BMC Bioinformatics

SOFTWARE Open Access

PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling

Richard Howey and Heather J Cordell*

Abstract

Background: Here we present two new computer tools, PREMIM and EMIM, for the estimation of parental and child genetic effects, based on genotype data from a variety of different child-parent configurations. PREMIM allows the extraction of child-parent genotype data from standard-format pedigree data files, while EMIM uses the extracted genotype data to perform subsequent statistical analysis. The use of genotype data from the parents as well as from the child in question allows the estimation of complex genetic effects such as maternal genotype effects, maternal-foetal interactions and parent-of-origin (imprinting) effects. These effects are estimated by EMIM, incorporating chosen assumptions such as Hardy-Weinberg equilibrium or exchangeability of parental matings as required.

Results: In application to simulated data, we show that the inference provided by EMIM is essentially equivalent to that provided by alternative (competing) software packages such as MENDEL and LEM. However, PREMIM and EMIM (used in combination) considerably outperform MENDEL and LEM in terms of speed and ease of execution.

Conclusions: Together, EMIM and PREMIM provide easy-to-use command-line tools for the analysis of pedigree data, giving unbiased estimates of parental and child genotype relative risks.

Keywords: Case/parent trio, Maternal-fetal interaction, Parent-of-origin, Genome-wide association study

Background

Genomewide association studies have popularized the use of the case/control design to detect effects associated with an individual's own genotype, however many diseases (especially those related to pregnancy outcomes) may in fact be due to more complex effects such as maternal genotype effects, maternal-fetal genotype interactions or parent-of-origin (imprinting) effects. To detect such effects it is necessary to collect genotype data from one or both parents of cases, in addition to genotyping the cases themselves. Two existing popular approaches analyse either genetic data from affected offspring and their mothers (case/mother duos), along with an appropriate control sample [1-3], or else analyse genetic data from affected offspring and both parents (case/parent trios), without use of controls [4-6]. In contrast, our software EMIM uses a multinomial modelling approach [7] that allows the simultaneous consideration of both case/mother duos and/or case/parent trios, with additional child and parent genotype data (such as individual cases and controls, case/father duos and control matings) included when available. The child-parent genotype data can be extracted from standard PLINK-format [8] pedigree files using our companion software PREMIM.

Full details and evaluation of the multinomial modelling approach used by EMIM have been described previously [7]. The early beta version of EMIM described in [7] allowed a more limited set of child-parent configurations than are supported in the current version, and did not include the current full range of optional likelihood assumptions (such as conditioning on parental genotypes (CPG) [6,9]). Most importantly, the companion program PREMIM was not available, limiting the ease with which EMIM could be applied to real data.

*Correspondence: heather.cordell@ncl.ac.uk
Institute of Genetic Medicine, Newcastle University, Central Parkway, Newcastle upon Tyne, NE1 3BZ, UK

PREMIM: Pedigree file conversion

For each SNP in turn, PREMIM performs a simple algorithm to select from each pedigree the most informative sub-unit of child-parent genotype data. Different pedigree sub-units are chosen in order of preference as listed in Table 1.

There are a number of options that may be given to PREMIM. In particular, it is possible to override the default choice of individuals by stating a *proband* subject for certain pedigrees. These proband subjects are then chosen as cases (with parents where available). This may be useful to avoid possible bias when larger pedigrees have been ascertained on the basis of a specific affected individual. For larger pedigrees, it is also possible to select multiple case/parent trios or multiple control matings from each pedigree, potentially increasing the power to detect genetic effects. This option does have the potential to generate bias (depending on the analysis options chosen [6,10]), and so results should be interpreted with caution, although we anticipate that most people will apply these types of method to small pedigrees such as child/parent trios, making this issue less of a concern in practice. (Alternative methods for dealing with larger pedigrees, valid under the assumptions of random mating and/or Hardy-Weinberg equilibrium (HWE), have been described by [10,11]).

EMIM methodology

The basic principle behind EMIM is simple: to test for the existence of (and estimate) genotype relative risk parameters that increase (or decrease) the probability that a child is affected. By default, PREMIM chooses the minor allele to be considered as the 'risk' allele, although this option can be overridden if required. We denote by R_1 (R_2) the factor by which an individual's disease risk is multiplied if they possess one (two) risk alleles at a given locus. We denote by S_1 (S_2) the factor by which an individual's disease risk is multiplied if their mother possesses one (two) risk alleles at that locus. We denote by I_m (I_p) the factor by which an individual's disease risk is multiplied if they inherit a risk allele from their mother (father). Lastly, to test for mother-child interactions, we denote by γ_{ij} the factor by which an individual's disease risk is multiplied if the mother carries i risk alleles and the child carries j risk alleles. A summary of these relative risk parameters is shown in Table 2. A variety of restrictions may be made on the parameters as desired. For example, a multiplicative model for the effects of the alleles in the mother ($S_2 = S_1^2$) or child ($R_2 = R_1^2$) may be imposed. In addition, EMIM also supports several alternative previously-proposed paramaterizations for the imprinting and interaction effects [4,5] (see [7] for more details).

As an example, denote the major and minor alleles by 1 and 2, then for a case/parent trio where the genotypes of the mother, father and child are 22, 11, 12, respectively, the penetrance is modelled as:

$$\mathrm{P}(\text{child diseased}|g_m = 22, g_f = 11, g_c = 12) = \alpha R_1 S_2 I_m \gamma_{21}$$

where α is the baseline probability of disease and g_m, g_f and g_c are the genotypes of the mother, father and child.

EMIM uses a multinomial model to estimate the relative risk parameters on the basis of observed counts of genotype combinations in case/parent trios as shown in Table 3. EMIM models the 15 different cell probabilities (corresponding to the 15 possible combinations of

Table 1 The order of preference of pedigree sub-units chosen by PREMIM for each SNP

Order	Pedigree sub-unit
1	case/parent trio
2	case/mother duo
3	case/father duo
4	case
5	case parental mating
6	case mother
7	case father
8	control parental mating
9	control/mother duo
10	control/father duo
11	control

Table 2 The relative risk parameters estimable by EMIM

Parameter	Description
R_1	Child has one minor allele (child genotype effect)
R_2	Child has two minor alleles (child genotype effect)
S_1	Mother has one minor allele (maternal genotype effect)
S_2	Mother has two minor alleles (maternal genotype effect)
γ_{11}	Mother has one minor allele and child has one minor allele (mother-child interaction effect)
γ_{12}	Mother has one minor allele and child has two minor alleles (mother-child interaction effect)
γ_{21}	Mother has two minor alleles and child has one minor allele (mother-child interaction effect)
γ_{22}	Mother has two minor alleles and child has two minor alleles (mother-child interaction effect)
I_m	The child receives a minor allele from the mother (maternally operating imprinting effect)
I_p	The child receives a minor allele from the father (paternally operating imprinting effect)

Table 3 Observed genotype combinations in case/parent trios

Genotypes[a]			Index of	Index of CEPG[b]	Index of CPG[c]	Observed
g_m	g_f	g_c	combination	parental mating type	parental mating type	count
22	22	22	1	1	1	n_1
22	12	22	2	2	2	n_2
22	12	12	3	2	2	n_3
12	22	22	4	2	3	n_4
12	22	12	5	2	3	n_5
22	11	12	6	3	4	n_6
11	22	12	7	3	5	n_7
12	12	22	8	4	6	n_8
12	12	12	9	4	6	n_9
12	12	11	10	4	6	n_{10}
12	11	12	11	5	7	n_{11}
12	11	11	12	5	7	n_{12}
11	12	12	13	5	8	n_{13}
11	12	11	14	5	8	n_{14}
11	11	11	15	6	9	n_{15}

