`libfbat`: a C++ library for family based association testing

Grégory Nuel[1,2], Yousri Slaoui[1] and Vincent Miele[1,3]

[1] Laboratoire Statistique et Génome, university of Evry, UMR CNRS 8071, UMR INRA 1152, Tour Evry II, 523 place des Terrasses, F-91000 Evry, France
yslaoui@genopole.cnrs.fr

[2] MAP5, University Paris Descartes, UMR CNRS 8145, 45 rue des Saints-Pères, F-75006 Paris, France
nuel@math-info.univ-paris5.fr

[3] Laboratoire Biometrie et Biologie Evolutive, UMR CNRS 5558, UCB Lyon 1, Bât. Grégor Mendel, 43 bd du 11 novembre 1918, F-69622 Villeurbanne, France
miele@biomserv.univ-lyon1.fr

Abstract: *We propose here a Family Based Association Testing (FBAT) programming library which is robust both to genotyping errors and missing genotypes. The underlying method consists in considering an incomplete data model where the true unobserved genotypes produce the observed one through an explicit genotyping error model. We propose an EM algorithm to estimate the parameters of this model and the posterior distribution of the true genotypes. This distribution may be used to detect and correct genotyping errors or to deal with missing genotypes in FBAT statistics. As a validation, we compare our method to GMCheck (dedicated software to detect genotyping errors in pedigree) on a complex pedigree including loops where our new approach seems to display similar or best performances which is very encouraging.*

Keywords: genetic epidemiology, genotyping errors, missing genotypes, EM algorithm.

1 Introduction

In genetic epidemiology it is frequent to consider families of individuals for which we have genotypes (ex: value for a bi-allelic Single Nucleotide Polymorphism – SNP) and phenotypes (ex: disease affection status, value of a phenotypic quantitative trait). The statistical challenge then consists in finding the genotypic markers that are significantly associated to the studied phenotypes.

A classical approach to this problem is the general framework of Family Based Association Testing (FBAT) [1,2,3], which is widely used in a form or another. The idea of FBAT is to combine both genotypes and phenotypes in a statistic (using a coding function for the genotypes) and then to test the association by comparing the observed value to the distribution under the null hypothesis where the genotypes are distributed conditionally to their ancestral ones.

This interesting approach faces two common difficulties: How to deal with genotyping errors[4] ? How to treat missing genotypes ? For the former question, a common approach consists in detecting Mendelian inconsistencies with a software like Pedcheck [4] and then to consider that the corresponding genotypes are missing. All missing genotypes (both those coming from Mendelian inconsistencies and the real ones) are then inferred in FBAT [5] through a Bayesian framework with uniform priors.

[4] Pedigree errors may also occur but this errors are usually quite easy to detect and correct when a large set of genotyping markers is used.

In this paper, we consider an incomplete data model where true (unobserved) genotypes produce the observed ones through an explicit genotyping error model. We propose an Expectation-Maximization (EM) algorithm [6] allowing to estimate the parameters of our model as well as producing the posterior distribution of the true genotypes (see §2). This distribution can be useful to detect genotypic error in pedigree (see §4). However, our objective is rather here to use this distribution to propose an implementation of FBAT which is robust both to genotyping errors and missing genotypes (see §3).

2 Parameter estimation

2.1 Notations

Pedigree structure Let $\mathcal{I} = \{1, \dots, n\}$ be a set of individuals. We denote by F_i (resp. M_i) the father (resp. mother) of individual $i \in \mathcal{I}$, if the father (resp. mother) is unkown, F_i (resp. M_i) take the value '?'. We then introduce the parent set $\mathcal{P}_i$ of individual $i \in \mathcal{I}$ which is recursively defined by and $\mathcal{P}_i = \{i\} \cup \mathcal{P}_{F_i} \cup \mathcal{P}_{M_i}$ (with the convention that $\mathcal{P}_? = \emptyset$). Two individuals $i, j \in \mathcal{I}$ then belong to the same family if only and only if $\mathcal{P}_i \cap \mathcal{P}_j \neq \{?\}$. Let $\mathcal{F}_1 \cup \ldots \cup \mathcal{F}_k$ be a partition of $\mathcal{I}$ in k disjoint families. An individual $i \in \mathcal{I}$ such as $F_i = M_i =?$ is called a founder. We assume that the parents of all individuals are known except for the founders.

