EM2: Enhanced Computational Algorithm for Haplotype-Based Association Analysis in Case-Control Studies

Isaac Amundson, Kevin M. Bradley, Brian L. Yaspan, Jeffrey R. Smith
Departments of Computer Science, Medicine, and Cancer Biology
Vanderbilt University

Overview

Haplotype-based association analysis within a case-control study is broadly anticipated to offer a powerful approach to identify genetic variants causing common disease. The Expectation-Maximization (EM) algorithm is commonly employed to estimate haplotype frequency in unphased data, such as case-control genotypes. Accurate assignment of phase to enable tests of heterogeneity between case and control haplotype profiles is computationally very demanding. Use of permutation testing to estimate exact P values compounds this demand. This can limit the feasibility of large studies.

We modified algorithms implemented in the Fortran 77 program of Fallin and Schork (AJHG 67:947-959, 2000) to facilitate large-scale studies. Daniele Fallin kindly provided the original source code for these modifications. The original program limits each run to a maximum of 10 SNPs and 2000 samples, and does not accommodate multi-allelic markers such as STRs. In general these restrictions favor application speed. We altered the program to enable use of multi-allelic as well as bi-allelic markers, and to place no intrinsic limit on the number of markers or samples. “EM2” attains significant speed improvements despite the more generalized implementation. Our enhanced version is written in C++ and is developed for both Linux and Windows platforms. The program core incorporates object-oriented design pattern reuse and utilizes the C++ standard template library functionality, ensuring efficient algorithm usage. EM2 also employs parallel-processing. We are currently modifying permutation test code and anticipate further significant speed gains. Improved computational algorithms will be crucial for large-scale genome-wide association studies using a case-control design.

Several runs may be performed consecutively in batch mode, allowing the user to create a sliding window (subset of markers) over a region of interest. This is accomplished by creating a batch file specifying the markers in the window and the extent of marker overlap of the sliding window. The user specifies the number of processors for use, as well as other run parameters such as number of restarts, maximum number of iterations, and number of case/control permutations. An option also exists to only display haplotypes that exist within the population above a specified frequency. Output can be presented in HTML or text format, and/or displayed directly to the terminal. EM2 runs from the command line, and can be employed as a stand-alone application or incorporated into a larger application (such as an interactive website).

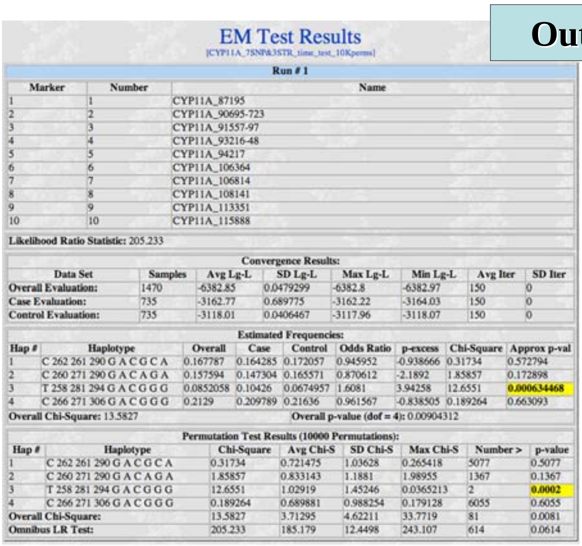
Tests

Program	Processors	Markers	
		10 htSNPs	7 htSNPs & 3 STRs
EM2	8	5 hr, 1 min	6 hr, 49 min
	1	25 hr, 48 min	-
Fallin & Schork	1	58 hr, 3 min	N/A

As a test of relative speed we employed a laboratory dataset of 735 breast cancer cases and 735 unaffected controls at 10 htSNPs and 3 STRs spanning 38 kb at the CYP11A gene. There was no missing experimental data for these samples. Missing data for a given sample in a specified window of markers would result in the sample being dropped from the analysis. SNP minor allele frequency ranged from 5.1% to 43.4%. STR heterozygosity ranged from 0.52 to 0.79. Run parameters were: 15 restarts, 150 maximum iterations, 10,000 permutations, rare haplotypes (<5%) grouped for overall chi-square test of heterogeneity, and a single window of 10 specified markers. Tests were executed on an SGI Altix 3000 configured with eight 1.3 GHz Itanium 2 processors and 16 GB of DDR RAM, running 64 bit Red Hat Linux.

