Hum Genet (2007) 122:159–174
DOI 10.1007/s00439-007-0387-2

ORIGINAL INVESTIGATION

Imprinting detection by extending a regression-based QTL analysis method

Olga Y. Gorlova · Lei Lei · Dakai Zhu · Shih-Feng Weng · Sanjay Shete · Yiqun Zhang · Wei-Dong Li · R. Arlen Price · Christopher I. Amos

Received: 16 January 2007 / Accepted: 25 May 2007 / Published online: 12 June 2007

Abstract We present an extension of a regression-based quantitative-trait linkage analysis method to incorporate parent-of-origin effects. We separately regressed total, paternal, and maternal IBD sharing on traits' squared sums and differences. We also developed a test for imprinting that indicates whether there is any difference between the paternal and maternal regression coefficients. Since this method treats the identity-by-descent information as the dependent variable that is conditioned on the trait, it can be readily applied to data from complex ascertainment processes. We performed a simulation study to examine the performance of the method. We found that when using empirical critical values, the method shows identical or higher power compared to existing methods for evaluation of parent-of-origin effect in linkage analysis of quantitative traits. Missing parental genotypes increase the type I error rate of the linkage test and decrease the power of the imprinting test. When the major gene has a low heritability, the power of the method decreases considerably, but the statistical tests still perform well. We also applied a permutation algorithm, which ensures the appropriate type I error rate for the test for imprinting. The method was applied to a data from a study of 6 body size related measures and 23 loci on chromosome 7 for 255 nuclear families. Multipoint identities-by-descent (IBD) were obtained using a modification of the SIMWALK 2 program. A parent-of-origin effect consistent with maternal imprinting was suggested at 99.67–111.26 Mb for body mass index, bioelectrical impedance analysis, waist circumference, and leptin concentration.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-007-0387-2) contains supplementary material, which is available to authorized users.

O. Y. Gorlova (✉) · L. Lei · D. Zhu · S.-F. Weng · S. Shete · Y. Zhang · C. I. Amos
Department of Epidemiology, MD Anderson Cancer Center,
University of Texas, Unit 1340, 1155 Pressler Street,
Houston, TX 77030, USA
e-mail: oygorlov@mdanderson.org

L. Lei
Amgen Inc., Thousand Oaks, USA

W.-D. Li · R. Arlen Price
Department of Psychiatry,
Center for Neurobiology and Behavior,
University of Pennsylvania,
Philadelphia, PA, USA

Introduction

The primary purpose of linkage analysis is to determine the chromosomal regions associated with diseases or traits. Model-free methods of linkage analysis often evaluate the relationship between the trait values and the proportion of alleles shared identical-by-descent (IBD) at a marker locus in a pedigree. For quantitative traits two major approaches have been developed, the Haseman–Elston (H–E) method (Haseman and Elston 1972) and the variance-components (VC) method (Amos 1994; Fulker and Cherny 1996), with many extensions to each. The advantages of a regression-based method like the H–E include robustness and computational feasibility, while the VC methods are often more powerful and more easily incorporate data from extended pedigrees.

Sham et al′s. (2002) proposed a regression-based procedure that has desirable properties from both of these major methods. The method achieves power comparable to variance components methods while maintaining the robustness to non-normality of the trait distribution that is a feature of the H–E method. Also, since this method treats the identity by descent information as the dependent variable that is conditioned on the trait, it can be readily applied to data from complex ascertainment processes. However, Sham et al's. (2002) method requires that the population parameters of the trait distribution be correctly specified in order to standardize the traits and apply it to ascertained samples.

Imprinting is an epigenetic alteration of genes in which primarily the maternally or paternally inherited copy is expressed (Tilghman 1999; Wilkins and Haig 2003). For example, 'maternal imprinting' means silencing of the maternal allele and expression of the paternal allele. This is also called a 'parent-of-origin' effect. Several hypotheses have been proposed to describe its origin, with the sex-conflict theory being the most popular one (Bartolomei and Tilghman 1997). Since the first observations of imprinting made in the 1970s, imprinted genes have been implicated in more than 20 human disorders, for example, the Prader–Willi (MIM 176270) and Angelman (MIM 105830) syndromes.

Due to the relatively high cost of molecular techniques like DNA methylation probing, genetic statistical methods that are powerful in detecting imprinting are desirable. Following detection of linkage, bioinformatics approaches may further reduce the number of candidate genes that require molecular screening (Greally 2002). However, since imprinting causes deviations from Mendelian law, conventional linkage analysis methods are not appropriate and require modification. Several methods have been developed to account for the parent-of-origin effect in linkage analysis of binary traits (Knapp and Strauch 2004; Strauch et al. 2000; Vincent et al. 2006; Wu et al. 2005). For quantitative traits, Hanson et al. (2001) and Shete and Amos (2002) partitioned the estimated IBD sharing proportion of marker alleles into parent-specific components, which were then used in H–E or variance components methods of linkage analysis. In evaluating evidence for imprinting, they calculated statistics for linkage between trait and marker loci derived from either or both parents and compared them to χ_1^2 or *t*-test. Gorlova et al. (2003) also used parent-specific IBD sharing proportions in H–E method to obtain parent-specific slope coefficients, while significance of the test for imprinting was evaluated using a permutation procedure.

Shete and Amos (2002) further showed that a joint test for linkage and imprinting is more powerful than a simple test for linkage as a first step in analyzing data in the presence of complete or substantial imprinting. In the absence of imprinting, test for linkage and imprinting can be less powerful than an usual linkage test. To ensure optimal power in applying linkage tests, the investigator could consider whether the trait under study has previously been found to show imprinting effects, if there is evidence from segregation analysis for imprinting, or if the trait under study is among those hypothesized to be influenced by imprinting, such as traits related to body size or embryonic development (Smith et al. 2006).

Notably, the methods evaluating parent-of-origin effects for quantitative traits are sensitive to ascertainment (Amos and de Andrade 2001; Iyengar et al. 1997) and may particularly suffer a loss of power or, in the case of variance components methods, may provide inaccurate estimates if a correction for nonrandom sampling is not provided. In contrast, the method proposed by Sham et al. is expected to provide a valid test even for samples selected for particular trait values.

In this paper, we apply Sham et al.'s linkage analysis method to evaluate not only total but also parent-specific linkage signals, using simulated (to evaluate its performance) as well as real data sets. In addition, we propose a test for imprinting that asymptotically has a standard normal distribution under the null. When applying the method to data on 6 obesity-related traits in 255 nuclear families, we detected a paternal effect in the proximal part of the studied segment of Chromosome 7.

Methods and data

General methodology

The regression equation proposed by Sham et al. (2002) is

$$\hat{\mathbf{\Pi}}_{\mathbf{C}} = \mathbf{\Sigma}'_{\mathbf{Y}\hat{\mathbf{\Pi}}}\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{Y}_{\mathbf{C}} + \mathbf{e},$$

where $\hat{\mathbf{\Pi}}_{\mathbf{C}}$ is the mean-centered vector of pairwise IBD sharing proportions calculated as $\hat{\mathbf{\Pi}}_{\mathbf{C}} = \hat{\mathbf{\Pi}} - E(\hat{\mathbf{\Pi}})$. $\mathbf{Y}_{\mathbf{C}}$ is the mean-centered vector of stacked pairwise squared sums and squared differences of standardized traits ($S_{ij} = (X_i + X_j)^2, D_{ij} = (X_i - X_j)^2$ for $i \neq j$). The vector of squared sums, **S**, and the vector of squared differences, **D**, are collinear, since each element of **S** and **D** is a linear combination of two squares and a cross-product, and there are *n* squares and *n*(*n*–1)/2 cross-products (overall *n*(*n* + 1)/2 elements), whereas there are *m* = *n*(*n*–1)/2 elements in each of the vectors **S** and **D** (corresponding to the number of pairs among *n* individuals) (Sham et al. 2002). To remove this collinearity between the vectors **S** and **D**, we trim the latter by removing the last *n*(*n*–3)/2 elements from it, retaining exactly *n* elements. This ensures that collinearity is removed, while each individual is represented at least once. The trimmed vector **D** is denoted **d**. Thus, vector **Y**, defined as $\mathbf{Y} = [\mathbf{S},\mathbf{d}]'$, has *m* + *n* elements, instead of 2*m*, due to the trimming of **D**; $\mathbf{Y}_{\mathbf{C}} = \mathbf{Y} - E(\mathbf{Y})$). $\mathbf{\Sigma}_{\mathbf{Y}}$ is the variance–covariance matrix of the vector $\mathbf{Y}$, and $\boldsymbol{\Sigma}_{\mathbf{Y}\hat{\boldsymbol{\Pi}}}$ is the covariance matrix of stacked $\boldsymbol{\Sigma}_{\mathbf{S}\hat{\boldsymbol{\Pi}}}$ and $\boldsymbol{\Sigma}_{\mathbf{d}\hat{\boldsymbol{\Pi}}}$. The dimensions of the arguments in the equation are

$$\underset{m\times 1}{\hat{\boldsymbol{\Pi}}_{\mathbf{C}}} = \underset{m\times(m+n)}{\boldsymbol{\Sigma}'_{\mathbf{Y}\hat{\boldsymbol{\Pi}}}}\ \underset{(m+n)\times(m+n)}{\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}}\ \underset{(m+n)\times 1}{\mathbf{Y}_{\mathbf{C}}} + \underset{m\times 1}{\mathbf{e}}.$$

The statistic used in the final linkage testing is denoted as T,

$$T = \hat{Q}\sum_{i=1}^{k}\left[\mathrm{B}'\hat{\Pi}_C\right]_i = \hat{Q}^2\sum_{i=1}^{k}\left[\mathrm{B}'\Sigma_{\hat{\Pi}}\mathrm{B}\right]_i,$$

where k is the number of pedigrees and $\hat{Q}$ is the phenotypic variance explained by the additive effects of the QTL, a scalar weighted across all pedigrees and calculated as

$$\hat{Q} = \frac{\sum_{i=1}^{k}\left[\mathrm{B}'\hat{\Pi}_C\right]_i}{\sum_{i=1}^{k}\left[\mathrm{B}'\Sigma_{\hat{\Pi}}\mathrm{B}\right]_i}.$$

Here $\mathbf{B} = \boldsymbol{H}\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{Y}_{\mathbf{C}},$ and $\mathbf{H}$ is a matrix composed of two blocks stacked horizontally, the first block being an m by m square matrix with diagonal elements 2 and off-diagonal elements 0, the second block being the first n columns of a similar square matrix with diagonal elements –2. $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Pi}}}$ is the variance–covariance matrix of the IBD sharing proportion vector $\hat{\boldsymbol{\Pi}}$.