Phenotypes and Genotypes Let us denote by $\varphi_i \in \mathbb{R}$ the phenotype of individual $i \in \mathcal{I}$ and by $g_{s,i} \in \mathcal{G}$ either the genotype of individual i at marker $s \in \mathcal{S} = \{1, \dots, N\}$, where $\mathcal{G}$ is the set of possible genotypes, or '?' if this data is missing. From now on, and for the sake of simplicity, we only consider the bi-allelic case $\mathcal{G} = \{\text{aa}, \text{aA}, \text{AA}\}$ (but we are not restricted to this particular cases).

2.2 Model

True genotypes Let $G^*_{s,i} \in \mathcal{G}$ for all $s \in \mathcal{S}$ and $i \in \mathcal{I}$ be the true random genotypes and $G_{s,i}$ (with $i \in \mathcal{I}$ such as $g_{s,i} \neq ?$) the observed ones. We assume that the true genotypes are independant in s and between distinct families. However, within each family, the founders' true genotypes are supposed independant but the offsprings are distributed conditionally to their parents. For simplification purpose we assume Hardy-Weinberg equilibrium (HWE) for the true genotypes of the founders.

Formally we get for all $i \in \mathcal{I}$, $s \in \mathcal{S}$ and $g \in \mathcal{G}$ that: $\mathbb{P}(G^*_{s,i} = g | G^*_{s,F_i} = f, G^*_{s,M_i} = m) = \text{offspring}(f, m, g)$ for the non-founders and $\mathbb{P}(G^*_{s,i} = g) = D_s(g)$ for the founders where D_s, the probability distribution function of the genotype of the marker $s \in \mathcal{S}$, is given by:

$$D_s(g) = \begin{cases} p_s^2 & \text{if } g = \text{aa} \\ 2(1-p_s)p_s & \text{if } g = \text{aA} \\ (1-p_s)^2 & \text{if } g = \text{AA} \end{cases}$$

with p_s the probability of allele 'a' for marker s, and, for all $f, m, g \in \mathcal{G}$, $\text{offspring}(f, m, g)$ is the conditional probability for parents with true genotypes f and m to have an child with genotype g (ex: $\text{offspring}(\text{aa}, \text{aa}, \cdot) = (1; 0; 0)$, $\text{offspring}(\text{aa}, \text{aA}, \cdot) = (1/2; 1/2; 0)$, $\text{offspring}(\text{aa}, \text{AA}, \cdot) = (0; 1; 0)$, $\text{offspring}(\text{aA}, \text{aa}, \cdot) = (1/2; 1/2; 0)$, $\text{offspring}(\text{aA}, \text{aA}, \cdot) = (1/4; 1/2; 1/4)$, etc.).

Let us note that the HWE assumption has for consequence that:

