CNVEM Manual

Author: Sebastian Zöllner; szoellne@umich.edu

Overview

CNVEM is a Bayesian Expectation-Maximization algorithm that infers carrier status of CNVs in large samples from SNP genotyping data. It can be applied to analyze data from genome-wide association studies. Using Bayesian computations the program calculates the posterior probability for carrier status of a known CNV in each individual of a sample by jointly analyzing genotype information and hybridization intensity. Signal intensity is modeled as a mixture of normal distributions, allowing for locus-specific and allele-specific distributions. Using an expectation maximization algorithm, these distributions are estimated and then used to infer the carrier status of each individual and the boundaries of the CNV.

The algorithm used here is described in
Zöllner S, Su G, Chen Y, McInnis MG, Burmeister M. EM Algorithm for scoring polymorphic deletions from SNP data and application to a common CNV on 8q24.
Please cite this paper if you consider the program useful.

Presently the program is still in beta testing, thus both the program and this manual are a little rough. Thus, if you see parts of the program or this manual that could be improved upon, please let me know.

Compiling the program

To compile the program type:

```
gcc ./cnvem2.6.c -lm -O3 -o cnvem
```

This will create an executable named cnvem in the same directory. In the following we assume that cnvem is the name of the executable.

Running the program

CNVEM can be run two different ways. Either you can assess samples for carrier status of a single CNV specified by the command line, or you can scan a chromosome for all CNVs in an input file. The basic data file (data.csv) is a comma delimited file containing genotypes and hybridization intensities for all individuals. Format for all input files is given below.

Option 1: Assuming that there is a single CNV in the dataset, the program analyzes a region set by the F and L option. Using this option, the command to run the program is

```
./cnvem -e20 -B4 -idata.csv -s200 –F5 –L22
```

This command will assess samples for a CNV spanning markers 5 – 22 (inclusive) with hybridization intensities in data.csv. Using this option, all output is printed on the screen and should be redirected to a file for analysis.

Option 2: This option restarts the algorithm for a set of CNVs. The program expects two additional input files, one file that contains the physical locations of the CNVs (cnvfile.txt) and one file that provides the location for all markers (mapfile.txt). Using this option, the command to run the program is

```
./cnvem -e20 -B4 -idata.csv -s200 –bcnvfile.txt –mmapfile.txt
```

The imputed carrier status for all CNVs is summarized in an output file named sumout.cnvfile.txt.

Parameter Choices

To run CNVEM, you must decide on the number of iterations for the EM part (e-option), the prior of the CNV (-P) and whether to fine-map the borders of a CNV (-B).

While the EM algorithm usually converges within 5 steps, the process is fast enough that there is no reason to select a number of iterations below 20.

Selecting a prior is only advisable if actual outside information about the frequency of the CNV is available or if you have reason to believe that your sample is not in Hardy-Weinberg equilibrium. Otherwise it makes more sense to let the program estimate the prior from genotype frequencies. Note that the prior is the probability to carry at least one allele of the CNV, so it's roughly twice the frequency of the minor allele.

Inferring the borders of a CNV is generally advisable, as most CNV-borders generated by hybridization methods with long probes are overestimates of the true length of the polymorphism. However, doing so may increase the computation time considerably. Setting a minimum length of 4 or 5 is advisable, as this reduces computation time, and inferences of CNVs below this length are not very reliable anyway.

Command line options

Generally command line options come in two flavors, flags (F) and input parameters (P). Flags indicate that the program should run in a certain way, for example model genotyping error. Parameter values are put in directly after the line option, without a space in-between.

The Dataset

i	P	Name of the datafile (data.csv). The file name should be shorter than 20 characters.
b	P	Name of input file for multiple CNV-borders. The file name should be shorter than 20 characters.
m	P	Mapfile for SNPs. The file name should be shorter than 20 characters.
p	P	Number of SNPs. This option is required when a single CNV is being specified on the command line (Option 1).
s	P	Sample Size
O	F	Only one hybridization intensity is in the datafile; for example if the log-likelihood ratio rather than the intensities of each marker are used to infer CNVs.
c	P	Number of columns with irrelevant info at the start of the file. For example for Illumina Beadstudio output this can be set to 5.

Analysis options:

e	P	Rounds of EM. This is the number of iterations performed to estimate the parameters underlying the signal distribution of the hybridization intensity.
B	P	Determine best boundaries; x is minimum length; If this parameter is set, the algorithm permutes through all possible pairs of start- and endpoints within the preset region
G	F	Model genotyping error. The error model used is described in Zöllner et al. (2008).
P	P	Prior probability for an individual to carry a CNV. If no prior is set, the program estimates deletion frequency from the genotype data.
T	P	Option to call carriers rather than to give a posterior probability. The parameter x gives the threshold; individuals with posterior probability >x are called as carriers, others as non-carriers.
F	P	Fist SNP to be considered in single-CNV analysis
L	P	Last SNP to be considered in single CNV analysis
D	F	Carriers with a higher mean intensity are removed and the program is restarted.
Q	P	Maximum proportion of missing data; markers with a higher proportion of missing data will be ignored.