In applying Sham et al. (2002) method for detection of parent-of-origin effect, we calculated three T statistics: T (overall linkage), T_{p} (paternal), and T_{m} (maternal), each from its individual regression model, using overall $\hat{\boldsymbol{\Pi}}$ and parent-specific $\hat{\boldsymbol{\Pi}}$s (paternal $\hat{\boldsymbol{\Pi}}_{\mathrm{p}}$ and maternal $\hat{\boldsymbol{\Pi}}_{\mathrm{m}}$), estimated by the IBD_FM program (Shete and Amos 2002). Under the null hypothesis, the test statistic T proposed by the authors asymptotically follows a 50:50 mixture of 0 and χ_1^2 because a negative $\hat{Q}$ is not biologically plausible as it corresponds to less sharing of alleles than expected by chance. Theoretically, therefore, the 95% quantile of T is $\chi_1^2(1-2\alpha) = \chi_1^2(0.9) \approx 2.71$. If there is no linkage and no imprinting, all three statistics should be similar in value and less than 2.71 with probability 0.95. If there is linkage but no imprinting, the overall T should be larger than both T_{p} and T_{m}, and exceed 2.71 (at the $\alpha = 0.05$ nominal level). In the case of linkage without imprinting, T_{p} and T_{m} will have similar values but are not expected to follow the 50:50 mixture of 0 and χ_1^2. If there is linkage and also full imprinting, which in our simulation is maternal (only the paternal allele is simulated to be expressed, while the maternal one is silenced), T_{p} should be larger than T_{m} as well as T, since T is not a good indicator of linkage under imprinting. T_{m}, on the other hand, should have a central Chi-square distribution. The three statistics considered together, therefore, could help determining the presence or absence of linkage and imprinting for a specific trait at a specific locus.

Test for imprinting

In addition to the T statistic proposed by Sham et al. (2002), we propose a statistical test for imprinting to further determine the presence or absence of imprinting, defined as follows: $I = \frac{\hat{Q}_{\mathrm{p}}-\hat{Q}_{\mathrm{m}}}{\sqrt{\mathrm{Var}(\hat{Q}_{\mathrm{p}})+\mathrm{Var}(\hat{Q}_{\mathrm{m}})},}$ where the subscript p or m indicates that $\hat{Q}$ is estimated with paternal or maternal IBD sharing proportions. $\hat{Q}_{\mathrm{p}}$ and $\hat{Q}_{\mathrm{m}}$ are assumed to be independent under the null hypothesis. Here, $\mathrm{Var}(Q) = 1/\sum[B'\Sigma_{\hat{\pi}}B]$, as follows from the Appendix C to Sham et al's paper (Sham et al. 2002). $\mathrm{Var}(\hat{Q}_{\mathrm{p}})$ and $\mathrm{Var}(\hat{Q}_{\mathrm{m}})$ are obtained by using corresponding paternal or maternal $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Pi}}}$ matrices. Under no imprinting, the distribution of I is expected to be asymptotically standard normal provided the parental genotypes are available and there is no selection. Violation of these requirements may lead to dependencies. Therefore, we suggest obtaining empirical P-values instead. Empirical critical values from data simulated under the null hypothesis can also be derived. There are two variants of null hypothesis for this situation: (1) linkage is present but there is no imprinting, and (2) there is no linkage. The distribution of I can be obtained for both of these scenarios in a simulation setting. When dealing with real data, the reference distribution of I can be obtained assuming no linkage, which, however, may result in a suboptimal test. A better way to obtain the reference distribution for I is a permutation technique similar to that described by Gorlova et al. (2003). We denote the permuted test as $I_{\mathrm{perm}}L$ since we are leaving intact the effects of linkage but permuting the imprinting effects within families. In that procedure, paternal and maternal components of IBD are permuted within blocks defined by all sib pairs belonging to the same family, with probability 0.5, independently for each family. Thus, the total identity-by-descent within families and hence total linkage is preserved, while any parent-specific linkage is removed by permuting the parental components.

In calculating the T and I statistics, both $\boldsymbol{\Sigma}_{\mathbf{Y}}$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Pi}}}$ can be obtained either using theoretical formulas (assuming multivariate normality) or imputation, or directly from the sample. The choice as to which one to use should be handled with caution. When the sample is uniform in structure and is large enough, the covariance matrices could be reliably estimated from the sample. However, if not all family sizes are well represented, then the sample estimates are likely to be biased and theoretical covariance matrices are preferred

(Sham, personal communication). In calculating theoretical $\mathbf{\Sigma}_{\mathbf{Y},}$ heritability and kinship coefficient are needed. To calculate imputed $\mathbf{\Sigma}_{\hat{\Pi},}$ one can use either the original or the alternative definition described in Sham et al.'s paper. The latter defines each element in the matrix as $Cov(\hat{\pi}_{ij,}\hat{\pi}_{kl}) = (\hat{\pi}_{ij} - 0.5)(\hat{\pi}_{kl} - 0.5)$ between pairs of sib-pairs in each family, where 0.5 is the expectation of IBD sharing proportions between sibs. Although this definition has some obvious problems (e.g. the estimates of the covariance can exceed the estimates of the variance), its advantage is that it can be implemented in a straightforward way, and is computationally easy. This, and the fact that its properties were not investigated in the original paper, served as a reason to choose this approach when implementing the method (Table 1). The performance of the method while using the sample-based estimate of $\mathbf{\Sigma}_{\hat{\Pi}}$ was not evaluated by Sham et al. as well, but is evaluated here (Table 2).

To increase the precision of the sample-based estimate of $\mathbf{\Sigma}_{\hat{\Pi},}$ we used information from both sib pairs and from larger sibships. In families with two children, there is only one sib pair and hence only $\pi_{1\ 2}$ exists. Thus, $\mathbf{\Sigma}_{\hat{\Pi}}$ matrix consists of just one element, the variance, for such families. However, in calculating this variance, we not only used families having exactly two children, but also families that included several independent pairs of children [for example, from a family with five children, we included (1,2) and (3,4) sibling pairs], since IBD sharing is pairwise as well as jointly independent for such sib pairs (Blackwelder and Elston 1985). When larger sibships were studied, the same approach of using all appropriate families was used for estimating variances and covariances. For example, when calculating $\mathbf{\Sigma}_{\hat{\Pi}}$ for triads (sets of three offspring), all families with one or more independent triads were used—there would be one triad from families with three, four, or five children (randomly chosen in case of more than three children available), and two triads from families with six to eight children and so on. From these triads (and larger clusters of sibs) variances and covariances between identities by descent were estimated.

Simulated data

The performance of the method was first evaluated on simulated data under seven different scenarios, one of which is shown in four different sample sizes. We simulated quantitative trait with the major gene variance of 0.25, the polygene variance of 0.50, and the residual environmental variance of 0.25. These settings were used in all scenarios except when otherwise indicated. In all settings we simulated the data using a disease allele frequency of 0.3. For the imprinting simulations we assumed complete maternal imprinting (maternal allele silencing) without dominance. Conditional on genotypes, a normally distributed major gene value was simulated. A normally distributed polygene value was simulated using a normal distribution with mean 0 and variance specified above. Finally, the residual was simulated based on the normal distribution. The major gene value, polygene value, and residual value were added to obtain the final phenotype value used in the analysis. We evaluated the performance of the method using both theoretical (imputed) and sample covariance matrices of π.

