CNVice User Guide

Version September 2016

Contact: maira.r.rodrigues@gmail.com

Overview

CNVice (Inbreeding Coefficients Estimation for CNV data) is a freely available R script for population genetics applications.

CNVice performs the following analyses:

1. Estimates allele frequencies for a given population assuming Hardy-Weinberg equilibrium (HWE), using the expectation maximization algorithm as implemented in CoNVEM (Gaunt *et al*. 2010), conditioning on the observed diplotype frequencies.
2. Estimates the population structure parameter f_{CNV} and allele frequencies, using the profiled likelihood function and the Expectation Maximization (EM) algorithm (Dempster *et al*. 1977)
3. Estimates the population genotype frequency, conditioning on the observed diplotype distribution and the estimated f_{CNV} and allele frequencies.
4. Uses trio information to improve the inference of an offspring's genotype, by considering the parents' diplotypes and the population genotype frequency.

Requirements

CNVice is implemented in R. Packages `aylmer` and `hwriter` are installed when CNVice is executed.

Input

The main input for CNVice is the distribution of observed diplotype frequencies of a given loci in a population. The distribution must always start with diplotype 0 (see the Examples section).

Running CNVice

The main function to be executed is:

```
executeCnvice(Nj,fpar,rept,document)
```

`Nj`	vector with number of individuals per copy number; starts with 0 copies (integer)
`fpar`	estimate parameter f or consider it 0 (True or False)
`rept`	number of repetitions (integer, default=100)
`document`	name of output file where the report will be saved (string)

The mandatory parameters are `Nj` and `fpar`. If `document` is not specified, then the output report (in MS word format) is not written.

For batch use, the function call should be as follows:

```
CNVice<-executeCnvice(Nj,T)
```

The object `CNVice` is a list that contains the following results:

1. Estimated allele frequencies matrix, accessed by `CNVice$allele_frequencies`
2. Estimated population genotype frequencies matrix, accessed by `CNVice$genotype_frequencies2`
3. Estimated population genotype frequencies matrix formatted for input into the `trio` function, accessed by `CNVice$genotype_frequencies`

The `genotype_frequencies` matrix contains the independent probabilities of each genotype, while the `genotype_frequencies2` matrix accounts for the fact that genotypes (X,Y) and (Y,X) correspond to the same phenotype and, thus, the probability of this phenotype is the sum of their individual probabilities.

To improve the inference of an **offspring's genotype by considering the parents' diplotypes**, the function is:

```
trio <- function(ft,mt,ch,matrix)
```

`ft`	Father's observed diplotype (integer)
`mt`	Mother's observed diplotype (integer)
`ch`	Child's observed diplotype (integer)
`matrix`	Population genotype frequencies matrix

The population genotype frequencies matrix is returned by the `executeCnvice` function (that is, `CNVice$genotype_frequencies`). Therefore, both functions should be used together (see the Examples section).

Output

The output of CNVice main function is a report containing (i) estimated allele frequencies, (ii) estimated population structure parameter f_{CNV} (if `fpar` is True), (iii) estimated population genotype frequencies, and (iv) estimated individual genotype frequencies. Statistical significance is provided for the estimated allele frequencies and parameter f_{CNV}.

Each part of the report is described in detail below, using as example `fpar` = True and the following distribution of observed diplotype frequencies: 78, 829, 2510, 756, 83, 9, 1.

The output of the trio function is the set of genotypic probabilities of the child (see the Examples section).

Estimated allele frequencies

Data Summary

j	Nj	expected.bin_1	freq.alel.est_1	expected.bin_2	freq.alel.est_2
0	78	74	0.132	74	0.132
1	829	839	0.745	840	0.746
2	2510	2502	0.119	2507	0.118
3	756	760	0.003	754	0.003
4	83	81	0.001	80	0.001
5	9	9	0	9	0
6	1	1	0	1	0

Where *j* is the observed diplotype, *Nj* is its frequency is the population, *freq.alel.est* is the estimated allele frequency and *expected.bin* is the expected diplotype frequency given the estimated allele frequencies. Suffixes in *expected.bin* and *freq.alel.est* columns (in this example, 1 and 2) indicate alternative results returned by the algorithm.

This information is also returned by the `CNVice<-executeCnvice(Nj,T)` function as `CNVice$allele_frequencies`.

Alternative estimations count:

1	2
86	14

This table shows the occurrence of each alternative result returned by the algorithm in 100 repeats (the default value of parameter `rept`). In our example, result 1 occurred more frequently (86 times) than result 2 (14 times).

P-values for Kolmogorov-Smirnov tests of alternative estimation results:

1	2
1	1

Here the hypothesis H0 is that observed frequency distribution = estimated frequency distribution.

Estimated population structure parameter *f*CNV

F values found for each estimation:

1	2
0.007	0

Shows the estimated f_{CNV} parameter, ranging from 0 to 1, for each alternative result, when `fpar` is True.