[a] g_m, g_f, g_c=genotypes of mother, father, child, respectively.
[b] CEPG= conditional on exchangeable parental genotypes.
[c] CPG= conditional on parental genotypes.

genotypes that are consistent with Mendelian inheritance) in terms of the desired genotype relative risk parameters ($R_1, R_2, S_1, S_2, I_m, I_p, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}$). A maximum of 7 parameters are estimable, meaning that not all of these parameters can be estimated simultaneously. Cordell et al. [12] suggested building up models from simpler to more complex via a series of nested hypothesis tests. Given a model for the penetrances in terms of the genotype relative risk parameters, the overall likelihood for the data in Table 3 may be written

$$\prod_{i=1}^{15} \left\{ \mathrm{P}\left(g_{m_i}, g_{f_i}, g_{c_i} | \text{child diseased}\right) \right\}^{n_i}$$

where $(g_{m_i}, g_{f_i}, g_{c_i})$ represent the genotypes of a mother, father and child in genotype combination i. The probabilities $\mathrm{P}(g_{m_i}, g_{f_i}, g_{c_i}|\text{child diseased})$ may be written in terms of the genotype relative risk parameters of interest and six nuisance parameters $\mu_1 - \mu_6$ (corresponding to mating type stratification parameters as indexed in Table 3, see [4,7,13] for details).

If any of the subjects are missing, we no longer have 15 genotype counts as shown in Table 3, but instead we must collapse together rows to express the data in terms of counts of observed genotype combinations. For example, given data for case/mother duos (i.e. all fathers missing), the 7 observable counts are as shown in Table 4. The likelihood for the data in this table may be written

$$\prod_{i=1}^{7} \left\{ \mathrm{P}(g_{m_i}, g_{c_i} | \text{child diseased}) \right\}^{m_i} = \prod_{i=1}^{7} \left\{ \sum_{g_f} \mathrm{P}(g_{m_i}, g_f, g_{c_i} | \text{child diseased}) \right\}^{m_i}$$

where (g_{m_i}, g_{c_i}) represent the genotypes of a mother and child in (Table 4) genotype combination i.

Table 4 Observed genotype combinations in case/mother duos

Genotypes[a]		Index of	Observed
g_m	g_c	combination	count
22	22	1	$m_1 = n_1 + n_2$
22	21	2	$m_2 = n_3 + n_6$
12	22	3	$m_3 = n_4 + n_8$
12	12	4	$m_4 = n_5 + n_9 + n_{11}$
12	11	5	$m_5 = n_{10} + n_{12}$
11	12	6	$m_6 = n_7 + n_{13}$
11	11	7	$m_7 = n_{14} + n_{15}$

[a] g_m, g_c=genotypes of mother, child, respectively.

In practice, at any given SNP, we observe genotype counts (some of which may equal 0) for the following types of unit: case/parent trios (15 possible genotype combinations); parents of cases (9 possible genotype combinations); case/mother duos (7 possible combinations); case/father duos (7 possible combinations); mothers of cases (3 possible combinations); fathers of cases (3 possible combinations); cases (3 possible combinations). The data for each unit creates a table corresponding to a (possibly collapsed) version of Table 3, and the overall likelihood to be maximized may be constructed as the product of the likelihoods for the individual tables. Similarly, we may add in data for controls (either unaffected individuals or population-based controls of unknown disease status) by further multiplying the likelihood by the product of the likelihoods for a similar set of control tables. EMIM makes use of the following types of control unit: parents of controls (9 possible genotype combinations); control/mother duos (7 possible combinations); control/father duos (7 possible combinations); controls (3 possible combinations). Furthermore, EMIM assumes that the frequencies of the different genotype combinations in control units correspond to those in the general population. This is equivalent to making a rare disease assumption, in the event that the controls are all genuinely unaffected.

By default, EMIM assumes 'mating symmetry' [13] (equivalent to a 'conditional on exchangeable parental genotypes' (CEPG) [12] model), which corresponds to assuming that parental matings ($g_m = i, g_f = j$) are as likely as matings ($g_m = j, g_f = i$). This results in the estimation of six mating type stratification parameters [13] $\mu_1 - \mu_6$ (see Table 3). Two more restricted (and therefore potentially more powerful) models are also available in EMIM:

1. A model that assumes parental allelic exchangeability (PAE) [2] (which corresponds in this context to assuming that $\mu_4 = \mu_3$)
2. A model that assumes Hardy-Weinberg equilibrium (HWE) and random mating, estimating a single allele frequency parameter in place of the six mating type stratification parameters.

In addition to these more restricted models, a *less* restricted 'conditional on parental genotypes' (CPG) [2,9,12] model (that results in the estimation of nine mating type stratification parameters $\mu_1 - \mu_9$, see Table 3) is also available. This model would be expected to be less powerful than the CEPG, PAE or HWE models, but should be more robust to any departure from mating symmetry, PAE or HWE.

EMIM reads in genotype data from input files created by PREMIM. In addition, there are two other files required by EMIM. Firstly, a file 'emimmarkers.dat', which provides the minor allele frequencies for each SNP (used as starting values in the maximization algorithm). These can optionally be estimated by PREMIM using the pedigree data, although other (e.g. population-based) sources for this information may be preferred where available. (See [7] for an investigation of EMIM's sensitivity to misspecification of the assumed or estimated allele frequencies). The other required file is a parameter file 'emimparams.dat', describing the type of analysis that EMIM should perform, which parameters to estimate, and which assumptions (such as HWE or PAE) should be made.

Implementation

PREMIM is written in C++ and for a binary pedigree file with 913 pedigrees, 1730 subjects and 45323 SNPs it takes 19 seconds to process on a Six-Core AMD Opteron^{TM} Processor with 2.6 GHz CPUs. EMIM is written in FORTRAN 77 and makes use of a subroutine MAXFUN, originally written as part of the S.A.G.E. [14] package. For these same data (pre-processed by PREMIM) on the same machine, EMIM takes 1 minute and 22 seconds to perform an analysis to test for multiplicative child genotype effects, assuming HWE. For larger data sets, EMIM and PREMIM have options that allow easy parallel processing by dividing the SNPs to analyse into different batches.

Results and discussion

Example analysis using simulated data

We used the program SimPed [15] to generate a single replicate of simulated data for 200 case/parent trios, 200 case/mother duos, 200 control/mother duos and 1000 unrelated controls at 8000 SNPs across a chromosome. We used a simplified linkage disequilibrium (LD) model that assumed LD operated in haplotype blocks, each of length 8 SNPs. We simulated child genotype effects ($R_1 = 1.5$ and $R_2 = 2.25$) at SNP 76 and maternal genotype effects ($S_1 = 2$ and $S_2 = 3$) at SNP 6004. We then used EMIM to test for maternal effects, with and without allowing for child genotype effects (Figure 1C, Figure 1A), and to test for child genotype effects, with and without allowing for maternal effects (Figure 1D, Figure 1B). In all four analyses, we see a strong signal at the correct location, with the high significance probably due to the relatively large effect sizes assumed.