1. *Base case scenario*. Our base case scenario included 100 nuclear families in each dataset, 6 family members (parents and four siblings) in each family.
2. *Misspecification of the population mean of the trait*. Sham's method requires that the population mean of the trait be correctly specified when standardizing the trait. To test the effect of the misspecification of the population mean, which might occur in analyses of real datasets, we analyzed data with mean estimate shifted by 1 SD.
3. *Missing parental genotypes*. In a real dataset, parental marker information usually is not complete; this happens especially often with the paternal genotype. To test the effect of missing parental genotypes, we randomly deleted a part of the parental markers in the simulated sample. We assumed that a paternal marker had a 50% chance and a maternal marker a 30% chance of being missing, independently. These proportions of missing genotypes are somewhat higher than (although comparable to) those observed in our real dataset. Thus, around 15% families had both parental genotypes missing. Absence of parental genotypes can be expected to reduce the power of imprinting tests.
4. *Low heritability*. To test the effect of the major gene component of genetic variance on the performance of the method, we set the major gene variance to 0.12, the polygenic variance to 0.50, and the residual environmental variance to 0.38.
5. *Effect of larger number of families (200) in the sample*. We also evaluated the effect of a larger number of families on the method's performance, assuming the sample size of 200 families with 4 children in each (1,200 sib pairs).
6. *Ascertainment*. We also explored the properties of the method when analyzing a selected sample. The selection criteria were chosen to mimic those by which the ascertainment was done in our real dataset. Namely, one of the sibs had to have the trait value that was greater than the mean by 1.3 SD; another sib had to exceed the mean by 0.25 SD; additionally, one more sib and at least one parent had to be at the mean value or below with respect to the trait.

7 a–d. *Presence of families with different sizes*. Finally, we simulated samples that were represented by a mixture of families with different sizes. The proportion of families of each size was as follows: 57% with 4 family members (two children), 28% with 5 members, 10% with 6 members and 5% with 7 members. These numbers approximately reflect the relative proportion of such families in the population (families with one child were excluded since no sib pair can be obtained from them; families with more than five children were also excluded because they are rare in the population). This analysis was performed with three different sample sizes: (a) 100 families as in the base case; (b) 150 families, to have about the same number of sibs as in the base case (which was 100 families with 4 children); (c) 240 families, to have about the same number of sib pairs as in the base case; (d) 1,000 families, to evaluate how well the larger sample size agrees with asymptotic properties of the statistics and how much the power is improved.

Analysis setup for the simulated data

To test the performance of the method in detecting linkage and imprinting, we simulated four datasets in each simulation run, as follows:

1. The trait locus is not imprinted and is not linked to the marker locus (the null hypothesis denoted as NI.NL);
2. The trait locus is imprinted but is not linked to the marker locus (another null hypothesis, I.NL);
3. The trait locus is not imprinted and is linked to the marker locus (alternative for linkage, null for imprinting in presence of linkage, NI.L); and
4. The trait locus is imprinted (maternal allele silenced) and is linked to the marker locus (alternative for both linkage and imprinting, I.L.).

We assumed a normally distributed trait and a genetic marker with eight alleles. Marker information was used as an input by the IBD_FM program. Its output (total and parent-specific $\hat{\mathbf{\Pi}}$s), along with the trait information, served as input for the program that implemented Sham et al.'s method in SAS. After 10,000 repeats of this simulation process we analyzed the four statistics, T, T_p, T_m and I. We expected T, T_p, *and* T_m to follow the 50:50 mixture of 0 and χ_1^2, and I—to asymptotically follow the standard normal distribution. These distributions were used to obtain theoretical type 1 error rates and power for each of these statistics. Empirical critical levels (the 95th percentiles of each linkage statistics) were obtained from the NI.NL output, since the distributions of the four statistics from the I.NL dataset were identical to those from NI.NL (since both are essentially the null). The empirical power to detect linkage was estimated for both the NI.L and I.L output.

We tested whether the distribution of I was standard normal under NI.NL, NI.L, and permuted imprinting I.L model, which we denote $I_\mathrm{perm}L$. The three models, NI.NL, NI.L, and $I_\mathrm{perm}L$ are all null for imprinting, with NI.L and $I_\mathrm{perm}L$ corresponding to the situation with linkage but no imprinting. Ten thousand permutations were performed to generate the I statistic's distribution under the $I_\mathrm{perm}L$ model. As mentioned, the theoretical type 1 error rate for the I statistic was estimated by comparison of its distributions under the no-imprinting, linkage models (NI.L and $I_\mathrm{perm}L$) to the standard normal distribution. The empirical type 1 error rate for the no-imprinting, linkage model (NI.L) was obtained using the critical value from the no-imprinting, no-linkage (NI.NL) model. The empirical type 1 error rate for the permutation-derived distribution of I (from the $I_\mathrm{perm}L$ model) (Gorlova et al. 2003) corresponds to its nominal level by definition. Also, the empirical power to detect imprinting was obtained from applying the critical values from the permutation-derived distribution of I to the I.L output. Since the permutation procedure results in a correct type 1 error, the use of critical values obtained from it provides a common level of type I error rate and, therefore, ensures a fair comparison of power across different scenarios (Table 3).

In addition to the conventional type 1 error rate and power, we also calculated the proportions of simulation runs that resulted in the expected ordering of T, T_p and T_m under the I.L model, which is $T_\mathrm{p} > T > T_\mathrm{m}$ (Table 2).

Real dataset: study participants

We applied the method to data that have been collected to study the genetics of body size. The dataset contained 255 nuclear families, with a variable number of siblings (Li et al. 2003). All family probands (extremely obese individuals with BMI > 40) had at least one obese sibling (BMI > 30) and at least one parent and one sibling of normal weight (BMI < 27). All subjects gave informed consent, and the protocol was approved by the Committee on Studies Involving Human Beings at the University of Pennsylvania. There were 891 siblings (1,341 sib pairs) overall. The median sibship size was 3. Data on six obesity-related traits, namely, body mass index (BMI), bioelectrical impedance analysis (BIA, a body fat percentage measure), waist–hip ratio (WHR), fasting glucose, plasma leptin concentration, and waist circumference, were obtained for each family member. The characteristics of the traits' distributions were presented previously (Li et al. 2003). Twenty-three markers on chromosome 7 (7q22.1–7q35), flanking but mostly 5′ of the leptin gene, located from 111.3 cM (99.67 Mb) to 155 cM (143.12 Mb) were

Table 1 Method's performance at the 0.05 significance level based on the simulated data, with the imputed covariance matrix of IBD

Base case	Test for linkage						Test for imprinting	
	T		T_p		T_m		I	
	5% critical value	Empirical power (NI.L, I.L)	5% critical value	Empirical power (NI.L, I.L)	5% critical value	Empirical type 1 error rate (I.L) or power (NI.L)	5% critical value	Empirical[a]/theoretical[b] type 1 error rate (NI.NL, NI.L) or empirical power (I.L)
NI.NL	2.260		2.279		2.117		2.142	0.050/0.073
NI.L	10.991	0.800	7.862	0.539	7.668	0.558	1.447	0.010/0.026
I.L	11.199	0.807	13.752	0.947	2.232	0.054	3.017	0.482[c]

[a] Based on the NI.NL-derived critical values

[b] Based on the standard normal distribution

[c] Based on the critical value from NI.L

genotyped for each individual (Table 4). The proportion of families with parental marker information completely missing varied from 4 to 9% for different markers. The genotype information on fathers was less than 50% missing, and on mothers less than 10% missing for all markers (Table 4). We identified 11 families for whom the parental genotype information was missing for all markers. These families were included when evaluating total linkage, but not when the analysis of parent-specific linkage and imprinting was performed (leaving 839 sibs and 1223 sib pairs available). This reduced the proportion of families without parental genotype information to at most 4.5% (for the marker –2,548). Five families were excluded because they had only one child.

Analysis setup for the real data

Overall and parent-specific IBD sharing proportions were calculated at each marker from the sample by a modification of the SIMWALK program that outputs multipoint parent-specific IBDs, which further reduces the impact of missing parental genotypes. Linkage and imprinting statistics T, T_p, T_m and I were calculated using the sample-derived $\boldsymbol{\Sigma}_{\hat{\Pi}}$ as described above, at each locus for each of the six traits, respectively. The use of the sample-derived $\boldsymbol{\Sigma}_{\hat{\Pi}}$ was justified because families of each size were well represented (we had 49, 90, 49, 31, and 20 families with 2, 3, 4, 5, and 6 children, respectively). For comparison, we performed the same analysis based on the imputed $\boldsymbol{\Sigma}_{\hat{\Pi},}$ but since the results were very similar (despite that the test for imprinting was likely less powerful), we only present those based on the sample-derived $\boldsymbol{\Sigma}_{\hat{\Pi}}$. To determine the significance of the statistics in each locus-trait combination, we simulated markers under the absence of linkage, using program SIMULATE (Terwilliger et al. 1993) for a sample that had the same family structure and trait values as the real sample (20,000 simulations were performed for each of the 6 traits). As to simulating the marker data, we realize that it would be appropriate to replicate patterns of missing data and allele frequencies in the original data. Alternatively, a set of assumptions that results in the most conservative null distribution can be obtained. We chose to use the setting corresponding to the marker –2,548, which is a diallelic marker with allele frequencies of 0.62 and 0.38 and for which 4.5% of families have both parental genotypes missing, because it produced the most conservative reference distribution. Empirical 95% critical values of the statistics were obtained from the simulated data after applying our method to them, and used to decide upon linkage significance. We applied the permutation technique as described above (Gorlova et al. 2003) to test for the presence of imprinting. This procedure was repeated 1,000 times for loci with evidence for linkage, where the parent-of-origin effect was probable in view of parent-specific linkage statistics.

