King's Research Portal

DOI:

[10.1515/sagmb-2017-0007](https://doi.org/10.1515/sagmb-2017-0007)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Sacco, C., Viroli, C., & Falchi, M. (2017). A statistical test for detecting parent-of-origin effects when parental information is missing. *Statistical applications in genetics and molecular biology*, 16(4), 275-289.
<https://doi.org/10.1515/sagmb-2017-0007>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

A statistical test for detecting parent-of-origin effects when parental information is missing

[1]Chiara Sacco*, [1] Cinzia Viroli, [2] Mario Falchi

July 28, 2017

[1] Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Via delle Belle Arti 41, 40126, Bologna, Italy, [2] Department of Twin Research and Genetic Epidemiology, King's College London, 3rd Floor, South Wing St Thomas' Hospital, Westminster Bridge Rd, SE1 7EH, London, United Kingdom

1 Introduction

Genome-Wide Association Studies (GWAS) are a powerful approach for identifying casual variants and genes for complex diseases (Hirschhorn and Daly, 2005; Bush and Moore, 2012). Usually, GWAS assume that the effect of a genetic variant is the same regardless of whether it is inherited from the mother or the father so that the two parental alleles are considered to be functionally equivalent. This assumption does not always hold, as it happens, for example, in the presence of *genomic imprinting* at the tested genetic variant (Lawson et al., 2013): indeed if the maternal (or paternal) allele is imprinted, its expression is partially or totally inactivated (Guilmatre and Sharp, 2012). This can generate functional haploidy that may increase the risk of expression of deleterious mutations. Several human syndromes such as the Beckwith-Wiedemann, PraderWilli and Angelman syndromes, and common diseases including obesity, diabetes, and psychiatric disorders have been associated with altered expression at imprinted genes (Peters, 2014). Parent-of-origin effects (POEs), if not carefully accounted for, could mask the potential associations between traits and genetic variants leading to underestimation of the proportion of the trait heritability explained by genetic variations, also called the “missing heritability” problem (see Eichler et al., 2010; Manolio et al., 2009, among the others).

Current methods for detecting POEs are based on the knowledge of the parental ancestry for each inherited allele. Data from cases and their parent genotype information can be analysed by likelihood-based test methods (Weinberg et al., 1998; Weinberg, 1999), linkage or association analysis of quantitative traits using linear mixed models (Hanson et al., 2001; Belonogova et al., 2010) and mixtures models for the POE analysis of inbred F2 designs (Cui et al., 2006).

In absence of parental information, an alternative and interesting approach to detect POEs in unrelated individuals has been developed by Hoggart et al. (2014). The key idea is to model the presence of POEs by a latent variable taking values 1 if an allele is inherited from the mother and value 0 if inherited from the father. The presence of POE at a particular genetic location is then assessed through the Brown-Forsythe test of the phenotypic variance between the homozygous and heterozygous groups, under the assumption that an increased variance in the heterozygous group arises because the heterozygous genotype group consists of two sub-populations depending on whether the reference allele is inherited from the mother or the father. This method is quite appealing for the assessment of POE at a genome-wide scale, though artefacts might result from the scale on which the phenotype is measured (Sun et al., 2013).

Here, we propose a novel test for POEs detection in absence of parental information, based on the interpretation of the POE problem as a model-based clus-

tering problem (McLachlan and Basford, 1988; Celeux et al., 2005; Wang et al., 2012). Indeed, we assume that the population Ω can be considered as the union of two disjoint sub-populations, Ω_M and Ω_P , such that $\Omega = \Omega_M \cup \Omega_P$, where Ω_M and Ω_P represent the group of individuals receiving a particular allele from the one parent (*e.g.* the mother, M) and from the other (*e.g.* the father, P), respectively. This is equivalent to define a mixture model with two components corresponding to the parental origin of a allele. We have developed the model for the POE analysis of twins data: more specifically, we define a mixture of linear mixed effects models (MLMMs) that simultaneously accounts for the correlation within twin pairs and for the unobserved heterogeneity among individuals due to parent-of-origin effects. We assume that the monozygotic twins (MZ), who share 100% of their genes, have to belong to the same mixture component, while each individual of a dizygotic twin pair (DZ), who share 50% of their genome, can belong to a different mixture component. Even if the proposed model is suited for the analysis of twin data, we will show that it is flexible enough to be easily applicable to the case of unrelated subjects.

The proposed model and its estimation will be presented in Sections 2 and 3. In Section 4 a statistical testing procedure for POEs detection is developed. In order to assess the method performance a simulation study taking into account several scenarios is presented in Section 5. The proposed method is finally applied to the MuTHER twin study data in Section 6.

2 Mixture Model for POE in Twin Data

Let y_{ij} denote a phenotypic response, or a trait of interest, observed in m unrelated twin-pairs, with $i = 1, \dots, m$ and $j = 1, 2$ to denote individual in the i -th twin pair. We are interested in measuring the association between the phenotype y and the alleles of a particular SNP and to check if the effect may depend on the parental origin of these alleles. We indicate the alleles of a bi-allelic SNP by “A” (reference) and “B” (alternative), so the possible genotype groups g_{ij} of the j -th individual of the i -th twin pair are AA, AB and BB. Following Hoggart et al. (2014) notation, the phenotype of any individual in the AA group is modelled by

$$y_{ij} = \alpha + \epsilon_{ij},$$

the phenotype of an individual in the BB genotype group consists of the additional effect of the maternal and paternal origins as

$$y_{ij} = \alpha + \beta_M + \beta_P + \epsilon_{ij},$$

while the phenotypic in the heterozygous group is a mixture of two distributions with component weights equal to $1/2$,

$$y_{ij} = \alpha + z_{ij}\beta_M + (1 - z_{ij})\beta_P + \varepsilon_{ij},$$

where z_{ij} is a latent variable distributed according to a Bernoulli with parameter $1/2$ that takes value 1 in case of maternal origin of the allele.