Output

S	F	Silent EM off; this gives intermediate results from each EM in the general output file.
A	F	Calculate distributions of hybridization intensity
h	F	Help – an abbreviated output of the input options

Input file format

In addition to the format described below, all input files can contain lines starting with '#'. These lines will be ignored by the program and can be used to include comments.

Hybridization intensities and genotype data (data.csv)

The data file is comma delimited text. The first line in the input file provides the names for all individuals followed by '.gt' for genotype, '.x' for x-hybridization intensity and

```
#Sample datafile for 4 individuals and 10 markers
54.GT,54.X,54.Y,55.GT,55.X,55.Y,56.GT,56.X,56.Y,57.GT,57.X,57.Y
AA,1.570,0.005,AA,1.644,0.003,AB,0.0,0.401,AA,1.60,0.007
AA,1.30,0.006,AA,1.32,0.005,AB,0.74,0.246,AA,1.364,0.002
AB,0.219,0.309,AA,0.569,0.010,AB,0.320,0.416,AA,0.569,0.01
AB,0.473,0.606,AA,1.27,0.022,AA,1.267,0.019,AA,1.226,0.013
BB,0.019,1.57,AB,0.636,0.777,AB,0.57,0.24,AB,0.519,0.940
AB,0.334,0.204,AB,0.361,0.240,BB,0.010,0.499,AB,0.397,0.215
AB,0.23,0.373,AB,0.277,0.317,BB,0.000,0.656,AB,0.27,0.333
AB,0.605,0.33,AB,0.603,0.360,AB,0.512,0.444,AB,0.615,0.35
AA,0.71,0.007,AB,0.39,0.653,AB,0.413,0.761,AB,0.465,0.772
AB,0.370,0.751,AB,0.32,0.710,AB,0.424,0.703,AB,0.403,0.697
```

‘.y’ for y-hybridization intensity. Each following line provides genotype and hybridization data for one SNP. The data is for each individual, the first two-letter combination provides genotype (AA, BB are the homozygotes, AB is the heterozygote and NC is missing data). The next floating point number provides the hybridization intensity for probes of the A allele, and the second floating point number provides the hybridization intensity for probes of the B allele. If you analyze a normalized joint measure of hybridization intensity such as the Log-R-ratio for Illumina genotyping, you may have only one intensity value. Indicate this by using the –O option.
Note that you may want to put additional columns in front of each line, for example containing the name of the markers or their position. To tell the program to ignore the first x columns, use the –cx option.

Borders of CNVs (cnvfile.txt)

The first line of the map file provides the number of lines in the file. It starts with a P and then the number of CNVs, without a white space in between.

```
#Example of a CNV border file
P7
S1,1,3
S1,1,12
S1,1,9
L1,137800000,140000000
L1,11,33
L1,137778489,137791241
S1,1,20
S1,8,14
```

After this, each line represents one CNV. The first letter in the line indicates the type of coordinates provided. An L indicates that the location of the CNV is given in physical coordinates (bp). An S indicates that the location is given relative to the provided SNP-map. The number after the letter indicates the type of CNV, a 1 indicates a polymorphic deletion. Finally the last two numbers provide the beginning and the end of the CNV. Hence `S1,1,9` indicates a polymorphic deletion starting at marker 1 and ending after marker 9, while `L1,13700000,140000000` indicates a deletion starting at bp 13700000 and ending after bp 140000000. Note that the file is comma-delimited.

Map of SNPs and other probes (mapfile.txt)

```
#Example of a map file
#containing positions of 10 SNPs
P10
137655983
137669835
137725337
137746857
137758484
137762820
137767017
137769680
137836926
137875956
```

The first line of the map file provides the number of lines in the file. It starts with a P and then the number of SNPs, without a white space in between. After that, every line gives the physical position of one marker.
The marker positions should be given in the same order as the markers are given in data.csv. Thus the markers have to be ordered, lowest to highest position.

Output format

Dependent on what options are chosen, the program generates one or two output files.

Screen output

The screen output contains a lot of intermediate results that are most likely not that useful except to diagnose the algorithm.

Output file

After finishing all calculations, the program generates an output file named sumout.cnvfile.txt if multiple CNVs have been analyzed (Option 2). Each line in this file represents the output for one CNV. The first column is the number of the CNV in the same order as given in cnvfile.txt. The next 3 columns repeat the information from cnvfile.txt: the type, the start and the end point of the CNV. The next two columns give the start point relative to the SNPs in the mapfile.txt and number of SNPs covered. The next two columns give the estimated start point and the estimated length. The next column provides the estimated number of carriers for this best boundary configuration, followed by the relative fit of this configuration. The column after this provides the a priori estimated frequency of the deletion as calculated from the genotype data. The next flag gives an OL flag if the imputed CNV carriers have a hybridization intensity that is higher than the baseline intensity. After this, one column is given for each individual in the dataset providing the posterior probability of that individual carrying the minor CNV allele.

Notes on analyzing genomewide data

- As markers in the input file need to be sorted by position, it is not possible to analyze more than one chromosome at a time.
- When analyzing Illumina data, we obtained better results from analyzing normalized intensities.