As was mentioned, Sham et al.'s method requires that the traits be standardized. In particular, the trait value should be population mean-centered. The population mean in many cases cannot be estimated from the sample due to ascertainment of the families. Thus, in our case of a highly selected sample, we relied upon the mean estimates available in literature. The mean estimates for BMI, waist-to-hip ratio and leptin were obtained from the paper by Ruhl and Everhart (2001); for waist circumference from Langenberg et al. (2003); for BIA from Chumlea et al. (2002). We used the midpoint of fasting glucose normal range, according to the American Diabetes Association, as its population mean. As for the standard deviations, since Sham et al.'s simulation study shows that the method is robust to misspecified variance, we just used the sample statistics. In obtaining the covariance matrices of squared sums and squared differences, the estimate of trait's heritability is required and also multivariate normality is assumed. We used the estimates of heritability from literature (Freeman et al. 2002; Luke et al. 2001). After standardization, the traits were adjusted for sex, age and

Table 2 Method's performance at the 0.05 significance level based on the simulated data, with the sample-derived covariance matrix of IBD: test for linkage

Simulation setting	Linkage and imprinting model	T			T_p			T_m			
		5% critical value	Empirical type 1 error rate (NI.NL) or power[a](NI.L, I.L)	Theoretical type 1 error rate (NI.NL)[b]	5% critical value	Empirical type 1 error rate (NI.NL) or power[a] (NI.L, I.L)	Theoretical type 1 error rate (NI.NL)[b]	5% critical value	Empirical type 1 error rate (NI.NL, I.L) or empirical power [a] (NI.L)	Theoretical type 1 error rate (NI.NL, I.L) [a]	% of $T_p > T > T_m^c$
1. Base case: 100 families, 4 children in each; major gene variance 25%	NI.NL	2.7183	0.0500	0.0503	2.8908	0.0500	0.0546	2.6834	0.0500	0.0490	
	NI.L	18.9457	0.7997		13.5689	0.5457		13.2618	0.5627		
	I.L	18.8417	0.8035		30.0247	0.9490		2.8147	0.0546	0.0536	90.47
	NI.NL mean	*0.4954*			*0.5193*			*0.5069*			
2. Misspecified mean (shifted by 1 SD)	NI.NL	2.6738	0.0500	0.0491	2.7976	0.0500	0.0528	2.7316	0.0500	0.0510	
	NI.L	16.8766	0.7416		12.2209	0.4901		12.1676	0.4975		
	I.L	16.7931	0.7362		26.9718	0.9146		2.9404	0.0568	0.0575	87.35
	NI.NL mean	*0.4927*			*0.5141*			*0.5021*			
3. Missing parental genotypes: paternal 50%, maternal 30%	NI.NL	2.6762	0.0500	0.0492	2.816	0.0500	0.0525	2.7447	0.0500	0.0515	
	NI.L	18.3408	0.7916		13.7713	0.5644		13.384	0.5625		
	I.L	18.5987	0.7995		28.707	0.9360		3.6298	0.0799	0.0819	88.22
	NI.NL mean	*0.4935*			*0.5106*			*0.5118*			
4. Lower major gene effect (major gene variance 12%)	NI.NL	2.7183	0.0500	0.0503	2.8908	0.0500	0.0546	2.6834	0.0500	0.0490	
	NI.L	10.1119	0.4232		8.0485	0.2645		7.7279	0.2806		
	I.L	8.7072	0.3384		12.4417	0.4949		2.7585	0.0520	0.0513	67.73
	NI.NL mean	*0.4954*			*0.5193*			*0.5193*			
5. 200 families	NI.NL	2.7305	0.0500	0.0505	2.8077	0.0500	0.0534	2.8253	0.0500	0.0528	
	NI.L	30.7709	0.9736		20.6338	0.8115		20.1661	0.8111		
	I.L	30.5849	0.9734		50.9539	0.9988		2.9734	0.0540	0.0581	97.43
	NI.NL mean	*0.5022*			*0.516*			*0.5071*			
6. Ascertainment	NI.NL	2.8215	0.0500	0.0533	2.8035	0.0500	0.0528	2.8698	0.0500	0.0556	
	NI.L	36.4183	0.9943		23.7737	0.8911		23.6671	0.8921		
	I.L	40.5745	0.9981		67.963	1.000		3.0199	0.0534	0.0585	99.26
	NI.NL mean	*0.5204*			*0.519*			*0.5193*			
7a. Varying family size, 100 families	NI.NL	2.9739	0.0500	0.0583	2.9413	0.0500	0.0574	2.8276	0.0500	0.0533	
	NI.L	11.2105	0.4683		8.3044	0.3075		8.4568	0.3106		
	I.L	10.8825	0.4686		15.3428	0.6963		2.9275	0.0537	0.0575	74.47
	NI.NL mean	*0.5226*			*0.5218*			*0.5211*			
7b. Varying family size, 150 families	NI.NL	2.7526	0.0500	0.0509	2.7724	0.0500	0.0525	2.8693	0.0500	0.0553	
	NI.L	13.9332	0.6367		10.3215	0.4196		10.2136	0.4110		
	I.L	13.7658	0.6348		20.2041	0.8460		3.0161	0.0542	0.0595	81.71
	NI.NL mean	*0.4908*			*0.4955*			*0.524*			

Table 2 continued

Simulation setting	Linkage and imprinting model	T			T_p			T_m			
		5% critical value	Empirical type 1 error rate (NI.NL) or power[a](NI.L, I.L)	Theoretical type 1 error rate (NI.NL)[b]	5% critical value	Empirical type 1 error rate (NI.NL) or power[a] (NI.L, I.L)	Theoretical type 1 error rate (NI.NL)[b]	5% critical value	Empirical type 1 error rate (NI.NL, I.L) or empirical power [a] (NI.L)	Theoretical type 1 error rate (NI.NL, I.L) [a]	% of $T_p > T > T_m^c$
7c. Varying family size, 240 families	NI.NL	2.4679	0.0500	0.0460	2.9659	0.0500	0.0590	2.5009	0.0500	0.0440	
	NI.L	18.7321	0.8112		12.9691	0.5456		13.153	0.5774		
	I.L	18.4304	0.8095		28.4155	0.9578		2.9025	0.0596	0.0564	89.68
	NI.NL mean	*0.4743*			*0.532*			*0.4672*			
7d. Varying family size, 1,000 families	NI.NL	2.6058	0.0500	0.0468	2.6743	0.0500	0.0487	2.7156	0.0500	0.0504	
	NI.L	50.6362	1.0000		31.0887	0.9735		31.2036	0.9761		
	I.L	48.9253	0.9997		84.4894	1.0000		3.0497	0.0596	0.0598	99.73
	NI.NL mean	0.4817			0.4943			0.4896			

[a] Empirical type 1 error rate and power are determined based on critical values empirically derived from the corresponding simulated null distribution (NI.NL) (by definition, there is 5% significance for the empirical critical value for the distribution from which it is derived)

[b] Theoretical type 1 error is based on the critical value for 50:50 mixture of 0 and χ^2_1, which corresponds to χ^2_1 (0.9) ≈ 2.71.

[c] This number indicates the percent of all simulations under the I.L model in which paternal T (T_p) exceeded overall T and simultaneously overall T exceeded maternal T (T_m), i.e. in which the three T statistics were ordered as expected

squared age; the residuals were then converted to approximate normality via the Box–Cox transformation. Race was not adjusted since it was never statistically significant when the other three covariates were in the model. While Sham's method remains valid even if the traits are not normally distributed, the power should be improved for more normally distributed data sets (Sham et al. 2002).

Results

Results of the simulation study

We first studied the performance of the method using the imputed covariance matrix $\boldsymbol{\Sigma}_{\hat{\Pi}}$ under the base case scenario. We observed that the distribution of T under the null model of no linkage deviated from the expected 50:50 mixture of 0 and χ^2_1 at the tail (Fig. 1a, b). The 95th percentiles of T, T_p, and T_m under the null were 2.26, 2.279, and 2.117, respectively, as compared to the theoretical value 2.71; the type 1 error rate for maternal linkage under the I.L model (maternal imprinting) was 0.054; the empirical power for the test for total linkage was 0.80 in both NI.L and I.L settings, and for the parent-specific linkage close to 0.55 under NI.L setting and 0.947 (paternal) for the I.L setting (Table 1). The distribution of I statistic deviated from the standard normal distribution under both NI.NL and NI.L models; the empirical power for the test for imprinting was 0.482 (Table 1). Due to the deviation of the linkage statistic distribution from expected, and to the fact that the test for imprinting was more powerful when the sample-derived estimates of $\boldsymbol{\Sigma}_{\hat{\Pi}}$ were used (see below), we decided to use the sample-derived, rather than imputed estimates of $\boldsymbol{\Sigma}_{\hat{\Pi},}$ in the rest of our simulations.

When using the sample-derived estimate for the $\boldsymbol{\Sigma}_{\hat{\Pi}}$ matrix under the base case scenario (Table 2, block 1), we noted that the distributions of T, T_p, and T_m under the NI.NL setting were very close to the expected 50:50 mixture of 0 and χ^2_1, as evidenced by both the mean values of the three statistics and by their 95% critical values (Table 2, block 1). Under the I.L setting (maternal imprinting), T_p had the highest power (94.9%) and T_m was significant in about 5.5% of simulations when using the 95% critical value from the NI.NL model, which was close to the correct type I error rate. With our simulation setting of eight alleles and high major gene effect of 0.25, the method gives a very high power (0.80) for overall T to test linkage (Table 2, block 1). When there is no imprinting, T_p and T_m have about the same power (close to 0.55), which is much lower than that of the overall T. The panel of T, T_p and T_m is helpful in determining the presence of imprinting: when imprinting is present, 90.5% simulations show the expected ordering of T, T_p and T_m, so that $T_p > T > T_m$.