The convergence results, as part of the maximization step, show log-likelihood information associated with the overall, case, and control data sets. Estimated frequencies are provided for observed haplotypes for overall, case, and control data sets. Haplotypes are represented by SNP alleles and STR lengths in base pairs. Chi square and p value is calculated for a given haplotype in case v.s. control groups, and calculated for overall haplotype profile across the two groups (χ^2 ndf). Permutation test results are also presented for each haplotype, estimating exact p values. For each permutation, the case/control status of each sample in the entire dataset is randomized using a random number generator and random seed value, and the EM algorithm re-run. Fallin and Schork’s omnibus likelihood ratio test is also presented, assessing the significance of overall haplotype profile difference between cases and controls.

Output



EM Test Results

[CYP11A_7SNP&3STR_time_test_10Kperms]

Run # 1

Marker	Number	Name
1	1	CYP11A_87195
2	2	CYP11A_90695-723
3	3	CYP11A_91557-97
4	4	CYP11A_93216-48
5	5	CYP11A_94217
6	6	CYP11A_106364
7	7	CYP11A_106814
8	8	CYP11A_108141
9	9	CYP11A_113351
10	10	CYP11A_115888

Likelihood Ratio Statistic: 205.233

Convergence Results:

Data Set	Samples	Avg Lg-L	SD Lg-L	Max Lg-L	Min Lg-L	Avg Iter	SD Iter
Overall Evaluation:	1470	-6382.85	0.0479299	-6382.8	-6382.97	150	0
Case Evaluation:	735	-3162.77	0.689775	-3162.22	-3164.03	150	0
Control Evaluation:	735	-3118.01	0.0406467	-3117.96	-3118.07	150	0

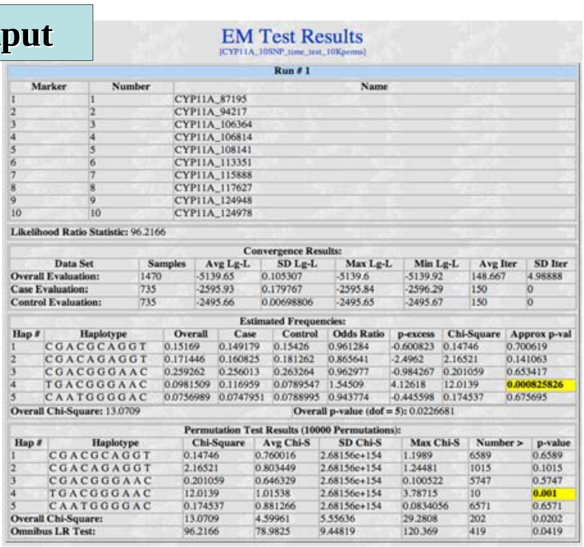
Estimated Frequencies:

Hap #	Haplotype	Overall	Case	Control	Odds Ratio	p-excess	Chi-Square	Approx p-val
1	C 262 261 290 G A C G C A	0.167787	0.164285	0.172057	0.945952	-0.938666	0.31734	0.572794
2	C 260 271 290 G A C A G A	0.157594	0.147304	0.165571	0.870612	-2.1892	1.85857	0.172898
3	T 258 281 294 G A C G G G	0.0852058	0.10426	0.0674957	1.6081	3.94258	12.6551	**0.000634468**
4	C 266 271 306 G A C G G G	0.2129	0.209789	0.21636	0.961567	-0.838505	0.189264	0.663093

Overall Chi-Square: 13.5827 **Overall p-value (dof = 4):** 0.00904312

Permutation Test Results (10000 Permutations):

Hap #	Haplotype	Chi-Square	Avg Chi-S	SD Chi-S	Max Chi-S	Number >	p-value
1	C 262 261 290 G A C G C A	0.31734	0.721475	1.03628	0.265418	5077	0.5077
2	C 260 271 290 G A C A G A	1.85857	0.833143	1.1881	1.98955	1367	0.1367
3	T 258 281 294 G A C G G G	12.6551	1.02919	1.45246	0.0365213	2	**0.0002**
4	C 266 271 306 G A C G G G	0.189264	0.689881	0.988254	0.179128	6055	0.6055
Overall Chi-Square:		13.5827	3.71295	4.62211	33.7719	81	0.0081
Omnibus LR Test:		205.233	185.179	12.4498	243.107	614	0.0614