Since twin data includes both MZ twins and DZ twins, we introduce two further latent variables respectively, in order to constrain MZ individuals to inherit the causal allele from the same parent, and to allow DZ individuals to have a different parental effect. By denoting with $\mathbb{1}_{MZ}(i)$ the indicator function of i , having value 1 for all MZ twins and value 0 for all DZ twins, the two latent variables are:

- when $\mathbb{1}_{MZ}(i) = 1$

$$z_i^{MZ} = \begin{cases} 0 & \text{if the } i\text{-th MZ pair} \in \Omega_P \\ 1 & \text{if the } i\text{-th MZ pair} \in \Omega_M \end{cases}$$

that is a realization of a random variable $Z_i^M \sim \text{Ber}(\frac{1}{2})$;

- when $\mathbb{1}_{MZ}(i) = 0$

$$z_{ij}^{DZ} = \begin{cases} 0 & \text{if the } j\text{-th individual of the } i\text{-th DZ pair} \in \Omega_P \\ 1 & \text{if the } j\text{-th individual of the } i\text{-th DZ pair} \in \Omega_M \end{cases}$$

that is a realization of a random variable $Z_{ij}^{DZ} \sim \text{Ber}(\frac{1}{2})$.

We also introduce two additional dichotomic variables, denoted by w_{ij}^{BB} and w_{ij}^{AB} , that are known and identify the genotype group g_{ij} :

$$w_{ij}^{BB} = \begin{cases} 1 & \text{if } g_{ij} = BB \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and

$$w_{ij}^{AB} = \begin{cases} 1 & \text{if } g_{ij} = AB \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Treating individuals as level-one units and the twin pairs as level-two units, the response y_{ij} for the i -th twin pair and the j -th individual is assumed to depend on fixed and random effects and eventual covariates X , as follows:

$$\begin{aligned} y_{ij} = & \alpha + (\beta_M + \beta_P)w_{ij}^{BB} + \beta_P w_{ij}^{AB} + \mathbb{1}_{MZ}(i)z_i^{MZ}(\beta_M - \beta_P)w_{ij}^{AB} \\ & + (1 - \mathbb{1}_{MZ}(i))z_{ij}^{DZ}(\beta_M - \beta_P)w_{ij}^{AB} + X_{ij}'\gamma + u_i + \varepsilon_{ij} \end{aligned} \quad (3)$$

where α is the intercept, β_M and β_P are the maternal and paternal effects of the “B” allele, respectively, $X_{ij} \in \mathbb{R}^p$ is a known covariate vector for fixed effects,

$\gamma \in \mathbb{R}^p$ is a vector of regression coefficients to be estimated. The random effects u_i , used to describe the correlation within each twin pair, is distributed as $N(0, \tau^2)$, independently from the error $\varepsilon_{ij} \sim N(0, \sigma^2)$. Since z_i^{MZ} and z_{ij}^{DZ} are Bernoulli latent variables, (3) is equivalent to a mixture model with two component densities, say $f_1(y_{ij}; \theta_1)$ and $f_2(y_{ij}; \theta_2)$, defined as

$$f(Y_{ij}|Z_i^{MZ} = 0) = f(Y_{ij}|Z_{ij}^{DZ} = 0) = \mathcal{N}(\mu_{1ij}, \sigma^2 + \tau^2) = f_1(y_{ij}; \theta_1) \quad (4)$$

where $\mu_{1ij} = \alpha + (\beta_M + \beta_P)w_{ij}^{BB} + \beta_P w_{ij}^{AB} + X_{ij}'\gamma$, $\theta_1 = \{\mu_{1ij}, \sigma^2, \tau^2\}$, and

$$f(Y_{ij}|Z_i^{MZ} = 1) = f(Y_{ij}|Z_{ij}^{DZ} = 1) = \mathcal{N}(\mu_{2ij}, \sigma^2 + \tau^2) = f_2(y_{ij}; \theta_2) \quad (5)$$

where $\mu_{2ij} = \alpha + (\beta_M + \beta_P)w_{ij}^{BB} + \beta_M w_{ij}^{AB} + X_{ij}'\gamma$, $\theta_2 = \{\mu_{2ij}, \sigma^2, \tau^2\}$.

Thus, the finite mixture density of mixed effects models with two components is given for the observation of the individual j of the twin pair i by

$$f(y_{ij}; \theta) = \sum_{k=1}^2 f_k(y_{ij}; \theta_k) = \frac{1}{2} \mathcal{N}(y_{ij}; \mu_{1ij}, \sigma^2 + \tau^2) + \frac{1}{2} \mathcal{N}(y_{ij}; \mu_{2ij}, \sigma^2 + \tau^2), \quad (6)$$

where $\theta = \{\alpha, \beta_M, \beta_P, \gamma, \sigma^2, \tau^2\}$ collectively contains the vector of parameters, that is $(5 + p) \times 1$ dimensional.

Notice that, in so doing, we have assumed that the MZ twin pair belong to the same mixture component, as done by Celeux et al. (2005) in a different context; on the contrary the DZ twin pair can belong to a different mixture component.

2.1 Modelling Independent Data

The previous model can be easily adapted to the general case of independent observations. Let y_i be the trait of interest observed in m independent observations. The linear model in (3) can be rewritten in the following way:

$$y_i = \alpha + (\beta_M + \beta_P)w_i^{BB} + \beta_P w_i^{AB} + z_i(\beta_M - \beta_P)w_i^{AB} + X_i'\gamma + \varepsilon_i \quad (7)$$

where z_i is a latent variable distributed according to a Bernoulli with parameter $1/2$ that takes value 1 if the i -th individual belongs to Ω_M and $\varepsilon_i \sim N(0, \sigma^2)$.

Thus, a finite mixture density of regression models with two components is defined by

$$f(y_i; \theta) = \sum_{k=1}^2 f_k(y_i; \theta_k) = \frac{1}{2} \mathcal{N}(y_i; \mu_{1i}, \sigma^2) + \frac{1}{2} \mathcal{N}(y_i; \mu_{2i}, \sigma^2). \quad (8)$$

where $\mu_{1i} = \alpha + (\beta_M + \beta_P) w_i^{BB} + \beta_P w_i^{AB} + X_i' \gamma$ and $\mu_{2i} = \alpha + (\beta_M + \beta_P) w_i^{BB} + \beta_M w_i^{AB} + X_i' \gamma$.