Table 3 Method's performance at the 0.05 significance level based on the simulated data, with the sample-derived covariance matrix of IBD—test for imprinting (*I* statistic)

Simulation setting	Linkage and imprinting model	Two-sided 5% critical value	Mean	SD	Empirical type 1 error rate (NI.NL, NI.L, $I_{\mathrm{perm}}L$) or empirical power[a](I.L)	Theoretical type 1 error rate (NI.NL, NI.L, $I_{\mathrm{perm}}L$) based on the critical values from the standard normal distribution
1. Base case: 100 families, 4 children in each; major gene variance 25%	NI.NL	1.9617	–0.0018	1.00	0.0500	0.0501
	NI.L	2.1665	0.0107	1.12	0.0800	0.0801
	$\mathrm{I_{perm}L}$[b]	2.2848	–0.0158	1.17	0.0500	0.0957
	I.L	4.2749			0.5885	
2. Misspecified mean (shifted by 1 SD)	NI.NL	1.9665	–0.0033	0.99	0.0500	0.0512
	NI.L	2.1942	0.0061	1.12	0.0793	0.0807
	$I_{\mathrm{perm}}L$	2.2670	–0.0099	1.16	0.0500	0.0893
	I.L	4.0920			0.5120	
3. Missing parental genotypes: paternal 50%, maternal 30%	NI.NL	1.9086	–0.0116	0.97	0.0500	0.0442
	NI.L	2.0978	0.0145	1.09	0.0765	0.0683
	$I_{\mathrm{perm}}L$	2.1985	–0.0089	1.14	0.0500	0.0837
	I.L	3.9901			0.5269	
4. Lower major gene effect (major gene variance 12%)	NI.NL	1.9617	–0.0018	1.00	0.0500	0.0501
	NI.L	2.0676	0.0108	1.07	0.0651	0.0651
	$I_{\mathrm{perm}}L$	2.1391	0.0087	1.09	0.0500	0.0731
	I.L	2.9164			0.1860	
5. 200 families	NI.NL	1.9529	0.0124	0.99	0.0500	0.0493
	NI.L	2.1817	0.0017	1.13	0.0837	0.0827
	$I_{\mathrm{perm}}L$	2.2888	–0.0026	1.18	0.0500	0.0957
	I.L	5.4282			0.9042	
6. Ascertainment	NI.NL	1.9608	0.0009	1.00	0.0500	0.0502
	NI.L	2.2031	–0.0048	1.12	0.0786	0.0786
	I.L(p)	2.3569	–0.0156	1.20	0.0500	0.1052
	I.L	6.1805			0.9803	
7a. Varying family size, 100 families	NI.NL	1.9622	–0.0068	1.01	0.0500	0.0504
	NI.L	2.0724	0.0037	1.06	0.0641	0.0646
	$I_{\mathrm{perm}}L$	2.1688	0.0021	1.10	0.0500	0.0757
	I.L	3.1844			0.2777	
7b. Varying family size, 150 families	NI.NL	1.9768	–0.0144	1.01	0.0500	0.0520
	NI.L	2.0627	–0.0041	1.06	0.0605	0.0626
	$I_{\mathrm{perm}}L$	2.1389	–0.0096	1.09	0.0500	0.0731
	I.L	3.5850			0.4116	
7c. Varying family size, 240 families	NI.NL	1.9372	0.0338	1.00	0.0500	0.0470
	NI.L	2.0935	–0.0081	1.07	0.0695	0.0662
	$I_{\mathrm{perm}}L$	2.1661	–0.0285	1.10	0.0500	0.0764
	I.L	4.119			0.6138	
7d. Varying family size, 1,000 families	NI.NL	1.9209	0.0086	0.99	0.0500	0.0459
	NI.L	2.1543	0.0058	1.10	0.0800	0.0740
	$I_{\mathrm{perm}}L$	2.2367	–0.0103	1.14	0.0500	0.0841
	I.L	6.9053			0.9978	

[a] Empirical type 1 error rate is 0.05 for the NI.NL model by definition (using the critical values obtained from the NI.NL itself). For the NI.L model, the empirical type 1 error is based on the NI.NL-derived critical value. For the $I_{\mathrm{perm}}L$[b] model, empirical type 1 error rate is 0.05 by definition (using the critical values obtained from the $I_{\mathrm{perm}}L$ itself). For the I.L model, the empirical power is based on the critical value from the $I_{\mathrm{perm}}L$ model

[b] $I_{\mathrm{perm}}L$ stands for the setting in which parent-specific values of IBD sharing (i.e. paternal vs. maternal) for sib pairs were permuted within families, to provide a null distribution for the test for imprinting, *I*, under the presence of linkage

Table 4 Genetic markers used in the study, their genetic and physical locations, and completeness of parental genotype information

No	Marker	Gen. location (cM)	Phys. location (Mb)	No of alleles	Both parents genotyped			Only mother missing			Only father missing			Both parents missing			
					N	% [a]	% [b]	N	% [a]	% [b]	N	% [a]	% [b]	N [a]	% [a]	N [b]	% [b]
1	D7S2480	111.3	99.67	14	110	45.1	43.1	18	7.4	7.1	109	44.7	42.7	7	2.9	18	7.1
2	D7S796	113.4	103.06	10	110	45.1	43.1	19	7.8	7.5	111	45.5	43.5	4	1.6	15	5.9
3	D7S2459	119.8	106.89	7	108	44.3	42.4	23	9.4	9	108	44.3	42.4	5	2	16	6.3
4	D7S692	121.4	107.9	11	111	45.5	43.5	22	9	8.6	111	45.5	43.5	0	0	11	4.3
5	D7S523	123	111.26	11	111	45.5	43.5	23	9.4	9	109	44.7	42.7	1	0.4	12	4.7
6	D7S643	125.2	120.29	14	110	45.1	43.1	22	9	8.6	109	44.7	42.7	3	1.2	14	5.5
7	D7S685	127.8	120.85	10	109	44.7	42.7	21	8.6	8.2	109	44.7	42.7	5	2	16	6.3
8	D7S2529	128.8	121.97	13	109	44.7	42.7	23	9.4	9	108	44.3	42.4	4	1.6	15	5.9
9	D7S514	130.2	126.58	11	110	45.1	43.1	22	9	8.6	111	45.5	43.5	1	0.4	12	4.7
10	D7S2501	130.7	127.07	11	111	45.5	43.5	22	9	8.6	109	44.7	42.7	2	0.8	13	5.1
11	D7S504	130.8	127.17	11	110	45.1	43.1	22	9	8.6	109	44.7	42.7	3	1.2	14	5.5
12	D7S1875	130.9	127.3	16	109	44.7	42.7	22	9	8.6	110	45.1	43.1	3	1.2	14	5.5
13	D7S1529	131	127.33	21	106	43.4	41.6	26	11	10	105	43	41.2	7	2.9	18	7.1
14	–2548	131.1	127.435	2	102	41.8	40	24	9.8	9.4	107	43.9	42	11	4.5	22	8.6
15	+19	131.102	127.435	2	111	45.5	43.5	20	8.2	7.8	112	45.9	43.9	1	0.4	12	4.7
16	D7S530	134.55	128.76	12	112	45.9	43.9	21	8.6	8.2	110	45.1	43.1	1	0.4	12	4.7
17	D7S649	136.1	130.27	8	109	44.7	42.7	22	9	8.6	111	45.5	43.5	2	0.8	13	5.1
18	D7S1804	137	131.7	17	109	44.7	42.7	23	9.4	9	108	44.3	42.4	4	1.6	15	5.9
19	D7S2452	138.3	132.84	14	109	44.7	42.7	24	9.8	9.4	109	44.7	42.7	2	0.8	13	5.1
20	D7S2438	138.42	133.33	15	108	44.3	42.4	23	9.4	9	111	45.5	43.5	2	0.8	13	5.1
21	D7S1837	142	136.12	7	109	44.7	42.7	23	9.4	9	110	45.1	43.1	2	0.8	13	5.1
22	D7S2202	149.9	139.19	9	104	42.6	40.8	25	10	9.8	107	43.9	42	8	3.3	19	7.5
23	D7S794	155	143.12	9	108	44.3	42.4	23	9.4	9	111	45.5	43.5	2	0.8	13	5.1

N Number of families

[a] Percentage among families not including those with both parental genotypes missing for all markers (11 excluded)

[b] Percentage among all 255 families

Under the base-case scenario, the test for imprinting based on index *I* (Table 3, block 1) follows the standard normal distribution perfectly under the no-linkage, no-imprinting model (as reflected by the mean, SD, and the critical value), resulting in the correct theoretical type 1 error rate. We also obtained the distribution of *I* under the NI.L model. We noted that the mean was close to zero, but the distribution was wider—its standard deviation was greater than that for the standard normal distribution (1.12 rather than 1). This resulted in an inflated type 1 error rate of *I* (both theoretical, i.e. based on the standard normal distribution, and empirical, i.e. based on the critical values from NI.NL) under NI.L, of about 0.08. We as well obtained the distribution of *I* using the permutation procedure