$$D_s(g) = \sum_{f,m \in \mathcal{G}} D_s(f) D_s(m) \text{offspring}(f, m, g) \quad \forall s \in \mathcal{S}, g \in \mathcal{G}.$$

```
$\varepsilon = 0.01$ and $p_s = 0.5$ for all $s \in \mathcal{S}$ // arbitrary initialization
while parameter has not converged do
  $\texttt{nerror} = 0$
  for all $s \in \mathcal{S}$ do
    $\texttt{nallele} = 0$
    for $j = 1, \dots, k$ do
      $\texttt{nlocalerror} = 0, \texttt{nlocalallele} = 0$, and $\texttt{normalization} = 0$
      for all possible values of $g^*_{s,i}$ for $i \in \mathcal{F}_j$ do
        compute $\texttt{proba} = P(\mathcal{F}_j)$
        $\texttt{normalization}+ = \texttt{proba}$
        $\texttt{nlocalerror}+ = \texttt{proba} \times \sum_{i \in \mathcal{F}'_j} \mathbb{I}\{g_{s,i} \neq g^*_{s,i}\}$
        $\texttt{nlocalallele}+ = \texttt{proba} \times \sum_{i \in \mathcal{F}_j} \left(2 \times \mathbb{I}\{g^*_{s,i} = \text{aa}\} + \mathbb{I}\{g^*_{s,i} = \text{aA}\}\right)$
      end for
      $\texttt{nerror}+ = \texttt{nlocalerror}/\texttt{normalization}$
      $\texttt{nallele}+ = \texttt{nlocalallele}/\texttt{normalization}$
    end for
    $p_s = \texttt{nallele}/(2 \times |\{i \in \mathcal{I}, F_i = M_i =?\}|)$
  end for
  $\varepsilon = \texttt{nerror}/(\sum_{j=1}^{k} |\mathcal{F}'_j|)$
end while
```

Algorithm 1: EM estimation of the model parameters.

Observed genotypes For all $i \in \mathcal{I}$ and $s \in \mathcal{S}$ such as $g_s(i) \neq ?$ we have:

$$\mathbb{P}(G_s(i) = g | G^*_s(i) = g^*) = (1 - \varepsilon)\mathbb{I}_{g=g^*} + \frac{\varepsilon}{2}\mathbb{I}_{g \neq g^*}$$

where $\varepsilon \in [0, 1]$ is the probability of a genotyping error to occur. More complex error models can obviously be defined (ex: error rate depending on g^*) but the simpler model we consider here is illustrative enough to present our method.

2.3 EM algorithm

For any SNP $s \in \mathcal{S}$ and for all family $\mathcal{F}_j$ we define:

$$\begin{aligned} P(\mathcal{F}_j) &= \mathbb{P}(G^*_{s,i} = g^*_{s,i} \text{ for } i \in \mathcal{F}_j \text{ and } G_{s,i} = g_{s,i} \text{ for } i \in \mathcal{F}'_j) \\ &= \mathbb{P}(G^*_{s,i} = g^*_{s,i} \text{ for } i \in \mathcal{F}_j) \prod_{i \in \mathcal{F}'_j} \mathbb{P}(G_{s,i} = g_{s,i} | G^*_{s,i} = g^*_{s,i}) \end{aligned}$$

with $\mathcal{F}'_j = \{i \in \mathcal{F}_j, g_{s,i} \neq ?\}$. One should note that the expression of $\mathbb{P}(G^*_{s,i} = g^*_{s,i} \text{ for } i \in \mathcal{F}_i)$ is complicated since it depends on the pedigree structure of the family and is computed through the conditional probabilities defined above.

The parameters of our model $(p_s)_{s \in \mathcal{S}}$ and ε. We estimate them through a classical EM framework [6] with Algorithm 1.

3 FBAT's statistics

Let $X : \mathcal{G}^h \to \mathbb{R}^d$ be a coding function corresponding to a given phenotypic model (ex: additive, genotypic, recessive, etc.). We first assume that $h = 1$ (only one marker at a time is considered)

```
$\texttt{stat} = 0, \texttt{expectation} = 0$, and $\texttt{variance} = 0$
for $j = 1, \ldots, k$ do
    $\texttt{localstat} = 0, \texttt{localexpectation} = 0$ and $\texttt{localvariance} = 0$
    $\texttt{normalization} = 0$
    for all possible values of $g^*_{s,i}$ for $i \in \mathcal{F}_j$ do
        compute $\texttt{proba} = \mathcal{P}(F_j)$
        $\texttt{normalization}+ = \texttt{proba}$
        $\texttt{localstat}+ = \texttt{proba} \times \left[\sum_{i \in \mathcal{I}' \cap \mathcal{F}_j} (\varphi_i - \text{offset}) \times X(g^*_{s,i})\right]$
        $\texttt{localexpectation}+ = \texttt{proba} \times \left(\sum_{i \in \mathcal{I}' \cap \mathcal{F}_j} (\varphi_i - \text{offset}) \times \mathbb{E}[X(\mathit{\Gamma}^*_{s,i})]\right)$
        $\texttt{localvariance}+ = \texttt{proba} \times \left(\sum_{i \in \mathcal{I}' \cap \mathcal{F}_j} (\varphi_i - \text{offset}) \times \mathbb{V}[X(\mathit{\Gamma}^*_{s,i})]\right)$
    end for
    $\texttt{stat}+ = \texttt{localstat}/\texttt{normalization}$
    $\texttt{expectation}+ = \texttt{localexpectation}/\texttt{normalization}$
    $\texttt{variance}+ = \texttt{localvariance}/(\texttt{normalization} \times \texttt{normalization})$
end for
```

