GlbPSs 1.0 Manual Part 13: Documentation of export_tables_07.0.pl

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

This program exports data from the database into several tab-delimited text files. If you have selected a subset of data with **data_selector**, **export_tables** exports the selected data. If not, it exports all data from the database. The program does not only export data from the database. It also re-analyses the data included in the selected subset before exporting them.

The program produces two files with genotypes and the sequencing depths of the alleles in these genotypes. Two additional files contain data about the loci in the exported dataset, such as a consensus sequence, the number of alleles, variable positions etc. One of these files describes properties of loci, as they are stored in the database. For example, the number of alleles is the number of all known alleles. The second file describes properties of loci WITHIN THE CURRENTLY SELECTED SUBSET OF DATA. For example, the number of alleles includes only alleles included in the selection, and the consensus sequence is the consensus of these alleles.

Usage

To start the program, enter

```
export_tables_07.0.pl
```
or
```
perl export_tables_07.0.pl
```

The program loads data, checks if a selection of data exists, if so, loads it and starts analyzing data and producing outfiles.

Outfiles

The program produces four outfiles in a directory named `export` within the main database directory. It creates this directory if it does not exist. The names of the outfiles are `genotypes.txt`, `reads.txt`, `locdata.txt`, `locdatasel.txt`. The program overwrites these files if they already exist.

`genotypes.txt`

Format: Table in a tab-delimited text file, first column: locus ID (poplocID), next columns: one per individual, first line: headers, next lines: one per locus. The line for a locus contains the poplocID followed by the genotypes of all individuals. Example:

```
poplocID    ind_01      ind_02      ind_03
1           consensus   consensus   -999
2           TCG         TCG/TTG     TCG/TTG/TTA
```

Locus 1: `consensus` means that only one allele is known from this locus in the whole database. This allele has been observed in individuals 01 and 02 but not in individual 03. The missing genotype is designated by `-999`.

Locus 2 has three variable positions. Alleles are designated by the characters at these variable positions. Different alleles in a genotype are separated by slashes. Individual 01 is homozygous. Individual 02 is heterozygous. Three alleles have been identified in individual 03.

Variable positions here means variable with respect to all known alleles of a locus in the database. Within our currently selected subset of data, only two positions are variable in locus 2. Nevertheless,

each allele is designated by three characters because the alignment of all known alleles of locus 2 in the database has three variable positions.

`reads.txt`

Format: Table in a tab-delimited text file, first column: locus ID (poplocID), next columns: one per individual, first line: headers, next lines: one per locus. The line for a locus contains the poplocID followed by the read depths of the alleles of each individual. The order of loci, individuals, and alleles corresponds to the order in file `genotypes.txt`. Example:

```
poplocID    ind_01      ind_02      ind_03
1           17          21          -999
2           19          14/7        8/12/3
```

As in file `genotypes.txt`, missing genotypes are designated by `-999` and alleles within a genotype are separated by slashes. Read depths are the depths of the alleles as determined by **indloc**. They are not always integers because **indloc** adds fractions of depths from certain discarded reads to certain retained reads. See the documentation of **indloc** for more details.

`locdatasel.txt`

This file contains data about the exported loci. It describes properties of the loci within the currently exported subset of data.

Format: Table in tab-delimited text file, first line: headers, next lines: one per locus:

poplocID: locus ID

sl: sequence length

cons: consensus sequence of the exported alleles of the locus with IUPAC ambiguity symbols at variable positions

nSNP: number of variable positions

varpos: variable positions, Example: `p:4/15` Positions 4 and 15 are variable.

n_all: number of alleles

n_Ind: number of individuals in which the locus has been genotyped by **indloc**

nIndloc: number of individual loci that have been assigned to this locus by **poploc**

ties: 0 if *n_Ind* = *nIndloc*. 1 if *n_Indloc* > *nInd.* In the latter case, several individual loci from the same individual have been assigned to this locus (tied) by **poploc**.

count_maj: count of the most frequent allele of this locus

The program re-analyzes the genetic data in the database before it produces this file. Example: The database contains five alleles of a locus. You selected a subset of data that excludes all genotypes with a depth of the minor allele below 6. The program reanalyzes the selected data and finds only three alleles. It aligns the alleles, determines variable positions, determines a new consensus sequence etc.

`locdata.txt`

This file has the same format and contains the same variables as `locdatasel.txt` except *count_maj*. The data in the file are valid with respect to the complete database. Examples: The consensus sequence of a locus is the consensus of all known alleles, even if not all of them are included in the exported dataset. The same holds true for the number of SNPs, variable positions etc.