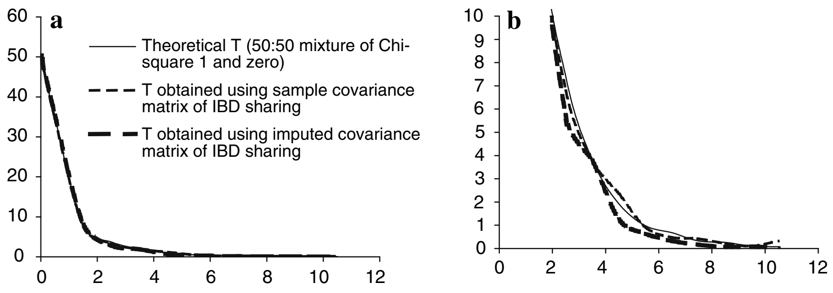

Fig. 1 **a** Distributions of *T* statistics calculated in three ways: (i) theoretically as a 50:50 mixture of χ_1^2 and zero (*solid line*), (ii) using the sample $\boldsymbol{\Sigma}_{\hat{\Pi}}$ (*dashed line*) and (iii) using the imputed $\boldsymbol{\Sigma}_{\hat{\Pi}}$ (*bold dashed line*). **b** Tails of the distributions of *T* statistics calculated theoretically, using the sample $\boldsymbol{\Sigma}_{\hat{\Pi}}$ and the imputed $\boldsymbol{\Sigma}_{\hat{\Pi}}$ to show differences in detail

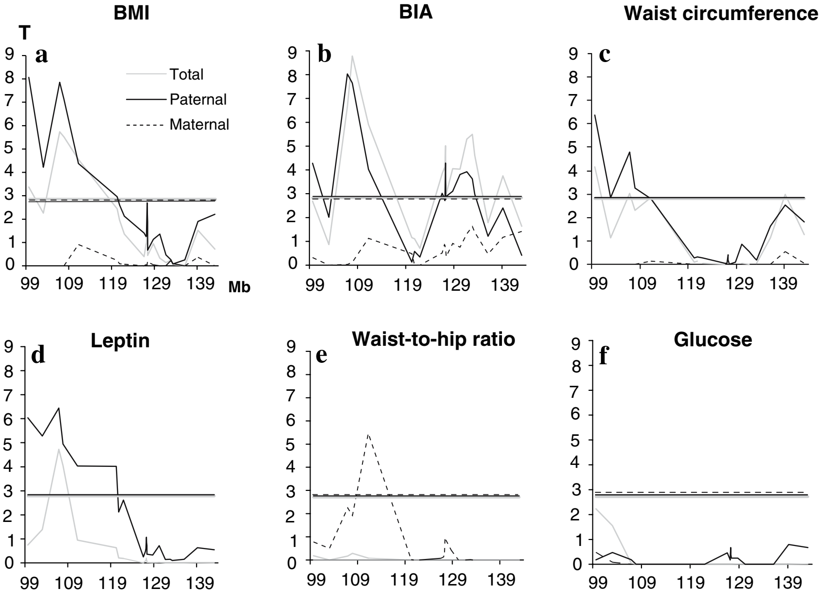

Fig. 2 Overall and parent-specific linkages to 6 obesity related traits along the segment of Chromosome 7 ranging from 99.67 to 143.12 Mb; **a** body mass index; **b** bioelectric impedance index; **c** waist circumference; **d** leptin concentration; **e** waist-to-hip ratio; **f** glucose concentration. *Horizontal lines* indicate critical values at the 0.05 significance level

(corresponding to linkage but no imprinting), observing that in this setting it was, again, wider than the standard normal distribution (mean –0.016, SD 1.17; Table 3 block 1). By definition, the permutation procedure ensures the correct type 1 error for a given significance level, yet we also wanted to evaluate the false-positive rate of the imprinting test associated with the use of standard normal distribution instead of permutation-based distribution (under this and other scenarios (below), line $I_{\text{perm}}L$). Our results (Table 3) suggest that under the base case scenario the use of critical values from the standard normal distribution would lead to a liberal test. The power for the test for imprinting was moderately high (0.589) based on the permutation-derived critical value.

When a biased population trait mean is used in standardization, both linkage and imprinting tests lose power, while the type 1 error rate is not greatly affected (Tables 2 and 3, block 2 in both).

A common problem with real datasets is unavailability of genetic information on parents, especially on fathers. However, the parental information is essential in determining parent-of-origin effect. When paternal information had a 50% chance and maternal information—a 30% chance of being missing, independently, resulting in 15% of families with totally missing parental information, the power of the tests for linkage did not decrease much, although the type 1 error rate for maternal linkage did increase to about 0.08 (Table 2, block 3). There was a moderate loss in power for the imprinting test as well (Table 3 block 3), while the type 1 error for this test did not suffer further inflation—it remained at about 0.08 level.

The effect of a lower major gene variance (0.25 vs. 0.12) on the method's performance (Tables 2 and 3, block 4) was substantial. The power of linkage tests dropped approximately by half and the power for the test for imprinting was only about 0.19, while the type 1 error rate for the tests for linkage and imprinting was similar to that for the base case.

Results in block 5 of Tables 2 and 3 were calculated for a sample of 200 families (parents and four children in each). A high power was observed for the linkage tests, and a considerable gain in power was noted for the test for imprinting (the power reached 0.90). Again, the type 1 error rate remained almost unchanged.

We also evaluated the effect of selected samples on the performance of our statistics (Tables 2 and 3, block 6). The particular ascertainment scheme used in the simulations resulted in a very high power for both linkage and imprinting tests. However, under this scenario the theoretical type 1 error for the test for imprinting was inflated to more than 0.1 (Table 3 block 6), indicating that empirical *P*-values should be obtained. The type 1 error rate for the linkage tests, however, was almost identical to that in the base case (between 0.05 and 0.06).

When the sample consisted of families non-uniform in size (Tables 2 and 3, blocks 7a–d), the theoretical type 1 error for the test for imprinting was modestly inflated (between 0.07 and 0.09, as it was in the base case), regardless of the number of families in the sample, making the empirical permutation-based critical values for the test for imprinting more appropriate. An inflation of the type 1 error for the tests for linkage was not observed in comparison to the base case. The power, as expected, depended on the number of families in the sample for both linkage and imprinting tests, exceeding 0.99 for the setting with 1,000 families per sample.

Table 5 Significant linkage and imprinting findings on Chromosome 7 (99.67–143.12 Mb)

No	Marker	Mb	Trait	T^{a}		T_{p}^{a}		T_{m}^{a}		I^{b}
				LOD	*P*-value	LOD	*P*-value	LOD	*P*-value	*P*-value
1	D7S2480	99.67	BMI	**0.73**[c]	**0.037**	**1.75**	**0.004**	0	>0.999	**0.007**
			BIA	0.58	0.052	**0.93**	**0.023**	0.0700	0.28	0.3216
			Leptin	0.16	0.190	**1.31**	**0.011**	0	>0.999	**0.002**
			Waist circumference	**0.9**	**0.023**	**1.39**	**0.007**	0.0005	0.47	0.0716
2	D7S796	103.06	BMI	0.49	0.069	**0.92**	**0.024**	0	>0.999	**0.0454**
			Leptin	0.30	0.120	**1.15**	**0.015**	0	>0.999	**0.0056**
			Waist circumference	0.25	0.143	0.62	0.051	0	>0.999	0.066
3	D7S2459	106.89	BMI	**1.25**	**0.010**	**1.71**	**0.005**	0	>0.999	**0.0212**
			BIA	**1.45**	**0.006**	**1.75**	**0.004**	0.0061	0.43	**0.0174**
			Leptin	**1.03**	**0.018**	**1.40**	**0.009**	0	>0.999	**0.0196**
			Waist circumference	**0.66**	**0.043**	**1.04**	**0.018**	0	>0.999	**0.0302**
4	D7S692	107.9	BMI	**1.2**	**0.012**	**1.55**	**0.007**	0	>0.999	**0.039**
			BIA	**1.91**	**0.002**	**1.66**	**0.004**	0.0261	0.36	**0.049**
			Leptin	**0.88**	**0.026**	**1.08**	**0.018**	0	>0.999	0.0558
			Waist circumference	0.50	0.066	**0.70**	**0.042**	0	>0.999	0.0566
5	D7S523	111.26	BMI	**0.99**	**0.020**	**0.95**	**0.023**	0.1976	0.17	0.3824
			BIA	**1.28**	**0.010**	**0.87**	**0.026**	0.2444	0.14	
			Waist-to-hip ratio	0.018	0.378	0	>0.999	**1.1884**	**0.01**	**0.0004**
			Leptin	0.21	0.164	**0.88**	**0.027**	0	>0.999	**0.0294**
			Waist circumference	**0.62**	**0.048**	0.61	0.052	0.0269	0.36	
6	D7S643	120.29	BMI	0.53	0.062	**0.64**	**0.048**	0.0514	0.31	
			Leptin	0.14	0.280	**0.87**	**0.027**	0	>0.999	**0.0294**
9	D7S514	126.58	BIA	**0.83**	**0.026**	**0.66**	**0.046**	0.1121	0.24	
10	D7S2501	127.07	BIA	**0.89**	**0.023**	0.61	0.051	0.1872	0.18	
11	D7S504	127.17	BIA	**0.87**	**0.024**	0.59	0.054	0.1806	0.19	
12	D7S1875	127.3	BIA	**1.09**	**0.014**	**0.93**	**0.023**	0.1404	0.21	
13	D7S1529	127.33	BIA	**0.77**	**0.032**	0.60	0.053	0.1370	0.22	
14	–2548	127.435	BIA	**0.73**	**0.036**	0.62	0.051	0.0773	0.28	
15	19	127.435	BIA	**0.65**	**0.046**	**0.63**	**0.050**	0.0805	0.27	
16	D7S530	128.76	BIA	**0.88**	**0.023**	**0.67**	**0.044**	0.1991	0.17	
17	D7S649	130.27	BIA	**0.88**	**0.023**	**0.83**	**0.029**	0.1646	0.20	
18	D7S1804	131.7	BIA	**1.15**	**0.012**	**0.85**	**0.027**	0.2870	0.13	
19	D7S2452	132.84	BIA	**1.19**	**0.011**	**0.77**	**0.032**	0.3528	0.10	
20	D7S2438	133.33	BIA	**1**	**0.016**	**0.65**	**0.047**	0.3012	0.12	
22	D7S2202	139.19	BIA	**0.82**	**0.028**	0.52	0.065	0.2548	0.14	
			Waist circumference	**0.65**	**0.045**	0.55	0.061	0.1153	0.23	

