Alma Mater Studiorum
Università di Bologna

DOTTORATO DI RICERCA IN
SCIENZE STATISTICHE
Ciclo XXIX

Settore Concorsuale di afferenza: 13/D1
Settore Scientifico disciplinare: SECS-S/01

Statistical Methods for the Detection of the Parent-of-Origin Effect in Genome-Wide Genotype Data

Presentata da: Chiara Sacco

Coordinatrice di Dottorato:	Relatori:
Prof.ssa Alessandra Luati	Prof.ssa Cinzia Viroli
	Co-Relatori:
	Dr. Leonardo Bottolo
	Dr. Mario Falchi

Esame Finale Anno 2017

Alma Mater Studiorum
Università di Bologna

DIPARTIMENTO DI SCIENZE STATISTICHE "PAOLO FORTUNATI"
DOTTORATO DI RICERCA IN

Author: Chiara Sacco
Title: Statistical Methods for the Detection
of the Parent-of-Origin Effect in Genome-Wide Genotype Data
Degree: Ph.D.
Convocation: January 2017

Abstract

Genomic imprinting is an epigenetic mechanism that leads to differential contributions of maternal and paternal alleles to offspring gene expression in parent-of-origin (POE) manner. Recently, parent-of-origin effects have attracted attention due to their potential contribution in the explanation of "missing heritability". We propose two different procedures for detecting the POEs in genome wide genotype data from related individuals (twins) when the parental origin cannot be inferred. In the first approach we suggest a multistep procedure to detect POEs based on a variance test to evaluate the gene expression heterogeneity between the homozygous and the heterozygous groups. The second method exploits a finite mixture of linear mixed models to propose a test for capturing the presence of POEs; the key idea is that in the case of POEs the population can be clustered in two different groups in which the reference allele is inherited by a different parent. The core advantage of the second method is that it is an integrated procedure developed specifically for the detection of the parental effects, while the first method a multistep procedure which results computationally faster. The performance of the proposed tests are evaluated through a wide simulation study. A discovery analysis on microarray gene expression data of the MuTHER study is performed by both methods.

Acknowledgements

First and foremost I would like to thank my advisor Prof. Cinzia Viroli. It has been an honor to be her Ph.D. student. I really appreciated all her contributions of time, ideas, to make my Ph.D. experience productive and stimulating and for allowing me to grow as a research scientist. Her guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank my two co-supervisors Dr. Mario Falchi and Dr. Leonardo Bottolo who gave me the opportunity to join their research team. Without their precious support it would not have been possible to conduct this research. I would like to thank them also for their constant help and their insightful comments that allowed me to write this Ph.D thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The context

One of the most important and challenging aims of the human genetics is the identification of genes involved in human diseases. Genome-wide association studies (GWAS) for common variants offer a powerful approach for identifying casual variants and genes for complex diseases (Bush and Moore, 2012; Hirschhorn and Daly, 2005). GWAS require the knowledge about common form of genetic variants, the single nucleotide polymorphisms (SNPs), and the possibility to genotype hundreds of thousands of SNPs in large patient samples.

Usually, GWAS assume that the effect of a genetic variant is the same regardless of whether they are inherited from the mother or the father; so in all such studies the two parental alleles are considered to be functionally equivalent. The existence of biological mechanism as genomic imprinting demonstrates that this assumption could be wrong (Lawson et al., 2013). Genomic imprinting results in a reduction of our genome to a functionally haploid state and is one of the epigenetic phenomena that can lead to the manifestation of parent-of-origin effects (POEs). In other words, if the allele inherited from the father is imprinted, it means that the allele is partially or totally inactivated and only the maternal allele is expressed (Guilmatre and Sharp, 2012), and viceversa. So, parent-of-origin phenomena is observed when the phenotypic effect of an allele depends on the sex of the parent from which it has been inherited. Indeed, it appears as a phenotypic difference between heterozygotes, where the paternal allele is different from the maternal one, depending on the allelic parent of origin.

Several human diseases, such as obesity, diabetes, Beckwith–Wiedemann, Prader–Willi and Angelman syndromes, and behavioural traits are demonstrated to be connected with imprinted genes (Peters, 2014). For identifying genes related to complex disease, the power of GWAS can be increased through the inclusion of the POEs information into the association model. While GWASs have been shown to be a successful tool for investigating the genetic architecture of complex diseases by finding a large number of significant variants, on the other hand, it has led to a hotly debated issue, indeed these variants typically account only for a minority part of the heritability,*i.e.* the portion of variance of a particular trait in a population that is due to genetic factors. For example, Visscher (2008) shows that a total of 54 variants are associated with the human height, that is a highly heritable quantitative trait (about the 80% of the height variation is explained by genetic factors). However, the 54 loci explain only 5% of the phenotypic variance, despite the variants have been identified using GWAS with hundreds of thousand of genetic marker on ~ 63000 people. This is known as the “missing heritability” problem and an important and widely debated issue is where the “missing heritability” of a complex disease might be found (Eichler et al., 2010; Manolio et al., 2009).

Several strategies have been suggested to investigate the “missing heritability” and a successful identification of parent-of-origin effects could also help in shedding light on it. Indeed these effects, if not carefully accounted for, could mask the associations and reduce the portion of explained heritability.

State of the art

Usually, GWASs include unrelated individuals only and the parental origin of the alleles cannot be inferred. In fact, current methods for detecting parent-of-origin effects are based on the identification of the parental ancestry for each inherited alleles. Data from cases and their parents can be analysed by likelihood-based test method (Weinberg, 1999; Weinberg et al., 1998); this approach produces estimates of the relative risks associated with a particular variant allele for imprinting. In the contest of linkage analysis of quantitative trait, Hanson et al. (2001) develop a method based on the estimation of the proportion of marker alleles shared identical by descent between siblings; to assess POEs, it is necessary to partition the proportion of identical copies of the same allele shared between the siblings into a component derived from the mother and a component derived from the father. These two different matrices can be used as kinship matrix in the linear mixed model to estimate the association. Clearly, the ability to partition alleles shared identical by descent into maternal and paternal components requires the genotype data on at least one parent. Cui et al. (2006) suggest to testing imprinted quantitive trait loci in inbred F_2 mice by a two-component mixture distribution for the heterozygous group and by a two component mixture for the homozygous group. In case of case-parent-triad data or pedigree data, Belonogova et al. (2010) introduce a two-steps approach for detecting POEs in association studies of quantitative traits. First, for each locus studied, the probability of an allele parental origin is estimated using multipoint haplotype reconstruction. Next, the parental origin of these alleles are included as a covariate in regression models during the second step of GRAMMAR, Genome-wide Rapid Association using Mixed Model And Regression (Aulchenko et al., 2007). If the genotype data of the parents is not available then it is not possible to apply the methods proposed in the current literature.

A novel interesting approach to detect POEs in genome-wide genotype data of unrelated individual is presented in Hoggart et al. (2014). The key idea is to verify the presence of POEs through a test on difference between the phenotypic variance of the heterozygous genotype group and the variance observed in the homozygous genotype group. The assumption is 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, each showing a different means. By considering a bi-allelic SNP, in which "A" is the reference allele and "B" is the alternative one, Hoggart et al. (2014) suggest that in the homozygous genotype groups (*i.e.*, "AA" and "BB") a phenotype y is distributed as a normal with mean equals to μ_{AA} or $\mu_{AA} + \beta_M + \beta_P$ according to the genotype group membership and with same variance: σ_E^2. The mean in the AA genotype group is denoted as μ^{AA}, whereas β_M and β_P stand for the maternal and the paternal effects of the B allele, respectively. The trait y in the heterozygous groups is modelled in the following way:

$$y^{AB} = \mu^{AA} + z\beta_M + (1-z)\,\beta_P + \varepsilon \tag{1.1}$$

where z is a Bernoulli random variable with parameter $\frac{1}{2}$, and ε is an individual level error with mean zero and variance σ_E^2. The key element of this model is the random variable z that allow to identify the two sub-population in the heterozygous genotype group and implicate an increase of the variance. Indeed, it is possible write the equation 1.1 in the following way:

$$y_{AB} = (\mu_{AA} + \beta_P) + z(\beta_M - \beta_P) + \epsilon \tag{1.2}$$

Thus, the variance of the trait y in the heterozygous group is given by

$$\begin{aligned} Var(y_{AB}) &= Var(z)(\beta_M - \beta_P)^2 + Var(\epsilon) \\ &= \frac{1}{4}(\beta_M - \beta_P)^2 + \sigma_E^2 \\ &= \sigma_{AB}^2 \end{aligned} \tag{1.3}$$

In presence of POEs the effect of the allele B inherited from the mother is different from the effect of the allele B inherited from the father, $\beta_M \neq \beta_P$; thus, $\sigma_{AB}^2 > \sigma_E^2$.

This results in a two-steps approach that can be applied in case of unrelated data. In the first step the standard GWAS is computed; in the second step, to investigate the presence of POEs, a non parametric test for the equality of the two variances is performed. This approach is different from those present in the state of the art because it can be used in case of unrelated data, so it is not necessary the parental genotype data, but on the other side the main limitation of this approach is that it can be applied only in case of unrelated data.

Thesis Objective, Solutions and Organization

We propose two different procedures for detecting the POEs in genome wide genotype data from related individuals (monozygous and dizygous twins) when the parental origin cannot be inferred. Firstly, an extension of the approach proposed by Hoggart is developed and it is described in the Chapter 2. The proposed procedure, that we named “Variance Approach”, is composed by two different phases: a first test for the detection of POEs and a second test for the association, the choice of the association model is related to the presence of POEs. A simulation study has been developed to show the performance of the proposed statistical procedures. The procedures has been applied on gene expression data of multiple human tissues of the MuTHER Study (Nica et al., 2011). The second contribution, discussed in the Chapter 3, explores a finite mixture model approach (McLachlan and Basford, 1988) and it is based on the idea that in case of POEs the population can be clustered in two different groups in which the reference allele is inherited by a different parent. Thus, our statistical model is developed within the context of the finite mixture of linear mixed model and we propose a statistical test for capturing the presence of the POEs. A large simulation study and a study on gene expression data (Nica et al., 2011) has been conducted.

Chapter 2

Contribution 1

2.1 Introduction

Several measures can be used to describe a quantitative trait and to infer a particular biological phenomena. The effect of a SNP on the mean values of a trait of interest is widely studied through genetic association studies which lead to the identification of genetic determinants. The use of differences in the variance of these traits per genotype is an important topic, because it is well known that a certain number of biological scenarios can lead to variance-heterogeneity across the genotype group of a SNP. For instance, an underlying interaction between two genetic markers, or between a SNP and an environmental factor, can result in variance heterogeneity of trait of interest (Deng et al., 2014; Forsberg et al., 2015; Struchalin et al., 2010). Variance heterogeneity can also be generated by aware and unaware transformations on a phenotype (Sun et al., 2013). Usually, aware transformations can happen for statistical purposes (e.g., log(Y)) whereas unaware transformations can occur due to the wrong selection of the phenotype measurement. Hoggart et al. (2014) have shown that it is possible to detect POEs by exploiting a variance test between the heterozygotes and homozygotes.

Several methods have been proposed to investigate the variance heterogeneity. The most popular one to evaluate the heterogeneity between k groups is the Levene's test (Levene, 1960) that has proven to be excellent in terms of power and robustness under non-normality. The Levene's test consists of applying the standard one-way analysis of variance (ANOVA) on the absolute deviations of the trait of interest from the group means. Since its introduction various modifications of the Levene's test have been proposed. Parra-Frutos (2009) proposes a large simulation study to analyse the robustness properties in terms of type I error and power of several heteroscedasticity tests based on the Levene's test, as the Brown-Forsythe test (Brown and Forsythe, 1974), in which a robust estimate of location (median and 10% trimmed mean) has been proposed as alternative to the mean in computing the absolute deviations, and the Levene's test modified by Keyes-Levy's adjustment (Keyes and Levy, 1997) and Satterthwaite's correction (Satterthwaite, 1946) for unbalance design, with and without bootstrap. All these tests work under the assumption of independence of observations. Our aim is to evaluate the variance heterogeneity in case of twin data, where individuals are correlated. Only few scale tests deal with correlated data. Iachine et al. (2010) proposed a specific method to test heteroscedasticity in twin studies, but it does not control the type I error in the presence of non-normal data or in the case of small and unequal group sizes.

Furthermore this topic has been abundantly studied in the econometric field too (Breusch and Pagan, 1979; Glejser, 1969; Goldfeld and Quandt, 1965; Koenker, 1981; White, 1980); in particular as diagnostic methods to verify the validity of the homoscedasticity assumption in regression models. In the presence of uncorrelated observations, the analysis of the residuals of a linear regression model is suggested to verify the homoscedasticity assumption. Essentially, the regression analysis is applied under the assumption that there is no heteroscedasticity, and the residuals or the squared residuals are used as proxy of the error. Graphical methods are largely used to verify the presence of some kinds of regular pattern in the residuals but, in the big data settings, they result computationally demanding. Breusch and Pagan (1979) propose a multi-step procedure based on the Lagrange Multiplier test as test for the heteroscedasticity, in which the key idea is to compute a auxiliary regression using the residuals as dependent variable. The Breusch-Pagan has shown to be sensitive to any violation of the normality assumption, a generalization of Breusch-Pagan test in the case of deviation from the normality has been proposed by Koenker (1981).

Our idea is to propose for the detection of the parent-of-origin effect a multi-step approach where the residuals of a linear mixed model are used as proxy of the error. The linear mixed models (LMMs) constitute the most popular alternative to analyse correlated data (Pinheiro and Bates, 2006) and they are widely used in the contest of twin data (Guo and Wang, 2002; Neale and Maes, 2004; Rabe-Hesketh et al., 2008) because they are able to represent properly such data structures. In the section 2.2 we will review three different type of residuals in linear mixed models; then, in the section 2.3 we will show the steps of the procedure for testing the presence of POEs. Finally, in section 2.4 and 2.5, we will present a simulation study to evaluate the performance of the procedure and the results of the analysis of the MuTHER data, respectively.

