Extreme Rank Selections for Linkage Analysis of Quantitative Trait Loci Using Selected Sib-Pairs

G. Zheng1,*, K. Ghosh2, Z. Chen3 and Z. Li2,4
1Office of Biostatistics Research, National Heart, Lung and Blood Institute, Bethesda, MD, USA
2Department of Statistics, George Washington University, Washington, DC, USA
3Department of Statistics and Applied Probability, National University of Singapore, Republic of Singapore
4Biostatistics Branch, DCEG, National Cancer Institute, Bethesda, MD, USA

Summary
It is well known that linkage analysis using simple random sib-pairs has relatively low power for detecting quantitative trait loci with small genetic effects. The power can be substantially increased by using samples selected based on their trait values. Usually, samples that are obtained by truncation selection consist of random samples from a truncated trait distribution. In this article we propose an alternative method using extreme ranks for linkage analysis with selected sib-pairs. This approach approximates the truncation selection. With similar screening sizes and the same sample size of selected sib-pairs, the extreme rank selection and truncation method have similar power performance, both of which are substantially more powerful than when using random sib-pairs. Simulation results on the comparison of powers between the truncation selection and the extreme rank selection and/or random selection for linkage analysis are reported.

Keywords: Extreme rank selection (ERS), Linkage, QTL, Ranked set sampling (RSS), Truncation selection (TS)

Introduction
In genetic linkage analysis for locating quantitative trait loci (QTL) using sib-pairs, all individuals in a simple random sample are usually genotyped (e.g., Haseman & Elston, 1972; Lange et al. 1976; Blackwelder & Elston, 1982). However, unless there is a single locus responsible for 50% or more of the genetic variation, the power of linkage analysis using random samples is relatively low. To detect a locus accountable for 10% to 20% genetic variation thousands of sib-pairs would need to be genotyped. When genotyping is much more time-consuming and costly than measuring the traits one cost-effective method to increase the power is to use selected sib-pairs based on their continuous trait values. Lander & Botstein (1989) studied selective sampling methods for plant genetics. In human genetics, Carey & Williamson (1991) demonstrated that the power of linkage analyses using selected sib-pairs are substantially higher when compared to the random sampling approach of Haseman & Elston (1972). The further development of statistical methods and applications in linkage analysis using selected sib-pairs have been extensively reported in the literature. For review and some recent research, see Forest & Feingold (2000), Freidlin et al. (2003), Szatkiewicz et al. (2003), Szatkiewicz & Feingold (2004, 2005), Wang (2005) and Chen et al. (2005).

The commonly used approach of selective sampling is screening and truncation; that is, individuals are randomly sampled and screened but only those who have trait values falling in certain extreme regions are selected for genotyping. This approach is also referred to as selective genotyping. In sib-pair analysis Carey & Williamson (1991) considered screening sib-pairs according to the trait value of only one sib (say the first one by some
natural order) and selecting those sib-pairs whose first sib trait value fell in the top 5% of the population distribution. Risch & Zhang (1995) considered screening according to the trait values of both sibs in a sib-pair and selecting discordant or concordant sib-pairs. Whenever the cost of screening is much less than the cost of genotyping, the approach of selective genotyping will be useful and cost-effective.

Prior research on selective genotyping has focused exclusively on truncation selection (TS). To apply the selective genotyping by TS, the cut-off points, e.g. the top 5% percentile of the trait distribution, must be known a priori, or otherwise they must be estimated using pilot data. For example, in Xu et al. (1999) about 40,000 individuals were first screened to estimate the percentiles of blood pressure and then more than 160,000 additional individuals were screened to select the extreme discordant sib-pairs using the estimated percentiles.

Recently Chen et al. (2005) applied an alternative approach, referred to as extreme rank selection (ERS), to linkage disequilibrium mapping of quantitative trait loci. ERS was shown to approximate TS as they both approach, referred to as extreme rank selection (ERS), to linkage disequilibrium mapping of quantitative trait loci. ERS was shown to approximate TS as they both

