GIbPSs 1.0 Manual Part 8: Documentation of data_selector_16.0.pl

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

This interactive tool enables you to select subsets of data from the database. You can select data according to different criteria at three levels: individuals, loci, and genotypes. A genotype here means the alleles of a given individual at a given locus. After having selected a subset of data with **data_selector**, you can export it from the database with the programs **export_tables**, **export_gt**, and **export_seq** or analyze it with the programs **pair_finder**, **indel_checker**, and **depth_analyzer**. Moreover, you can create a subset of your subset of data by using **data_selector** again.

Usage

To start the program, enter

```
data_selector_16.0.pl
```
or
```
perl data_selector_16.0.pl
```

The program loads data, checks if a previous selection exists, if so, loads it and displays the main menu:

```
DATA SELECTION START MENU

1 select individuals
2 select loci
3 select genotypes
4 quit

Please select an option (number):
```

Within one session, you can select data at one of the three levels – individuals, loci, genotypes. In order to select data at several levels, you must run **data_selector** several times. When you run **data_selector** the first time, it applies your selection criteria to the whole database. When you run the program again, it loads your previous selection, reanalyzes the genetic data within your previously selected subset of data, and then applies your new selection criteria to this subset. Re-analysis means that the program updates certain properties of loci based on your previous selection, e.g. the number of alleles, number of SNPs, number of individuals with existing data.

A simple example: Some loci in your database are monomorphic, others are variable, i.e. have at least one SNP / at least two alleles. You wish to export genotypes from a subset of individuals and you wish to include only those loci that are variable WITHIN THIS SUBSET of individuals. In order to do so, you run data_selector a first time and select your individuals. Then you run the program again and select variable loci (several options to do so, see below). Finally, you can export the selected data with e.g. **export_gt** or analyze the selected subset of data with another program.

You can run **data_selector** as often as you wish and apply most selection criteria in any order to define subsets of data with specific properties. Moreover, you can combine **data_selector** with other programs. For example, you could

run **data_selector** to select loci with a minimum length of 32,

run **indel_checker** to identify loci that could be indel variants of each other,

run **depth_analyzer** to analyze the depth of loci and clusters of loci found by **indel_checker**,

run **data_selector** to exclude loci that are indel variants of other loci,

run **pair_finder** to identify pairs of loci,

run **data_selector** to select one locus of each pair,

run **data_selector** to select loci based on the results from **depth_analyzer** and further criteria.

run **data_selector** to select genotypes according to various criteria...

Outfiles

data_selector creates four outfiles in a directory `export` within the main database directory. The program creates this directory if it does not exist.

The following files contain your selection of individuals, loci and genotypes: `sel_ind.txt`, `sel_loc.txt`, `sel_gt.txt`. Please do not modify these files. If you delete one of them, delete them all (and delete the fourth file, `sel_rep.txt`). These three files represent the current selection of data, which will be used by **data_selector**, **export_tables**, **export_gt**, **export_seq**, **pair_finder**, **indel_checker** and **depth_analyzer**. Delete or rename the directory `export` when you wish to restart data selection or analyses from the whole database.

The fourth file, `sel_rept.txt`, contains a report about the selections made up to now, the original dimensions of the dataset and its dimensions after each selection step. Each time you use **data_selector**, it appends information to this file. You could store a copy of this file together with an exported dataset to document how you selected the data.

Select individuals

When you select "`1 select individuals`" in the main menu, the program asks you for the name of a text file with a list of selected individuals:

```
SELECT INDIVIDUALS:

I need a list of selected individuals in a text file.
Format: one individual ID per line, no spaces, no blank lines.

Please enter filename:
```

You can store this text file in the main database folder and simply enter its name. You can create a subdirectory in the main database directory store it therein and enter a path, e.g. `user_files/my_individuals.txt`. Finally, you can store it elsewhere and enter the full path to the file.

Select loci

When you select “2 select loci” in the main menu, the program displays this menu:

```
SELECT LOCI:

1  Paired loci: select one of each pair        off
2  List of loci in a file                      off
3  Sequence length                             off
4  Number of individuals                       off
5  Number of alleles                           off
6  Number of SNPs                              off
7  Numb. of alleles in allele-richest ind.     off
8  Ties                                        off
9  Split loci                                  off
10 Selection based on depth analysis           off
11 Merging conflicts                           off
12 Used fraction of read depth                 off

13 Run
14 Quit

Please select a number to activate/inactivate an option:
```

You can select one or several of these options. Before you can use options 1 and 10, you must analyze your database with **pair_finder** and **depth_analyzer**. Once you have selected an option and defined a criterion, the menu displays your criteria, e.g.:

```
6  Number of SNPs                              on min: 1 max: 3
```

When you have finished the definition of criteria, enter `13` to apply them.

Select loci: options in detail:

1 Paired loci: select one of each pair

This option is useful when you have paired reads from a GBS library and used the *rc-duplicate* strategy for *rc-locus-pairs*. Before you can use it, you must identify pairs of loci with **pair_finder**. **data_selector** reads in the output from **pair_finder** in directory `pairs` and selects one locus of each pair. If you activate this option, **data_selector** automatically activates option `9 no split loci` as well because **pair_finder** ignores split loci.