[a] LOD is obtained by dividing the corresponding value of T by 2ln(10)

[b] *P*-values for I are only presented when the parent-of-origin effect was suggested by the panel of T, T_{p}, and T_{m}

[c] LOD scores significant at the 0.05 level, along with the corresponding *p*-values, are shown in bold

Due to the high major gene effect incorporated in this simulation setting of the datasets, the power of overall linkage test (and, accordingly, the T statistic values) was almost identical for the datasets with and without imprinting (I.L and NI.L scenarios, 0.80 power and T close to 19). On the other hand, when the major gene effect was lower—12% (block 4 of the Table 2), the linkage signal provided by overall T was higher when there was no imprinting.

Overall, the method performs relatively well under different scenarios, is not very sensitive to a modest amount of missing parental genotype information or to ascertainment, has an acceptable type 1 error rate for the tests for linkage, and possesses a high power to detect both linkage and imprinting provided that the permutation-based critical values are obtained for the test for imprinting. The critical values from the standard normal distribution can still be used, although with caution, for the preliminary screening of the results.

The corresponding type 1 errors and powers for the 0.01 and 0.001 significance levels are presented in the Supplementary Tables 1a and 1b.

Results of the analysis of families ascertained through extremely obese probands

Figure 2a–f shows linkage for the 6 obesity related traits. In each graph, the horizontal axis is the physical location of the 23 loci on chromosome 7, ranging from 99.67 to 143.12 Mb (Table 4). The vertical axis is the T statistics scale. Three curves for the T, T_p, and T_m are plotted in each graph, and the three horizontal reference lines represent the empirical 95% critical values (obtained from the SIMULATE's output) for the three T's, respectively to show significance. As expected, these three values are very close, resulting in overlapping reference lines.

For four traits, BMI, BIA, leptin, and waist circumference, there was a broad region of linkage, paternal and total, but not maternal, located mainly in the proximal part of the studied segment of Chromosome 7. It ranged approximately between 99.67 and 111.26 Mb. Only a single peak of maternal linkage at 111.26 Mb (marker D7S523) was found for waist-to-hip ratio, while no linkage was detected for fasting glucose. Empirical P-values for the above-referenced peaks are shown in Table 5. For loci showing linkage in the proximal part for BMI, BIA, waist circumference, and leptin, the T_p is the greatest, followed by overall T and then by an insignificant T_m, indicating possible existence of maternal imprinting (silencing of the maternal allele). For loci showing linkage for BIA more distally (128–132 Mb), overall T is greater in value than either T_p or T_m, suggesting linkage without imprinting.

The test for imprinting was significant for several loci (Table 5). In particular, locus D7S2480 at 99.67 Mb showed maternal imprinting (silencing) for BMI and leptin, while statistical significance for waist circumference was not reached. Maternal imprinting was also detected at locus D7S796 for BMI and leptin, at locus D7S2459 for BMI, BIA, leptin, and waist circumference, at locus D7S692 for BMI and BIA, and at loci D7S523 and D7S643 for leptin only. In addition, we observed maternal linkage only and paternal imprinting at locus D7S523 for the waist-to-hip ratio, albeit no total linkage was noted.

Discussion

In this study we described a method to detect parent-of-origin effect based on the regression approach proposed by Sham et al. (2002). We implemented this method in SAS and applied it to both simulated and real data.

We explored the performance of the method when using an alternative definition of the imputed $\mathbf{\Sigma}_{\hat{\Pi}}$ matrix (Sham et al. 2002). We found that although this definition allows easier implementation and does not require intense computations, it results in a distribution for T that deviates at the tail from the 50:50 mixture of 0 and χ^2_1 under the null (Table 1; Fig. 1a, b). On the other hand, our implementation also allows using sample-derived covariance matrices of estimated IBD sharing and squared sums and squared differences of the trait. In this case, the distribution of the linkage test statistics is a 50:50 mixture of χ^2_1 and zero under the null. Thus, in case when the marker information is complete, sample $\mathbf{\Sigma}_{\hat{\Pi}}$ should be a good estimate of the real covariance matrix and no simulated dataset is needed to determine the critical value for the test for linkage. We compared the T statistics (for total linkage) from our program with the LOD score obtained from MERLIN-REGRESS (the LOD score was multiplied by 4.6 to bring it to the same scale as T) under the null of no linkage. The two statistics had almost identical distribution and were highly correlated with the coefficient 0.93 when we used the sample-derived $\mathbf{\Sigma}_{\hat{\Pi}}$ matrix (the slight difference in output may be largely explained by the fact that MERLIN-REGRESS removes uninformative families). However, when the imputed $\mathbf{\Sigma}_{\hat{\Pi}}$ matrix was used, the correlation coefficient was 0.85.

Another advantage of the sample-derived $\mathbf{\Sigma}_{\hat{\Pi}}$ is that it provides higher power to detect imprinting, compared to the imputed $\mathbf{\Sigma}_{\hat{\Pi}}$ (Table 1 and 2).

The test for imprinting proposed here follows the standard normal distribution under the null model of no linkage. However, its distribution becomes wider than the standard normal under the null model of linkage but no imprinting. This results in an inflated type 1 error rate if one uses the critical values from the standard normal distribution, which constitutes a limitation of the proposed method. Although the critical values from the standard normal distribution can still be used for the initial analysis, the empirical exact P-values should be obtained by permutation.

We were particularly interested to evaluate the performance of the test for imprinting when one parent suffered from more missing genotype data than the other parent, because there might have been an inflation of the type I error for this test in the presence of linkage. We did not observe such behavior of the test for imprinting: for the model with missing parental genotypes, the null distribution resulting from the permutation (again, this is the only empirical distribution which maintains linkage, but not imprinting, that can be obtained when analyzing real data) was very similar to such distribution obtained under the scenario of complete parental genotype information (Table 3).

This method can be applied to data from ascertained families, since the trait data are conditioned upon. Although ascertainment can cause a deviation of the trait distribution from normality, we in our simulation setting did not observe a strong effect on the type 1 error for any of the linkage statistics. In contrast, when applying variance components methods to non-randomly selected samples, an ascertainment correction is generally required in order to obtain unbiased parameter estimates (Amos and de Andrade 2001). The method that we have been studying here is applicable to non-normally distributed trait observations, since they are conditioned upon, given a highly polymorphic marker (Sham et al. 2002). However, to reduce the impact of extreme observations that may become influential, one could apply a transformation. In a separate study of variance components methods we have found that first ranking the data and then applying an inverse normal transformation to the ranks preserved power and did not lead to an excess of false positive results (Peng et al. 2007).

We compared the performance of this method to the extensions of variance components and H–E methods (Hanson et al. 2001) that incorporate parent-of-origin effects in the linkage analysis of quantitative traits. Hanson et al. (2001) simulated data for the sample of 263 families with 956 sibs overall (2 to 11 children per family), allowing for 20% of maternal and 48% of paternal genotypes to be missing, for a range of the major gene variance, under the null model of no linkage, and under the alternative models of linkage with and without imprinting. For the test of imprinting, assuming the major gene variance of 0.1, they observed the type 1 error of 0.034 and the power of 0.196 for the variance components method, and the type 1 error of 0.072 and the power of 0.212 for the H–E method. This may be compared to our simulation setting presented in Table 3 (block 4), where, for a slightly higher major gene variance of 0.12 and a considerably smaller sample size we observe the nominal type 1 error of 0.05 and the similar power of 0.186 for the test of imprinting using the permutation procedure. Assuming the major gene variance of 0.3, the authors observed the type 1 error of 0.054 and the power of 0.904 for the variance components method and the type 1 error of 0.102 and power of 0.855 for the H–Elston method. Our comparable setting is shown in Table 3, block 5, corresponding to a lower major gene variance of 0.25 and 200 families with 4 children in each, for which the power was 0.904, with the nominal type 1 error of 0.05 (as assessed by permutation). These comparisons indicate that the method described here tends to perform identically or better than existing methods for evaluation of parent-of-origin effect in linkage analysis of quantitative traits. Unfortunately, the authors (Hanson et al. 2001) did not present results for selected samples; thus, we cannot evaluate whether there was an inflation of type 1 error and what the power would be for imprinting detection for the variance components method or for the H–E method, whereas the presented method is extremely powerful in that setting while maintaining the type 1 error rate.