2.2 Residuals in linear mixed models

The general specification of LMMs can be written as:

$$y_i = X_i \beta + Z_i a_i + \varepsilon_i \tag{2.1}$$

where y_i is $(n_i \times 1)$ a vector of response variable measured on the group/family i, β is a $p \times 1$ vector of the fixed effects coefficients, X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ design matrices, respectively, a_i is a $(q \times 1)$ vector of random effects and ε_i is a $(n_i \times 1)$ vector of measurement errors. a_i and ε_i are independent and normally distributed:

$$a_i \sim \mathcal{N}_q\left(0, \sigma_A^2 G\right) \text{ and } \varepsilon_i \sim \mathcal{N}_{n_i}\left(0, \sigma_E^2 R_i\right), \text{ for } i = 1, ..., m \tag{2.2}$$

where m denotes the number of group/family and G and R_i are $(q \times q)$ and $(n \times n)$ positive definite matrices. We can write the model (2.1) in standard matrix notation:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{\varepsilon} \tag{2.3}$$

that implies

$$Var\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_A^2 \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_E^2 \boldsymbol{\Sigma} \end{bmatrix} \tag{2.4}$$

where $\boldsymbol{D} = I_m \otimes G$ and $\boldsymbol{\Sigma} = \oplus_{i=1}^m R_i$, with $\otimes$ denoting the Kronecker product, $\oplus$ denoting the direct sum and I_m the identity matrix of order m.

The classic definition of residual, introduced by Cox and Snell (1968), allows for a single source of variability. Hilden-Minton (1995), Verbeke and Lesaffre (1996) and Pinheiro and Bates (2006) define three types of residual that manage the presence of the extra source of variability:

- Marginal residuals, $\hat{\xi} = y - X\hat{\beta}$ that predict the marginal errors, given by $Za + \varepsilon$;
- Conditional residuals $\hat{\varepsilon} = y - X\hat{\beta} - Z\hat{a}$ that predict the conditional errors;
- BLUP, $Z\hat{a}$, that predicts the random effects.

Santos Nobre and Da Motta Singer (2007) summarize which type of residuals can be used for the diagnostic of LMM, *e.g.* assessing normality and homoscedasticity of residuals, checking linearity of the effects and checking for outliers. For the evaluation of the normality and the homoscedasticity of the LMM Pinheiro and Bates (2006) consider a plot of the elements of $\frac{\hat{\varepsilon}}{\hat{\sigma}_E}$, where $\hat{\sigma}_E$ is an estimate of σ_E, versus the predict value $\hat{y} = X\hat{\beta} - Z\hat{a}$. Similar proposals to check for homoscedasticity are suggested by Oman (1995); Weiss and Lazaro (1992).

Thus, the conditional residuals are the more appropriate proxy of the error for the evaluation of the variance heterogeneity between the genotype groups. Furthermore, the conditional residuals are widely used in the GWASs. Indeed, Aulchenko et al. (2007) proposes a genome-wide rapid association using mixed model and regression (GRAMMAR) implemented in the well-known GenABEL software where first the residuals from a LMM are estimated under the null model (no SNP effect) and then they are treated as dependent variable (phenotype) for a genome-wide analysis by a standard linear model. Clearly, this approximated approach reduces substantially the computational time per-SNP.

2.3 Detection of POEs

2.3.1 Hoggart's test

Let us denote the alleles of a bi-allelic SNP as "A" (reference) and "B" (causal), so the possible genotypes are AA, AB and BB. In the GWAS study the major interest is to test the association between a particular SNP and the trait of interest, y; genetic association studies assume that the effect of a casual allele is the same regardless of its origin is paternal or maternal. Hoggart et al. (2014) showed that in presence of POEs the heterozygous genotype group is split into two subgroups, depending on the parental origin of the A and B alleles. As discussed in the Introduction, Hoggart et al. (2014) assumed that a trait y of any individual i in each genotype group $j =$ AA, AB, BB , can be modelled in the following way:

$$y_i = \begin{cases} \mu_{AA} + \varepsilon_i & if\ g_i \in AA \\ \mu_{AA} + z_i\beta_M + (1 - z_i)\,\beta_P + \varepsilon_i & if\ g_i \in AB \\ \mu_{AA} + \beta_M + \beta_P + \varepsilon_i & if\ g_i \in BB \end{cases} \tag{2.5}$$

where μ_j denotes the mean of the genotype group j, g_i indicates the genotype group for the individual i, z_i is a Bernoulli random variable with parameter $\frac{1}{2}$, β_M and β_P are the maternal and the paternal effects of the B allele respectively, and ε is an individual level error with mean zero and variance σ_E^2. The average of the maternal- and the paternal effects ($i.e.$ $\frac{\beta_M+\beta_P}{2}$) is equal to the association effect size β, the fixed effect of a SNP codified in a numerical form (0,1,2) denoting the number of the minor alleles B.

In the equation (1.3), we observe that the variance in the heterozygous group increases because the random variable z, that, instead is not present in the case of the homozygous genotype groups.

Hoggart et al. (2014) propose a robust version of the Brown-Forsythe test in which he computes the deviation from the median of the phenotype in each group:

$$\tilde{y}_i = \begin{cases} y_i - \mu_{AA} & \text{if } g_i = AA \\ y_i - \mu_{AB} & \text{if } g_i = AB \\ y_i - \mu_{BB} & \text{if } g_i = BB \end{cases} \tag{2.6}$$

where the general μ_j is the median of the genotype group j. The absolute deviation $|\tilde{y}_i|$ is regressed on a dummy variable, that assumes value 1 for heterozygous and 0 for homozygous individuals, in order to estimate the POE effect size, b. The POE test statistic, for large n, is:

$$\frac{\hat{b}}{SE_b} \sim \mathcal{N}(0,1) \tag{2.7}$$

where $\hat{b}$ is the estimated POE effect size and SE_b is the corresponding standard error that can be write in the following way

$$SE_b^2 = \frac{RSS}{n-1}\bigg/\frac{n_{AB}\left(n_{AA}+n_{BB}\right)}{n}$$

where

$$RSS = \sum_{i=1}^{n}\left(\tilde{y}_i - \alpha\right)^2 - \hat{b}^2 n_{AB}\frac{\left(n_{AA}+n_{BB}\right)}{n} \text{ and } \alpha = \sum_{i=1}^{n}\frac{|\tilde{y}_i|}{n}$$

2.3.2 Proposed method

In this section, our aim is to expose the procedure proposed for the detection of the POEs in the case of related subjects using twin data. As shown in the Sections 2.1 and 2.2, the use of the residuals for the study of the heteroscedasticity is very common. For the analysis of twin data, we suggest to use the conditional residuals of a LMMs where the random effect a of the equation (2.1) represents the covariance structure of the related individuals. In particular, in genetic models in case of related data, the total variance of the trait y is decomposed in two components (Falconer et al., 1996; Fisher, 1919): additive genetic and a unique environmental effects (ε), that correspond to the random effects a and the measurement error ε in the equation (2.1). The additive genetic variance is estimated using only within-family differences.

Thus, by exploiting the LMMs, we can write the model in the equation (2.5) for the related model in the following way:

$$y_{ij} = \begin{cases} \mu^{AA} + a_i + \varepsilon_{ij} & if\ g_{ij} \in AA \\ \mu^{AA} + z_{ij}\beta_M + (1 - z_{ij})\,\beta_P + a_i + \varepsilon_{ij} & if\ g_{ij} \in AB \\ \mu^{AA} + \beta_M + \beta_P + a_i + \varepsilon_{ij} & if\ g_{ij} \in BB \end{cases} \tag{2.8}$$

where g_{ij} defines the genotype group of the jth individual of the ith twin pair, with $j = 1, 2$ and $i = 1, ..., m$. Since a_i, ε_{ij} and z_{ij} are assumed to be mutually independent, the variance of the trait in the homozygous groups, $(i.e.\ y^{hom})$, is

$$Var\left(y^{hom}\right) = Var\,(a) + Var\,(\varepsilon) = \sigma_A^2 + \sigma_E^2 \tag{2.9}$$

and in the heterozygous group $(i.e.\ y_{AB})$ is

$$\begin{aligned} Var\left(y^{AB}\right) &= Var\,(z)\,(\beta_M - \beta_P)^2 + Var\,(a) + Var\,(\varepsilon) & (2.10) \\ &= \frac{1}{4}\,(\beta_M - \beta_P)^2 + \sigma_A^2 + \sigma_E^2 & (2.11) \end{aligned}$$

In presence of POEs $\beta_M \neq \beta_P$, therefore $Var\left(y^{AB}\right) > Var\left(y^{hom}\right)$. Thus, as proposed by Hoggart et al. (2014), we can detect POE via the increased trait variance in the heterozygous group relative to the homozygous groups. To test the POEs we propose a “Variance Approach” in a three steps procedure to avoid spurious POEs caused by population structure, unequal relatedness among individuals in a given cohort and covariates as the age. Then, our procedure is completed by a further step to test the association. The “Variance Approach” can be summarized in the following steps:

STEP-1 Estimating a linear mixed model under the null model (without SNP effect, see Section 2.5);

STEP-2 Computing the conditional residuals of the linear mixed model ;

STEP-3 Computing the test statistic (2.7) using the residuals obtained by the STEP 2.

STEP-4 Estimating in presence of POEs a linear heteroscedastic mixed model, otherwise a linear homoscedastic model.

Le us stress that the presence of POEs leads to heteroscedasticity in the trait y. To detect association and to take into account the heteroscedasticity caused by the POEs, we can relax the assumption of homoscedasticity using a variance function model for the within-group errors in the linear mixed model framework (Davidian and Giltinan, 1995; Pinheiro and Bates, 2006). By adding a stratum-specific variance parameter, $\boldsymbol{\delta}_{s_l}$, we can impose that the variance of the error is given by $\sigma_E^2 \boldsymbol{\delta}_{s_l}^2$, $l = 1, 2$, thus we can estimate different variances for the different levels of the stratification variable S. In presence of POEs, the variance of the homozygous group is different from the one of the heterozygous group, thus we introduce as stratification variable S a 0-1 coded genotype group identifier (1 for heterozygous and 0 for homozygous individual).

2.4 Simulation study

We explored the power and the efficiency of the “residual approach” to detect the POEs through a simulation study. We simulated a continuous trait according to the model in equation (2.8); we imposed $m = 500$ twin pairs, divided into MZ and DZ, the 30% and 70%, respectively. In this way, we are able to simulate complete and partial POE. Thus, the variance of the simulated trait is given by the sum of σ^2_β, the variance explained from the main allelic effect β, σ^2_A, the additive variance, and σ^2_E, the environmental variance, where $\sigma^2_\beta + \sigma^2_A + \sigma^2_E = 1$. σ^2_E has been fixed to three different values (0.70, 0.50, 0.30), corresponding to a total heritability, h^2, (*i.e.* sum of σ^2_β and σ^2_A), of 0.30, 0.50, 0.70. The main allelic effect β has been derived from the variance explained σ^2_β that has been fixed to 1%, 10%, 20% and 30% of the total trait variation. As illustrated in the Appendix (A.2), the explained variance can be written in the following way:

$$\sigma^2_\beta = 4\beta^2 \left(2k^2 - 2k + 1\right) pq. \tag{2.12}$$

The parameter k indicates the proportion of the main allelic effect explained from the allele B inherited from the parents "M", in other words, we can say that k denotes the absence/presence of POE and its intensity. It is possible to say that the POE are not observed only when $k = 0.5$. In all other situations, the maternal and paternal alleles contribute to the POE in a percentage equal to k e 1-k. In the extreme cases, where k is equal to 1 or 0, we face a paternal POE (the paternally derived allele is completely silenced) or a maternal POE (the maternally derived allele is completely silenced), respectively.

Under the null hypothesis, we have that $\beta_M = \beta_P$, thus we impose $k = 0.5$. 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. Let us note that the equation (2.12) highlights the possibility of underestimate the h^2 when a missed POE identification occurs. Indeed, from the equation (2.12) is it possible to observe that, by keeping constant β and p, the explained variance in the case of $k \neq 0.5$ is always lower than the explained variance in presence of POEs. In particular when k is equal to 1 or 0 the variance explained by the main allelic effect is doubled than the variance explained under the null hypothesis. This is the reason why a successful identification of parent-of-origin effects and an exact estimation of the variance explained by the genetic marker could help in shedding light on "missing heritability" problem.

Summing up, to evaluate the power and the type I error of our procedure we have simulated 36 and 12 different scenarios, respectively. Each scenarios is replicated for four levels of minor frequency allele (MAF = 0.1, 0.2, 0.3, and 0.5).

2.4.1 Results of simulation study

Type I error rates at a significance level of 0.05 for the evaluated methods are shown in Figure 2.2 where we can see that the type I error of the test is in good agreement with the nominal 5% level for low values of heritability, h^2, and for high values of MAF. In case of $h^2 = 0.7$ and low MAF, the test in less conservative; indeed in case of MAF = 0.2 and MAF = 0.1 the type I error rate is approximately the 8% and the 10%, respectively. Moreover, from the error bars we can observe that in case of MAF = 0.1, the type I error is more variable; in other words, for different levels of variance explained by the main allelic effect the type I error assumes different values.

Regarding the power of the test, from the Figure 2.1, we can immediately observe that the power of the test increases as the variance explained σ^2_β increases. Indeed, if the proportion of the variance explained by the marker is $\sim$ 1%, the power of the test is nearly 0. On the contrary, the power approaches to 1 when the marker explains the 20% or 30% of the total trait variation, in particular in the case of a 100% maternal POE. Moreover, we note that the power increases when the MAF becomes smaller. It occurs because, as the MAF decreases the coefficient β has to assume high value to explain the same proportion of the total trait variation (as shown in the Appendix A.2).

We can conclude that in general, under the null hypothesis, the probability to detect POE is equivalent to the nominal value, and that the test is inflated only in the case of high heritability and low MAF. On the other hand, the performances in terms of power are good only in case of high values of the variance explained by the main allelic effect. Usually, in real data, a SNP explains around the 1% - 5% of the total trait variation when the analysed trait is a phenotype, while in the case of gene expression data the proportion of the variance explained by the genetic component is higher, $\sim$ 10% - 40%.