3 Model estimation

The log-likelihood of the proposed model is given by

$$\mathcal{L}(\theta) = \sum_{i=1}^m \sum_{j=1}^2 \ln f(y_{ij}; \theta) = \sum_{i=1}^m \sum_{j=1}^2 \ln \left(\sum_{k=1}^2 \frac{1}{2} f_k(y_{ij}; \theta_k) \right). \quad (9)$$

The direct maximization of the log-likelihood function, $\mathcal{L}(\theta)$, is complicated; a popular method for finding maximum likelihood estimators of the parameters in finite mixture models is the Expectation-Maximization algorithm (Dempster et al., 1977; McLachlan and Basford, 1988). The EM algorithm is an iterative procedure composed by two steps, the E-step and the M-step, that maximizes the conditional expectation of the so-called complete log-likelihood, that is the likelihood we should maximize if we knew the values of the latent or missing data. In the proposed model there are two types of missing data: the latent allocation variables, z_i^{MZ} and z_{ij}^{DZ} , and the random effects, u_i . The log-likelihood function associated to the complete data can be defined by

$$\begin{aligned} \mathcal{L}_C(\theta) = \sum_{k=1}^2 \sum_{i=1}^m \sum_{j=1}^2 \left\{ \mathbb{1}_{MZ}(i) \left[z_{ik}^{MZ} \ln \frac{1}{2} + z_{ik}^{MZ} \ln f_k(y_{ij}, u_i | z_{ik}^{MZ}; \theta_k) \right] \right. \\ \left. + (1 - \mathbb{1}_{MZ}(i)) \left[z_{kij}^{DZ} \ln \frac{1}{2} + z_{kij}^{DZ} \ln f_k(y_{ij}, u_i | z_{kij}^{DZ}; \theta_k) \right] \right\} \end{aligned} \quad (10)$$

where $\ln f_k(\cdot)$ is the log-density function of the joint distribution of y_{ij} and u_i conditionally to the component k from which it arises, and it is given by

$$\ln f_k(y_{ij}, u_i | \cdot; \theta_k) \propto -\frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y_{ij} - \mu_{kij} - u_i)^2 - \frac{1}{2} \ln \tau^2 - \frac{1}{2\tau^2} u_i^2. \quad (11)$$

At iteration $r > 0$, the E-step consists of computing the expectation of the complete log-likelihood function given the observed data, Y , and the current values of the

parameters $\theta^{(r)}$. Thus, the expectation of the complete log-likelihood is defined by

$$\begin{aligned} \mathcal{Q}(\theta, \theta^{(r)}) &= \mathbb{E}_{\theta^{(r)}} \{ \mathcal{L}_C(\theta) | Y \} \\ &\propto \sum_{k=1}^2 \sum_{i=1}^m \sum_{j=1}^2 \left\{ \mathbb{1}_{MZ}(i) \tau_k^{MZ}(\mathbf{y}_i; \theta_k^{(r)}) \mathbb{E}_{\theta^{(r)}} [\ln f_k(\mathbf{y}_i, u_i | z_{ik}^{MZ}; \theta_k) | \mathbf{y}_i] \right. \\ &\quad \left. + (1 - \mathbb{1}_{MZ}(i)) \tau_k^{DZ}(y_{ij}; \theta_k^{(r)}) \mathbb{E}_{\theta^{(r)}} [\ln f_k(y_{ij}, u_i | z_{kij}^{DZ}; \theta_k) | y_{ij}] \right\}. \end{aligned} \quad (12)$$

where we assume that the MZ and DZ twin pair are conditioned to $\mathbf{y}_i$ and to y_{ij} , respectively, and τ_k are the posterior probabilities given by

$$\tau_k^{MZ}(\mathbf{y}_i; \theta_k^{(r)}) = f(z_{ik}^{MZ} | \mathbf{y}_i; \theta_k^{(r)}) = \frac{\frac{1}{2} \prod_{j=1}^2 \mathcal{N}(y_{ij}; \mu_{kij}, \tau^2 + \sigma^2)}{\sum_{k=1}^2 \frac{1}{2} \prod_{j=1}^2 \mathcal{N}(y_{ij}; \mu_{kij}, \tau^2 + \sigma^2)} \quad (13)$$

$$\tau_k^{DZ}(y_{ij}; \theta_k^{(r)}) = f(z_{kij}^{DZ} | y_{ij}; \theta_k^{(r)}) = \frac{\frac{1}{2} \mathcal{N}(y_{ij}; \mu_{kij}, \tau^2 + \sigma^2)}{\sum_{k=1}^2 \frac{1}{2} \mathcal{N}(y_{ij}; \mu_{kij}, \tau^2 + \sigma^2)}. \quad (14)$$

The details of the E- and M-steps for the parameter estimation are described in the Appendix.

3.1 Computational Issues

A well known problem of the EM solutions is that they can be highly dependent on the starting values of the algorithm and can get stuck in a local optimum. To increase the chance to converge to a global optimum, it is recommended to perform multiple short runs of the EM algorithms, starting from a different random initialization (McLachlan and Peel, 2004) and choosing the one with highest likelihood value (Biernacki et al., 2003). However, using random initial values can often not solve the problem of finding bad local optima. For the proposed model, we suggest to initialize the EM from $B = 10$ starting points obtained by fitting a linear mixed model, using the R package nlme (Pinheiro and Bates, 2006), on a random sub-sample of the data.

Another important issue with mixture estimation is the label switching (Stephens, 2000). That problem arises because the likelihood of a mixture model can be invariant to permutations of the components labels. In other words, the values of the parameters β_M and β_P are exchangeable and lead to the same value of the likelihood function in (10). In order to take into account this problem, we impose an identifiability constraint on the parameters β_M and β_P : $\beta_M > \beta_P$. Therefore, we assume that the first component, denoted by “M”, always represents the group of individuals where the allele “B” has the largest effect size.

The estimation algorithm, whose details are described in the Appendix, has been implemented in R, and it is available from the authors upon request.