Likelihood ratio (LR) test:
Value of statistic found by LR: 0.25369
Reject H0, i.e., F different than ZERO?
No

The Likelihood ratio (LR) test is used for hypothesis testing of estimated f_{CNV}.

Estimated population genotype frequencies

Population genotype probability:

	0	1	2	3	4	5	6
0	0.018	0.194	0.031	0.001	0	0	0
1		0.557	0.176	0.004	0.001	0	0
2			0.015	0.001	0	0	0
3				0	0	0	0
4					0	0	0
5						0	0
6							0

Population genotype probabilities conditioned on the observed diplotype distribution, the estimated f_{CNV}, and estimated allele frequencies. First column and first row are alleles.

This information is also returned by the `CNVice<-executeCnvice(Nj,T)` function as `CNVice$genotype_frequencies2`.

Estimated individual genotype frequencies

Individual genotype probability (given the individual's copy number)

	00
Prob	1

	01	10
Prob	0.5	0.5

	02	11	20
Prob	0.026	0.947	0.026

Individual genotype probabilities conditioned on the observed individual diplotype distribution, the estimated f_{CNV}, and estimated allele frequencies.

Examples

Estimating allele and population genotype frequencies assuming Hardy-Weinberg equilibrium (HWE):

Consider a population with the following observed diplotype distribution of a particular loci: 0 individuals with 0 copies, 8 individuals with 1 copy, 10 individuals with 2 copies, 10 with 3 copies, and 1 individual with 4 copies. To estimate allele frequencies and population genotype frequencies assuming Hardy-Weinberg equilibrium, with the default 100 repetitions:

```
>Nj = c(0,8,10,10,1)

>executeCnvice(Nj,F,document="output-report-HW")
```

Output: population allele frequency estimation and population genotype frequency probabilities.

Estimating the population structure parameter fCNV, allele and population genotype frequencies:

Considering the same population as above, to estimate allele frequencies and population genotype freqecies assuming a departure from the HWE (f parameter not zero), with 10 repetitions:

```
>Nj = c(0,8,10,10,1)

>executeCnvice(Nj,T,document="output-report-F")
```

Output: population allele frequency estimation, f parameter estimation, and population genotype frequency probabilities.

Batch use Example:

```
>Nj = c(0,8,10,10,1)
```

```
>CNVice<-executeCnvice(Nj,T)
```

```
>CNVice$allele_frequencies
```

```
  j Nj expected.bin_1 freq.alel.est_1
1 0  0              1           0.171
2 1  8              6           0.589
3 2 10             12           0.240
4 3 10              8           0.000
5 4  1              2           0.000
```

```
>CNVice$genotype_frequencies2
```

```
      [,1]  [,2]  [,3] [,4] [,5]
[1,] 0.000 0.276 0.066    0    0
[2,] 0.276 0.279 0.345    0    0
[3,] 0.066 0.345 0.034    0    0
[4,] 0.000 0.000 0.000    0    0
[5,] 0.000 0.000 0.000    0    0
```

```
>CNVice$genotype_frequencies
```

```
      [,1]   [,2]   [,3] [,4] [,5]
[1,] 0.000 0.1380 0.0330    0    0
[2,] 0.138 0.2790 0.1725    0    0
[3,] 0.033 0.1725 0.0340    0    0
[4,] 0.000 0.0000 0.0000    0    0
[5,] 0.000 0.0000 0.0000    0    0
```

Using trio information to improve the inference of an offspring's genotype:

For estimating the genotype probabilities of an offspring with 3 copies, given that his father has 4 copies and his mother 3 copies of the same loci, and considering the same population as above, we have:

```
>Nj = c(78,829,2510,756,83,9,1)
```

```
>CNVice<-executeCnvice(Nj,T)
```

```
>trio(4,3,3,CNVice$genotype_frequencies)
```

Output:

Father	Mother	Offspring	Prob.Father	Prob.Mother	Prob.Offspring
04	03	03	0.0000000	0.0056497	0.0000000
13	03	30	0.2105263	0.0056497	0.0011947
13	12	12	0.2105263	0.9943503	0.2102748
22	12	21	0.7894737	0.9943503	0.7885305

The resulting table of probabilities does not show duplicated values or duplicated combinations for mother and father's diplotypes. That is why there is only diplotype 04 for the father and not 40. However, all diplotypes are considered in the probability calculation.

References

Dempster, A. P., Laird, N. M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, v. 39, n. 1, p.1–38, 1977.

Gaunt, T., et al. An Expectation-Maximization Program for Determining Allelic Spectrum from CNV Data (CoNVEM): Insights into Population Allelic Architecture and Its Mutational History. Human Mutation 2010;31(4):414-420.

Severini, T.A. Likelihood Methods in Statistics. Oxford University Press; 2000.