By this method, we detected a parent-of-origin effect consistent with maternal imprinting at several markers on chromosome 7 for several body-size and obesity related traits. The proportions of missing genotype data were low and should not have led to a significant power loss based on the simulation study, especially with the use of multipoint IBDs. The detected parent-of-origin effect is not likely to be an artifact due to sex differences in the frequency and distribution of crossover exchanges. For that to happen, the sex difference in recombination must reach five-to tenfold, as assessed by simulation studies (Hanson et al. 2001) and by analytical approach (Shete and Amos 2002). However, in this region of chromosome 7 the sex difference in recombination never exceeds fourfold.

Four out of six traits, namely BMI, BIA, leptin, and waist circumference, showed a very similar pattern of linkage and of parent-of-origin effect, suggesting that they are controlled by the same gene or genes located in this region. On the other hand, fasting glucose did not show any linkage, while waist-to-hip ratio only showed maternal linkage at a single locus (D7S523), with a significant test for imprinting (Table 5). However, as the signal for total linkage was very weak, this result cannot be viewed as convincing.

Interestingly, two clusters of imprinted genes were identified in this region, one containing *PEG*10 (7q21–31) and the other—it PEG1/MEST (MIM 601029), *MESTIT1* (MIM 607794), and *COPG2* (MIM 604355) (7q32) (Kobayashi et al. 1997; Nakabayashi et al. 2002; Okita et al. 2003; Yamasaki et al. 2000). Most of markers that show paternal linkage in our study are located between these two identified clusters. In particular, Kobayashi et al. (1997) demonstrated that the *PEG1/MEST* gene located near D7S649 (one of the markers used in this analysis) is an imprinted gene expressed from a paternal allele. In our study, we observed a significant total and paternal (but not maternal) linkage at this marker for BIA, but the pattern of linkage was more consistent with total linkage than with imprinting, although there might be partial imprinting. The *PEG1/MEST* gene might be relevant to the body size-related traits. In experiments with mice it was shown that when the null allele of the *PEG1/MEST* gene is paternally transmitted, the offspring exhibit severe intrauterine growth retardation (Ferguson-Smith et al. 1991).

A possible limitation of this study is that both definitions of $\mathbf{\Sigma}_{\hat{\Pi}}$ have some problems. The calculation of the imputed matrix, under certain circumstances, can result in estimates of the covariance greater than that of the variance. The sample based matrix, on the other hand, does not allow for family-specific marker information, which is not appropriate with heterogeneous sample, but is a good estimate when families in the sample are similar in contributing marker informativeness.

A possible future extension that could further improve power would be an extension of the method to non-nuclear families while allowing for the parent-of-origin effect. Shete et al. (2003) showed how to extend variance components and H–E tests for extended families and documented a substantial increase in power when extended pedigrees are studied.

Acknowledgments The work was supported by the National Institutes of Health grant ES09912. We thank Drs. Pak C. Sham and Goncalo R. Abecasis for their help regarding the regression model and its implementation in Merlin-Regress, and Mrs. Wei Chen for her help with the simulation setup.

Appendix

Electronic-Database Information

The URLs for data presented herein are as follows: American Diabetes Association, http://www.diabetes.org/about-diabetes.jspOnline Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/. The software developed in this study is available from the author free of charge upon request.

References

Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535–543

Amos CI, de Andrade M (2001) Genetic linkage methods for quantitative traits. Stat Methods Med Res 10:3–25

Bartolomei MS, Tilghman SM (1997) Genomic imprinting in mammals. Annu Rev Genet 31:493–525

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85–97

Chumlea WC, Guo SS, Kuczmarski RJ, Flegal KM, Johnson CL, Heymsfield SB, Lukaski HC, Friedl K, Hubbard VS (2002) Body composition estimates from NHANES III bioelectrical impedance data. Int J Obes Relat Metab Disord 26:1596–1609

Ferguson-Smith AC, Cattanach BM, Barton SC, Beechey CV, Surani MA (1991) Embryological and molecular investigations of parental imprinting on mouse chromosome 7. Nature 351:667–670

Freeman MS, Mansfield MW, Barrett JH, Grant PJ (2002) Heritability of features of the insulin resistance syndrome in a community-based study of healthy families. Diabet Med 19:994–999

Fulker DW, Cherny SS (1996) An improved multipoint sib-pair analysis of quantitative traits. Behav Genet 26:527–532

Gorlova OY, Amos CI, Wang NW, Shete S, Turner ST, Boerwinkle E (2003) Genetic linkage and imprinting effects on body mass index in children and young adults. Eur J Hum Genet 11:425–432

Greally JM (2002) Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. Proc Natl Acad Sci USA 99:327–332

Hanson RL, Kobes S, Lindsay RS, Knowler WC (2001) Assessment of parent-of-origin effects in linkage analysis of quantitative traits. Am J Hum Genet 68:951–962

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19

Iyengar S, Calafell F, Kidd KK (1997) Detection of major genes underlying several quantitative traits associated with a common disease using different ascertainment schemes. Genet Epidemiol 14:809–814

Knapp M, Strauch K (2004) Affected-sib-pair test for linkage based on constraints for identical-by-descent distributions corresponding to disease models with imprinting. Genet Epidemiol 26:273–285

Kobayashi S, Kohda T, Miyoshi N, Kuroiwa Y, Aisaka K, Tsutsumi O, Kaneko-Ishino T, Ishino F (1997) Human PEG1/MEST, an imprinted gene on chromosome 7. Hum Mol Genet 6:781–786

Langenberg C, Hardy R, Kuh D, Brunner E, Wadsworth M (2003) Central and total obesity in middle aged men and women in relation to lifetime socioeconomic status: evidence from a national birth cohort. J Epidemiol Community Health 57:816–822

Li WD, Li D, Wang S, Zhang S, Zhao H, Price RA (2003) Linkage and linkage disequilibrium mapping of genes influencing human obesity in chromosome region 7q22.1–7q35. Diabetes 52:1557–1561

Luke A, Guo X, Adeyemo AA, Wilks R, Forrester T, Lowe W Jr, Comuzzie AG, Martin LJ, Zhu X, Rotimi CN, Cooper RS (2001) Heritability of obesity-related traits among Nigerians, Jamaicans and US black people. Int J Obes Relat Metab Disord 25:1034–1041

Nakabayashi K, Bentley L, Hitchins MP, Mitsuya K, Meguro M, Minagawa S, Bamforth JS, Stanier P, Preece M, Weksberg R, Oshimura M, Moore GE, Scherer SW (2002) Identification and characterization of an imprinted antisense RNA (MESTIT1) in the human MEST locus on chromosome 7q32. Hum Mol Genet 11:1743–1756

Okita C, Meguro M, Hoshiya H, Haruta M, Sakamoto YK, Oshimura M (2003) A new imprinted cluster on the human chromosome 7q21-q31, identified by human–mouse monochromosomal hybrids. Genomics 81:556–559

Peng B, Yu R, DeHoff K, Amos C (2007) Normalizing a large number of quantitative traits using empirical normal quantile transformation. BMC Genetics (in press)

Ruhl CE, Everhart JE (2001) Leptin concentrations in the United States: relations with demographic and anthropometric measures. Am J Clin Nutr 74:295–301

Sham PC, Purcell S, Cherny SS, Abecasis GR (2002) Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet 71:238–253

Shete S, Amos CI (2002) Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. Am J Hum Genet 70:751–757

Shete S, Zhou X, Amos CI (2003) Genomic imprinting and linkage test for quantitative-trait Loci in extended pedigrees. Am J Hum Genet 73:933–938

Smith FM, Garfield AS, Ward A (2006) Regulation of growth and metabolism by imprinted genes. Cytogenet Genome Res 113:279–291

Strauch K, Fimmers R, Kurz T, Deichmann KA, Wienker TF, Baur MP (2000) Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. Am J Hum Genet 66:1945–1957

Terwilliger JD, Speer M, Ott J (1993) Chromosome-based method for rapid computer simulation in human genetic linkage analysis. Genet Epidemiol 10:217–224

Tilghman SM (1999) The sins of the fathers and mothers: genomic imprinting in mammalian development. Cell 96:185–193

Vincent Q, Alcais A, Alter A, Schurr E, Abel L (2006) Quantifying genomic imprinting in the presence of linkage. Biometrics 62:1071–1080

Wilkins JF, Haig D (2003) What good is genomic imprinting: the function of parent-specific gene expression. Nat Rev Genet 4:359–368

Wu CC, Shete S, Amos CI (2005) Linkage analysis of affected sib pairs allowing for parent-of-origin effects. Ann Hum Genet 69:113–126

Yamasaki K, Hayashida S, Miura K, Masuzaki H, Ishimaru T, Niikawa N, Kishino T (2000) The novel gene, gamma2-COP (COPG2), in the 7q32 imprinted domain escapes genomic imprinting. Genomics 68:330–335