4 The Statistical Test for POEs

The proposed method gives directly an estimate of the parental effects by the values of the parameters β_M and β_P , as shown in equations (28) and (29), respectively. The parent-of-origin phenomena is observed when the effect of the allele “B” inherited from the mother is different from the effect of the allele “B” given from the father. This occurs when the allele inherited from one of the parents is completely or partially silenced. In order to verify the presence of POEs, we are interested in evaluating the equality between β_M and β_P . Thus, the null hypothesis of our test can be represented in the following way:

$$H_0 : \beta_M = \beta_P. \quad (15)$$

Since the EM estimators are maximum likelihood estimators (MLE) and considering the constraint imposed to avoid the identifiability problem discussed in the previous section, we have that $\hat{\beta}_M - \hat{\beta}_P \geq 0$. Therefore, $\hat{\beta}_M - \hat{\beta}_P$ is distributed according to Half Normal distribution with scale parameter equal to the variance of the difference between the parameters, $Var(\hat{\beta}_M - \hat{\beta}_P)$.

Thus, the test-statistics under the null hypothesis, is asymptotically distributed according a χ^2 distribution with 1 degree of freedom:

$$\left(\frac{\hat{\beta}_M - \hat{\beta}_P}{\sqrt{Var(\hat{\beta}_M - \hat{\beta}_P)}} \right)^2 \Big|_{H_0} \sim \chi_1^2. \quad (16)$$

where $\hat{\beta}_M$ and $\hat{\beta}_P$ are the EM-estimates. To compute the test statistic, we need to estimate the variance $Var(\hat{\beta}_M - \hat{\beta}_P)$ that is equal to $Var(\hat{\beta}_M) + Var(\hat{\beta}_P) - 2Cov(\hat{\beta}_M, \hat{\beta}_P)$. One of the criticisms of the EM algorithm is that it does not automatically provide the covariance matrix of the MLE. The asymptotic covariance matrix of $\hat{\theta}$, the vector of ML estimators, can be obtained by parametric bootstrap or by numerical approximation of the Fisher Information. In the latter case, we can compute the robust sandwich covariance matrix estimator by approximating the Fisher Information with the empirical information matrix, $\mathcal{I}_g(\hat{\theta}) = \sum_{j=1}^n [\mathbf{q}_j(\hat{\theta}) \mathbf{q}_j(\hat{\theta})']$, and with the observed information matrix, $\mathcal{I}_H(\hat{\theta}) = -\sum_{j=1}^n [\mathbf{Q}_j(\hat{\theta})]$, where q and Q denote the gradient and Hessian of the likelihood function in equation (9) in the maximum point, $\hat{\theta}$. Thus the asymptotic sandwich variance of $\hat{\theta}$ is given by:

$$\hat{\mathcal{I}}_H(\theta)^{-1} \hat{\mathcal{I}}_g(\theta) \hat{\mathcal{I}}_H(\theta)^{-1}. \quad (17)$$

5 Simulation Study

Simulation studies have been performed under different scenarios to investigate the convergence of the empirical type I error to the nominal level and the power of the test for the null hypothesis (15).

A quantitative continuous trait was simulated according to (3) in 500 twin pairs, 30% of which were identical MZ twins.

Thus, the variance of the simulated trait is given by the sum of σ_β^2 , the variance explained from the main allelic effect β , where β indicates the mean of the maternal and paternal effects, τ^2 , the variance of the random effects, and σ^2 , the variance of the error, where $\sigma_\beta^2 + \tau^2 + \sigma^2 = 1$.

The main allelic effect β can be derived from the variance explained σ_β^2 that is a proportion of the total trait variation. The explained variance can be written in the following way:

$$\sigma_\beta^2 = 4\beta^2 (2k^2 - 2k + 1) pq \quad (18)$$

where p represents the minor frequency allele and $q = 1 - p$. The parameter k indicates the proportion of the main allelic effect explained by the “B” allele inherited from the parent “M”, which denotes the absence/presence of POE and its intensity. In the absence of POE, $k = 0.5$, otherwise the maternal and paternal alleles contribute to the POE with a percentage equal to k e $1-k$. In the extreme cases, where k is equal to 1 or 0, the POE (paternal or maternal) is complete and one of the parental allele is epigenetically silenced.

In the first simulation study we assess the adequateness of the statistical procedure evaluating the convergence of the empirical type I error to the nominal value as the number of replicates increases and for different levels of significance of the test ($\alpha = 0.1, 0.05$ and 0.01). We have simulated a quantitative trait under the null hypothesis, $\beta_M = \beta_P$ thus we impose $k = 0.5$, controlled by one SNP, simulated under Hardy Weinberg Equilibrium, with minor allele frequency of 0.5. In this simulation we assume that, under the null hypothesis, the trait variance is given by a SNP effect, that explain 30% of the total trait variation, by $\tau^2 = 0.4$, and by a normally distributed error with a variance of $\sigma^2 = 0.3$. Figure 1 (a,c,e) shows the behaviour of the type I error with the increase of the number of simulations. For all levels of significance the empirical type I error approaches to the corresponding nominal value. Figure 1 (b,d,f) depicts the distribution of the p-values under the null hypothesis in the case of one thousand of replicates. The uniform distribution of the p-values under the null hypothesis at the nominal significance level and the

convergence of the test statistic at the increasing of the simulations number ensure the capability of controlling the type I error.

In the second simulation study we evaluate the performance of the proposed test in terms of power and type I error under several scenarios, comparing our test with the one proposed by Hoggart et al. (2014). σ^2 has been fixed to three different values (0.70, 0.50, 0.30), corresponding to a total narrow heritability, h^2 , (*i.e.* $\frac{\sigma_\beta^2 + \tau^2}{\sigma_\beta^2 + \tau^2 + \sigma^2}$), of 0.30, 0.50, 0.70. Under the alternative hypothesis, we assume $\beta_M \neq \beta_P$ and we analyse three different levels of POEs, $k = 1, 0.9$ and 0.8 . Summing up, to evaluate the power and the type I error of our procedure we have simulated 36 and 12 different scenarios, respectively. Each scenario is replicated using four different minor allele frequencies for the B allele (MAF = 0.1, 0.2, 0.3, and 0.5). Since the variance test proposed by Hoggart et al. (2014) works under assumption of independence among observations, therefore we randomly sampled a single individual for each twin pair.