2.5 Application to MuTHER Study Data

In this study, we analyse the gene expression microarray data of the MuTHER study in a diverse set of human tissues, skin and fat, and the lymphoblastoid cell lines, LCL. The data are described in the details in the Appendix A.1. Briefly, the sample is composed by a total of 856 female twins (of which 154 monozygotic and 84 singletons). The gene expression is measured on $\sim$ 48000 probes and, after quality control, the genotype data is constituted by $\sim$ 1 million of SNPs. For the detection of the parent-of-origin effects, we selected SNPs in eQTLs in cis within a 1Mb window of the gene transcripts at a FDR of 5%. The linear mixed model of the STEP 1 was adjusted for age, experimental batch and BMI (only for skin and fat), as fixed effects, and for twin pairing as random effects.

In the Figure 2.3 and 2.4, we can observe the p-values of the POE test for the three different tissues in the case of the Bonferroni correction and the FDR adjustment. We identified one gene with POE p-values statistically significant with both the multiple testing correction, in more of 20 loci that are overlapped in the three tissues in the chromosome 14 (region 14q23.3, gene CHURCH1), although the signals are higher in more loci in the lymphoblastoid cell lines. As explained in the Introduction of this chapter, the

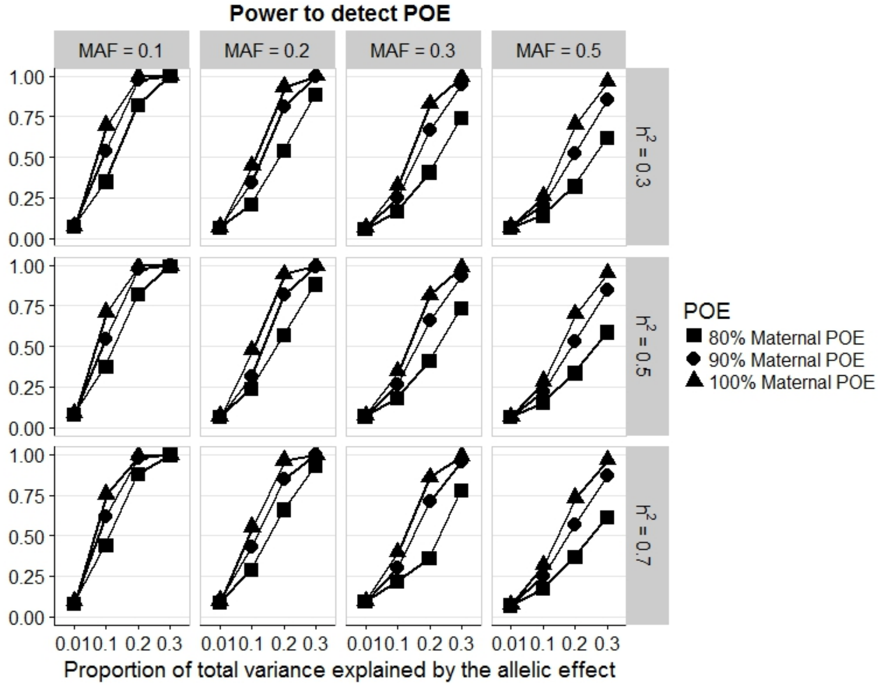


Figure 2.1: Power to detect POEs by the "Variance Approach". The columns show power for different MAF; the power achieved under three heritability model is shown in rows. The y axis of each panel shows power, whereas the x axis shows the proportion of total variance explained by the allelic effect. POE size was set to 80%, 90% and 100% of the allelic effect size.

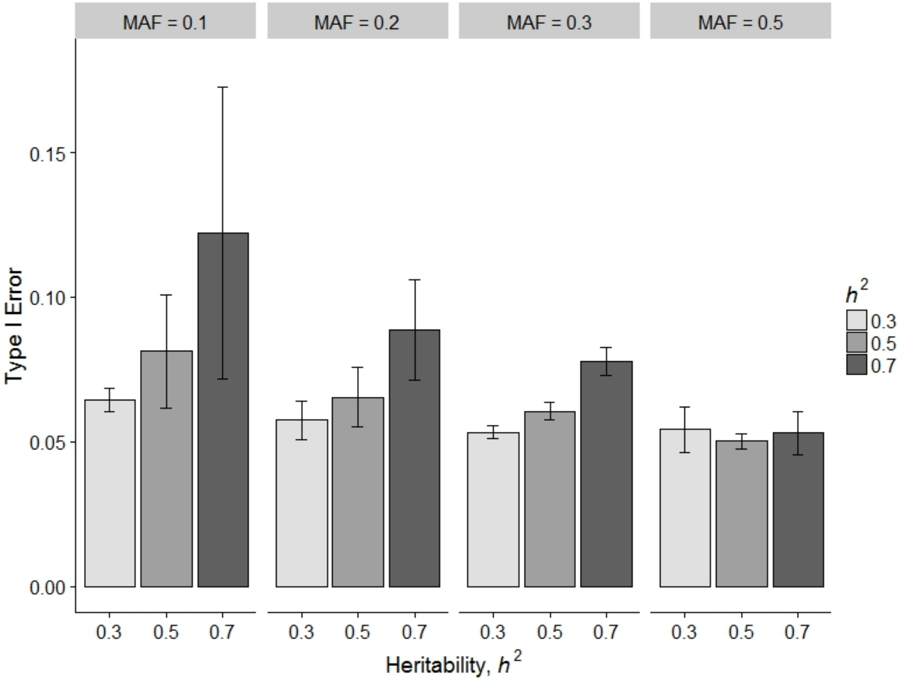


Figure 2.2: Type I Error for the proposed test of POEs . The columns show type I error under four different level of MAF. The y axis of each panel shows type I error, whereas Different colours indicates the level of the heritability (0.3, 0.5, 0.7). Keeping constant the MAF and the heritability, the type I errors corresponding to the different levels of explained variance is synthesized in error bars at 95% confidence interval. intervals.

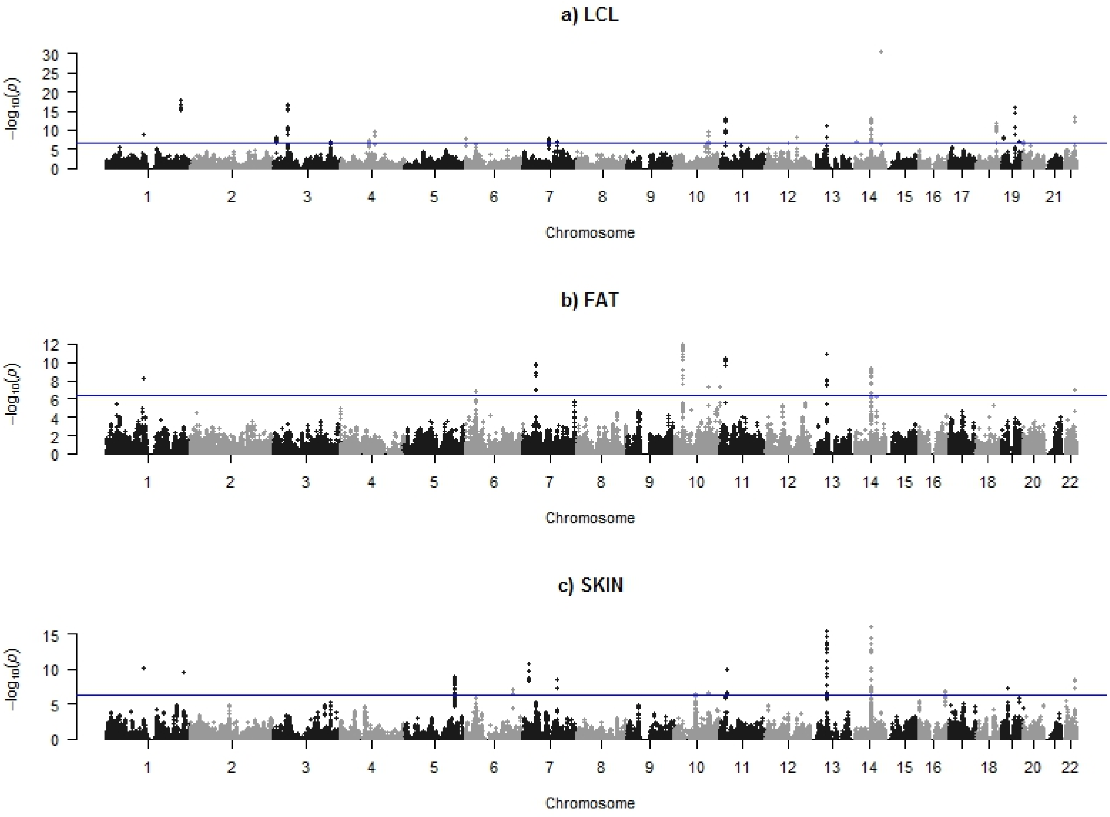


Figure 2.3: Manhattan Plot of POE p-values, the blue horizontal line is the Bonferroni correction

variance heterogeneity between the genotype groups can be the result of other biological phenomena. For this reason, all these results are validated checking the assumption of homoscedasticity in the homozygous groups thought the classic Levene's test. The regional signal plots of the CHURCH1 locus for the three tissues are in the Figure 2.5. The region 14q23.3 is a deletion region, that regulates neurological development and seems to be associated with some neuronal disorder as the autism (Griswold et al., 2011). The analysis conducted by Hu et al. (2015) revealed the father is a carrier for the 14q23.3 deletion.

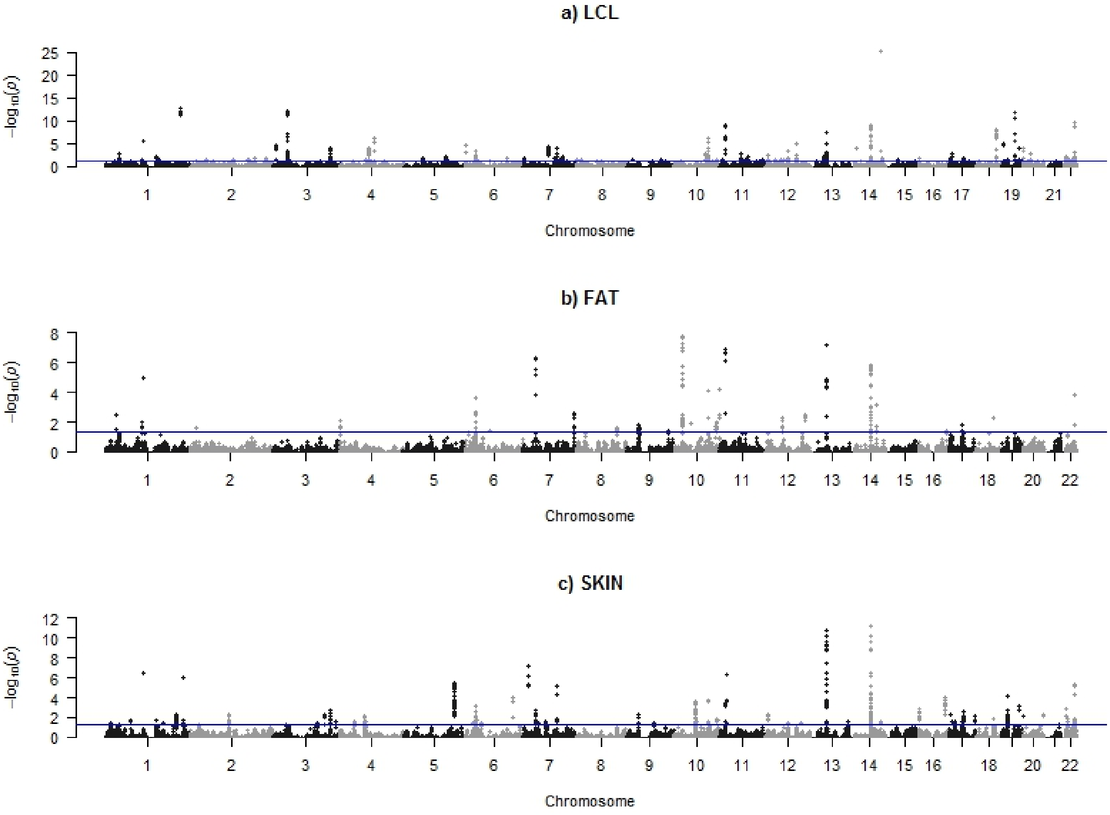
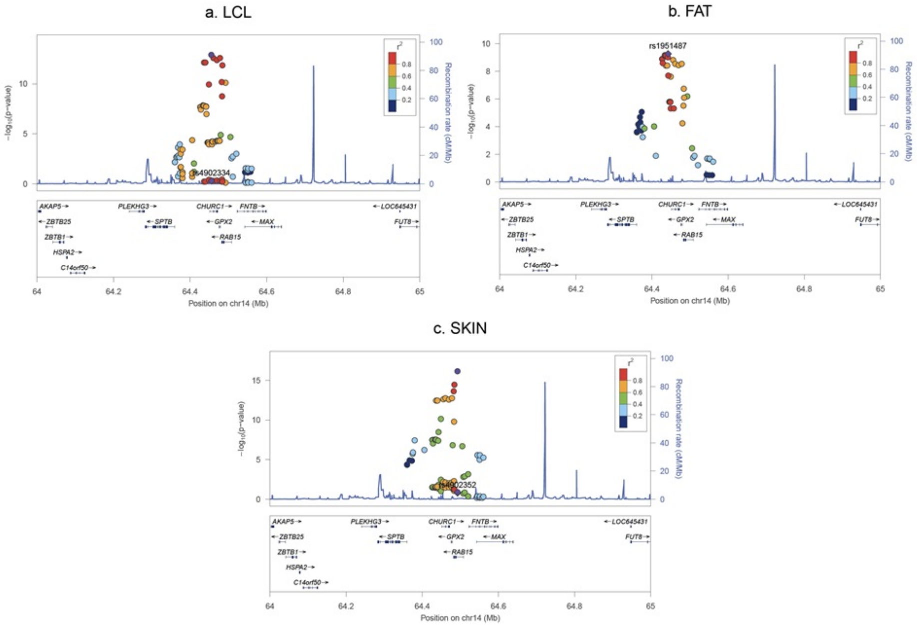


