GIbPSs 1.0 Manual Part 14: Documentation of export_gt_13.0.pl

5/22/2015 Andreas Hapke,
Institute of Anthropology, Johannes Gutenberg University Mainz, Germany, ahapke2@gmail.com

This program exports selected genotypes from the database into several popular formats that population genetic and other programs can read: STRUCTURE, Genepop, Arlequin, FASTA, Nexus, PHYLIP, binary FASTA, binary Nexus, and binary PHYLIP. You can export genotypes comprising the alleles identified by **indloc** and **poploc** or, alternatively export “unlinked SNPs”. With the latter option, **export_gt** defines alleles based on the first variable position in each locus and ignores other SNPs in the locus. Variable means variable within your currently selected subset of data. You can include all unlinked SNP loci or restrict the export to biallelic unlinked SNPs. With the latter option, you can export genotypes as binary data in FASTA and Nexus format. You can export binary data as two sequences per individual with 0 and 1 for the two alleles or export integer data in one sequence per individual with 0 and 2 for the two homozygous genotypes and 1 for heterozygotes. The Nexus file in the latter format can, for example, be analyzed by the SNAPP package implemented in BEAST2.

Usage

The program must reside within the main database directory together with the other programs of the GIbPSs package. Before you export genotypes with **export_gt**, make a selection of data with **data_selector**. Your selection must fulfill one condition: It may not contain any genotype with more than two alleles. Use this option in **data_selector** to meet the condition: `2 select loci: 7 Numb of alleles in allele-richest ind.: min: 1 max: 2`. Before and/or after you use this option, you may use any further selection criteria offered by **data_selector**. Note that **data_selector** always analyzes alleles, as they are stored in the database, i.e. considering all known SNPs in a locus. That means it treats a genotype like AC/AT/GT as comprising three alleles. **export_gt** in turn would treat this genotype as comprising two alleles when you export unlinked SNPs.

Call **export_gt** as follows to export genotypes in STRUCTURE, Genepop, Arlequin, FASTA, Nexus, and PHYLIP format with default settings:

```
export_gt_13.0.pl
```
or
```
perl export_gt_13.0.pl
```

The following command flags enable you to modify the output formats and to inactivate production of specific outfiles. Defaults are given in parentheses.

Command flags

```
-uS (0)    0: export alleles as identified by indloc and poploc
           2: export biallelic unlinked SNPs
           3: export unlinked SNPs including those with more than 2 alleles
-S  (1)    produce outfile for STRUCTURE, 0: don’t
-Sc ()     name of a user file to modify the STRUCTURE outfile
-Sm (-9)   value for a missing data point in the STRUCTURE outfile
-G  (1)    produce outfile for Genepop, 0: don’t
-Gp ()     name of a user file to modify the Genepop outfile
-A  (1)    produce outfile for Arlequin, 0: don’t
-Ap ()     name of a user file to modify the Arlequin outfile
-Fv (1)    export genotypes in FASTA format
-Fmis (N)  character for missing data in FASTA file
-Nxv (1)   export genotypes in Nexus format
-Nxgap (-) gap character for Nexus.(Missing data are encoded as gaps.)
-Pv (1)    export genotypes in PHYLIP format
-Pmis (N)  character for missing data in PHYLIP file
-bin (0)   0: DNA characters in FASTA, Nexus, and PHYLIP, 2 seqs per ind
           1: binary data recoded as integer data: one seq per ind with 012
           2: binary data: two seqs per ind with 01
           with -bin 1 or -bin 2, program automatically activates -uS 2
-indf ()   name of a user file to add information to individual IDs
           in Genepop, FASTA, Nexus, and PHYLIP outfiles
-rs ()     random number seed
```

Each command flag must be followed by one space and an appropriate value, example:

```
export_gt_13.0.pl -Sc user_files/structure_format.txt -G 0 -A 0
```