All the results are shown in Figures 2 and 3. The power of the proposed POE test, denoted by MLMM, is nearly 1 in all scenarios when the proportion of the total trait variation explained by the simulated marker (σ_β^2) is equal at least to 10%. When $\sigma_\beta^2 = 0.01$ the power ranges from 0.2 to 0.5. In the case of 100% POE the power assumes clearly higher values; when β_M represents the 80%-90% of the SNP effect and $\sigma_\beta^2 = 0.01$, we have an higher probability of detecting the POEs in correspondence to small MAF. Under all scenarios, the power of the test based on the MLMM is always higher than the Hoggart's test.

Regarding the type I error, with only one hundred replicates, we can say that the proposed test statistic is not inflated. The type I error, under different scenarios, results in good agreement with the nominal significance 5% level for both tests.

6 Study on imprinting genes

6.1 MuTHER Study

The MuTHER project (Multiple Tissue Human Expression Resource) is a resource of genomic and transcriptome data generated on a sample of 856 female twins (154 monozygotic twin pairs, 232 dizygotic twin pairs and 84 singletons) from the Twin-sUK adult registry. Detailed information about these data can be found in Grundberg et al. (2012). Briefly, twins were aged between 40 and 87 years old (mean age 62) at the time of visit. Peripheral blood samples were collected to generate lymphoblastoid cell lines (LCL). Expression profiling of the LCL samples were performed using Illumina Human HT-12 V3 BeadChips (Illumina) including more

than 48,000 probes. Log2-transformed expression signals were normalized using quantile normalization across the replicates of each individual followed by quantile normalization across all individuals as described in Nica et al. (2011). Uniquely mapping probes with no mismatches and either an Ensembl or RefSeq ID were kept for analysis. Probes mapping to genes of uncertain function and those encompassing a common SNP were further excluded, leaving 23,552 probes to be used in the analysis. Genotyping of TwinsUK dataset was done with a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1M-Duo, and 1.2MDuo 1M) and imputation was performed with the IMPUTE software package (v2) (Howie et al., 2009) as previously described (Consortium et al., 2011). eQTL data identified in the MuTHER dataset (Consortium et al., 2011) were downloaded from the MuTHER web site <http://www.muther.ac.uk/Data.html>.

6.2 Analysis and Results

In order to illustrate the proposed method, we selected 97 known imprinted genes in humans as catalogued in the Imprinted Gene Database (<http://www.geneimprint.com>) and on 28 genes recently that have been suggested to be imprinted through the analysis of allelic expression data in primary tissue samples from the Genotype Tissue Expression (GTEx) project (Baran et al., 2015). In the MuTHER dataset 110 transcripts on 68 imprinted genes were available after quality control (Consortium et al., 2011). We focused the tests on a subset of known imprinted genes for which SNP eQTLs have been previously identified in the MuTHER study at FDR level of 1% within a 1Mb window both upstream and downstream of each transcript. This leads to the selection of 13 genes (14 transcripts) and 366 SNPs (eQTL for these transcripts), for a total of 372 tests; the number of test performed for each transcript is reported in the fourth column of Table 1. In correspondence of each transcript and each SNP the mixture model has been fitted and the proposed statistical test (denoted by MLMM) has been computed for the detection of the POEs. Since the only available test for POE in absence of parental information is the one developed by (Hoggart et al., 2014), we have additionally performed their variance test; we have applied the Hoggart’s test as explained in Section 5. Age and experimental batch were added as covariates, and the first principal component calculated on the GWAS SNP data were included in the fixed effects to adjust for potential population stratification (Price et al., 2006). To control the total first error in multiple comparisons, all the obtained p-values were adjusted following the procedure of Benjamini and Hochberg. In order to investigate the power of both methods to detect the parent-of-origin effects, we measured the number of genes declared in POEs at the

confidence level of 0.05 (Table 1). The MLM confirmed in POEs 8 out of 13 imprinting genes, while Hoggart method validated only 2 genes (MEST, CPA4). The median (1st - 3d quantile) distance between SNPs showing a POE effects and the affected transcripts was 23kb (13kb - 40kb). Table 2 shows the proportion of variance explained for the set of SNPs selected in eQTL for each transcript. We can observe that the number of the SNPs declared in POEs depends on the level of the variance explained from each SNPs. Indeed, in the simulation study we have highlighted the relation between the power of the proposed test and the proportion of variance explained from the main allelic effect, the power to detect POEs is approximately equal to 1 for a proportion of variance explained $\geq 10\%$. This result is confirmed on the real data by observing Table 3. For a level of variance explained $\geq 5\%$ the proportion of SNPs declared in POEs at confidence level of 0.05 ranges from 91% to 100%. On the other hand, for very low percentage of variance explained, from 0% to 5%, the power ranges between 6% and 19%.

7 Conclusions

Several approaches for POE identification have been proposed in the statistical literature but they are not suitable in case of related data without information about the alleles parental origins. We propose here a novel test for POEs detection in absence of parental information, that has been developed for the special case of twin data. The idea of using a latent variable to model the unobserved heterogeneity among the individuals due to the parental origin of the transmitted alleles was already suggested by Hoggart et al. (2014). Here, we formalize this concept by exploiting a mixture model. In particular, in the case of twin data, in order to account for the heterogeneity due to the two sub-populations generated by the parental origin of the transmitted allele and the correlation among the individuals, we use a mixture of linear mixed model to estimate the effects of inheriting the allele from the mother and from the father. The estimation of these coefficients allows us to test directly the difference between these two parameters to detect the POEs when the parental information is missing. The simulation study shows that our approach is consistent in terms of parameter estimation and hypothesis testing, the type I error is controlled. In the simulation study, we demonstrate that the power of proposed test is nearly 1 when the proportion of the total trait variation explained by the simulated marker is equal at least to 10%. This result has been confirmed by the analysis on known imprinted genes, indeed, for a level of variance explained by the main allelic effect $\geq 5\%$ the proportion of SNPs declared in POEs at a confidence level of 0.05 ranges from 91% to 100%. Therefore the proposed test is a powerful tool for the identification of the POE in absence of parental information, in a genome wide

discovery analysis we suggest to prioritise the top hits obtained by applying the POE test using independent family data and finally supported by lab experiments. In this paper, we focus on the twins data but the test can be adapted to other family structure and to unrelated samples, as shortly shown in the Section 2.1, in an easy manner.