Figure 2.4: Manhattan Plot of POE FDR-adjusted p-values, the blue horizontal line is significance level at 5%

Moreover we identified in LCL other three most likely imprinted genes (Figure 2.6): C20orf194 in 30 loci (chr 20), SERPINB10 in 29 loci (chr18) and SERPIND1 in 49 loci (chr 22). To assess the validity of the genetic results, we have verified that in the identified regions the gene expression variation is not caused by the presence of the copy number variations (CNVs). The CNVs consist in deletions or duplications of chromosomal segments and they constitute the major source of variation between the individuals. From a fast analysis of the CNVs, from the TwinsUK genotype data, we have observed that in these region only one individual show a CNV in the chromosome

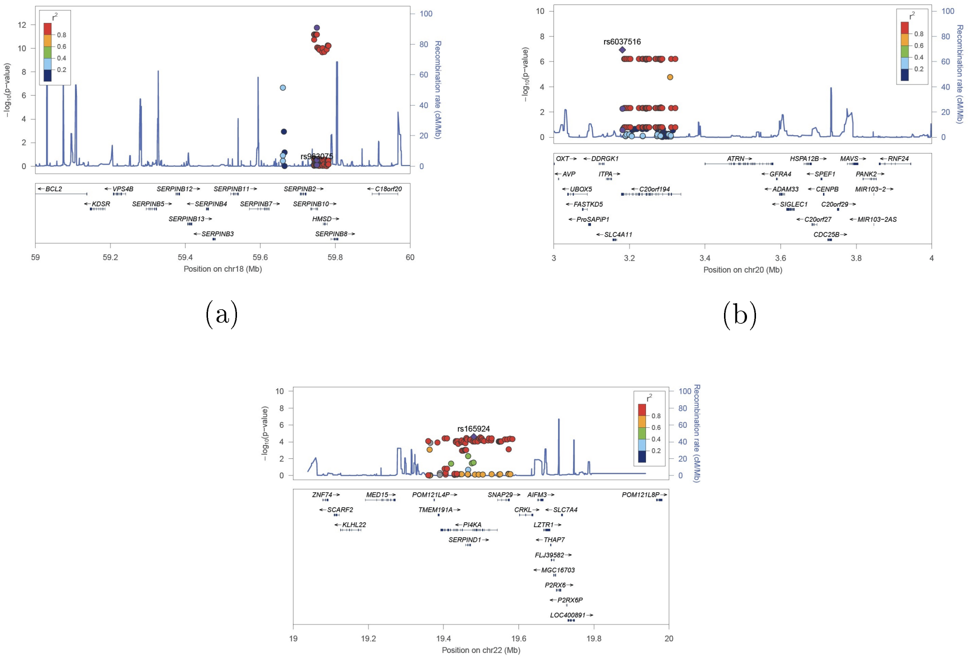

a. LCL

b. FAT

c. SKIN

Figure 2.5: Regional signal plots at the CHURCH1 locus.

22 close by the gene SERPIND1. Moreover, one of the mechanism underlying the allele-specific expression is the DNA methylation, the addition of a methyl group to DNA, that controls gene expression. The allele-specific methylation of differentially methylated regions (DMRs) is the primary epigenetic mechanism of imprinting, controlling monoallelic expression (Skaar et al., 2012). In order to validate the results, we will analyse the methylation data as suggested by Baran et al. (2015) and Joshi et al. (2016). Furthermore, an additional validation of the genetic results will be proposed in the Section 3.6.1.



(a)

(b)

(c)

Figure 2.6: Regional POE signal plots at locus (a) SERPINB10, (b) C20orf194 and (c) SERPIND1. LCL.

2.5.1 Spurious association due to data trasformation

One crucial step in the analysis of microarray data is the normalization. Normalization aims at adjusting microarray data for effects which arise from variations in the technology rather than from biological differences between the RNA samples or between the printed probes (Quackenbush, 2002). Indeed, normalization is necessary to eliminate low-quality measurements, to adjust the measured intensities and to allow direct array-to-array comparisons. A Log2 transformation is applied on gene expression data to improve the symmetry of the distribution and to make the distribution more Gaussian-like. Then, the Log2-transformed expression signals were normalized separately by quantile normalization across individuals.

Sun et al. (2013) state that the scales on which we measure interval-scale quantitative traits are man-made and have little intrinsic biological relevance. Therefore, Sun et al. (2013) argue that can be difficult to detect and interpret differences in trait variances among SNP-specific genotype. Before claim a biological interpretation for genotype differences in variance, we should be sure that no monotonic transformation of the data can reduce/eliminate or amplify/generate these differences.

Before the Log2 transformation we can assume that the gene expression can be approximated as a gamma distribution, $\tilde{y} \sim \mathit{\Gamma}\left(k, \theta\right)$. Let μ be the trait mean in the heterozygous group and δ the marginal effect of the SNP (on the original scale) and let $g\left(\right)$ denotes a Log2 transformation. By using a first order Taylor expansion, the variance of the transformed trait in the heterozygous group can be approximates as follows:

$$\begin{aligned} Var\left(g\left(\tilde{y}|G=AB\right)\right) &\simeq g^{'}\left(\mu\right)^{2} Var\left(\tilde{y}|G=AB\right) \\ &= \left(\frac{1}{\mu log2}\right)^{2} k\theta^{2} \\ &= \left(\frac{1}{k\theta log2}\right)^{2} k\theta^{2} \\ &= \frac{1}{2klog2} \end{aligned}$$

Similarly, it is possible estimate the variance in AA genotype group. Thus, $Var\left(g\left(\tilde{y}|G=AB\right)\right)-Var\left(g\left(\tilde{y}|G=AA\right)\right)$ is approximately equal to:

$$\begin{aligned} g^{'}\left(\mu\right)^{2}Var\left(\tilde{y}|G=AB\right)-g^{'}\left(\mu-\delta\right)^{2}Var\left(\tilde{y}|G=AA\right) \\ =\frac{1}{2klog2}-\frac{k\theta^{2}}{\left(k\theta-\delta\right)^{2}2log2} \end{aligned} \tag{2.13}$$

Therefore, in the presence of a strong marginal association, ($i.e.$ $\delta\to\infty$), the Log2 transformation can implicate a spurious association using variance tests. For this reason, we recommend to validate the top hits obtained by our POE test by excluding all the SNPs that violate the homoscedasticity assumption in the homozygous groups, by comparing the results in the case of transformed and untransformed trait, if it is possible, and by analysing the methylation data as suggested by Baran et al. (2015) and Joshi et al. (2016).

Chapter 3

Contribution 2

3.1 Introduction

Twin designs are commonly used to study and measure the environmental and genetic effects on a trait of interest (Falconer et al., 1996; Fisher, 1919; Neale and Maes, 2004). A major advantage of twin studies is the possibility to control and to establish the environmental effect on an observed trait because twins share the same genes. Indeed, monozygotic (MZ) twins share all of their genes and consequently become more similar than dizygotic (DZ) twins that share only about 50% of them. As proposed by Guo and Wang (2002) and Rabe-Hesketh et al. (2008), we handle the twin structure within the contest of linear mixed models. To the best of our knowledge, in the literature there are no models to detect the POEs in case of correlated data but in absence of the parental genetic information. The aim of this work is to propose a statistical approach for testing the POEs in case of twin studies. To achieve this goal we interpret the problem of the POEs detections as a model-based clustering problem. Indeed, we assume that, in presence of POEs, 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 that receive the causal allele "B" from the first parent M and from the second parent P, respectively. We account for the correlation between twins and we model the unobserved heterogeneity among individuals, due to the presence of POEs, at the same time. The method proposed for detecting the POEs is developed in the framework of mixtures of linear mixed effects models (MLMMs). Since their introduction by Pearson (1894), finite mixture models (FMMs) have been widely considered in statistical modelling in several research areas. FMMs provide a useful and powerful tool to deal with population heterogeneity (McLachlan and Basford, 1988) . In the last decades FMMs have been extended in several ways; contributions relative to the their estimation and application have increased thanks to the development of Expectation-Maximization (EM) algorithm (Dempster et al., 1977), that has reduced the complexity of the maximum likelihood (ML) estimation. Celeux et al. (2005) have extended FMM to linear mixed model context in order to introduce a model-based cluster analysis method for repeated data for the clustering of time-course gene expression data. Ng et al. (2006) and Wang et al. (2012) have incorporated multilevel and nested random effects and autoregressive random effects in the MLMM , respectively. In these works to take into account the correlation between the correlated observations and to model the heterogeneity between the components, they assume that the correlated measures of a statistical unit belong to the same mixture component, however, this assumption can be too much restrictive to describe some phenomena. Indeed, in the contest of the twin data and to meet genetic requirements, our model have to be more flexible: we assume that the monozygotic twin brothers have to belong to the same mixture component, on the contrary each individual of the dizygotic twin pair can belong to a different mixture component.

In Section 3.2, we introduce the proposed model and in Section 3.3 we describe the EM algorithm for the ML estimation and we highlight how we solve some of the main issues of the EM algorithm. Next, in Section 3.4, we derive the statistical test proposed for the detection of the POEs. Finally, we show a description of the simulation study results, in section 3.5, and we apply the proposed method on the gene expression data of the MuTHER study, (section 3.6).

3.2 Mixture Model for Twin data

Suppose we observe m unrelated twin-pairs, with $i = 1, \ldots, m$, with twins $j = 1, 2$ for each pair. Data contain two types of genetically related individuals, MZ and DZ twins; we denote $\mathbb{1}_{MZ}(i)$ the indicator function of i, having value 1 for all MZ twins and value 0 for all DZ twins.

We indicate the alleles of a bi-allelic SNP by "A" (reference) and "B" causal, so the possible genotype groups g_{ij} of the j-th individual of the i-th twin pair are AA, AB and BB. Let w_{ij}^{BB} and w_{ij}^{AB} be two dichotomic variables that identify the genotype group g_{ij} :

$$w_{ij}^{BB} = \begin{cases} 1 & if\ g_{ij} = BB \\ 0 & otherwise \end{cases} \tag{3.1}$$

and

$$w_{ij}^{AB} = \begin{cases} 1 & if\ g_{ij} = AB \\ 0 & otherwise \end{cases} \tag{3.2}$$

As proposed by Hoggart et al. (2014), in presence of POEs the heterozygous genotype group, AB, consists of two sub-populations depending on the reference allele is inherited from the father, Ω_P, or from the mother, Ω_M. In the contest of FMM, we take into account the unobserved heterogeneity due to the POEs, assuming that the mixture is composed by two components, where the component weights are known and equal to $\frac{1}{2}$.

Treating individuals as level-one units and the twin pairs as level-two units, we denote by y_{ij} the random variable that represents the trait of interest for the i-th twin pair and and the j-th individual. The trait is assumed to depend on fixed and random effects as follows:

$$\begin{aligned} y_{ij} &= \alpha + (\beta_M + \beta_P)\, w_{ij}^{BB} + \beta_P w_{ij}^{AB} + \mathbb{1}_{MZ}(i)\, z_i^{MZ} (\beta_M - \beta_P)\, w_{ij}^{AB} \\ &+ (1 - \mathbb{1}_{MZ}(i))\, z_{ij}^{DZ} (\beta_M - \beta_P)\, w_{ij}^{AB} + X_{ij}'\boldsymbol{\gamma} + u_i + \varepsilon_{ij} \end{aligned} \tag{3.3}$$

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, $\boldsymbol{\gamma} \in \mathbb{R}^p$ is a vector of fixed effects to be estimated. The random effects u_i, used to describe the correlation within each twin pair, is distributed as

$N\left(0,\tau^2\right)$, independently from the error $\varepsilon_{ij} \sim N\left(0,\sigma^2\right)$. z_i^{MZ} and z_{ij}^{DZ} are the latent component variables of the mixture. Indeed, depending on the zygosity of the i-th twin pair the two component mixture of mixed linear models is determined from a different latent variable:

- if $\mathbb{1}_{MZ}(i) = 1$, we introduce

$$z_i^{MZ} = \begin{cases} 0 & if\, the\, \text{i}th\; MZ\, pair\; \in \varOmega_P \\ 1 & if\, the\, \text{i}th\; MZ\, pair\; \in \varOmega_M \end{cases} \tag{3.4}$$

  that is a realization of a random variable $Z_i^M \sim Ber\left(1/2\right)$;

- if $\mathbb{1}_{MZ}(i) = 0$, we introduce

$$z_{ij}^{DZ} = \begin{cases} 0 & if\, the\, \text{j}th\; individual\; of\; the\, \text{i}th\; DZ\, pair\; \in \varOmega_P \\ 1 & if\, the\, \text{j}th\; individual\; of\; the\, \text{i}th\; DZ\, pair\; \in \varOmega_M \end{cases} \tag{3.5}$$

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

So to take into account the correlation between the twin pair and under the genetic assumption, we assume that the monozygotic twin pair belong to the same mixture component, as assumed by Celeux et al. (2005) in a different contest; on the contrary the dizygotic twin pair can belong to a different mixture component.

The components of the model proposed in equation (3.3) are the following:

$$f\left(Y_{ij}|Z_i^{MZ} = 0\right) = f\left(Y_{ij}|Z_{ij}^{DZ} = 0\right) = \mathcal{N}\left(\mu_{1_{ij}}, \sigma^2 + \tau^2\right) = f_1\left(y_{ij};\theta_1\right) \tag{3.6}$$

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

$$f\left(Y_{ij}|Z_i^{MZ}=1\right) = f\left(Y_{ij}|Z_{ij}^{DZ}=1\right) = \mathcal{N}\left(\mu_{2_{ij}}, \sigma^2+\tau^2\right) = f_2\left(y_{ij};\theta_2\right) \quad (3.7)$$

where $\mu_{2_{ij}} = \alpha + (\beta_M + \beta_P) w_{ij}^{BB} + \beta_M w_{ij}^{AB} + X'_{ij}\gamma$, $\theta_2 = \left\{\mu_{2_{ij}}, \sigma^2, \tau^2\right\}$. We define $\theta = \{\alpha, \beta_M, \beta_P, \gamma, \sigma^2, \tau^2\}$ the vector of parameters that is $(5+p) \times 1$ dimensional.