A tutorial for this example (with a listing of the required commands) is available on the PREMIM and EMIM website: http://www.staff.ncl.ac.uk/richard.howey/emim/example.html

Comparison of HWE, PAE, CEPG and CPG likelihoods

The power to detect genetic effects can vary depending on the assumptions made. As a demonstration, we

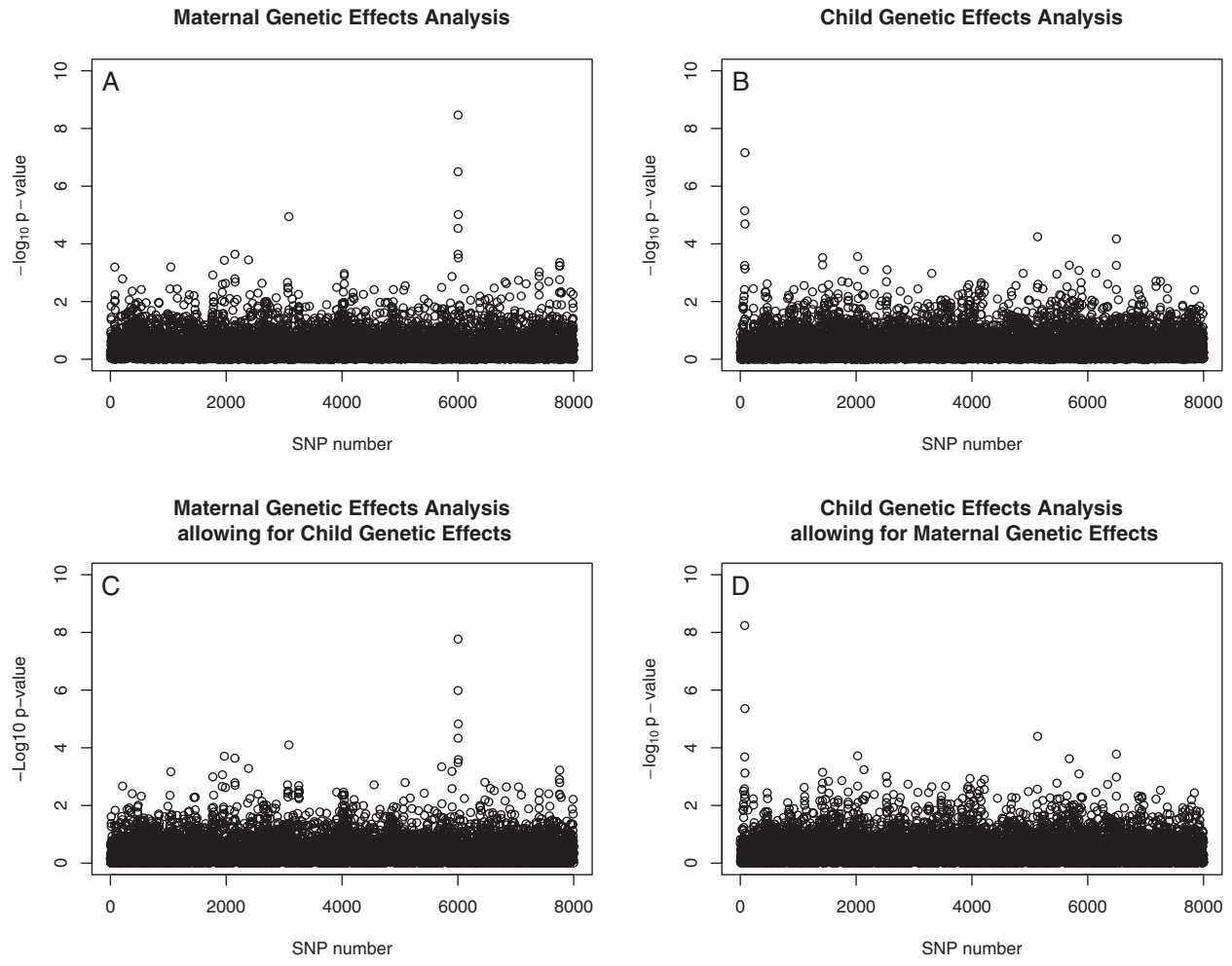

Figure 1 Genetic Effects. Plot of the $-\log_{10}$ p-values for each SNP given by EMIM to detect: (**A**) child genetic effects; (**B**) maternal genetic effects; (**C**) maternal genetic effects whilst allowing for child effects; and (**D**) child genetic effects whilst allowing for maternal effects.

simulated 1000 replicates of data at a single SNP for a sample consisting of 50 of each of the following units: case/parent trios, case/mother duos, case/father duos, control matings, control/mother duos and control/father duos. We assumed either a child genotype effect ($R_2 = 2$), a maternal genotype effect ($S_2 = 2$), or a maternal imprinting effect ($I_m = 1.8$). PREMIM and EMIM were used to estimate the parameters R_1, R_2, S_1, S_2 and I_m for each different likelihood assumption and for each set of simulated data. Figure 2(A-C) shows that the power to detect the relevant effect decreases as one makes less restrictive (but potentially more robust) assumptions, while Figure 2(D-E) shows that unbiased parameter estimation is achieved using the most restrictive assumption (HWE) (provided that assumption is correct). Similar unbiased parameter estimation is achieved for the other likelihood assumptions, when they are met (data not shown).

Effect of missing data on power

As a demonstration of the effect that missing data has on the power, we performed analyses at a single SNP using simulated data (10,000 replicates, each replicate consisting of 100 case/parent trios and 100 control/parent trios) and assuming a range of probabilities of missing genotype data. We assumed a maternal genotype effect ($S_1 = 1.5$, $S_2 = 2.25$). The expected proportion of pedigree units of different types remaining in the analysis are shown in Figure 3A and Figure 3B respectively. The trios are all present when there is no missing data, but the expected proportion quickly decreases when the probability of missing genotype data is increased. The expected proportion of the other pedigree types then increases, but subsequently decreases and converges to 0 as the probability of missing data approaches 1. The power to detect the maternal genetic effects (when correctly modelled) also decreases with increasing proportion of missing data