Algorithm 2: FBAT's statistics computation.

and $d = 1$ then we will discuss the extension of our result to more complex coding functions. For example, one can consider the following coding function:

$$X(\text{aa}) = 2 \quad X(\text{aA}) = 1 \quad X(\text{AA}) = 0$$

which is closely related to the additive model (but not exactly since this particular function depend on the choice of allele 'a').

We define the FBAT's statistic $t(s)$ of the marker s by:

$$t(s) = \sum_{g^*_s} \left[\sum_{i \in \mathcal{I}'} (\varphi_i - \text{offset}) \times X(g^*_{s,i}) \right] \mathbb{P}(G^*_s = g^*_s | G_s = g_s)$$

where offset $\in \mathbb{R}$ is a constant (ex: offset $= 0$), $\mathcal{I}' = \{i \in \mathcal{I}, \varphi_i \neq ?\}$, $g^*_s = \{g^*_{s,i}, i \in \mathcal{I}\}$ (similar notation for G^*_s) and $g_s = \{g_{s,i}, i \in \mathcal{I}'\}$ (similar notation for G_s).

In order to use this statistic in a testing framework, we want to compare $t(s)$ to the distribution of

$$T(s) = \sum_{g^*_s} \left[\sum_{i \in \mathcal{I}'} (\varphi_i - \text{offset}) \times X(\mathit{\Gamma}^*_{s,i}) \right] \mathbb{P}(G^*_s = g^*_s | G_s = g_s)$$

where all $\mathit{\Gamma}^*_{s,i}$ are independant and distributed according to:

$$\mathbb{P}(\mathit{\Gamma}^*_{s,i} = \gamma) = \begin{cases} \text{offspring}(g^*_{s,F_i}, g^*_{s,M_i}, \gamma) & \text{if } F_i \neq ? \text{ and } M_i \neq ? \\ D_s(\gamma) & \text{if } F_i = ? \text{ and } M_i = ? \end{cases} .$$

We then define the normalized Z-score z_s by:

$$z_s = \frac{t(s) - \mathbb{E}[T(s)]}{\sqrt{\mathbb{V}[T(s)]}}$$

which can be computed with Algorithm 2.

These result naturally extends to more complex coding function ($h > 1$ or $d > 1$) by resulting in a multidimensional FBAT's statistic. In such a case however, it is then necessary to compute both expectation and covariance matrix of the multidimensional statistic in order to perform a classical chi-square normalization instead of the Gaussian normalization which is done above.

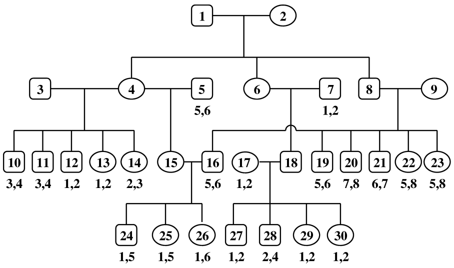

Figure 1. A complex pedigree (including loops) with genetic markers with 8 alleles and missing data (source: [8]) .

4 Application

We want here to compare our Algorithm 1 to other methods especially designed to detect (and correct) genotyping errors in pedigree like Pedcheck [4], Merlin [7] or GMCheck [8]. This last software is the most recent one and is supposed to outperform the previous ones both in terms of reliability and in its ability to consider complex pedigree (including loops for example). We hence consider in Figure 1 the same pedigree problem than in [8].