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

$$f\left(y_{ij};\theta\right) = \sum_{k=1}^{2} f_k\left(y_{ij};\theta_k\right) = \frac{1}{2}\mathcal{N}\left(y_{ij};\mu_{1ij},\sigma^2+\tau^2\right) + \frac{1}{2}\mathcal{N}\left(y_{ij};\mu_{2ij},\sigma^2+\tau^2\right). \quad (3.8)$$

Finite mixture models allow to estimate the posterior probability of belonging to each component. For the MZ twins, $\mathbb{1}_{MZ}(i) = 1$, we estimate the posterior probability that the i-th pair belongs to the k component:

$$\begin{aligned}
\tau_k^{MZ}\left(\boldsymbol{y}_i;\theta_k\right) &= \Pr\left(i th\ MZ\, pair\ \in \Omega_k|\boldsymbol{y}_i\right) \\
&= \Pr\left(Z_{ik}^{MZ}=1|\boldsymbol{y}_i\right) \\
&= \frac{\Pr\left(\boldsymbol{y}_i|Z_{ik}^{MZ}=1\right)\Pr\left(Z_{ik}^{MZ}=1\right)}{\Pr\left(\mathbf{y}_i\right)} \\
&= \frac{\frac{1}{2}\prod_{j=1}^{2} f_k\left(y_{ij};\theta_k\right)}{\sum_{k=1}^{2}\frac{1}{2}\prod_{j=1}^{2} f_k\left(y_{ij};\theta_k\right)}
\end{aligned}$$

For the DZ twins, $\mathbb{1}_{MZ}(i) = 0$, we estimate the posterior probability that the j-th individual of the i-th couple of twins belongs to the k component:

$$\begin{aligned}
\tau_k^{DZ}(y_{ij};\theta_k) &= \Pr\left(the\, \mathrm{j}th\ individual\ of\ the\, \mathrm{i}th\ DZ\ pair \in \Omega_k | y_{ij}\right) \\
&= \Pr\left(Z_{kij}^{DZ} = 1 | y_{ij}\right) \\
&= \frac{\Pr\left(y_{ij} | Z_{kij}^{DZ} = 1\right) \Pr\left(Z_{kij}^{DZ} = 1\right)}{\Pr(y_{ij})} \\
&= \frac{\frac{1}{2} f_k(y_{ij};\theta_k)}{\sum_{k=1}^{2} \frac{1}{2} f_k(y_{ij};\theta_k)}
\end{aligned}$$

3.3 Estimation with EM algorithm

Let $\theta = \{\alpha, \beta_M, \beta_P, \gamma, \sigma^2, \tau^2\}$ be the vector of model parameters. The log-likelihood of the model is given by

$$\mathcal{L}(\theta) = \sum_{i=1}^{m}\sum_{j=1}^{2} \ln f(y_{ij};\theta) = \sum_{i=1}^{m}\sum_{j=1}^{2} \ln\left(\sum_{k=1}^{2} \frac{1}{2} f_k(y_{ij};\theta_k)\right). \tag{3.9}$$

The direct maximization of the log-likelihood function, $\mathcal{L}(\theta)$, is complicated; a general technique 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 to compute MLEs in the contest of incomplete-data problems. Each iteration of the EM algorithm is composed by two steps: the E-step and the M-step and they are repeated until convergence. The aim of the EM algorithm is the maximization of the conditional expectation of the log-likelihood of the so called complete data given the observable data. The complete log-likelihood is the joint density of the observable and missing data of the model. 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

$$\begin{aligned}\mathcal{L}_C\left(\theta\right) = \sum_{k=1}^{2}\sum_{i=1}^{m}\sum_{j=1}^{2}\Bigg\{ & \mathbb{1}_{MZ}(i)\left[z_{ik}^{MZ}\ln\frac{1}{2} + z_{ik}^{MZ}\ln f_k(y_{ij}, u_i|z_{ik}^{MZ};\theta_k)\right] \\ & + \left(1 - \mathbb{1}_{MZ}(i)\right)\left[z_{kij}^{DZ}\ln\frac{1}{2} + z_{kij}^{DZ}\ln f_k(y_{ij}, u_i|z_{kij}^{DZ};\theta_k)\right]\Bigg\} \qquad (3.10)\end{aligned}$$

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

$$\ln f_k(y_{ij}, u_i|\cdot;\theta_k) \propto -\frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left(y_{ij} - \mu_{kij} - u_i\right)^2 - -\frac{1}{2}\ln\tau^2 - \frac{1}{2\tau^2}u_i^2 \qquad (3.11)$$

At the 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 $\theta^{(r)}$, where the r represents the iteration index. Thus, the expectation of the complete log-likelihood is defined by

$$\begin{aligned}\mathcal{Q}(\theta, \theta^{(r)}) &= \mathbb{E}_{\theta^{(r)}}\{\mathcal{L}_C\left(\theta\right)|Y\} \\ &\propto \sum_{k=1}^{2}\sum_{i=1}^{m}\sum_{j=1}^{2}\Bigg\{\mathbb{1}_{MZ}(i)\tau_k^{MZ}(\boldsymbol{y}_i;\theta_k^{(r)})\mathbb{E}_{\theta^{(r)}}\left[\ln f_k(\boldsymbol{y}_i, u_i|z_{ik}^{MZ};\theta_k)|\boldsymbol{y}_i\right] \\ &\quad + \left(1 - \mathbb{1}_{MZ}(i)\right)\tau_k^{DZ}(y_{ij};\theta_k^{(r)})\mathbb{E}_{\theta^{(r)}}\left[\ln f_k(y_{ij}, u_i|z_{kij}^{DZ};\theta_k)|y_{ij}\right]\Bigg\}. \qquad (3.12)\end{aligned}$$

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

$$\tau_k^{MZ}(\boldsymbol{y}_i;\theta_k^{(r)}) = f(z_{ik}^{MZ}|\boldsymbol{y}_i;\theta_k^{(r)}) = \frac{\frac{1}{2}\prod_{j=1}^{2}\mathcal{N}(y_{ij};\mu_{kij},\tau^2+\sigma^2)}{\sum_{k=1}^{2}\frac{1}{2}\prod_{j=1}^{2}\mathcal{N}(y_{ij};\mu_{kij},\tau^2+\sigma^2)} \tag{3.13}$$

$$\tau_k^{DZ}(y_{ij};\theta_k^{(r)}) = f(z_{kij}^{DZ}|y_{ij};\theta_k^{(r)}) = \frac{\frac{1}{2}\mathcal{N}(y_{ij};\mu_{kij},\tau^2+\sigma^2)}{\sum_{k=1}^{2}\frac{1}{2}\mathcal{N}(y_{ij};\mu_{kij},\tau^2+\sigma^2)}. \tag{3.14}$$

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

For this variant of 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

$$\Sigma_{u_i}^{MZ} = \left(\frac{1}{\tau^2}+\frac{2}{\sigma^2}\right)^{-1} \tag{3.15}$$

and the conditional mean is given by

$$\mu_{u_i,k}^{MZ} = \Sigma_{u_i}\frac{1}{\sigma^2}\mathbb{1}'(\boldsymbol{y}_i-\boldsymbol{\mu}_{ik})^2 \tag{3.16}$$

where $\boldsymbol{y}_i$ is the observed data vector 2×1 dimensional of the i-th twin pair and $\boldsymbol{\mu}_{ik}$ 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

$$\Sigma_{u_i,k}^{DZ} = (\tau^2+\sigma^2)^{-1}\sigma^2\tau^2 \tag{3.17}$$

and the conditional mean is given by

$$\mu_{u_i,kj}^{DZ} = (1+\sigma^2\tau^{-2})^{-1}(y_{ij}-\mu_{kij})^2 \tag{3.18}$$

The computation of conditional mean and conditional variance of u_i for both cases are reported in the appendix B.1.

The M-step consists of determining the values maximizing the equation (3.12) where

$$\begin{aligned}\mathbb{E}\left[\ln f(\boldsymbol{y}_i, u_i|z_{ik}^{MZ};\theta_k)\right] \propto & -\ln\sigma^2 - \frac{1}{2}\left[\left(\boldsymbol{y}_i - \boldsymbol{\mu}_{ik} - \mu_{u_i,k}^{MZ}\right)^2 + \Sigma_{u_i,k}^{MZ}\right] \\ & -\frac{1}{2}\ln\tau^2 - \frac{1}{2\tau^2}\left(\mu_{u_i,k}^{MZ} + \Sigma_{u_i,k}^{MZ}\right)\end{aligned} \tag{3.19}$$

and

$$\begin{aligned}\mathbb{E}\left[\ln f(y_{ij}, u_i|z_{ijk}^{DZ};\theta_k)\right] \propto & -\frac{1}{2}\ln\sigma^2 - \frac{1}{2}\left[\left(y_{ij} - \mu_{kij} - \mu_{u_i,kj}^{DZ}\right)^2 + \Sigma_{u_i,k}^{DZ}\right] \\ & -\frac{1}{2}\ln\tau^2 - \frac{1}{2\tau^2}\left(\mu_{u_i,kj}^{DZ} + \Sigma_{u_i,k}^{DZ}\right).\end{aligned} \tag{3.20}$$

It follows that

$$\begin{aligned}\mathcal{Q}(\theta,\theta^{(r)}) \propto & \sum_{k=1}^{2}\sum_{i=1}^{m}\sum_{j=1}^{2}\Bigg\{-\frac{1}{2}\ln\sigma^2 - \frac{1}{2}\ln\tau^2 \\ & -\frac{1}{2}\mathbb{1}_{MZ}(i)\left[\frac{\left(y_{ij}-\mu_{ijk}-\mu_{u_i,k}^{MZ}\right)^2+\Sigma_{u_i,k}^{MZ}}{\sigma^2} - \frac{1}{\tau^2}\left(\mu_{u_i,k}^{MZ}+\Sigma_{u_i,k}^{MZ}\right)\right]\Bigg\} \\ & -\frac{1}{2}(1-\mathbb{1}_{MZ}(i))\left[\frac{\left(y_{ij}-\mu_{ijk}-\mu_{u_i,kj}^{DZ}\right)^2+\Sigma_{u_i,k}^{DZ}}{\sigma^2} - \frac{1}{\tau^2}\left(\mu_{u_i,kj}^{DZ}+\Sigma_{u_i,k}^{DZ}\right)\right]\Bigg\}\end{aligned} \tag{3.21}$$

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, (3.21), with respect to parameters α, β_M, β_P, $\boldsymbol{\gamma}$, τ^2 and σ^2, and setting them to zero. Thus we obtain:

$$\hat{\alpha} = \frac{1}{N} \sum_{k=1}^{2} \sum_{i=1}^{m} \sum_{j=1}^{2} \tilde{y}_{kij} \left(\mathbb{1}_{MZ}(i) \tau_k^{MZ} + (1 - \mathbb{1}_{MZ}(i)) \tau_k^{DZ} \right) \tag{3.22}$$

where $N = 2m$ and

$$\tilde{y}_{kij} = \begin{cases} y_{ij} - (\beta_M + \beta_P) w_{ij}^{BB} - \beta_P w_{ij}^{AB} - X'_{ij}\boldsymbol{\gamma} - \mu_{u_i,1} & k = 1 \\ y_{ij} - (\beta_M + \beta_P) w_{ij}^{BB} - \beta_M w_{ij}^{AB} - X'_{ij}\boldsymbol{\gamma} - \mu_{u_i,2} & k = 2 \end{cases}, \tag{3.23}$$

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

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

$$\hat{\beta}_M = \frac{1}{n_{BB} + \sum_{i=1}^{m} \sum_{j=1}^{2} \tau_2 w_{ij}^{AB}} \left\{ \sum_{k=1}^{2} \sum_{i=1}^{m} \sum_{j=1}^{2} w_{ij}^{BB} \left(y_{ij} - \alpha - X_{ij}\boldsymbol{\gamma} - \beta_P - T_k \right) \right.$$
$$\left. + \sum_{i=1}^{m} \sum_{j=1}^{2} \tau_2 w_{ij}^{AB} \left(y_{ij} - \alpha - X_{ij}\boldsymbol{\gamma} - \mu_{u_i,2} \right) \right\} \tag{3.24}$$

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

$$\hat{\beta}_P = \frac{1}{n_{BB} + \sum_{i=1}^{m}\sum_{j=1}^{2} \tau_1 w_{ij}^{AB}} \Bigg\{ \sum_{k=1}^{2}\sum_{i=1}^{m}\sum_{j=1}^{2} w_{ij}^{BB}\left(y_{ij} - \alpha - X_{ij}\boldsymbol{\gamma} - \beta_M - T_k\right) + \sum_{i=1}^{m}\sum_{j=1}^{2} \tau_1 w_{ij}^{AB}\left(y_{ij} - \alpha - X_{ij}\boldsymbol{\gamma} - \mu_{u_i,1}\right)\Bigg\} \tag{3.25}$$

The covariate coefficients $\boldsymbol{\gamma}$ are

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{m}\sum_{j=1}^{2} X_{ij}X_{ij}'\right)^{-1} \sum_{k=1}^{2}\sum_{i=1}^{m}\sum_{j=1}^{2} \tau_k X_{ij}(\ddot{y}_{kij}) \tag{3.26}$$

where

$$\ddot{y}_{kij} = \begin{cases} y_{ij} - \alpha - (\beta_M + \beta_P)w_{ij}^{BB} - \beta_P w_{ij}^{AB} - \mu_{u_i,1} & k = 1 \\ y_{ij} - \alpha - (\beta_M + \beta_P)w_{ij}^{BB} - \beta_M w_{ij}^{AB} - \mu_{u_i,2} & k = 2 \end{cases}. \tag{3.27}$$

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

$$\tau^2 = \frac{1}{N}\sum_{k=1}^{2}\sum_{i=1}^{m}\sum_{j=1}^{2} \tau_k\left(\mu_{u_i,k} + \Sigma_{u_i,k}\right) \tag{3.28}$$

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

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{2}\sum_{i=1}^{m}\sum_{j=1}^{2} \tau_k e_{ijk}^2 \tag{3.29}$$

where

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

Computational Issues

A well known problem of the EM solutions is that can be highly dependent on its starting values, θ^0, 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 each run from a different random starting points (McLachlan and Peel, 2004) and choosing the one with maximum likelihood value. Short run means that the algorithm is stopped after a limited number of iteration without waiting for the convergence (Biernacki et al., 2003). However, using random initial values can often not solve the problem of finding bad local optimum. 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. 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, (3.10). In order to take into account for this problem, we impose an identifiability constraint on the parameters β_M and β_P: $\beta_M > \beta_P$. Due to that, we assume the first component represents always the group of individuals where the allele "B" received from the parent, denoted by "M", has the largest effect size whereas the second component, denoted by "P", represents the group of individuals that receive the allele "B" from the other parent.