$\mu = a + (p - q) + 2pqd$ and $\mu - a$, corresponding to the genotypes $MM$, $MN$ and $NN$, with probabilities $p^2$, $2pq$ and $q^2$, respectively. The mean and variance of $g$ are given, respectively, by $\mu_g = \mu + a(p - q) + 2pqd$ and $\sigma^2_g = \sigma^2_\mu + \sigma^2_d$ where $\sigma^2_\mu = 2pq[a - d(p - q)]^2$ and $\sigma^2_d = 4p^2q^2d^2$. The genetic variance $\sigma^2_d$ is decomposed to $\sigma^2_\mu$ and $\sigma^2_d$, which are referred to as the additive and dominant components, the BMI can be easily calculated and used to rank and select individuals rather than using the DEXA measurements. Once the samples are selected based on the ranks of BMI, the DEXA measurements of fat can be obtained only from these selected samples. A similar example is to evaluate the effect of selective phenotypes in response to gene and low fat diet interaction in overweight and obesity interventions using young twins. The ERS can be applied by ranking the BMI of the twins. Formal measures of obesity, e.g. the DEXA measurement of fat, can be obtained only on those selected twins. Another example is in genetic studies of personality traits using selected samples. The personality trait values are difficult to measure. However, when applying ERS some questionnaires can be used to rank and select individuals and the personality trait values are only obtained from the individuals selected by ERS. Compared to TS ERS is more cost-effective for traits that are difficult and/or costly to measure.

In this article we apply ERS to linkage analysis using sib-pairs, and compare the power performance of ERS, TS and/or random samples of sib-pairs by simulation studies. For a sib-pair two selections using ERS are studied. The first selection is based on the trait of the first sib using the regression model of Carey & Williamson (1991). In the second selection, the traits of both sibs are used and the robust statistic of Szatkiewicz & Feingold (2004) is applied.

**Background**

**Linkage Analysis Using Random Sib-Pairs**

Let $X$ be the quantitative trait of concern. Assume that a putative QTL has two alleles $M$ and $N$ with allele frequencies $p$ and $q = 1 - p$, respectively. Let $g$ be the genotypic value of the QTL. Under the assumption of Hardy–Weinberg equilibrium (HWE), $g$ is a random variable taking values $\mu + a$, $\mu + d$ and $\mu - a$, corresponding to the genotypes $MM$, $MN$ and $NN$, with probabilities $p^2$, $2pq$ and $q^2$, respectively. The mean and variance of $g$ are given, respectively, by $\mu_g = \mu + a(p - q) + 2pqd$ and $\sigma^2_g = \sigma^2_\mu + \sigma^2_d$ where $\sigma^2_\mu = 2pq[a - d(p - q)]^2$ and $\sigma^2_d = 4p^2q^2d^2$. The genetic variance $\sigma^2_d$ is decomposed to $\sigma^2_\mu$ and $\sigma^2_d$, which are referred to as the additive and dominant components,
respectively. The trait value \( X \) of an individual can be expressed as \( X = g + \epsilon \), where \( \epsilon \) is a random variable with zero mean and variance \( \sigma^2_\epsilon \), independent of \( g \), accounting for environmental and other unidentifiable effects.

Based on the above genetic model Haseman & Elston (1972) developed an important regression model for linkage analysis using sib-pairs. Let \( Y = (X_1 - X_2)^2 \) where \( X_1 \) and \( X_2 \) are the trait values of the sibs in a sib-pair. Let \( \pi_m \) denote the proportion of genes identical by descent (IBD) shared by the sibs at a marker locus and \( \hat{\pi}_m \) its estimate. Haseman & Elston (1972) proved that, if \( d = 0 \),

\[
E(Y|\hat{\pi}_m) = \alpha + \beta \hat{\pi}_m,
\]  

(1)

where \( \alpha \) and \( \beta \) are parameters. When \( d \neq 0 \) they showed that (1) holds asymptotically. The coefficient \( \beta \) is specifically given by \( \beta = -2(1 - \theta)^2 \sigma^2_\epsilon \) where \( \theta \) is the recombination fraction between the marker and the putative QTL. Hence, to test the existence of a QTL linked to the marker \( (\theta < 1/2) \), the hypotheses can be formulated as \( H_0: \beta = 0 \) against \( H_1: \beta < 0 \). The linkage analysis can then be carried out through a simple linear regression model based on (1).

**TS Using One Sib and Carey-Williamson’s Regression Models**

To increase the power of linkage analysis using sib-pairs Carey & Williamson (1991) proposed a TS approach. Let \( t \) be the value such that \( \text{Pr}(X > t) = s \), where \( s \) is pre-specified, say \( s = 0.05 \). In TS a sib-pair is selected and genotyped if the trait value of the first sib (determined by some natural order) exceeds \( t \). The truncation point \( t \) may be known or may have to be estimated after screening a large number of individuals. The selection is independent of the trait value of the other sib, their genotypes at the marker locus, and the IBD sharing proportion \( \pi_m \).