Appendix: EM algorithm

For the sake of brevity we denote the posterior probability in case of MZ twins as $\tau_k^{MZ}(\mathbf{y}_i)$ and $\tau_k^{DZ}(y_{ij})$, for DZ twin pairs.

In the E-step, in order to compute $\mathcal{Q}(\theta, \theta^{(r)})$, we require the conditional variance and the conditional mean. For the MZ twin pair, the conditional variance is

$$\Sigma_{u_i}^{MZ} = \left(\frac{1}{\tau^2} + \frac{2}{\sigma^2} \right)^{-1} \quad (19)$$

and the conditional mean is given by

$$\mu_{u_i,k}^{MZ} = \Sigma_{u_i} \frac{1}{\sigma^2} \mathbf{1}' (\mathbf{y}_i - \mu_{ik})^2 \quad (20)$$

where $\mathbf{y}_i$ is the observed data vector 2×1 dimensional of the i -th twin pair and μ_{ik} is the mean vector 2×1 dimensional of the i -th twin pair of the k -th component.

If $\mathbb{1}_{MZ}(i) = 0$, we obtain that the conditional variance is defined by

$$\Sigma_{u_i,k}^{DZ} = (\tau^2 + \sigma^2)^{-1} \sigma^2 \tau^2 \quad (21)$$

and the conditional mean is given by

$$\mu_{u_i,kj}^{DZ} = (1 + \sigma^2 \tau^{-2})^{-1} (y_{ij} - \mu_{kij})^2. \quad (22)$$

The M-step consists of determining the values maximizing the equation (12) where

$$\begin{aligned} \mathbb{E} [\ln f(\mathbf{y}_i, u_i | z_{ik}^{MZ}; \theta_k)] &\propto -\ln \sigma^2 - \frac{1}{2} \left[(\mathbf{y}_i - \mu_{ik} - \mu_{u_i,k}^{MZ})^2 + \Sigma_{u_i,k}^{MZ} \right] \\ &\quad - \frac{1}{2} \ln \tau^2 - \frac{1}{2\tau^2} (\mu_{u_i,k}^{MZ} + \Sigma_{u_i,k}^{MZ}) \end{aligned} \quad (23)$$

and

$$\begin{aligned} \mathbb{E} [\ln f(y_{ij}, u_i | z_{ijk}^{DZ}; \theta_k)] &\propto -\frac{1}{2} \ln \sigma^2 - \frac{1}{2} \left[(y_{ij} - \mu_{kij} - \mu_{u_i,kj}^{DZ})^2 + \Sigma_{u_i,k}^{DZ} \right] \\ &\quad - \frac{1}{2} \ln \tau^2 - \frac{1}{2\tau^2} (\mu_{u_i,kj}^{DZ} + \Sigma_{u_i,k}^{DZ}). \end{aligned} \quad (24)$$

It follows that

$$\begin{aligned} \mathcal{Q}(\theta, \theta^{(r)}) \propto & \sum_{k=1}^2 \sum_{i=1}^m \sum_{j=1}^2 \left\{ -\frac{1}{2} \ln \sigma^2 - \frac{1}{2} \ln \tau^2 \right. \\ & - \frac{1}{2} \mathbb{1}_{MZ}(i) \left[\frac{(y_{ij} - \mu_{ijk} - \mu_{u_i,k}^{MZ})^2 + \Sigma_{u_i,k}^{MZ}}{\sigma^2} - \frac{1}{\tau^2} (\mu_{u_i,k}^{MZ} + \Sigma_{u_i,k}^{MZ}) \right] \Big\} \\ & - \frac{1}{2} (1 - \mathbb{1}_{MZ}(i)) \left[\frac{(y_{ij} - \mu_{ijk} - \mu_{u_i,k}^{DZ})^2 + \Sigma_{u_i,k}^{DZ}}{\sigma^2} - \frac{1}{\tau^2} (\mu_{u_i,k}^{DZ} + \Sigma_{u_i,k}^{DZ}) \right] \Big\}. \end{aligned} \quad (25)$$

For this model the parameters can be determined in closed form by solving the equations derived by computing the derivatives of the expected complete likelihood, (25), with respect to parameters α , β_M , β_P , γ , τ^2 and σ^2 , and setting them to zero. Thus we obtain:

$$\hat{\alpha} = \frac{1}{N} \sum_{k=1}^2 \sum_{i=1}^m \sum_{j=1}^2 \tilde{y}_{kij} (\mathbb{1}_{MZ}(i) \tau_k^{MZ} + (1 - \mathbb{1}_{MZ}(i)) \tau_k^{DZ}) \quad (26)$$

where $N = 2m$ and

$$\tilde{y}_{kij} = \begin{cases} y_{ij} - (\beta_M + \beta_P) w_{ij}^{BB} - \beta_P w_{ij}^{AB} - X'_{ij} \gamma - \mu_{u_i,1} & k = 1 \\ y_{ij} - (\beta_M + \beta_P) w_{ij}^{BB} - \beta_M w_{ij}^{AB} - X'_{ij} \gamma - \mu_{u_i,2} & k = 2 \end{cases}, \quad (27)$$

where $\mu_{u_i,k} = \mathbb{1}_{MZ}(i) \mu_{u_i,k}^{MZ} + (1 - \mathbb{1}_{MZ}(i)) \mu_{u_i,k}^{DZ}$.