3.4 The Statistical Test for POEs

One of the major advantage of the proposed method is the possibility to obtain the estimated values of the parameters β_M and β_P, as shown in equations (3.24) and (3.25), respectively. The parent-of-origin phenomena is observed when the effect of the allele "B" inherited from mother is different from the effect of the allele "B" given from the father. This occurs because only the allele inherited from one parents is expressed, the other one is silenced (completely or partially). 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:

$$H_0 : \beta_M - \beta_P = 0 \tag{3.30}$$

Since the EM estimators are maximum likelihood estimators and considering the constraint imposed to avoid the identifiability problem, we have that $\beta_M - \beta_P$ is always greater than 0 and is distributed according an Half Normal distribution with scale parameter equal to the variance of the difference between the parameters, $Var(\beta_M - \beta_P)$.

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

$$\left(\frac{\hat{\beta}_M - \hat{\beta}_P}{\sqrt{Var(\hat{\beta}_M - \hat{\beta}_P)}} \right)^2 \Bigg| H_0 \sim \chi^2_1. \tag{3.31}$$

where $\hat{\beta_M}$ and $\hat{\beta_P}$ are the EM-estimators. For computing 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 provide automatically the covariance matrix of the MLE. The asymptotic covariance matrix of $\hat{\theta}$, the vector of ML estimators, can be obtained in several ways. First of all we propose to compute the robust sandwich covariance matrix estimator that is obtained using Fisher information approximated by the empirical information matrix, $\mathcal{I}_{\mathcal{G}}(\hat{\theta}) = \sum_{j=1}^{n}[\mathbf{q}_j(\hat{\theta})\mathbf{q}_j(\hat{\theta})']$, and by the observed information matrix, $\mathcal{I}_{\mathcal{H}}(\hat{\theta}) = -\sum_{j=1}^{n}[\mathbf{Q}_j(\hat{\theta})]$, where q and Q denote the gradient and Hessian of the likelihood function in equation (3.9) in the maximum point, $\hat{\theta}$. Thus the asymptotic sandwich variance of $\hat{\theta}$ is given by:

$$\hat{\mathcal{I}}_{\mathcal{H}}(\theta)^{-1}\hat{\mathcal{I}}_{\mathcal{G}}(\theta)\hat{\mathcal{I}}_{\mathcal{H}}(\theta)^{-1}. \tag{3.32}$$

The standard error estimated by the observed information matrix can be negative due numerical problem. In this case the standard errors are fixed equal to 0.001. Moreover the standard errors are estimated through a parametric and a non parametric bootstrap procedure.

3.5 Simulation Study

Here, simulation studies are performed under different scenarios to investigate the type I error and the power of the test.

We simulate $m = 500$ families, each family composed by a twin pair. We assume that around the 30% of the twin pairs is identical (MZ) and that the remaining part is constituted by fraternal (DZ) twins.

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 level of significance of the test ($\alpha = 0.1$, 0.05 and 0.01). We have simulated a quantitative trait under the null hypothesis 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 the 30% of the total trait variation (see Appendix A.2), of a polygenic additive variance, $\tau^2 = 0.4$, and a of a normally distributed environmental effect with a variance of $\sigma^2 = 0.3$. The Figure 3.1 shows the behaviour of the type I error when the simulations number increases. For all levels of significance the empirical type I error approaches to the corresponding nominal value. In the second column of the Figure 3.1 it is represented 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 first-type error.

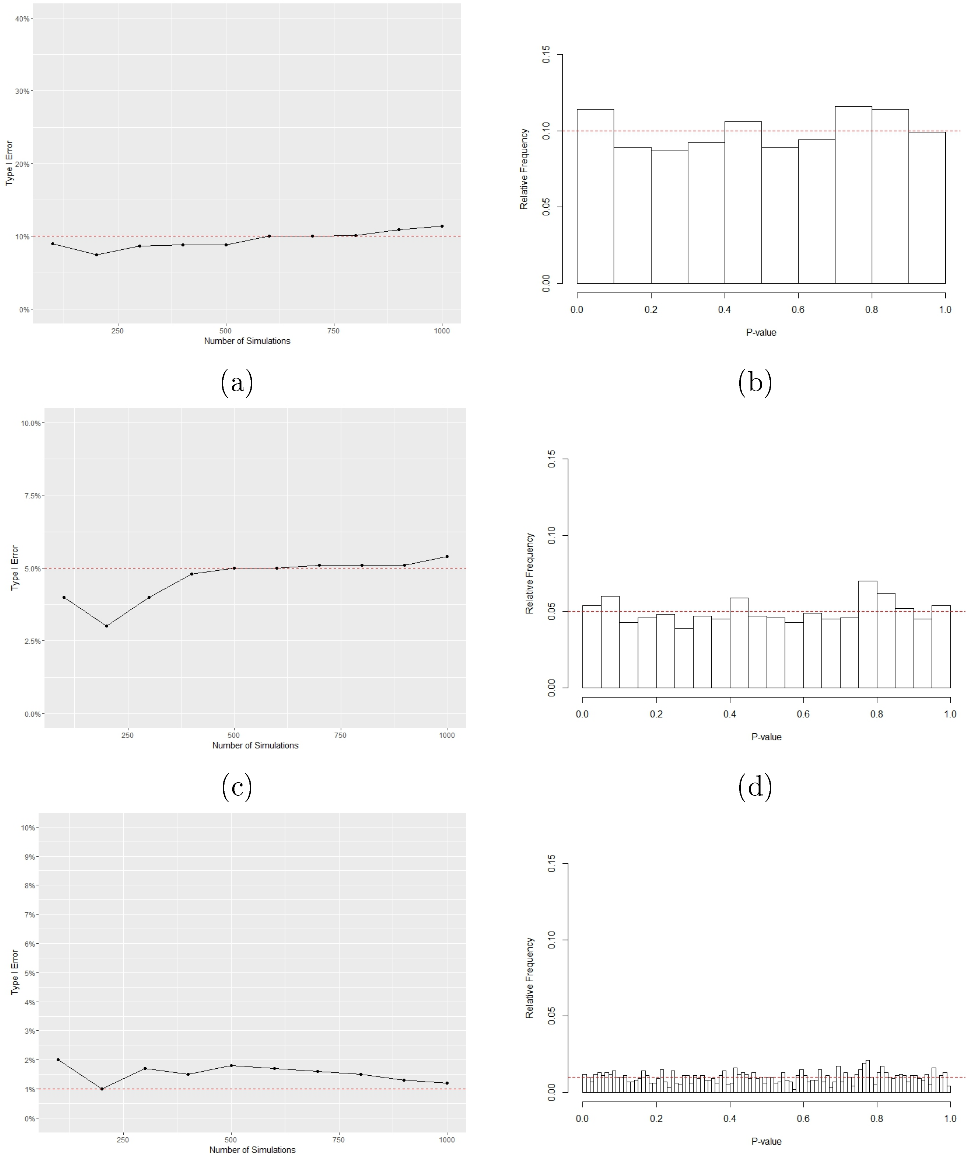
In the second simulation study we evaluate the performance of the proposed test in terms of power and type I error under several scenarios. In particular, the structure of the simulation study proposed here is the same used in the Section 2.4; the only difference is that, in this case, the number of replicates is equal to one hundred. All the results are represented in the Figures 3.2 and 3.3. Let us note that the power of the POE test is nearly 1 in all scenarios in correspondence of a proportion of the total trait variation explained by the simulated marker (σ_{β^2}) equal at least to 10%. When $\sigma_{\beta^2} = 0.01$ the power ranges from 0.5 to 0.20. In the case of 100% maternal 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 detect the POEs in correspondence of small MAF.

Regarding the first-type error, with only one hundred of replicate, we can say that the test statistic is not inflated. The type I error, under different scenarios, results in good agreement with the nominal significance 5% level.

3.6 Application to MuTHER Study Data

3.6.1 Validation of the gene CHURCH1

The discovery analysis proposed in the Section 2.5 has leaded to the identification of several regions in imprinting. In particular, we identified the CHURCH1 gene in the region 14q23.3 in correspondence of 20 loci in the three tissues. As explained in the Section 2.5.1, the Log2 transformation applied on the gene expression data to normalize the data can implicate a



(a) (b)

(c) (d)

(e) (f)

Figure 3.1: 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) $\alpha = 0.01$

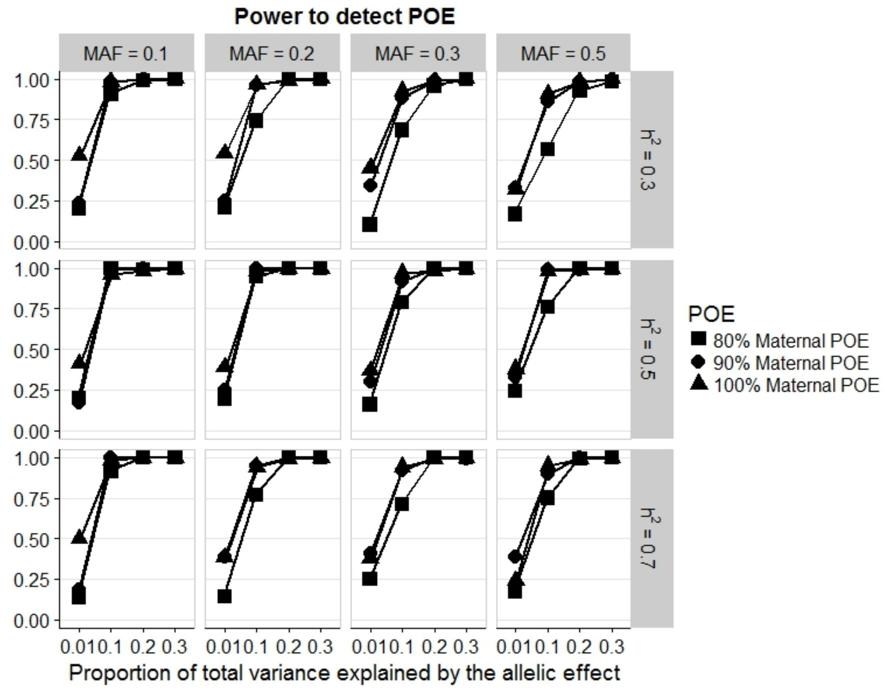


Figure 3.2: Power to detect POEs by the the test on the difference between β_M and β_P. The columns show power for different MAF; the power achieved under three heritability model is shown in rows. The y axis of each panel shows power, whereas the x axis shows the proportion of total variance explained by the allelic effect. POE size was set to 80%, 90% and 100% of the allelic effect size.

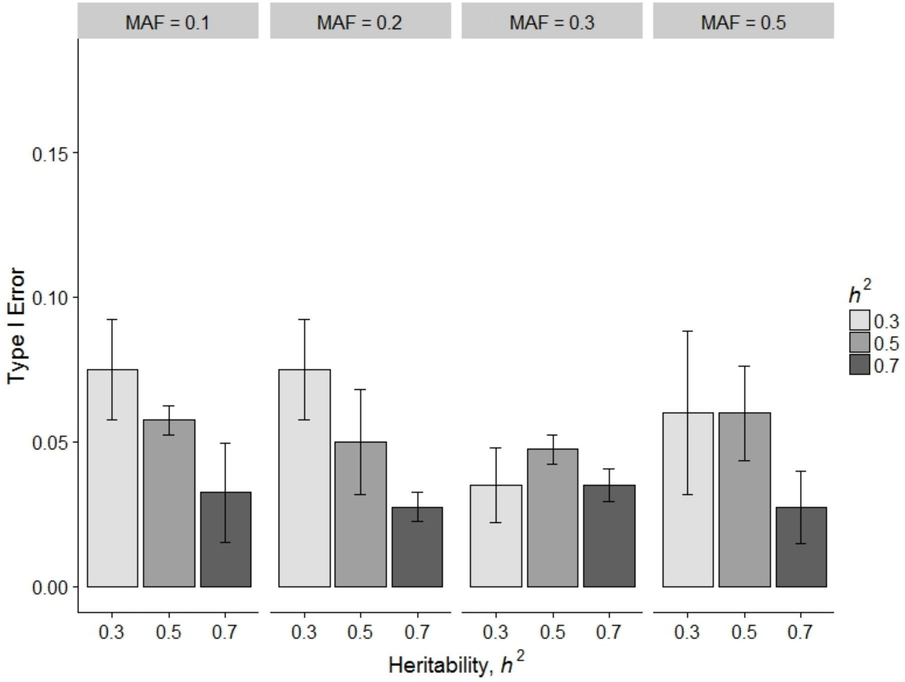


Figure 3.3: Type I Error for the proposed test of POEs. The columns show type I error under four different level of MAF. The y axis of each panel shows type I error, whereas different colours indicates the level of the heritability (0.3, 0.5, 0.7). Keeping constant the MAF and the heritability, the type I errors corresponding to the different levels of explained variance is synthesized in error bars at 95% confidence interval.

spurious association using the variance test in presence of a strong marginal association. For this reason, we validate the founded results fitting the mixture model approach.

The tables 3.1, 3.2 and 3.3 shows the list of the confirmed SNPs in correspondence of each tissues. In each table the estimated effects β_M and β_P, the corresponding standard errors computed by using a parametric bootstrap procedure and the by the robust sandwich covariance matrix estimator are reported. For each kind of standard error, the statistical test and the corresponding p-value are shown.

