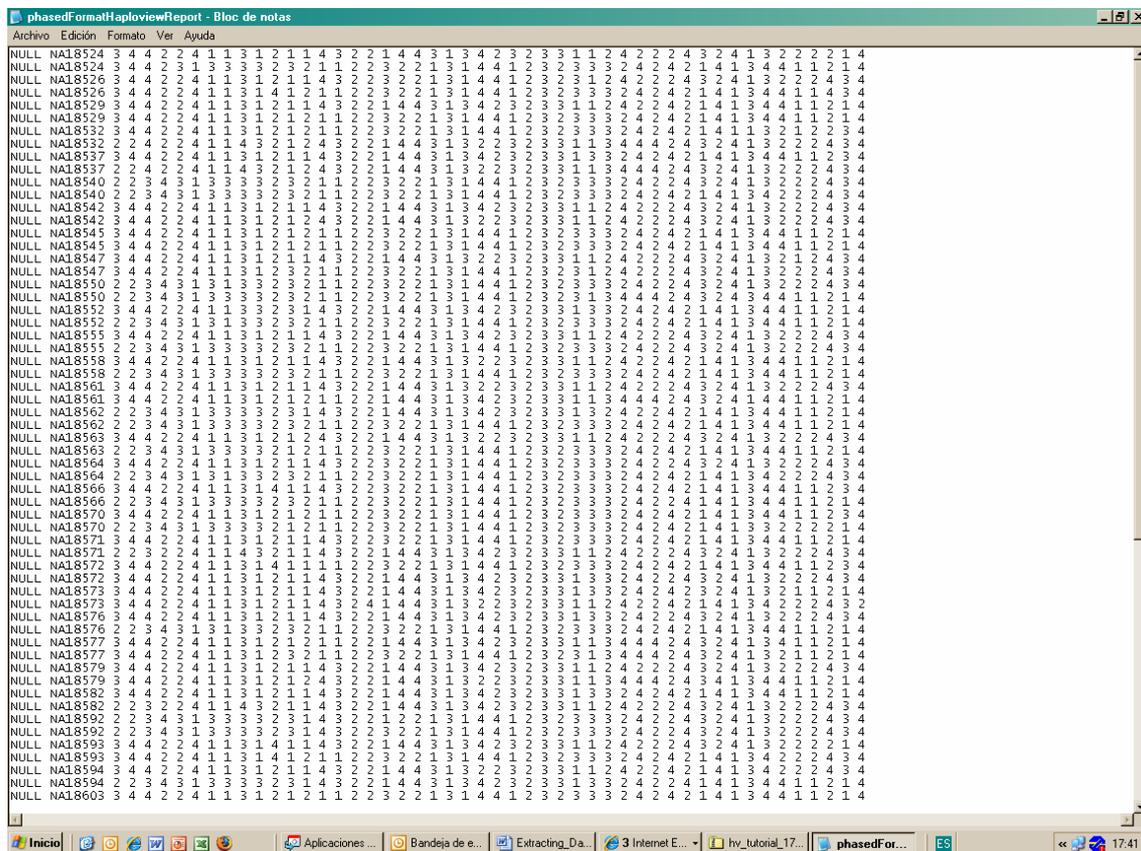
- **snps.txt**

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