To compare the power under TS with that under simple random sampling (SRS), Carey & Williamson (1991) considered two different regression models based on the expression of \( E(X_2|X_1, \pi_m) \) under TS and SRS, where \( X_1 \) and \( X_2 \) denote the trait values of the first and second sib in the same sib-pair. To distinguish between TS and SRS, we denote paired trait values under SRS by \((X_1, X_2)\) and their counterpart under TS by \((X^*_1, X^*_2)\). Carey & Williamson (1991) found that, under SRS,

\[
E(X_2|X_{i1}, \pi_{mi}) = \beta_1 + \beta_3 \delta_i + \beta_4 \gamma_i + \beta_5 X_{i1} + \beta_6 \delta_i X_{i1} + \beta_7 \gamma_i X_{i1},
\]

\[i = 1, \ldots, n,\]

(2)

and, under TS,

\[
E(X^*_2|X^*_{i1}, \pi_{mi}) = \beta_1^* + \beta_3^* X^*_{i1} + \beta_4^* \delta_i + \beta_7^* \gamma_i,
\]

\[i = 1, \ldots, n,\]

(3)

where \( \delta_i \) and \( \gamma_i \) are defined as follows: \( \delta_i = -1/2, 0, 1/2 \) and \( \gamma_i = 1/4, -1/4, 1/4 \) according as \( \pi_{mi} = 0, 1/2, 1 \). They proved that, under model (2), the least squares estimator (LSE) of \( \beta_3 \) has an asymptotic expression \( \hat{\beta}_3 = (1 - 2\theta)^2 \sigma^2_\epsilon / (\sigma^2_\epsilon + \sigma^2_\gamma) \), and that, under model (3), the LSE of \( \beta_3^* \) has an asymptotic expression \( \hat{\beta}_3^* = (1 - 2\theta)^2 \sum_{l=0}^{2} f_l g_l \), where \( f_l \) is the conditional probability given \( X > t \) that the genotype of the putative QTL has \( l \) \( N \) alleles and \( g_l \) is the deviation of the corresponding genotypic value from \( \mu_{gc} \). Note that if the putative QTL is in fact not a QTL then \( \sum_{l=0}^{2} f_l g_l = 0 \), since the conditional probabilities in this case are the same as the unconditional probabilities, that is \( f_0 = p^2, f_1 = 2pq \) and \( f_2 = q^2 \). Therefore, to test the existence of a QTL linked to the marker, the hypotheses can be formulated as \( H_0: \beta_3 = 0 \) against \( H_1: \beta_3 > 0 \) under model (2). Note that for model (3) the alternative hypothesis was not stated in Carey & Williamson (1991). From \( \hat{\beta}_3^* = (1 - 2\theta)^2 \sum_{l=0}^{2} f_l g_l \), it is not straightforward whether the one-side or two-sided alternative should be used. Prior simulation studies under various practical situations indicate that under the alternative \( \beta_3^* > 0 \) for model (3). Thus the one-sided alternative will be used for model (3) in our simulation studies. In their simulation studies Carey & Williamson (1991) demonstrated that the power in most situations using TS is dramatically increased compared to using SRS.

**TS Using Both Sibs and the Robust Statistic**

To select a discordant sib-pair (DSP), denoted as \((X_1, X_2)\), by TS a sib-pair is screened such that \( \text{Pr}(X_1 > t_1) = s \) and \( \text{Pr}(X_2 < t_2) = s \) are satisfied, where \( s \) is a small percentile. Thus sib-pairs are screened until a pre-specified number of DSPs are obtained. The
screening probability of DSP is \( \Pr (X_1 > t_1, X_2 < t_2) \) which is usually less than \( \xi^2 \). The greater the correlation between two sibs’ traits the smaller the screening probability. For example, assuming \((X_1, X_2)\) follow a bivariate normal distribution with correlation \( \rho \), the screening probabilities for a moderate DSP (top/bottom 33%) and an extreme DSP (top/bottom 5%) are 0.098 (\( \rho = 0.1 \)), 0.0722 (\( \rho = 0.3 \)) and 0.00158 (\( \rho = 0.1 \)), 0.00046 (\( \rho = 0.3 \)), respectively.

