seqEM documentation

to run:

Open a command prompt in Linux/Unix or Windows and type:
seqem name_of_SeqEM_control_file

The typical control file looks like:

```
Header_files:     8
rv1.txt
rv2.txt
rv3.txt
rv4.txt
vv1.txt
vv2.txt
rr1.txt
rr2.txt
HWE:              0
Debug:            0
MAX_iterations:   100
EM_threshold:     0.000001
Cov_matrix:       1
Min_read_depth:   10
Fix_error:        0
Initial_error:    0.01
Threads:          1
Outfile:          result.out
Force_pval:       -1
```

Header_files:
: The number of data files. One individual per file. The file names are listed one per line following the “Header_files:” identifier.

HWE:
: Assume Hardy-Weinberg equilibrium? 0 = no, 1 = yes.

Debug:
: Turns off/on 0/1 several debugging features.

MAX_iterations:
: The maximum number of iterations in the EM algorithm.

EM_threshold:
: The threshold when to stop EM iteration.

Cov_matrix:
: Print out covariance matrices? 0/1
  See documentation of algorithm for details.

Min_read_depth:
: Minimum number of reads for any individual to be considered in calculations. Default is set to 1.

Fix_error:
: Allows user to fix the error estimate. It will be fixed to the value specified under “Initial_error”.

Initial_error:
: Allows the user to set the initial error estimate. The default is 0.01.

Threads:
: How many threads should the program run? The default is 1.

Outfile:
: The name of the file that holds the results from SeqEM.

Force_pval:

Force the initial MAF to this value. Omit or enter -1 to let the program decide the initial MAF. It will select p(variant read).

To compile:

Extract source code and makefile to a directory. Make sure you have GNU's g++ compiler version 4.1.2 (or later) installed. Type make at the command prompt and you're done. A SeqEM executable for Windows is included in the win32 directory that came with this download. It was compiled form the same source code using Microsoft's Visual Studio 2008 compiler version 9.0.21022.8 RTM. You might have to update your .NET Framework which is available at Micorsoft's web site. You also need to place the included pthreadVC2.dll file into the same directory as the SeqEM executable.

The Input Format:

>chr1 rs0 7873730 T C 10 0 0

The order of the columns is:

- chromosome number preceded by ">chr". Yes, those 4 chrarcters are critical.
- name, any string without tabs or whitespace will do
- position or unique ID number. That's how it determines the SNP id.
- reference nucleotide
- variant nucleotide (the value is not used but should be a string)
- total number of reads/ read depth
- number of variant reads
- true genotype (optional), 0 = rr, 1 = rv, 2 = vv. SeqEM doesn't read this column but the summary script will to determine SeqEM's accuracy. In fact SeqEM will only read the first 7 columns of each line. The rest of each line will be discarded.

Running Sample Data:

Please take a look at the "sample_data" directory. You can run this data by typing:

<name of seqEM executable> one.ctrl

Your output should look similar to the in sample_result.txt. We refer to the alleles as:

r = reference

v = variant

The individuals are ordered the same way as the input files in the control file. So:

[0] = one_0.out
[1] = one_1.out
[2] = one_2.out
[3] = one_3.out
[4] = one_4.out
[5] = one_5.out
[6] = one_6.out
[7] = one_7.out
[8] = one_8.out
[9] = one_9.out

The summary of each position/variant is listed at the beginning:

0. 385856 iter= 24 error= 0.05329 p= 0.47001 hwe-chisq= -1.000000
P(rr)= 0.21950 p(rv)= 0.50101 p(vv)= 0.27949

0. the 0th position

385856 SNP id number
iter= 24 number of iterations for the EM-alogrithm to converge
error= 0.05329 estimated nucleotide read error
p= 0.47001 reference allele frequency
hwe-chisq= -1.000000 Gives test statistic for HWE equilibrium if a minimum 10 individuals are typed otherwise gives the value is -1.00000

P(rr)= 0.21950 p(rv)= 0.50101 p(vv)= 0.27949 estimated posterior probabilities of genotypes

Indvidual results for [0]:
genotype[0]= rv props(rr,rv,vv)= 0.000 0.944 0.056 reads= 10 , variants= 7

Most likely genotype "rv" (one with highest posterior probability).
Probabilities of genotypes rr, rv and vv are 0.000 0.944 0.056 respectively.
The read depth (= 10) and the number of variant reads (=7) are listed last.