We have that the parental effect of the "B" allele are equal, respectively,

$$\begin{aligned} \hat{\beta}_M = & \frac{1}{n_{BB} + \sum_{i=1}^m \sum_{j=1}^2 \tau_2 w_{ij}^{AB}} \left\{ \sum_{k=1}^2 \sum_{i=1}^m \sum_{j=1}^2 w_{ij}^{BB} (y_{ij} - \alpha - X_{ij} \gamma - \beta_P - T_k) \right. \\ & \left. + \sum_{i=1}^m \sum_{j=1}^2 \tau_2 w_{ij}^{AB} (y_{ij} - \alpha - X_{ij} \gamma - \mu_{u_i,2}) \right\} \end{aligned} \quad (28)$$

where $n_{BB} = \sum_{i=1}^m \sum_{j=1}^2 w_{ij}^{BB}$, $\tau_k = \mathbb{1}_{MZ}(i) \tau_k^{MZ}(\mathbf{y}_i) + (1 - \mathbb{1}_{MZ}(i)) \tau_k^{DZ}(y_{ij})$ and $T_k = \tau_k \mu_{u_i,k}$ and

$$\begin{aligned} \hat{\beta}_P = & \frac{1}{n_{BB} + \sum_{i=1}^m \sum_{j=1}^2 \tau_1 w_{ij}^{AB}} \left\{ \sum_{k=1}^2 \sum_{i=1}^m \sum_{j=1}^2 w_{ij}^{BB} (y_{ij} - \alpha - X_{ij} \gamma - \beta_M - T_k) \right. \\ & \left. + \sum_{i=1}^m \sum_{j=1}^2 \tau_1 w_{ij}^{AB} (y_{ij} - \alpha - X_{ij} \gamma - \mu_{u_i,1}) \right\}. \end{aligned} \quad (29)$$

The covariate coefficients γ are

$$\hat{\gamma} = \left(\sum_{i=1}^m \sum_{j=1}^2 X_{ij} X'_{ij} \right)^{-1} \sum_{k=1}^2 \sum_{i=1}^m \sum_{j=1}^2 \tau_k X_{ij} (\ddot{y}_{kij}) \quad (30)$$

where

$$\ddot{y}_{kij} = \begin{cases} y_{ij} - \alpha - (\beta_M + \beta_P) w_{ij}^{BB} - \beta_P w_{ij}^{AB} - \mu_{u_i,1} & k = 1 \\ y_{ij} - \alpha - (\beta_M + \beta_P) w_{ij}^{BB} - \beta_M w_{ij}^{AB} - \mu_{u_i,2} & k = 2 \end{cases}. \quad (31)$$

Finally, the variance parameters of the model are defined by:

$$\tau^2 = \frac{1}{N} \sum_{k=1}^2 \sum_{i=1}^m \sum_{j=1}^2 \tau_k (\mu_{u_i,k} + \Sigma_{u_i,k}) \quad (32)$$

where $\Sigma_{u_i,k} = \mathbb{1}_{MZ}(i) \Sigma_{u_i,k}^{MZ} + (1 - \mathbb{1}_{MZ}(i)) \Sigma_{u_i,k}^{DZ}$, and

$$\sigma^2 = \frac{1}{N} \sum_{k=1}^2 \sum_{i=1}^m \sum_{j=1}^2 \tau_k e_{ijk}^2 \quad (33)$$

with

$$e_{ijk}^2 = \mathbb{1}_{MZ}(i) (y_{ij} - \mu_{ijk} - \mu_{u_i,k}^{MZ})^2 + (1 - \mathbb{1}_{MZ}(i)) (y_{ij} - \mu_{ijk} - \mu_{u_i,kj}^{DZ})^2.$$

References

- Baran, Y., M. Subramaniam, A. Biton, T. Tukiainen, E. K. Tsang, M. A. Rivas, M. Pirinen, M. Gutierrez-Arcelus, K. S. Smith, K. R. Kukurba, et al. (2015): “The landscape of genomic imprinting across diverse adult human tissues,” *Genome research*, 25, 927–936.
- Belonogova, N. M., T. I. Axenovich, and Y. S. Aulchenko (2010): “A powerful genome-wide feasible approach to detect parent-of-origin effects in studies of quantitative traits,” *European Journal of Human Genetics*, 18, 379–384.
- Biernacki, C., G. Celeux, and G. Govaert (2003): “Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models,” *Computational Statistics & Data Analysis*, 41, 561–575.
- Bush, W. S. and J. H. Moore (2012): “Genome-wide association studies,” *PLoS Comput Biol*, 8, e1002822.

- Celeux, G., O. Martin, and C. Lavergne (2005): "Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments," *Statistical Modelling*, 5, 243–267.
- Consortium, M. et al. (2011): "Identification of an imprinted master trans regulator at the *klf14* locus related to multiple metabolic phenotypes," *Nature genetics*, 43, 561–564.
- Cui, Y., Q. Lu, J. M. Cheverud, R. C. Littell, and R. Wu (2006): "Model for mapping imprinted quantitative trait loci in an inbred f 2 design," *Genomics*, 87, 543–551.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, 1–38.
- Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau (2010): "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Reviews Genetics*, 11, 446–450.
- Grundberg, E., K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T.-P. Yang, E. Meduri, A. Barrett, et al. (2012): "Mapping cis-and trans-regulatory effects across multiple tissues in twins," *Nature genetics*, 44, 1084–1089.
- Guilmatre, A. and A. Sharp (2012): "Parent of origin effects," *Clinical genetics*, 81, 201–209.
- Hanson, R. L., S. Kobes, R. S. Lindsay, and W. C. Knowler (2001): "Assessment of parent-of-origin effects in linkage analysis of quantitative traits," *The American Journal of Human Genetics*, 68, 951–962.
- Hirschhorn, J. N. and M. J. Daly (2005): "Genome-wide association studies for common diseases and complex traits," *Nature Reviews Genetics*, 6, 95–108.
- Hoggart, C. J., G. Venturini, M. Mangino, F. Gomez, G. Ascari, J. H. Zhao, A. Teumer, T. W. Winkler, N. Tšernikova, J. Luan, et al. (2014): "Novel approach identifies snps in *slc2a10* and *kcnk9* with evidence for parent-of-origin effect on body mass index," *PLoS Genet*, 10, e1004508.
- Howie, B. N., P. Donnelly, and J. Marchini (2009): "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genet*, 5, e1000529.
- Lawson, H., J. Cheverud, and J. Wolf (2013): "Genomic imprinting and parent-of-origin effects on complex traits," *Nature Reviews Genetics*, 14, 609–617.
- Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, et al. (2009): "Finding the missing heritability of complex diseases," *Nature*, 461, 747–753.
- McLachlan, G. and D. Peel (2004): *Finite mixture models*, John Wiley & Sons.