EM Test Results

[CYP11A_10SNP_time_test_10Kperms]

Run # 1

Marker	Number	Name
1	1	CYP11A_87195
2	2	CYP11A_94217
3	3	CYP11A_106364
4	4	CYP11A_106814
5	5	CYP11A_108141
6	6	CYP11A_113351
7	7	CYP11A_115888
8	8	CYP11A_117627
9	9	CYP11A_124948
10	10	CYP11A_124978

Likelihood Ratio Statistic: 96.2166

Convergence Results:

Data Set	Samples	Avg Lg-L	SD Lg-L	Max Lg-L	Min Lg-L	Avg Iter	SD Iter
Overall Evaluation:	1470	-5139.65	0.105307	-5139.6	-5139.92	148.667	4.98888
Case Evaluation:	735	-2595.93	0.179767	-2595.84	-2596.29	150	0
Control Evaluation:	735	-2495.66	0.00698806	-2495.65	-2495.67	150	0

Estimated Frequencies:

Hap #	Haplotype	Overall	Case	Control	Odds Ratio	p-excess	Chi-Square	Approx p-val
1	C G A C G C A G G T	0.15169	0.149179	0.15426	0.961284	-0.600823	0.14746	0.700619
2	C G A C A G A G G T	0.171446	0.160825	0.181262	0.865641	-2.4962	2.16521	0.141063
3	C G A C G G G A A C	0.259262	0.256013	0.263264	0.962977	-0.984267	0.201059	0.653417
4	T G A C G G G A A C	0.0981509	0.116959	0.0789547	1.54509	4.12618	12.0139	**0.000825826**
5	C A A T G G G G A C	0.0756989	0.0747951	0.0788995	0.943774	-0.445598	0.174537	0.675695

Overall Chi-Square: 13.0709 **Overall p-value (dof = 5):** 0.0226681

Permutation Test Results (10000 Permutations):

Hap #	Haplotype	Chi-Square	Avg Chi-S	SD Chi-S	Max Chi-S	Number >	p-value
1	C G A C G C A G G T	0.14746	0.760016	2.68156e+154	1.1989	6589	0.6589
2	C G A C A G A G G T	2.16521	0.803449	2.68156e+154	1.24481	1015	0.1015
3	C G A C G G G A A C	0.201059	0.646329	2.68156e+154	0.100522	5747	0.5747
4	T G A C G G G A A C	12.0139	1.01538	2.68156e+154	3.78715	10	**0.001**
5	C A A T G G G G A C	0.174537	0.881266	2.68156e+154	0.0834056	6571	0.6571
Overall Chi-Square:		13.0709	4.59961	5.55636	29.2808	202	0.0202
Omnibus LR Test:		96.2166	78.9825	9.44819	120.369	419	0.0419

Note that the highlighted haplotype of our example includes the same CYP11A allele originally associated with breast cancer risk in the full study population of 1193 cases and 1310 controls. In the published study, a lone STR marker comprised of a [$(TAAAA)_n$] repeat of the distal promoter region was analyzed by conditional logistic regression (Zheng, W., Gao, Y., Shu, X., Wen, W., Cai, Q., Dai, Q., and Smith, J. (2004) *Cancer Epidemiology, Biomarkers & Prevention* **13**, 709-714). In that study, allele 281 of STR CYP11A_91557-97 was in significant excess in cases (12.56% freq) vs. controls (8.46% freq), p <0.0001, OR 1.6 (1.3-1.9). This data illustrates that specific alleles of each STR marker and htSNP CYP11A_87195 each mark the disease-associated haplotype. Note that at this gene, alleles of the STRs as well as htSNPs efficiently tag the study population haplotypes. EM2 identifies ten htSNP-STR and six htSNP-only haplotypes of 1-5% frequency, which were grouped for analysis and not displayed per our run option.