The program searches for an existing selection of data in directory `export` in the main database directory and reads it in if it finds one. It produces outfiles in directory `export`. With default settings, they have these names: `export_gt_settings.txt`, `in_structure.txt`, `in_genepop.txt`, `in_arlequin.arp`, `varpos.txt`, `varfas.fas`, `varnx.nex`, `varph.phy`. When you export unlinked SNPs including those with more than two alleles (`–uS 3`), outfilenames start with `uS_`, e.g. `uS_varnx.nex`. When you export biallelic unlinked SNPs (`–uS 2`), outfilenames start with `uSb_`, e.g. `uSb_varnx.nex`. When you export binary data as two sequences of 0 and 1 for each individual (–`bin 2`), the last four outfiles have these names: `uS_bin2_varpos.txt`, `uSbin2_varfas.fas`, `uSbin2_varnx.nex`, `uSbin2_varph.phy`. When you use -bin 1, their names start with `uSbin1_`.

The program overwrites these files when they already exist.

Your selection may contain invariable loci. When you export unlinked SNPs (`–uS 2 or –uS 3`) or when the production of FASTA, Nexus or Phylip outfiles is activated, **export_gt** gives a message and excludes all invariable loci.

STRUCTURE, Genepop, and Arlequin files with -uS 0, -uS 2, and -uS 3

The program encodes alleles as integers in the outfiles in STRUCTURE, Genepop, and Arlequin format. The following example explains the meaning of these numbers: A locus has three SNPs and five known alleles. The characters of each allele at the three SNPs designate an allele (popallvar):

popall_ID	popallvar	
1022	GCA	not selected
1023	ACG	selected
1024	ACA	selected
1025	ATA	selected
1026	ATG	selected

You have filtered your data with **data_selector**. Only the last four alleles occur in the selected subset. An individual has genotype `ACG ATA`. Per default (`-uS 0`), **export_gt** looks up all selected alleles (popall_IDs) in the database and renumbers them separately for each locus in ascending popall_ID order starting from 1. Accordingly, our individual has genotype `1 3`. When you export unlinked SNPs (`-uS 2` or `-uS 3`), the program identifies the first SNP that is variable within the selected alleles and encodes characters as follows: A=1, C=2, G=3, T=4. The second SNP in our example is variable and our individual has genotype `C T`, which evaluates to `2 4`. Some SNPs may have more than two alleles. With `-uS 3`, they are included in the output. With `-uS 2`, they are excluded. The outfile `varpos.txt` contains a list of the finally included loci.

Important note: When you activate the export of binary data `(-bin 1)`, **export_gt** automatically activates `-uS 2`. All outfiles then contain biallelic unlinked SNPs, and loci with more than two alleles are excluded. Apart from that, the numbers in the STRUCTURE, Genepop, and Arlequin files do not change. The program exports binary data in FASTA, Nexus and PHYLIP format only.

Format of the outfile for STRUCTURE (-S)

With default settings, the program produces an outfile for STRUCTURE in the following format. The example contains data from three individuals at four loci:

```
ind01 1      1      2      1
ind01 2      2      3      1
ind02 1      2      2      1
ind02 1      2      2      1
ind03 3      2      -9     1
ind03 3      2      -9     1
```

There are two lines for each individual and one column for each locus. Columns are separated by tabs and each line ends with a line ending. The first column contains individual IDs in ascending order. The following columns contain the genotypes in ascending order of the selected poplocIDs. Missing genotypes are encoded with `-9` as for individual ind03 at the second locus. You can change the value for missing genotypes with the command flag `-Sm`. You can add additional columns between the individual ID column and the genotype columns with the aid of a user file. Provide the name or path of this file with the command flag `-Sc`.