- McLachlan, G. J. and K. E. Basford (1988): “Mixture models. inference and applications to clustering,” *Statistics: Textbooks and Monographs*, New York: Dekker, 1988, 1.
- Peters, J. (2014): “The role of genomic imprinting in biology and disease: an expanding view,” *Nature Reviews Genetics*, 15, 517–530.
- Pinheiro, J. and D. Bates (2006): *Mixed-effects models in S and S-PLUS*, Springer Science & Business Media.
- Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006): “Principal components analysis corrects for stratification in genome-wide association studies,” *Nature genetics*, 38, 904–909.
- Stephens, M. (2000): “Dealing with label switching in mixture models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 795–809, URL <http://dx.doi.org/10.1111/1467-9868.00265>.
- Sun, X., R. Elston, N. Morris, and X. Zhu (2013): “What is the significance of difference in phenotypic variability across snp genotypes?” *The American Journal of Human Genetics*, 93, 390–397.
- Wang, K., S. K. Ng, and G. J. McLachlan (2012): “Clustering of time-course gene expression profiles using normal mixture models with autoregressive random effects,” *BMC bioinformatics*, 13, 1.
- Weinberg, C., A. Wilcox, and R. Lie (1998): “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,” *The American Journal of Human Genetics*, 62, 969–978.
- Weinberg, C. R. (1999): “Methods for detection of parent-of-origin effects in genetic studies of case-parents triads,” *The American Journal of Human Genetics*, 65, 229–235.

(a)

(b)

(c)

(d)

(e)

(f)

Figure 1: Type I Error behaviour at the increasing of simulations number and p-value distribution for different levels of the tests: (a-b) $\alpha = 0.1$ (c-d) $\alpha = 0.05$ and (e-f) $\alpha = 0.01$

Figure 2: Power to detect POEs by the the test on the difference between β_M and β_P . The columns show power for different MAF; the power achieved under three heritability model is shown in rows. The y axis of each panel shows power, whereas the x axis shows the proportion of total variance explained by the allelic effect. POE size was set to 80%, 90% and 100% of the allelic effect size.

Figure 3: Type I Error for the proposed test of POEs. The columns show type I error under four different level of MAF. The y axis of each panel shows type I error, whereas different colours indicates the level of the heritability (0.3, 0.5, 0.7). Keeping constant the MAF and the heritability, the type I errors corresponding to the different levels of explained variance is synthesized in error bars at 95% confidence interval.

Table 1: Number of SNPs declared in POEs at the confidence level of 0.05 and top adjusted p-value for both compared methods (* Genes for which POE has been suggested in Baran et al. (2015). The remaining genes have been selected from the Imprinted Gene Database.)

Probe ID	Chr	Gene	Nr. Tested SNPs	Nr. Identified SNPs		Top SNP MLM		Top SNP Hoggart	
				MLMM	Hoggart	SNP name	p-value	SNP name	p-value
ILMN_1664878	19	<i>NLRP2</i>	56	52		rs10416380	5.88E-302		
ILMN_1669479	7	<i>MEST*</i>	46	5	10	rs11767461	1.02E-03	rs12112550	2.13E-03
ILMN_1669617	7	<i>GRB10*</i>	2						
ILMN_1674620	7	<i>SGCE</i>	59	6		rs2374735	6.59E-04		
ILMN_1675612	20	<i>BLCAP</i>	1						
ILMN_1679301	8	<i>ZFAT</i>	78	2		rs1442454	3.90E-03		
ILMN_1692058	15	<i>NDN*</i>	2	2		rs7497973	3.14E-26		
ILMN_1740711	2	<i>KIAA1571</i>	17						
ILMN_1773307	4	<i>NAPIL5</i>	14	3		rs3017905	3.16E-05		
ILMN_1779917	10	<i>INPP5F</i>	1						
ILMN_1784294	7	<i>CPA4*</i>	76	16	22	rs6946506	2.65E-66	rs968404	3.22E-04
ILMN_1786429	13	<i>P2RY5</i>	14	2		rs198604	4.93E-18		
ILMN_2278590	7	<i>MEST*</i>	5						
ILMN_2382505	11	<i>SLC22A18</i>	1						

Table 2: Proportion of variance explained from SNPs within 1Mb window both upstream and downstream of each transcript

Probe ID	Chr	Gene	Variance Explained (%)					
			Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max
ILMN_1664878	19	<i>NLRP2</i>	3.56	10.82	11.91	11.59	12.38	20.71
ILMN_1669479	7	<i>MEST</i>	0.05	0.12	0.15	0.17	0.20	0.44
ILMN_1669617	7	<i>GRB10</i>	0.03	0.03	0.03	0.03	0.03	0.03
ILMN_1674620	7	<i>SGCE</i>	0.61	0.98	1.48	1.41	1.96	2.00
ILMN_1675612	20	<i>BLCAP</i>	0.18	0.18	0.18	0.18	0.18	0.18
ILMN_1679301	8	<i>ZFAT</i>	0.06	0.07	0.08	0.11	0.17	0.22
ILMN_1692058	15	<i>NDN</i>	2.65	2.67	2.68	2.68	2.70	2.71
ILMN_1740711	2	<i>KIAA1571</i>	0.04	0.05	0.05	0.05	0.05	0.06
ILMN_1773307	4	<i>NAP1L5</i>	0.26	0.30	0.32	0.31	0.33	0.34
ILMN_1779917	10	<i>INPP5F</i>	0.04	0.04	0.04	0.04	0.04	0.04
ILMN_1784294	7	<i>CPA4</i>	0.47	1.27	1.71	1.66	2.05	2.74
ILMN_1786429	13	<i>P2RY5</i>	0.68	0.75	0.77	0.81	0.90	0.97
ILMN_2278590	7	<i>MEST</i>	0.04	0.04	0.05	0.05	0.05	0.05
ILMN_2382505	11	<i>SLC22A18</i>	0.55	0.55	0.55	0.55	0.55	0.55

Table 3: Power of POE test respect to different levels of variance explained by each SNPs

Variance Explained	Power (%)
0.03% - 0.30%	6%
0.30% - 1%	15%
1% - 5%	19%
5% - 10%	100%
10% - 20%	91%
>20%	100%