If you analyzed all loci in the database with **pair_finder**, and if nothing went wrong with your analysis of *rc-duplicates*, each group of loci identified by **pair_finder** will contain two loci. Under the recommended workflow of GIbPSs, you use **indel_checker** and **data_selector** to remove loci that could be indel variants of each other before you run **pair_finder**. It can then happen that some groups found by **pair_finder** contain only one locus. You should remove these single loci. **data_selector** will automatically do that for you when you now select one locus of each pair. It selects one locus of each group with two loci and unselects all other loci.

2 List of loci in a file

This option enables you to define a list of loci to be selected. Prepare a text file that contains one locus ID per line, no spaces and no blank lines. **data_selector** will ask you for the name (or full path) of this file. Use this option for example to select loci that are not indel variants of each other. The program **indel_checker** produces an appropriate list of these loci in file `indelcheck/no_indel_loc.txt`.

3 Sequence length

You can select loci in a given range of sequence lengths. The program will ask you for a minimum and maximum.

4 Number of individuals

It happens that a given locus is detected and genotyped in a part of your individuals but not in all individuals. Here, you can select loci with available data from certain numbers of individuals. **data_selector** will ask you for a minimum and maximum.

5 Number of alleles

You can select loci with certain numbers of alleles. The program will ask you for a minimum and maximum.

6 Number of SNPs

You can select loci with certain numbers of SNPs. The program will ask you for a minimum and maximum.

7 Numb. of alleles in allele-richest ind.

It may happen that **indloc** identifies more than two alleles of a given locus in a given individual. This could happen because of high sequencing error or assignment of paralogous reads to the same locus. This option is useful to select only those loci that have a maximum of two alleles in each individual. The program will ask you for a minimum and maximum for the number of alleles in the individual with the greatest number of alleles of a locus (allele-richest ind.).

8 Ties

A tie means that **poploc** assigned two or more individual loci from one individual to the same population locus. Genotypes of such loci may be incorrect, because **indloc** corrected errors, identified alleles and determined their sequences without being able to compare all available reads from the locus. This option enables you to select only loci without ties (in any individual) or only loci with ties.

9 Split loci

It can happen that **poploc** assigns different alleles of one individual locus to different population loci, i.e. splits up an individual locus. This option enables you to exclude population loci containing alleles from split individual loci (no split loci) or to select only such loci (split loci only).

10 Selection based on depth analysis

Before you can use this option, you must analyze the sequencing depth of the loci in the database with **depth_analyzer**. **depth_analyzer** analyzes the depth of each locus in each individual and determines a scaled depth and a depth percentile. It then calculates the minimum, median and maximum of these variables across all individuals (with available data) for each locus. See the documentation of **depth_analyzer** for more details. Selection based on depth-analysis is useful when you wish to exclude loci with extreme sequencing depth compared to most loci. Extreme depth can indicate that a locus is a repetitive element.

When you activate selection based on depth-analysis, a dialog starts that asks you to define a sequence length range. Next, you can select one of these criteria:

`1 min_scaldep max:`	minimum of scaled depth across individuals
`2 med_scaldep max:`	median of scaled depth across individuals
`3 max_scaldep max:`	maximum of scaled depth across individuals
`4 min_dep_perc min:`	minimum of depth percentile across individuals
`5 med_dep_perc min:`	median of depth percentile across individuals
`6 max_dep_perc min:`	maximum of depth percentile across individuals

The program asks you for a maximum for options 1-3 and for a minimum for options 4-6.

Next, the program asks you if you wish to define another sequence length range. You can define several non-overlapping ranges and you can select different criteria/values for different ranges.

depth_analyzer does not analyze the depth of split loci. Split loci will remain selected when you select loci based on a depth analysis. Activate “Split loci: no split loci” to exclude them.

11 Merging conflicts

poploc merges overlapping sequence fragments in those loci that **indloc** identified based on concatenated sequences from f and r reads. It merges all alleles of a locus with the longest overlap that is possible (and allowed by your settings) in all alleles. A merging conflict means that the longest possible overlap differs between the alleles of a locus. This can happen when a locus contains repeat motifs that vary slightly between alleles. It can also be due to genotyping errors. Finally, it can happen by chance when you used a very small minimum overlap for merging. This option enables you to select loci without merging conflicts or to select loci with merging conflicts. The latter option is useful when you wish to have a closer look at the loci with merging conflicts, which could help you to define better settings for merging with **poploc**, e.g. an appropriate minimum overlap.

12 Used fraction of read depth

The depth of a genotype is the sum of depths of its alleles as determined by **indloc**. It can include the depth of certain discarded reads. The total read depth of a genotype is the total number of reads that **indloc** has assigned to a locus and includes all retained and discarded reads. Used fraction of read depth means depth of genotype / total read depth of genotype. The program asks you for a minimum, which must be a decimal between 0 and 1. It determines the used fraction of read depth for each selected locus in each selected individual. The program excludes a locus when the used fraction of read depth is below your minimum in at least one individual.