Format of the user file for STRUCTURE (-Sc):

STRUCTURE can use additional information that you can provide in a specific format and order in additional columns before the genotype columns. Please refer to the documentation of STRUCTURE for more information. You can provide the name/path of a user file with the flag `-Sc`. The following example user file will instruct **export_gt** to add five additional columns:

```
ind03 3      1      2      2      B
ind01 1      1      1      1      A
ind02 2      0      1      2      A
```

The file is a text file with one line per individual. Columns are separated by tabs. Each line must contain the same number of columns. You should avoid empty cells. Use an appropriate text editor that shows all non-printable characters to verify that. The first column must contain the individual IDs. Individuals may appear in any order. In the outfile, they will appear in the same order as in your user file. This can be useful when you wish to have your individuals in a specific order in outfiles of STRUCTURE. All selected individuals in the database must appear in the file and all individuals in the file must be in the current selection. You can use a copy of file `export/sel_ind.txt` to prepare this file. The example user file will instruct **export_gt** to produce an outfile in this format:

```
ind03 3      1      2      2      B      3      2      -9     1
ind03 3      1      2      2      B      3      2      -9     1
ind01 1      1      1      1      A      1      1      2      1
ind01 1      1      1      1      A      2      2      3      1
ind02 2      0      1      2      A      1      2      2      1
ind02 2      0      1      2      A      1      2      2      1
```

Format of the outfile in Genepop format (-G)

This file is in Genepop-2-digit format. With default settings, **export_gt** will print the genotype data above as follows:

```
Title
1230
1456
1512
1732
Pop
ind01 , 0102 0102 0203 0101
ind02 , 0101 0202 0202 0101
ind03 , 0303 0202 0000 0101
```

The first line contains the word “Title”. The next lines contain the poplocIDs of the selected loci in ascending order. The keyword “Pop” indicates that the following lines contain genotype data from one population. The next line with a keyword Pop starts the next population. Per default, **export_gt** assigns all individuals to one population. You can assign your individuals to several different populations with the aid of a user file (see below). Genotype lines: There is one line per individual, which contains the individual ID, a space, a comma, a space and then the genotypes at all loci in ascending order of poplocIDs and separated by spaces. Four digits designate a genotype: two for the first and two for the second allele. 0000 designates a missing genotype. Each line ends with a line ending.

This format has one restriction: The number of alleles at a given locus may not be greater than 99. The program produces a message and inactivates the production of this format when any locus has more than 99 alleles. You can unselect these loci with the aid of **data_selector** (`Select loci: Number of alleles`).

Format of the user file for Genepop (-Gp)

You can provide the name/path of a user file for Genepop with the flag `-Gp`. This file must be a text file with a tab-delimited table with two columns and one line per individual. Column 1 contains the individual ID, column 2 contains an integer that designates a population. Individuals may appear in any order. The following user file will instruct **export_gt** to assign your individuals to two populations:

```
ind01 1
ind02 2
ind03 1
```

The resulting outfile in Genepop format will look like this:

```
Title
1230
1456
1512
1732
Pop
ind01 , 0102 0102 0203 0101
ind03 , 0303 0202 0000 0101
Pop
ind02 , 0101 0202 0202 0101
```

The populations appear in ascending order according to your user file. The individuals appear in ascending ID order.

Format of the outfile for Arlequin (-A)

With default settings, export_tables produces an outfile for Arlequin in the following format:

```
[Profile]
Title="Title"

NbSamples=1
DataType=STANDARD
GenotypicData=1
LocusSeparator=WHITESPACE
GameticPhase=0
RecessiveData=0
MissingData="?"

[Data]
[[Samples]]
SampleName="Sample1"
SampleSize=3
SampleData= {
ind01 1  01 01 02 01
         02 02 03 01
ind02 1  01 02 02 01
         01 02 02 01
ind03 1  03 02  ? 01
         03 02  ? 01
}
[[Structure]]

StructureName="One group for all samples"
NbGroups=1
Group={
"Sample1"
}
```

Please refer to the manual of Arlequin for details about the format. The data section contains the genotypes. Per default, **export_tables** assigns all individuals to one sample with name “Sample1”. You can assign your individuals to several different populations with the aid of a user file (see below). The genotypes are encoded in two digit format with two lines per individual and spaces as separators between loci. A “?” designates missing data. The first genotype line of an individual contains the individual ID followed by value 1, which means that this genotype occurs once in the sample. At the end of the file, the “Structure” section describes a simple structure where all samples belong to one group of samples. You can easily edit this structure to assign your samples to different groups when you have several samples.