We have here a total of 8 free parameters to estimate: 7 for the 8 alleles and one for the genotyping error probability ε. Using Algorithm 1 we get $\varepsilon = 0.0476$ and

a	1	2	3	4	5	6	7	8
$\mathbb{P}(\text{allele} = a)$	0.2302	0.2155	0.0725	0.0748	0.1310	0.1310	0.0725	0.0725

Like in [8] we found that this pedigree have at least one genotyping error and that this error should only appear among individuals $5, 7, 14, 16$ and 28. Like in [8] we consider the two individuals with the higher error probability: 28 and 14. Here is the posterior distribution of genotypes for these two individuals (GMCheck values are given in parenthesis, only genotypes with probability higher than 2% are displayed): $\mathbb{P}(g^*_{28} = 1,1) = 0.218$ (0.258), $\mathbb{P}(g^*_{28} = 1,2) = 0.307$ (0.364), $\mathbb{P}(g^*_{28} = g_{28} = 2,4) = 0.288$ (0.157), $\mathbb{P}(g^*_{28} = 2,2) = 0.089$ (na) and $\mathbb{P}(g^*_{14} = g_{14} = 3,2) = 0.935$ (0.922).

We can see that both our method and GMCheck seems to give similar results. However, more intensive validation work (simulation studies for example) would be necessary to backup properly this assertion.

One should note that the computations are here performed assuming that at most one genotyping error may arise in the pedigree. This assumption may seems very restrictive, but we also have performed the same computations allowing two genotyping errors instead of one and the results only differ slightly with an absolute error of at most 0.005 (data not shown). This is an evidence that our approximated results allowing at most one error is very similar to the unconstrained one. One should not that for bi-allelic markers or smaller pedigree the full unconstrained computation remain tractable.

5 Conclusion

The method we propose here both allows to detect genotyping errors and to produce FBAT statistics robust to these errors and missing genotypes. Despite the fact that it is not its main purpose, our algorithm seems to displays similar performance than a state-of-the-art dedicated software like GM-Check [8] for the problem of detecting and correcting genotyping errors, which is an encouraging result. However, the real purpose of this new method is to be applied to FBAT. On this matter, a full comparison with classical FBAT implementation will be done soon.

Let us add that our methods are implemented into a programming library called `libfbat` which is written in ANSI `C++` and developed on x86 GNU/Linux systems with GCC 4.1.3. Compilation and installation are compliant with the GNU standard procedure. The library is free and will be soon available on the web. R bindings will also be developed. `libfbat` is licensed under the GNU General Public License [9].

Acknowledgements

We would like to thanks the member of the *Laboratoire Statistique et Génome* and especially Bernard Prum for our countless useful and constructive discussions.

References

[1] N. Laird, S. Horvath and X. Xu, Implementing a unified approach to family based tests of association. *Genet Epidemiol*, 19(Suppl 1):S36-S42, 2000.

[2] C. Lange, E. K. Silverman, X. Xu, S. T. Weiss and N. M. Laird, A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics*, 4(2):195-206, 2003.

[3] X. Xu, C. Rakovski, X. Xu, N. Laird, An efficient family-based association test using multiple markers. *Genet Epidemiol*, 30:620-626, 2006.

[4] J. R. O'Connell and D. E. Weeks, PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet*, 63(1):259-266, 1998.

[5] FBAT web page, http://www.biostat.harvard.edu/ fbat/, 2008.

[6] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Stat. Society. Series B*, 39(1):1-38, 1977.

[7] G. R. Abecasis, S. S. Cherny, W. O. Cookson and L. R. Cardon, Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics*, 30:97-101, 2001.

[8] A. Thomas, GMCheck: Bayesian error checking for pedigreegenotypes and phenotypes. *Bioinformatics*, 21(14):3187-3188, 2005.

[9] GPL version 3.0, http://www.gnu.org/licenses/licences.html, 2008.