It can happen that **indloc** calls several SNPs at an individual locus, which are not in phase and define more than two, possibly multiple alleles. It is further possible that **indloc** finally accepts only one or two of them, which represent only a small fraction of all reads. (Under the binomial likelihood ratio method for SNP calling, **indloc** accepts the two most frequent alleles. Under the frequency method, you can apply a frequency filter (-a) in **indloc**, which accepts only alleles with a certain fraction of read depth.) This problem will arise when you have extremely high sequencing error at a locus or when the sequences assigned to it are paralogs. Use filter `Select loci: 12 Used fraction of read depth` to ensure that **indloc** assigned a high fraction (say 0.9) of all reads to the finally accepted alleles. You can even set the minimum to 1 and combine this filter with this one:

`Select loci: Numb. of alleles in allele-richest ind.: min 1 max 2`

This way you can exclude any conflict between called SNPs in all selected loci.

Select genotypes

When you select “3 select genotypes” in the main menu, the program displays this menu:

```
SELECT GENOTYPES:

1 Depth of genotype                off
2 Total read depth of genotype     off
3 Used fraction of read depth      off
4 Depth of minor allele            off
5 Number of alleles                off
6 List of genotypes in a file      off

7 Run
8 Quit

Please select a number to activate/inactivate an option:
```

A genotype means the alleles of a locus in a given individual. You can select one or several of these options. Once you have selected an option and defined a criterion, the menu displays your criteria, e.g.:

```
4 Depth of minor allele       on min: 6
```

When you have finished the definition of criteria, enter `7` to apply them.

Select genotypes: options in detail:

1 Depth of genotype

You can select genotypes with a certain depth – the program will ask you for a minimum. The depth of a genotype here is the sum of depths of its alleles, which **indloc** has determined during the identification of alleles. The depth of a genotype can include fractions of depth from certain discarded reads. Nevertheless, it can be smaller than the total read depth of an individual locus, which includes all retained and discarded reads assigned to it. See the documentation of **indloc** for more details.

2 Total read depth of genotype

You can select genotypes with a certain total read depth - the program will ask you for a minimum. The total read depth includes all retained and discarded reads that **indloc** assigned to a locus. You should exclude split loci before you use this filter (`Select loci: 9 Split loci`).

3 Used fraction of read depth

Used fraction of read depth means depth of genotype / total read depth of genotype (see above, options 1 and 2). The program will ask you for a minimum. It can happen that **indloc** calls several SNPs at an individual locus, which are not in phase and define more than two, possibly multiple alleles. It is further possible that **indloc** finally accepts only one or two of them, which represent only a small fraction of all reads. (Under the binomial likelihood ratio method for SNP calling, **indloc** accepts the two most frequent alleles. Under the frequency method, you can apply a frequency filter (-a) in **indloc**, which accepts only alleles with a certain fraction of read depth.) This problem will arise when you have extremely high sequencing error at a locus or when the sequences assigned to it are paralogs. Use filter `Select genotypes: 3 Used fraction of read depth` to ensure that indloc assigned a high fraction (say 0.9) of all reads to the finally accepted alleles. You can even set the minimum to 1 and combine this filter with one of the following:

```
Select genotypes: Number of alleles: min 1 max 2
```

```
Select loci: Numb. of alleles in allele-richest ind.: min 1 max 2
```

This way you can exclude any conflict between called SNPs in all selected genotypes.

4 Depth of minor allele

Here, the program will ask you for a minimum for the depth of the minor allele. The minor allele of a genotype is the allele with the smallest sequencing depth as determined by **indloc**. In a homozygous genotype, the depth of the minor allele is half the depth of the single allele.

5 Number of alleles

You can select genotypes with certain numbers of alleles. The program will ask you for a minimum and maximum.

6 List of genotypes in a file

You can define your own selection of genotypes. To do so, prepare a tab-delimited text file with this format: no header line, no blank lines, one genotype per line: poplocID TAB ind_ID (see example file list_of_genotypes.txt):

poplocID: ID of a population locus determined by **poploc**

ind_ID: ID of an individual.

The program will ask you for the name (or full path) of this file.

Infiles

The program needs the following infiles, some of them only when certain options are active. It issues an error message when it cannot open a file.

Infile	produced by
individuals.txt	indloc
*_alleles.txt	indloc
*_loci.txt	indloc
poploc.txt	poploc
popall.txt	poploc
*_indpoploc.txt	indpoploc
*_indpopall.txt	indpoploc
split_loci.txt	indpoploc
depth_analysis/da_all_loc.txt	depth_analyzer
pairs/loc_groups.txt	pair_finder
pairs/self_match.txt	pair_finder
pairs/pairs_rep.txt	pair_finder
list of individual IDs (any name)	user
list of locus IDs (any name)	user
list of genotypes (any name)	user
export/sel_ind.txt	data_selector
export/sel_loc.txt	data_selector
export/sel_gt.txt	data_selector
export/sel_rep.txt	data_selector

*_ stands for an individual ID