This format has one restriction because **export_gt** uses two digits two encode an allele: The number of alleles at a given locus may not be greater than 99. The program produces a message and inactivates the production of this format when any locus has more than 99 alleles. You can unselect these loci with the aid of **data_selector** (`Select loci: Number of alleles`).

Format of the user file for Arlequin (-Ap)

You can provide the name/path of a user file for Arlequin with the flag `-Ap`. This file must be a text file with a tab-delimited table with two columns and one line per individual. Column 1 contains the individual ID. Column 2 contains the ID of a population. Individuals may appear in any order. The following user file will instruct export_gt to assign your individuals to two populations with names "Ana7" and "Ber45":

```
ind01 Ana7
ind02 Ber45
ind03 Ana7
```

The outfile in Arlequin format will look like this:

```
[Profile]
Title="Title"

NbSamples=2
DataType=STANDARD
GenotypicData=1
LocusSeparator=WHITESPACE
GameticPhase=0
RecessiveData=0
MissingData="?"

[Data]
[[Samples]]
SampleName="Ana7"
SampleSize=3
SampleData= {
ind01 1  01 01 02 01
         02 02 03 01
ind03 1  03 02  ? 01
         03 02  ? 01
}
SampleName="Ber45"
SampleSize=3
SampleData= {
ind02 1  01 02 02 01
         01 02 02 01
}


[[Structure]]

StructureName="One group for all samples"
NbGroups=1
Group={
"Ana7"
"Ber45"
}
```

Genotypes in FASTA, Nexus and PHYLIP format

Per default, each of these files contains two sequences per individual. Per default (`-uS 0`), the two sequences contain the characters at variable positions of the two alleles of each locus. Example: an individual has genotypes ACA/ACT and GGA/GGA at the first and second locus. The two sequences start as follows:

```
ACAGGA
ACTGGA
```

The outfile `varpos.txt` contains the start and end positions of each locus. When you export unlinked SNPs (`-uS 2` or `-uS 3`), there is only one position per locus in the FASTA file. Loci are ordered in ascending poplocID order. When you export binary data with `-bin 2`, alleles are encoded as `0` or `1` with `1` designating the more frequent allele.

In all of these cases, **export_gt** randomly assigns the two alleles of a locus to the first or second sequence. The random phasing is identical in all three files in FASTA, Nexus, and PHYLIP format. If you wish to use the program a second time and to get the same sequences, you must use flag `-rs` and provide the random number seed that the program used the first time. You find it in file `export_gt_settings.txt`.

When you export integer data with `-bin 1`, the two sequences of an individual are collapsed into one sequence. Genotypes are encoded as `0` (homozygote for allele A), `1` (heterozygote) and `2` (homozygous for allele B).

Genotypes in FASTA format (-Fv, -Fmis, -bin)

Use `-Fv 1` to activate this format. Per default, the program uses `N` for missing data. Use `-Fmis` to select another character.

Example of a FASTA file with DNA data: (`-Fv 1 -Fmis N -bin 0`)

```
>ind01_1
TCAAGCCCCC
>ind01_2
CCAGACTCTC
>ind02_1
TTAAGACCNC
>ind02_2
TCGGGCCTNG
```

Example of a FASTA file with binary data (`-Fv 1 -Fmis - -bin 2`):

```
>ind01_1
1111111101
>ind01_2
0110010111
>ind02_1
10111011-1
>ind02_2
11001110-0
```

The same data with these settings: (`-Fv 1 -Fmis - -bin 1`):

```
>ind01
1221121212
>ind02
21112121-1
```

Genotypes in Nexus format (-Nxv, -Nxgap, -bin)

Use `-Nxv 1` to activate this format. The program encodes missing data as gaps. (This seems to be helpful for importing the data into BEAUti for an analysis with SNAPP). Per default, it uses "-" as gap character. You can select another character with flag `-Nxgap`. You can also easily change the format section of the file and replace "`gap`" with "`missing`" to encode missing data as such.