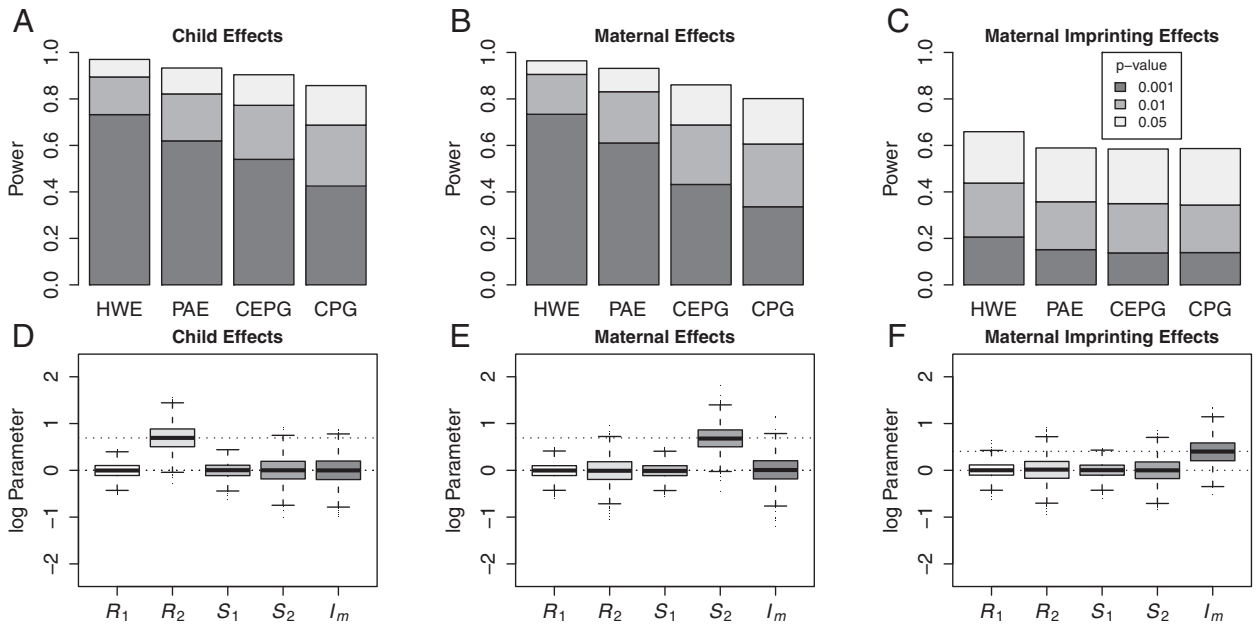

Figure 2 Comparison of Likelihood Assumptions in EMIM. Results from simulated data. (**A-C**) The power of likelihood ratio tests to achieve significance levels (p-values) of 0.001, 0.01 and 0.05 for different simulated effects and likelihood assumptions: HWE - Hardy-Weinberg Equilibrium; CEPG - Conditional on Exchangeable Parental Genotypes; CPG - Conditional on Parental Genotypes; PAE - Parental Allelic Exchangeability. (**D-F**) Box plots on a log-scale of parameter estimates for R_1, R_2, S_1, S_2 and I_m, assuming HWE. Dotted lines show the true parameter values.

(Figure 3C). An advantage of the EMIM framework is that it makes efficient use of data from all possible available individuals, allowing one to recover information even from incompletely genotyped trios.

Buyske [16] pointed out that maternal genotype effects can masquerade as child genotype effects, if analysed as such. If the maternal genetic effects are incorrectly modelled as child genetic effects (Figure 3D), we find limited power to detect these effects even when there is no missing data. Increasing the proportion of missing data has little effect on the power of this analysis, until the probability of missing genotype data becomes very large (e.g. more than 80%).

Comparison with MENDEL

Several other software packages exist that allow testing and estimation of genotype relative risk parameters similar to those tested in EMIM. One such package is MENDEL [17]. MENDEL most easily allows the estimation and testing of mother-child interaction effects via the maternal-fetal genotype incompatibility (MFG) test [5], although a "Generalized Risk" analysis that allows implementation of more complex user-defined paramaterizations (through the imposition of various parameter restrictions) is also available.

We used computer simulations (500 replicates each with 200 case parent trios) to compare the performance of MENDEL and EMIM under three different comparable models:

1. **Model 1**. This model has been used to test for RhD incompatibility [18] and estimates the relative risk corresponding to the mother having no risk alleles and the child one risk allele. MENDEL was used to estimate this one relative risk parameter by setting the sex-specific effects (parameters MFG_M and MFG_F in MENDEL) to be equal. The equivalent single parameter γ_{01} (corresponding to the parametrization of [5,18]) was estimated in EMIM. The data were simulated assuming $\gamma_{01} = 2$.
2. **Model 2**. This model has been used to test for non-inherited maternal antigens (NIMA) on rheumatoid arthritis (RA) [19] and consists of three parameters (ignoring sex-specific MFG testing): a relative risk parameter (γ_{10}) for MFG when the mother has one risk allele and the child has no risk alleles, and two parameters for child effects when the child has one or two risk alleles. In order to compare EMIM with MENDEL under this model, we used PREMIM to reassign which allele should be considered as the risk allele by EMIM. A model equivalent to MENDEL's NIMA model can then be fit in EMIM by estimating parameters (with respect to the reassigned allele) R_1, R_2 and γ_{12}. Data were simulated assuming an MFG

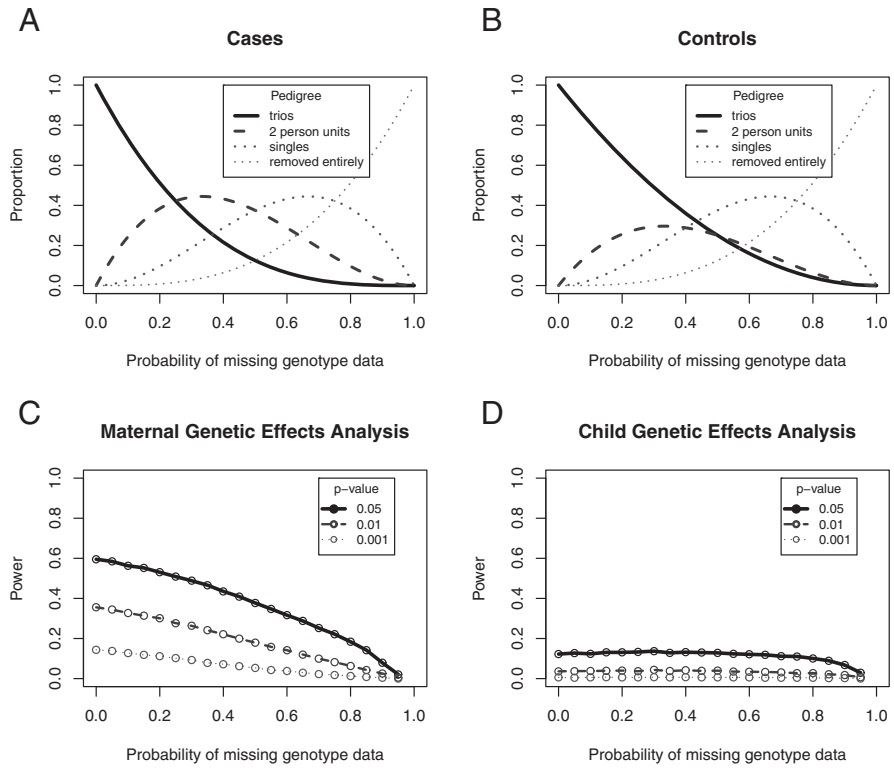

Figure 3 Effect of Missing Genotype Data. Plots showing the effect as the probability of missing genotype data is increased for data simulated with a maternal genetic effect. **A**: The expected proportions of different types of pedigree unit output by PREMIM from a set of case/parent trios, for different probabilities of missing genotypes. **B**: The expected proportions of different types of pedigree unit output by PREMIM from a set of control trios, for different probabilities of missing genotypes. **C**: Power of EMIM to detect maternal genetic effects (by estimating parameters S_1 and S_2). **D**: Power of EMIM to detect maternal genetic effects masquerading as child genetic effects (by estimating parameters R_1 and R_2).

effect $\gamma_{10} = 2$. The power to detect the the MFG effect in either MENDEL or EMIM was calculated by considering twice the difference between the negative log likelihood from a model that includes all three parameters (R_1, R_2 and the MFG parameter) and that from a model where the MFG parameter has been removed.