3.6.2 Study on the imprinting genes

We have analysed the gene expression microarray data of the MuTHER study in the lymphoblastoid cell lines, LCL. The data are described in the details in the Appendix B.1. We focused on the known imprinted genes using the Imprinted Gene Database (http://www.geneimprint.com/site/genes-by-species) gathered from the NCBI (National Center for Biotechnology Information). We selected 110 transcripts on 68 imprinted genes. Furthmore, we considered only the transcript associated at FDR level of 1% with the all the SNPs located within a 1Mb window of the gene transcript. Thus, we focused only on 12 genes for a total of 212 tests; the number of test performed for each transcripts is reported in the fourth column of the table 3.4. In correspondence of each transcript and each SNP the mixture model has been fitted and for the detection of the POEs the statistical test (denoted by *MLMM*) proposed in the Section 3.4 has been computed. Several approaches

Table 3.1: Validated SNPs in POE in the region of the gene CHURCH1 for the tissue FAT

SNP	Effect		Sandwich estimator				Parametric Bootstrap estimator			
	β_m	β_p	s.e. β_m	s.e. β_p	Test Statistic	P-Value	s.e. β_m	s.e. β_p	Test Statistic	P-Value
rs10139595	1,450	0,567	0,033	0,089	74,228	6,96E-18	0,039	0,032	158,040	3,03E-36
rs10151701	1,451	0,568	0,033	0,089	74,574	5,84E-18	0,032	0,069	76,184	2,58E-18
rs11158568	0,983	0,889	0,061	0,066	38,812	4,67E-10	0,079	0,118	5,819	0,0158553
rs11848113	1,045	0,874	0,057	0,067	41,815	1,00E-10	0,116	0,161	14,708	0,0001255
rs12433639	1,045	0,874	0,057	0,067	41,857	9,82E-11	0,116	0,161	14,708	0,0001255
rs17102298	1,450	0,567	0,033	0,089	74,221	6,99E-18	0,039	0,032	158,040	3,03E-36
rs1957421	1,045	0,874	0,057	0,067	41,779	1,02E-10	0,116	0,161	14,708	0,0001255
rs4902332	1,450	0,567	0,033	0,089	74,276	6,79E-18	0,039	0,032	158,040	3,03E-36
rs4902336	1,453	0,586	0,033	0,095	64,892	7,91E-16	0,021	0,001	1484,645	0,00E+00
rs7143432	1,453	0,586	0,033	0,095	64,896	7,89E-16	0,021	0,001	1484,645	0,00E+00

Table 3.2: Validated SNPs in POE in the region of the gene CHURCH1 for the tissue LCL

SNP	Effect		Sandwich estimator				Parametric Bootstrap estimator			
	β_m	β_p	s.e. β_m	s.e. β_p	Test Statistic	P-Value	s.e. β_m	s.e. β_p	Test Statistic	P-Value
rs10498518	1,230	1,210	0,063	0,065	25,085	5,49E-07	0,017	0,011	5,586	1,81E-02
rs11848113	1,236	1,212	0,061	0,063	26,182	3,11E-07	0,018	0,012	4,722	2,98E-02
rs12433639	1,236	1,212	0,062	0,064	26,430	2,73E-07	0,018	0,012	4,722	2,98E-02
rs12590986	1,230	1,210	0,063	0,065	25,272	4,98E-07	0,017	0,011	5,586	1,81E-02
rs1951487	1,234	1,216	0,062	0,064	24,823	6,28E-07	0,017	0,011	6,318	1,19E-02
rs1951488	1,234	1,216	0,062	0,064	25,184	5,21E-07	0,017	0,011	6,318	1,19E-02
rs1957421	1,236	1,212	0,062	0,064	26,404	2,77E-07	0,018	0,012	4,722	2,98E-02
rs743264	1,230	1,210	0,063	0,065	25,231	5,09E-07	0,017	0,011	5,586	1,81E-02
rs11158568	1,212	1,170	0,063	0,067	25,512	4,40E-07	0,025	0,016	4,767	2,90E-02
rs1957425	1,226	1,162	0,068	0,076	19,369	1,08E-05	0,021	0,015	14,834	1,17E-04
rs2296327	0,814	0,717	0,115	0,123	21,637	3,30E-06	0,021	0,014	26,522	2,60E-07
rs2649196	0,616	0,567	0,150	0,154	25,003	5,72E-07	0,026	0,019	23,012	1,61E-06
rs3759681	0,954	0,850	0,098	0,109	15,613	7,77E-05	0,161	0,136	22,091	2,60E-06

Table 3.3: Validated SNPs in POE in the region of the gene CHURCH1 for the tissue SKIN

SNP	Effect		Sandwich estimator				Parametric Bootstrap estimator			
	β_m	β_p	s.e. β_m	s.e. β_p	Test Statistic	P-Value	s.e. β_m	s.e. β_p	Test Statistic	P-Value
rs10139595	1,347	0,565	0,093	0,017	60,651	6,81E-15	0,033	0,051	131,374	2,05E-30
rs10151701	1,347	0,565	0,093	0,017	61,127	5,35E-15	0,032	0,051	132,222	1,34E-30
rs11158568	0,971	0,802	0,080	0,071	29,134	6,76E-08	0,038	0,059	4,433	0,035242
rs17102298	1,347	0,565	0,093	0,017	60,625	6,91E-15	0,033	0,051	131,374	2,05E-30
rs2649196	0,492	0,469	0,142	0,140	20,011	7,70E-06	0,027	0,019	5,147	0,023283
rs4902332	1,347	0,565	0,093	0,017	60,666	6,76E-15	0,033	0,051	131,374	2,05E-30
rs4902336	1,347	0,550	0,089	0,020	69,844	6,42E-17	0,034	0,061	96,243	1,02E-22
rs7143432	1,347	0,550	0,089	0,020	69,874	6,32E-17	0,034	0,061	96,243	1,02E-22
rs7153990	1,347	0,565	0,093	0,017	60,629	6,89E-15	0,033	0,051	131,374	2,05E-30
rs7154406	1,347	0,565	0,093	0,017	60,663	6,77E-15	0,033	0,051	131,374	2,05E-30

Table 3.4: Number of SNPs in POEs

Probe ID	Gene	Chr	Nr. Tested SNPs	Nr. Identified SNPs	
				VarT	*MLMM*
ILMN_1740711	KIAA1571	2	10		
ILMN_1815121	PLAGL1	6	4		
ILMN_1669479	MEST	7	6		
ILMN_1674620	SGCE	7	31	2	4
ILMN_1784294	CPA4	7	21	2	4
ILMN_1679301	ZFAT	8	111		2
ILMN_1664099	GLIS3	9	2		
ILMN_1693427	GLIS3	9	2	1	
ILMN_1688569	KCNQ1	11	1		
ILMN_2263086	NTM	11	2		
ILMN_2382505	SLC22A18	11	1		
ILMN_1786429	P2RY5	13	3		1
ILMN_1664878	NLRP2	19	18		18

have been proposed in the statistical literature for the POE identification but they are not suitable in case of related data without information about the alleles parental origin. Thus, for comparative purposes we have performed the variance test presented in Hoggart et al. (2014), although the proposed test works under the assumption of independence of observations; in order to apply the modified version of the Brown-Forsythe test proposed by Hoggart (denoted by *VarT*), we sampled for each twin pair a single individual. The Table 3.4 shows the number of SNPs identified with the two methods.

Chapter 4

Conclusion

We proposed two novel frameworks for the detection of the parent-of-origin effects in related samples when information on the parental generation is missing, in particular for the special case of twin data. 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. The two proposes methods address this issue. They stemmed from the method proposed by Hoggart et al. (2014) that identifies POE effects in the absence of parental information, but it is only applicable to unrelated samples. The first method that we propose extends the idea of the variance test to detect POEs proposed by Hoggart et al. (2014) to the case of twins data related samples and it is based on the use of the residuals of a linear mixed model, resulting in a multistep procedure. On the contrary, the second method is a unified approach developed specifically for the problem of the POEs detection . In our second approach we use a mixture of linear mixed models to estimate the effect of the causal allele inherited from the mother and the effect of the casual allele inherited from the father. The estimation of these coefficients through the EM algorithm allows us to test directly the difference between these parameters and to identify the POEs. The estimate of these coefficients is not computed in the first procedure and so far the only way to obtain the estimate of parental effects was by exploiting family data. On the other side the "Variance Approach" results clearly to be computationally faster than "Mixture Approach" (see Table 4.1), indeed, it is well known that one of the criticisms of the EM algorithm is that its convergence may be quite slow.

Table 4.1: Computational times comparison

Method	user	system	elapsed
"Variance Approach"	0.63	0.00	8.50
"Mixture Approach"	30.08	0.07	44.94

Furthermore, the inflated variance can be caused by several phenomena. Indeed "Variance Approach" major weakness is that it can lead to spurious association produced by the combination of the transformation of the scale on which the trait is measured and a strong association with the marker. For this reason a validation of the results is always required. Instead, the "Mixture Approach" overcomes this problem, because the proposed test for identifying the POEs is not affected by trait transformation. Moreover, by using mixture model, if the trait of interest is not normally distributed, it is always possible to relax the normality assumption working with a mixture of different density functions.

By comparing the simulation study in the Section 2.4, Figures 2.1 and 2.2, and the one in the Section 3.5, Figures 3.2 and 3.3, we observe that the power of the “Mixture Approach” is always higher than the power of the “Variance Approach”. Moreover the statistical test proposed in the Section 3.4 results to be consistent in terms of parameters estimation and the type I error is controlled at the selected significance levels.

From a genetic point of view, the application proposed in Section 2.5 has leaded to interesting results because the regions identified from the “Variance Approach” are deletion-type regions. The CHURCH1 gene has been validated through the “Mixture Approach”. A study on the known imprinted genes of the MuTHER data has been conducted to compare the performance of the “Mixture Approach” and the test proposed by Hoggart.

Appendices

Appendix A

A.1 Description of MuTHER Data

A.1.1 Sample Collection

The MuTHER (Multiple Tissue Human Expression Resource, Grundberg et al. (2012); Nica et al. (2011)) includes a total of 856 female twins (154 monozygotic twin pair and 84 singletons) of European descent aged between 40 and 87 years old (mean age 62) recruited from the TwinsUK Adult twin registry (Spector and Williams, 2006). Skin punch biopsies were taken from a relatively photoprotected area adjacent and inferior to the umbilicus. The fat tissue was then carefully dissected from the same skin biopsy. Peripheral blood samples were collected to generate lymphoblastoid cell lines (LCL).

A.1.2 Gene expression measurements and genotyping

Expression profiling of the samples, of skin and adipose tissues and LCLs, were performed using Illumina Human HT-12 V3 BeadChips (IlluminaInc) including more that 48,000 probes. All samples were randomized prior to array hybridization and the technical replicates were always hybridized on different BeadChips.

Genotyping of TwinsUK dataset was done with a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1M-Duo, and 1.2MDuo 1M). For samples, the following exclusion criteria were utilized: (i) sample call rate 98%; (ii) heterozygosity across all SNPs ≤ 2 standard deviations from the sample mean; (iii) evidence of non-European ancestry are assessed by principal component analysis; (iv) observed pairwise identity by descent probabilities suggestive of sample identity errors. Instead, for the SNPs, the exclusion criteria are the followed: (i) Hardy-Weinberg $P < 10^{-6}$, assessed in a set of unrelated samples; (ii) MAF $< 1\%$, assesed in a set of unrelated samples; or (iii) SNP call rate $< 97\%$.

A.1.3 Post experimental normalization of gene expression data

Log2-transformed expression signals were normalized separately per tissue as follows: quantile normalization was performed across the replicates of each individual followed by quantile normalization across all individuals as previously described (Nica et al., 2011).

A.2 Main allelic addittive effect

In our simulation study we suppose an additive effects of the locus on the trait of interest. Under the null hypothesis, the main allelic addittive effects, β, is derived from the explained variance of the effect, denoted by σ^2_β, according to the biometrical model shown in the table (A.1).

Table A.1: Biometrical addittive model for a single biallelic SNP

Genotype	AA	AB	BB
Effect	0	β	2β
Frequency	q^2	$2pq$	p^2

The genotypic effect of the homozygotes AA and BB are 0 and 2β, respectively. The genotypic effect of the heterozygote AB is β. The gene frequency of the allele A and B are denoted by q and p, respectvely, where $p+q=1$. Thus we have that, under the null hypothesis, the variance explained of the main allelic effect is given by

$$\sigma^2_\beta = \mathbb{E}\left(\text{Effect}^2\right) - \mathbb{E}\left(\text{Effect}\right)^2 \tag{A.1}$$

where

$$\mathbb{E}\left(\text{Effect}\right) = 2\beta p \tag{A.2}$$

and

$$\mathbb{E}\left(\text{Effect}^2\right) = 2\beta^2 p\left(1+p\right) \tag{A.3}$$

In this way, we can write

$$\begin{aligned}\sigma_\beta^2 &= 2\beta^2 p\left(1+p\right) - \left[2\beta p\right]^2 \\ &= 2\beta^2 pq\end{aligned}$$

From the equation A.4 we can obtain easily the value of the parameter β. It is clear that, with the same level of explained variance, the allelic effect β increases as the minor frequecy allele decreases.

In presence of POEs, the biometrical model of the table A.1 is not valid. Indeed the effect of the herozygous group AB is different from the effect of heterozygous BA. Thus, under the alternative hypothesis, we assume the biometrical model in the table (A.2).