Example of a Nexus file with DNA data (`-Nxv 1 -Nxgap - -bin 0`):

```
#NEXUS

Begin data;
      Dimensions ntax=4 nchar=10;
      Format datatype=dna gap=-;
      Matrix
ind01_1 CCAAGCCCCC
ind01_2 CCAGACTCTC
ind02_1 TTAAGACC-C
ind02_2 TCGGGCCT-G
      ;
End;
```

Example of a Nexus file with binary data (`-Nxv 1 -Nxgap - -bin 2`):

```
#NEXUS

Begin data;
      Dimensions ntax=4 nchar=10;
      Format datatype=binary symbols="01" gap=-;
      Matrix
ind01_1 0111111101
ind01_2 0110010111
ind02_1 10111011-1
ind02_2 11001110-0
      ;
End;
```

The same data with these settings: (`-Nxv 1 -Nxgap - -bin 1`):

```
#NEXUS

Begin data;
      Dimensions ntax=2 nchar=10;
      Format datatype=integerdata symbols="012" gap=-;
      Matrix
ind01 0221121212
ind02 21112121-1
      ;
End;
```

Genotypes in PHYLIP format (-Pv, -Pmis, -bin)

Use `-Pv 1` to activate this format. Per default, the program uses `N` for missing data. Use `-Pmis` to select another character. Usually, individual IDs may not have more than 10 characters in this format. **export_gt** pads individual IDs with spaces to a length of 10 or to the length of the longest occurring ID if it is longer than 10 characters. The program does not truncate IDs. Your IDs may be longer than 10 characters, particularly when you use flag `-indf` to add information to IDs. Avoid long IDs when you plan to use a program that requires IDs with a maximum length of 10. Following to the ID, there are three spaces, then the sequence without any spaces. The first line contains the number of sequences and the number of characters in each sequence.

Example of a PHYLIP file with DNA data: (`-Pv 1 -Pmis N -bin 0`)

```
 4 10
ind01_1      TCAAGCCCCC
ind01_2      CCAGACTCTC
ind10102_1   TTAAGACCNC
ind10102_2   TCGGGCCTNG
```

Example of a PHYLIP file with binary data (`-Pv 1 -Pmis - -bin 2`):

```
 4 10
ind01_1      1111111101
ind01_2      0110010111
ind10102_1   10111011-1
ind10102_2   11001110-0
```

The same data with these settings: (`-Pv 1 -Pmis - -bin 1`):

```
 2 10
ind01        1221121212
ind10102     21112121-1
```

Flag -indf

This flag is useful when you wish to add a group ID or something else to the individual IDs in the Genepop, FASTA, Nexus, and PHYLIP outfiles. This can be helpful if you wish to use BEAST. BEAUti can then extract species or group assignments out of your Nexus file.

Provide the name (or path) of a file with this format: one line per individual, two entries separated by TAB: individual ID (as in the database), one string. The string contains your group ID and may not contain whitespace. Example:

```
ind01 group1
ind02 group2
```

The program appends the string to the individual ID with an underscore as separator:

Genepop:

```
ind01_group1 , 0102 0102 0203 0101
ind02_group2 , 0101 0202 0202 0101
```

FASTA: `>ind01_group1_1`

Nexus/PHYLIP: `ind01_group1_1`

Flag -rs

Use this flag when you wish to repeat a data export with identical phasing of alleles in FASTA, Nexus, and PHYLIP files. Lookup the previously used random seed in file `export_gt_settings.txt` and provide it with flag `-rs`.

Infiles

The program reads the following infiles. Some of them are optional. The program produces an error message and stops execution when it cannot open a file that it needs.

Infile	produced by
user file for STRUCTURE format	user
user file for Genepop format	user
user file for Arlequin format	user
user file for FASTA, Nexus, PHYLIP formats	user
`individuals.txt`	indloc
`popall.txt`	poploc
`*_indpopall.txt`	indpoploc
`*_indpoploc.txt`	indpoploc
`export/sel_ind.txt`	data_selector
`export/sel_loc.txt`	data_selector
`export/sel_gt.txt`	data_selector

*_ stands for an individual ID