3. **Model 3.** This MENDEL model is a general MFG test consisting of one relative risk parameter for each of the 7 mother/child genotype combinations. The relative risk parameter denoted U_00 in the MENDEL documentation (corresponding to the situation where the mother and child have no risk alleles) was set to 1 and not estimated to avoid over-parametrization. The other 6 parameters, U_22, U_21, U_12, U_11, U_10, U_01, were estimated. The 6 parameters estimated by EMIM were R_1, R_2, S_1, S_2, γ_{11} and γ_{22}. These parameters are not indvidually equivalent to the 6 MENDEL parameters, but the models as a whole can be shown to be equivalent. Data for this comparison were simulated assuming $R_1 = S_1 = \gamma_{11} = \gamma_{22} = 1.5$ and $R_2 = S_2 = 2.25$.

Figures 4 and 5 show a comparison of the null model (no estimated parameters) and the full model log likelihoods from EMIM and MENDEL, for Models 1 and 2 respectively. EMIM was set to assume HWE (since MENDEL assumes HWE by default). We see that the null and full model log likelihoods from the two programs are are very similar (Figures 4(A), 4(B), 5(A), 5(B)), resulting in approximately equal powers and parameter estimates (Figures 4(C), 4(D), 5(C), 5(D)). For Model 3, EMIM and MENDEL similarly gave approximately equal log likelihoods and powers (results not shown).

One difference between EMIM and MENDEL was the time taken to perform the analysis, with EMIM performing considerably quicker than MENDEL. For example, the time to run model 3 (with 200 case/parent trios) showed that PREMIM and EMIM combined took 0.0257 seconds and MENDEL took 6.45 seconds (averaged over 300 runs). This shows that PREMIM and EMIM combined were approximately 250 times faster than MENDEL in this example. The same analysis with 400 case/parent trios gave times of 0.0302 seconds for PREMIM and EMIM combined and 14.3 seconds for MENDEL (averaged over

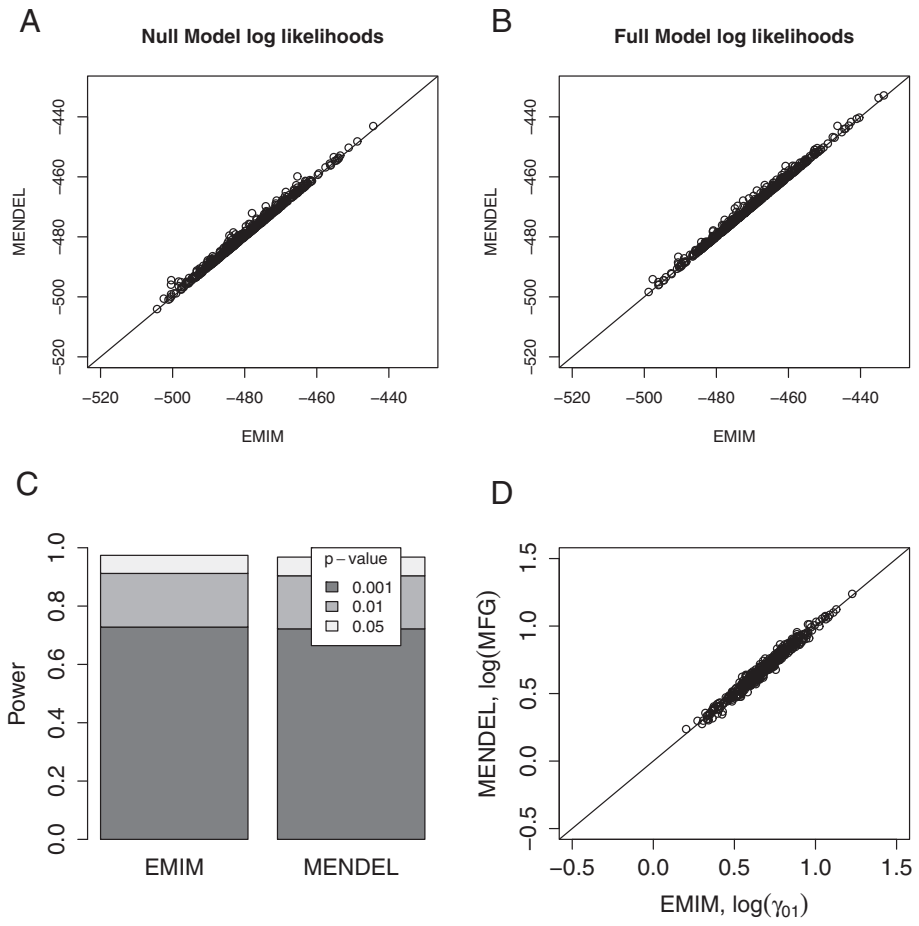

Figure 4 Mother-Child Interaction Effects Comparison with MENDEL, RHD. Plots showing the comparison of EMIM and MENDEL - "option 26, Model 1: RHD" using simulated data. **A**: Plot of the null model log likelihood values calculated using EMIM and MENDEL. **B**: Plot of the full (alternative) log likelihood values calculated using EMIM and MENDEL. **C**: The power to detect a genetic effect for p-values of 0.05, 0.01 and 0.001. **D**: Plot of the MFG parameter estimates calculated using EMIM and MENDEL.

300 runs), showing PREMIM and EMIM to be approximately 472 times faster then MENDEL. A possible reason for the difference in running times is the fact that the extended MFG model [11] implemented in MENDEL is a slightly more complicated model than the parent/offspring trio model implemented in EMIM (thus providing MENDEL with the ability to analyse larger pedigrees).

Comparison with LEM

Another program with the capability to analyse complex genetic effects (most notably mother/child/imprinting effects) is LEM [20]. LEM is a Windows-based log-linear modelling program designed primarily to be used via a graphical user interface, although it is possible to run it from the DOS command line, in order to implement scripts that allow the analysis of large numbers of loci or replicates. LEM takes an input parameter file which defines the model, the parameters to be estimated and the name of the input data file. We created input parameter and data files based on examples provided by the authors of LEM [20] for case/parent trios and by [21] for case/mother and control/mother duos.

1. **Case/parent trios.** SimPed [15] was used to simulate a single replicate of data at 8000 SNPs across a chromosome for 4000 case/parent trios. Child effects ($R_1 = 1.5, R_2 = 2.25$) were simulated at SNP number 1004 and maternal effects ($S_1 = 2, S_2 = 3$) were simulated at SNP number 6004. In both EMIM and LEM we tested for maternal effects while allowing for child and maternal imprinting effects (i.e. we compared an alternative 5-parameter model (R_1, R_2, S_1, S_2, I_m) with a null 3-parameter model (R_1, R_2, I_m)). We calculated the p-value for LEM on the basis of the reported log likelihoods by using the Wald