Table A.2: Biometrical addittive model for a single biallelic SNP in presence of POE

Genotype	AA	AB	BA	BB
Effect	0	β_M	β_P	2β
Frequency	q^2	pq	pq	p^2

Similarly, we obtain that

$$\mathbb{E}\left(\text{Effect}\right) = 2\beta p \tag{A.4}$$

and

$$\begin{aligned}\mathbb{E}\left(\text{Effect}^2\right) &= \beta_M^2 pq + \beta_P^2 pq + (2\beta)^2 p^2 \\ &= (\beta_M^2 + \beta_P^2)(pq + p^2)\end{aligned} \tag{A.5}$$

Under the null and alternative hypothesis, the expectation of the effect is the same, whereas the the expectation of the squared effect is different. Thus, under the alternative hypothesis, the variance explained from the main allelic effect is given by:

$$\sigma_\beta^2 = \left(\beta_M^2 + \beta_P^2\right) pq \tag{A.6}$$

If we assume that $\beta_M = k\beta$ and $\beta_P = (1-k)\beta$, where k denote the proportion of the main allelic effect explained from the allele B inherited from the first parent M, we can generalize the equation (A.4) and (A.6) in the following way:

$$\sigma_\beta^2 = 4\beta^2 \left(2k^2 - 2k + 1\right) pq \tag{A.7}$$

It is clear that in case of $k = 0.5$, we have that $\beta_M = \beta_P$ and the equation (A.6) is equal to the variance explained in (A.4). In presence of POE $k \neq 0.5$, indeed for $k = 1$ we have a paternal POE, the parternally derived allele is silenced, and for $k = 0$ the contrary.

Appendix B

B.1 Random effects as missing data

For this variant of the E-step, we need to compute the mean, $\mu^k_{u_i}$, and the variance, $\Sigma^k_{u_i}$, of the random effects u_i conditional on the current parameter estimates and the observed data. In case of monozygotic twins, $\mathbb{1}_{MZ}(i) = 1$, we have to condition on the vector 2×1 dimensional of the observed data $\boldsymbol{y}_i$; whereas if $\mathbb{1}_{MZ}(i) = 0$ we have to condition to the observed data y_{ij}.
In the next sections we will show the derivation of the conditional moments for the identical and fraternal twins.

B.1.1 Case MZ twin pair

We have that

$$\begin{bmatrix} \boldsymbol{y}_i \\ u_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{ik} \\ 0 \end{bmatrix}, \begin{bmatrix} V & \mathbb{1}\tau^2 \\ \mathbb{1}'\tau^2 & \tau^2 \end{bmatrix} \right) \tag{B.1}$$

where $\mathbb{1}$ is a 2×1 unit vector and V is the variance matrix (2×2) of $\boldsymbol{y}_i$,

$$V = \begin{bmatrix} \tau^2 + \sigma^2 & \tau^2 \\ \tau^2 & \tau^2 + \sigma^2 \end{bmatrix}. \tag{B.2}$$

From the well known results of the multivariate theory, we obtain that the conditional variance of u_i is given by

$$\begin{aligned} \mathbb{V}_{\theta_r}(u_i|\boldsymbol{y}_i) &= \tau^2 - \tau^2 \mathbb{1}' V^{-1} \mathbb{1} \tau^2 \\ &= \tau^2 - (\sigma^2\tau^{-2} + 2)2\tau^2 \\ &= \sigma^2(\sigma^2\tau^{-2} + 2)^{-1} \\ &= \left(\frac{1}{\tau^2} + \frac{2}{\sigma^2}\right)^{-1} \\ &= \Sigma_{u_i}^{MZ}. \end{aligned} \tag{B.3}$$

The conditional mean of u_i is defined by

$$\begin{aligned} \mathbb{E}_{\theta_r}(u_i|\boldsymbol{y}_i) &= \mathbb{1}' V^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_{ik})^2 \\ &= \left(\frac{1}{\tau^2} + \frac{2}{\sigma^2}\right)^{-1} \frac{1}{\sigma^2} \mathbb{1}' (\boldsymbol{y}_i - \boldsymbol{\mu}_{ik})^2 \\ &= \Sigma_{u_i} \frac{1}{\sigma^2} \mathbb{1}' (\boldsymbol{y}_i - \boldsymbol{\mu}_{ik})^2 \\ &= \mu_{u_i,k}^{MZ} \end{aligned} \tag{B.4}$$

where $\boldsymbol{y}_i$ is the observed data vector 2×1 dimensional of the i-th twin pair and $\mu_{u_i}^k$ is the mean vector 2×1 dimensional of the i-th twin pair of the k-th component.

B.1.2 Case DZ twin pair

In case of DZ twin pair, we have

$$\begin{bmatrix} y_{ij} \\ u_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{ijk} \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2+\sigma^2 & \tau^2 \\ \tau^2 & \tau^2 \end{bmatrix} \right) \tag{B.5}$$

Thus, the conditional mean of u_i is

$$\begin{aligned} \mathbb{E}_{\theta_r}(u_i|y_{ij}) &= \tau^2(\tau^2+\sigma^2)^{-1}(y_{ij}-\mu_{ijk})^2 \\ &= (1+\sigma^2\tau^{-2})^{-1}(y_{ij}-\mu_{ijk}) \\ &= \mu^{DZ}_{u_i,kj} \end{aligned} \tag{B.6}$$

and the conditional variance is given by

$$\begin{aligned} \mathbb{V}_{\theta_r}(u_i|y_{ij}) &= \tau^2-\tau^2(\tau^2+\sigma^2)^{-1}\tau^2 \\ &= (\tau^2+\sigma^2)^{-1}\{\tau^2(\tau^2+\sigma^2)-\tau^4\} \\ &= (\tau^2+\sigma^2)^{-1}\sigma^2\tau^2 \\ &= \Sigma^{MZ}_{u_i} \end{aligned} \tag{B.7}$$

Bibliography

Yurii S Aulchenko, Dirk-Jan De Koning, and Chris Haley. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177(1):577–585, 2007.

Yael Baran, Meena Subramaniam, Anne Biton, Taru Tukiainen, Emily K Tsang, Manuel A Rivas, Matti Pirinen, Maria Gutierrez-Arcelus, Kevin S Smith, Kim R Kukurba, et al. The landscape of genomic imprinting across diverse adult human tissues. *Genome research*, 25(7):927–936, 2015.

Nadezhda M Belonogova, Tatiana I Axenovich, and Yurii S Aulchenko. A powerful genome-wide feasible approach to detect parent-of-origin effects in studies of quantitative traits. *European Journal of Human Genetics*, 18 (3):379–384, 2010.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575, 2003.

Trevor S Breusch and Adrian R Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pages 1287–1294, 1979.

Morton B Brown and Alan B Forsythe. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367, 1974.

William S Bush and Jason H Moore. Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 2012.

Gilles Celeux, Olivier Martin, and Christian Lavergne. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, 5(3):243–267, 2005.

David R Cox and E Joyce Snell. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 248–275, 1968.

Yuehua Cui, Qing Lu, James M Cheverud, Ramon C Littell, and Rongling Wu. Model for mapping imprinted quantitative trait loci in an inbred f 2 design. *Genomics*, 87(4):543–551, 2006.

Marie Davidian and David M Giltinan. *Nonlinear models for repeated measurement data*, volume 62. CRC press, 1995.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Wei Q Deng, Senay Asma, and Guillaume Paré. Meta-analysis of snps involved in variance heterogeneity using levene's test for equal variances. *European Journal of Human Genetics*, 22(3), 2014.

Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450, 2010.

Douglas S Falconer, Trudy FC Mackay, and Richard Frankham. Introduction to quantitative genetics (4th edn). *Trends in Genetics*, 12(7):280, 1996.

Ronald A Fisher. Xv. the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, 52 (02):399–433, 1919.

Simon KG Forsberg, Matthew E Andreatta, Xin-Yuan Huang, John Danku, David E Salt, and Örjan Carlborg. The multi-allelic genetic architecture of a variance-heterogeneity locus for molybdenum concentration in leaves acts as a source of unexplained additive genetic variance. *PLoS Genet*, 11 (11):e1005648, 2015.

Herbert Glejser. A new test for heteroskedasticity. *Journal of the American Statistical Association*, 64(325):316–323, 1969.

Stephen M Goldfeld and Richard E Quandt. Some tests for homoscedasticity. *Journal of the American statistical Association*, 60(310):539–547, 1965.

Anthony J Griswold, Deqiong Ma, Stephanie J Sacharow, Joycelyn L Robinson, James M Jaworski, Harry H Wright, Ruth K Abramson, Helle Lybæk, Nina Øyen, Michael L Cuccaro, et al. A de novo 1.5 mb microdeletion on chromosome 14q23. 2-23.3 in a patient with autism and spherocytosis. *Autism Research*, 4(3):221–227, 2011.

Elin Grundberg, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, et al. Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature genetics*, 44(10):1084–1089, 2012.

A Guilmatre and AJ Sharp. Parent of origin effects. *Clinical genetics*, 81(3): 201–209, 2012.

Guang Guo and Jianmin Wang. The mixed or multilevel model for behavior genetic analysis. *Behavior genetics*, 32(1):37–49, 2002.

Robert L Hanson, Sayuko Kobes, Robert S Lindsay, and William C Knowler. Assessment of parent-of-origin effects in linkage analysis of quantitative traits. *The American Journal of Human Genetics*, 68(4):951–962, 2001.

James Andrew Hilden-Minton. *Multilevel diagnostics for mixed and hierarchical linear models*. PhD thesis, University of California Los Angeles, 1995.

Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

Clive J Hoggart, Giulia Venturini, Massimo Mangino, Felicia Gomez, Giulia Ascari, Jing Hua Zhao, Alexander Teumer, Thomas W Winkler, Natalia Tšernikova, Jian'an Luan, et al. Novel approach identifies snps in slc2a10 and kcnk9 with evidence for parent-of-origin effect on body mass index. *PLoS Genet*, 10(7):e1004508, 2014.

Jie Hu, Malini Sathanoori, Sally Kochmar, Meron Azage, Susan Mann, Suneeta Madan-Khetarpal, Amy Goldstein, and Urvashi Surti. A novel maternally inherited 8q24. 3 and a rare paternally inherited 14q23. 3 cnvs in a family with neurodevelopmental disorders. *American Journal of Medical Genetics Part A*, 167(8):1921–1926, 2015.

Ivan Iachine, Hans Chr Petersen, and Kirsten O Kyvik. Robust tests for the equality of variances for clustered data. *Journal of Statistical Computation and Simulation*, 80(4):365–377, 2010.

Ricky S Joshi, Paras Garg, Noah Zaitlen, Tuuli Lappalainen, Corey T Watson, Nidha Azam, Daniel Ho, Xin Li, Stylianos E Antonarakis, Han G Brunner, et al. Dna methylation profiling of uniparental disomy subjects provides a map of parental epigenetic bias in the human genome. *The American Journal of Human Genetics*, 99(3):555–566, 2016.

Tim K Keyes and Martin S Levy. Analysis of levene's test under design imbalance. *Journal of Educational and Behavioral Statistics*, 22(2):227–236, 1997.

Roger Koenker. A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, 17(1):107–112, 1981.

Heather A Lawson, James M Cheverud, and Jason B Wolf. Genomic imprinting and parent-of-origin effects on complex traits. *Nature Reviews Genetics*, 14(9):609–617, 2013.

Howard Levene. Robust tests for equality of variances1. *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, 2:278–292, 1960.

Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Geoffrey J McLachlan and Kaye E Basford. Mixture models. inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988*, 1, 1988.

Michael Neale and Hermine H. Maes. *Methodology for genetic studies of twins and families*. Richmond, VA: Virginia Commonwealth University, Department of Psychiatry, 2004.

Shu-Kay Ng, Geoffrey J McLachlan, Kui Wang, L Ben-Tovim Jones, and S-W Ng. A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, 22(14):1745–1752, 2006.

Alexandra C Nica, Leopold Parts, Daniel Glass, James Nisbet, Amy Barrett, Magdalena Sekowska, Mary Travers, Simon Potter, Elin Grundberg, Kerrin Small, et al. The architecture of gene regulatory variation across multiple human tissues: the muther study. *PLoS Genet*, 7(2):e1002003, 2011.

Samuel D Oman. Checking the assumptions in mixed-model analysis of variance: a residual analysis approach. *Computational statistics & data analysis*, 20(3):309–330, 1995.

Isabel Parra-Frutos. The behaviour of the modified levene's test when data are not normally distributed. *Computational Statistics*, 24(4):671–693, 2009.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

Jo Peters. The role of genomic imprinting in biology and disease: an expanding view. *Nature Reviews Genetics*, 15(8):517–530, 2014.

Jose Pinheiro and Douglas Bates. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2006.

John Quackenbush. Microarray data normalization and transformation. *Nature genetics*, 32:496–501, 2002.

Sophia Rabe-Hesketh, Anders Skrondal, and Hakon K Gjessing. Biometrical modeling of twin and family data using standard mixed model software. *Biometrics*, 64(1):280–288, 2008.

Juvêncio Santos Nobre and Julio Da Motta Singer. Residual analysis for linear mixed models. *Biometrical Journal*, 49(6):863–875, 2007.

Franklin E Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114, 1946.

David A Skaar, Yue Li, Autumn J Bernal, Cathrine Hoyo, Susan K Murphy, and Randy L Jirtle. The human imprintome: regulatory mechanisms, methods of ascertainment, and roles in disease susceptibility. *ILAR journal*, 53(3-4):341–358, 2012.

Tim D Spector and Frances MK Williams. The uk adult twin registry (twinsuk). *Twin Research and Human Genetics*, 9(6):899–906, 2006.

Maksim V Struchalin, Abbas Dehghan, Jacqueline CM Witteman, Cornelia van Duijn, and Yurii S Aulchenko. Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. *BMC genetics*, 11(1):92, 2010.

Xiangqing Sun, Robert Elston, Nathan Morris, and Xiaofeng Zhu. What is the significance of difference in phenotypic variability across snp genotypes? *The American Journal of Human Genetics*, 93(2):390–397, 2013.

Geert Verbeke and Emmanuel Lesaffre. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221, 1996.

Peter M Visscher. Sizing up human height variation. *Nature genetics*, 40(5): 489–490, 2008.

Kui Wang, Shu Kay Ng, and Geoffrey J McLachlan. Clustering of time-course gene expression profiles using normal mixture models with autoregressive random effects. *BMC bioinformatics*, 13(1):1, 2012.

Clarice R Weinberg. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *The American Journal of Human Genetics*, 65(1):229–235, 1999.

CR Weinberg, AJ Wilcox, and RT Lie. 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(4):969–978, 1998.

Robert E Weiss and Carlos G Lazaro. Residual plots for repeated measures. *Statistics in Medicine*, 11(1):115–124, 1992.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980.