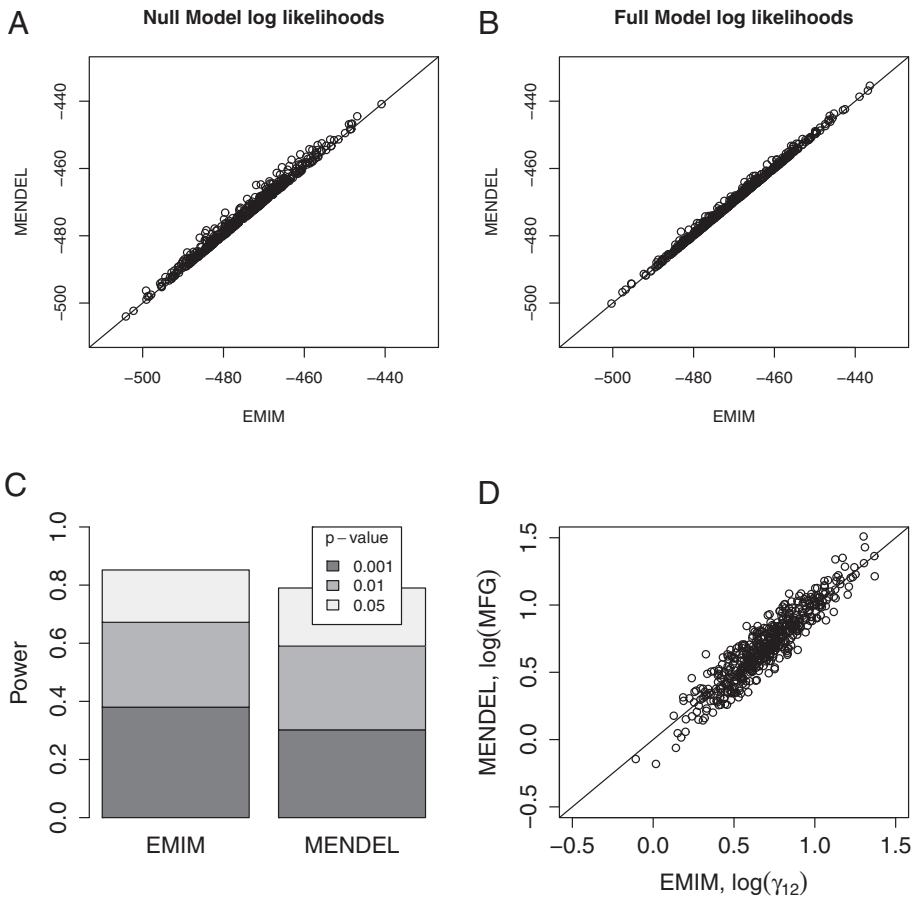

Figure 5 Mother-Child Interaction Effects Comparison with MENDEL, NIMA. Plots showing the comparison of EMIM and MENDEL - "option 26, Model 2: NIMA" using simulated data. **A**: Plot of the null model log likelihood values (with child effects fitted, but no MFG interaction effect) calculated using EMIM and MENDEL. **B**: Plot of the full (alternative) log likelihood values (fitting both child effects and MFG effect) calculated using EMIM and MENDEL. **C**: The power to detect the MFG effect for p-values of 0.05, 0.01 and 0.001. **D**: Plot of the MFG parameter estimates calculated using EMIM and MENDEL.

statistic as a χ^2 value with 2 degrees of freedom. (The p-value reported by LEM was not suitable as it is only given to 3 decimal places, which was insufficient for SNPs with p-values less than 10^{-3}).

2. **Case/mother duos and control/mother duos.** Again, data were simulated at 8000 SNPs but this time for 2000 case/mother duos and 2000 control/mother duos. Child effects ($R_1 = 1.5$, $R_2 = 2.25$) were simulated at SNP number 1000 and maternal effects ($S_1 = 2, S_2 = 3$) were simulated at SNP number 6004. In both EMIM and LEM we tested for maternal and child effects i.e. we compared a null model with no fitted parameters to an alternative model with parameters (R_1, R_2, S_1, S_2).

A comparison of EMIM versus LEM for the case/mother and control/mother duos is shown in Figure 6. Figure 6(A) and 6(B) show that the p-values across the chromosome appear to be indistinguishable, and Figure 6(G) shows that the p-values for each SNP from the two programs are indeed approximately equal. Figures 6(C) and 6(E) show that the estimates of R_1 and S_1 are approximately equal and Figures 6(D) and 6(F) show that R_2 and S_2 are also approximately equal, but with more variability.

Figure 7 shows the same plots for the case/parent trios, but with the addition of estimates for the extra parameter I_m. We see that the p-values and parameter estimates provided by the two programs are virtually indistinguishable.

These results indicate that the inference provided by LEM and EMIM is essentially identical. This is as expected given the mathematical equivalence [7,22] between the multinomial model fit by EMIM and the log linear model fit by LEM. The main difference between the programs is the time taken to perform the analysis, with EMIM performing considerably quicker than LEM. For example, the

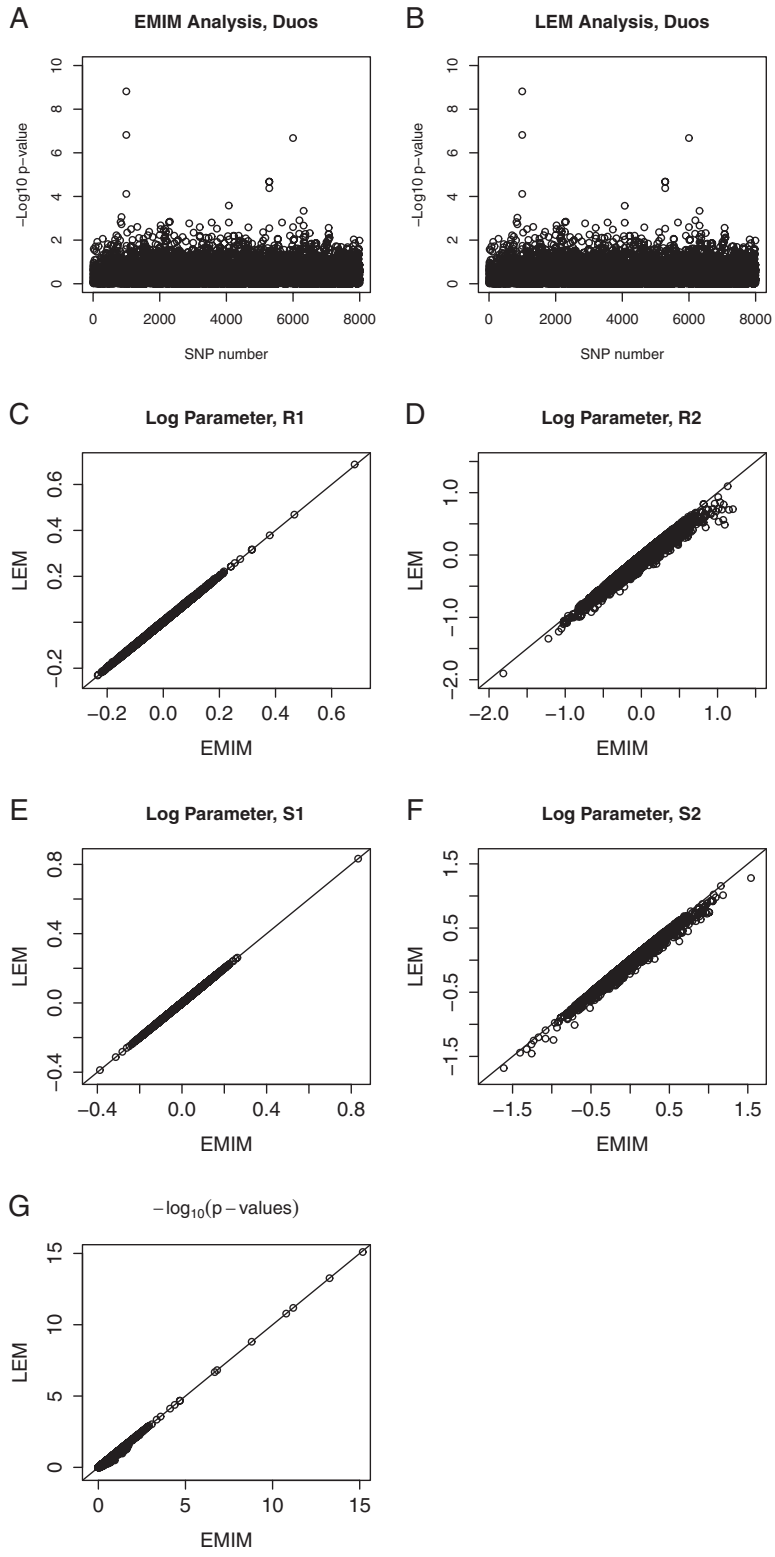

Figure 6 Comparison of EMIM and LEM for Child/Mother Duos. Plots showing the comparison of EMIM and LEM using simulated data for 2000 case/mother duos and 2000 control/mother duos, assuming $R_1 = 1.5$ and $R_2 = 2.25$ at SNP number 1000 and $S_1 = 2$ and $S_2 = 3$ at SNP number 6004. Plots of the $-\log_{10}$ p-values for each SNP to detect child and maternal effects by: **A**: EMIM and **B**: LEM. Plots of the log parameters values for: **C**: R_1; **D**: R_2; **E**: S_1 and **F**: S_2. **G**: Plot of the $-\log_{10}$ p-values for the alternative versus the null model calculated using EMIM and MENDEL.

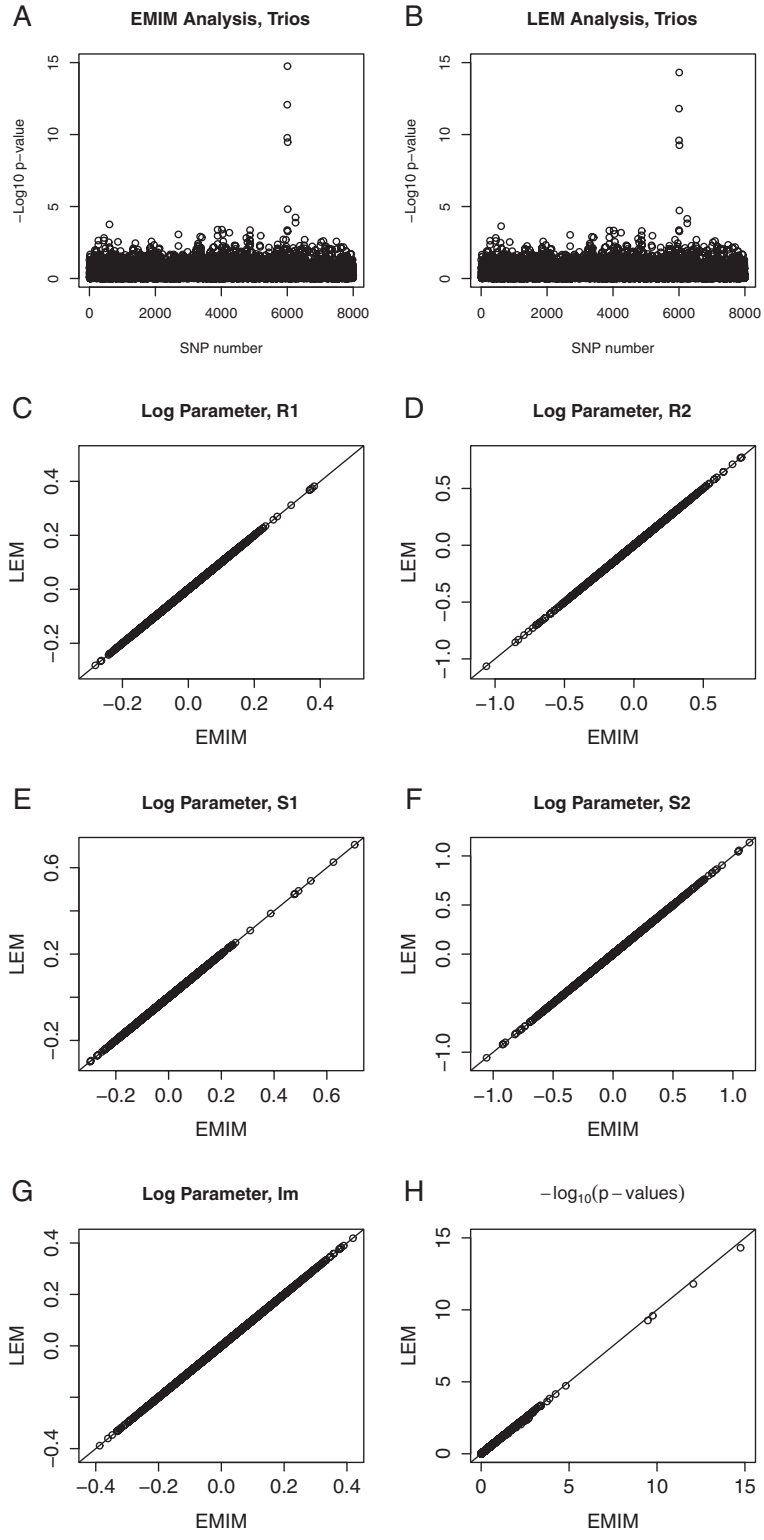

Figure 7 Comparison of EMIM and LEM for Case/Parent Trios. Plots showing the comparison of EMIM and LEM using simulated data for 4000 case/parent trios, assuming $R_1 = 1.5$ and $R_2 = 2.25$ at SNP number 1004 and $S_1 = 2$ and $S_2 = 3$ at SNP number 6004. Plots of the $-\log_{10}$ p-values for each SNP to detect maternal effects given child and imprinting effects by: **A**: EMIM and **B**: LEM. Plots of the log parameters values for: **C**: R_1; **D**: R_2; **E**: S_1; **F**: S_2; **G**: I_m. **H**: Plot of the $-\log_{10}$ p-values for the alternative (5-parameter) model versus the null (3-parameter) model calculated using EMIM and MENDEL.

time taken to run the case/mother and control/mother duos analysis across 8000 SNPs in PREMIM/EMIM was 1 minute 21 seconds on a Linux machine (6-Core AMD Opteron^{TM} Processor with 2.6 GHz CPUs) or 2 minutes 4 seconds on Windows (using a 2-core Intel^{TM} Processor with 2.93 GHz CPUs), whereas the same analysis in LEM took 16 hours, 52 minutes and 8 seconds on Windows (via the DOS command line). The difference in speed between the two programs for the case/parent trios analysis was not as extreme, with PREMIM/EMIM taking 3 minutes 7 seconds on Linux or 4 minutes 49 seconds on Windows, versus LEM's time of 63 minutes 58 seconds on Windows. The improved speed for the LEM trios analysis was most likely due to the fact that it took fewer steps than the duos analysis during the likelihood maximization process (possibly on account of the fact that the example parameter file we were using requested the program to switch to using a Newton-Raphson algorithm following 10 iterations of an EM algorithm). It is possible that differences between maximization algorithms and convergence criteria could account for some of the differences in speed between PREMIM/EMIM and LEM; we found it difficult to determine how to obtain precise control over such factors in LEM and were forced to use input files that very closely matched the examples provided by [20,21]. Another factor influencing speed could be the fact that LEM does not (as far as we are aware) allow the input of multiple SNPs simultaneously, meaning that we had to create and read into LEM a separate input file for each SNP analysed.

Conclusions

Here we have presented two new computer tools, PREMIM and EMIM, for the estimation of parental and child genetic effects, based on genotype data from a variety of different child-parent configurations. The current version of EMIM improves upon the early beta version described in [7] by allowing a larger set of possible child-parent configurations, a larger range of optional likelihood assumptions, and by the development of the companion program, PREMIM, for generating the required input files from standard PLINK-format files, considerably improving the ease with which EMIM can be applied to real data.

In application to simulated data, we have shown that the inference provided by EMIM is essentially equivalent to that provided by alternative (competing) software packages such as MENDEL and LEM. EMIM does have the advantage of allowing easy implementation of a wider class of models than are most easily implemented in MENDEL and LEM, although the expert MENDEL/LEM user could probably achieve the same model flexibility through judicious choice of parameter restrictions. However, PREMIM and EMIM (used in combination) considerably outperform MENDEL and LEM in terms of speed of execution, an advantage that is likely to be all the more important when applying these approaches to large-scale data sets such as those generated in genome-wide association studies. To allow further increases in speed, PREMIM and EMIM also have the advantage of allowing easy parallel processing (e.g. on a computer cluster) by dividing the SNPs to analyse into different batches.

Limitations of PREMIM and EMIM include the fact that larger pedigrees are divided into case/parent or control/parent trios (or smaller sub-units) prior to analysis, and the fact that SNPs are analysed one at a time, without borrowing information from neighbouring markers (e.g. on the basis of regional linkage disequilibrium patterns). Methods for dealing with larger pedigrees, valid under the assumptions of random mating and/or Hardy-Weinberg equilibrium (HWE), have been described by [10,11], while [23] present an approach that models haplotypes rather than individual SNPs, allowing the borrowing of information (including information on parent-of-origin or missing genotype data) across neighbouring SNPs. Both of these features would be valuable additions to future releases of our software. Nevertheless, the current versions of EMIM and PREMIM provide easy-to-use command-line tools for the analysis of pedigree data, allowing testing and estimation of a variety of parental and child genotype relative risks.

Availability and requirements

Project name: EMIM and PREMIM

Project home page: http://www.staff.ncl.ac.uk/richard.howey/emim/

Operating systems: Windows and Linux executables; FORTRAN and C++ source code

Programming language: FORTRAN and C++

Other requirements: None

Licence: GNU General Public License

Any restrictions to use by non-academics: None

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RH developed the PREMIM software, performed computer simulations and drafted the manuscript. HJC conceived the experiment, developed the EMIM software and revised the manuscript. Both authors read and approved the final manuscript.

Author's information

HJC is Professor of Statistical Genetics and a Wellcome Senior Fellow at the Institute of Genetic Medicine, Newcastle University, UK. RH is a Research Associate at the Institute of Genetic Medicine, Newcastle University, UK.

Acknowledgements

This work was supported by the Wellcome Trust (Grant reference 087436) and by the European Community's 7th Framework Programme contract ('CHeartED') HEALTH-F2-2008-223040. Some of the results of this paper were obtained by using the program package S.A.G.E., which was supported by a U.S. Public Health Service Resource Grant (RR03655) from the National Center for Research Resources.

Received: 14 February 2012 Accepted: 9 June 2012
Published: 27 June 2012

References

1. Weinberg CR, Umbach DM: **A hybrid design for studying genetic influences on risk of diseases with onset early in life.** *Am J Hum Genet* 2005, **77:**627–636.
2. Shi M, Umbach DM, Vermeulen SH, Weinberg CR: **Making the most of case-mother/control-mother studies.** *Am J Epidemiol* 2008, **168:**541–547.
3. Li S, Lu Q, Fu W, Romero R, Cui Y: **A regularized regression approach for dissecting genetic conflicts that increase disease risk in pregnancy.** *Stat Appl Genet Mol Biol* 2009, **8:**45.
4. Weinberg CR: **Methods for detection of parent-of-origin effects in genetic studies of case-parents triads.** *Am J Hum Genet* 1999, **65:**229–235.
5. Sinsheimer JS, Palmer CG, Woodward JA: **Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test.** *Genet Epidemiol* 2003, **24:**1–13.
6. Cordell HJ: **Properties of case/pseudocontrol analysis for genetic association studies: Effects of recombination, ascertainment, and multiple affected offspring.** *Genet Epidemiol* 2004, **26:**186–205.
7. Ainsworth HF, Unwin J, Jamison DL, Cordell HJ: **Investigation of maternal effects, maternal-foetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring.** *Genet Epidemiol* 2011, **35:**19–45.
8. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker, P I, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559–575.
9. Healy J, Bourgey M, Richer C, Sinnett D, Roy-Gagnon MH: **Detection of fetomaternal genotype associations in early-onset disorders: evaluation of different methods and their application to childhood leukemia.** *J Biomed Biotechnol* 2010, **2010:**369534.
10. Childs EJ, Sobel EE, Palmer CG, Sinsheimer JS: **Detection of intergenerational genetic effects with application to HLA-B matching as a risk factor for schizophrenia.** *Hum Hered* 2011, **72:**161–172.
11. Childs EJ, Palmer CG, Lange K, Sinsheimer JS: **Modeling maternal-offspring gene-gene interactions: the extended-MFG test.** *Genet Epidemiol* 2010, **34:**512–521.
12. Cordell HJ, Barratt BJ, Clayton DG: **Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions and parent-of-origin effects.** *Genet Epidemiol* 2004, **26:**167–185.
13. Weinberg CR, Wilcox AJ, Lie RT: **A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting.** *Am J Hum Genet* 1998, **62:**969–978.
14. SAGE: **Statistical analysis for genetic epidemiology, release 2.2. computer program package obtained from the Department of Biometry and Genetics, LSU Medical Center, New Orleans.** 1994.
15. Leal SM, Yan K, Müller-Myhsok B: **SimPed: a simulation program to generate haplotype and genotype data for pedigree structures.** *Hum Hered* 2005, **60:**119–122.
16. Buyske S: **Maternal genotype effects can alias case genotype effects in case-control studies.** *Eur J Hum Genet* 2008, **16:**784–785.
17. Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, Sobel E: **Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets.** *Am J Hum Genet* 2001, **69:**A504.
18. Palmer CG, Turunen JA, Sinsheimer JS, Minassian S, Paunio T, Lönnqvist J, Peltonen L, Woodward JA: **RHD maternal-fetal genotype incompatibility increases schizophrenia susceptibility.** *Am J Hum Genet* 2002, **71:**1312–1319.
19. Hsieh HJ, Palmer CG, Harney S, Newton JL, Wordsworth P, Brown MA, Sinsheimer JS: **The v-MFG test: investigating maternal, offspring and maternal-fetal genetic incompatibility effects of disease and viability.** *Genet Epidemiol* 2006, **30:**333–347.
20. van DenOord, E J, Vermunt JK: **Testing for linkage disequilibrium, maternal effects, and imprinting with (in)complete case-parent triads, by use of the computer program LEM.** *Am J Hum Genet* 2000, **66:**335–338.
21. Weinberg CR, Shi M: **The genetics of preterm birth: Using what we know to design better association studies.** *Am J Epidemiol* 2009, **170:**1373–1381.
22. Baker SG: **The multinomial–Poisson transformation.** *The Statistician* 1994, **43:**495–504.
23. Gjessing HK, Lie RT: **Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes.** *Ann Hum Genet* 2006, **70:**382–396.

doi:10.1186/1471-2105-13-149
Cite this article as: Howey and Cordell: PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. *BMC Bioinformatics* 2012 **13**:149.