Let the trait of \( n \) DSPs obtained by TS be \((X_{i1}, X_{i2}), i = 1, \ldots, n\), and \( \mu \) and \( \sigma \) be the mean and standard deviation of the population trait. Suppose \( \pi_i \) is the estimated proportion of alleles shared IBD at the marker locus in the \( i \)th sib-pair. The robust statistic proposed by Szatkiewicz & Feingold (2004) can be written as

\[
Z = \frac{- \sum_{i=1}^{n} (X_{i1} - X_{i2})^2 (\pi_i - \frac{1}{2})}{\left\{ \frac{1}{n} \sum_{i=1}^{n} (X_{i1} - X_{i2})^4 \left[ \sum_{i=1}^{n} (\pi_i - \frac{1}{2})^2 \right] \right\}^{1/2}},
\]

which has an asymptotic standard normal distribution under the null. Linkage is detected for a strong positive value of \( Z \). As one does not need to estimate \( \mu \) and \( \sigma \) the test statistic (4) is easy to apply.

A Brief Review on Ranked Set Sampling

The ERS is motivated from ranked set sampling (RSS) which is briefly reviewed here. The summary of research on RSS can be found in Chen et al. (2004). RSS applies if the measurement of the variable of interest is costly and/or time consuming, but sampling units can be easily obtained and ranked with negligible cost by a certain means without measuring the variable of interest. The procedure of RSS goes as follows. To obtain the first observation on the variable of interest \( k \) units are randomly drawn, and their latent values on the variable of interest, denoted by \( Y_1, \ldots, Y_k \), are ranked in an ascending order without actual measurement as \( Y_{[1]} \leq \cdots \leq Y_{[k]} \). Let \( r \) be a pre-specified rank, say \( r = 1 \), then \( Y_{[1]} \) is actually measured. To get the second observation on the variable of interest the procedure is repeated; if the pre-specified rank is \( r = 2 \) this time, \( Y_{[2]} \) is then actually measured. The scheme continues this way and each time a rank is pre-specified. If the ranks from \( r = 1 \) to \( r = k \) are specified an equal number of times the scheme is called a balanced RSS. Otherwise it is called an unbalanced RSS. In particular if only the extreme rank \( r = k \) (or \( r = 1 \)) is specified in the whole process the scheme is called an extreme RSS. The ranking in RSS can be done by various means depending on the real situations. It can be done by visual inspection, auxiliary information or measurement on easily obtainable closely correlated concomitant variables. The rankings need not be perfect. As long as the ranking is not completely arbitrary efficiency will be gained through the RSS procedure (see Chen et al. 2004, Chapter 2).

The procedure of RSS is analogous to that of stratified sampling. Instead of pre-stratifying the population the RSS post-stratifies the population based on the random ranks of the individuals. With set size \( k \) the RSS virtually stratifies the population into \( k \) strata. In the case of extreme RSS with set size \( k \), the distribution of the sampled extreme rank samples mimics the upper \( 1/k \) proportion of the original population. The connection between RSS and stratified sampling makes it clear that the extreme RSS can be applied to selective genotyping in addition to TS. In principle, an upper 10% TS in Carey & Williamson (1991) can be replaced by an extreme RSS with set size \( k = 10 \). The trait can be considered as a concomitant variable which is used for ranking in RSS.

Linkage Analysis Using Extreme Rank Selections

ERS for Carey-Williamson’s Regression Models

By using ERS similar to Chen et al. (2005) the selection of sib-pairs is carried out as follows. To ascertain one sib-pair for genotyping, first \( k \) randomly chosen sib-pairs are screened by measuring their trait values \((X_{ij}, X_{j2}), j = 1, \ldots, k\), where \( X_{ij} \) and \( X_{j2} \) denote the trait values of the first and second sib, respectively, in the \( j \)th sib-pair. Here the order of the sibs in a sib-pair is determined by certain natural ordering independent of their trait values. The trait values of the first sibs in these \( k \) sib-pairs are ranked in ascending order as \( X_{1(1)} \leq X_{1(2)} \leq \cdots \leq X_{1(k)} \). Finally, the sib-pair whose first sib has trait value \( X_{1(k)} \) is selected for genotyping. If \( n \) sib-pairs are to be ascertained the above procedure is repeated \( n \) times. For
the convenience of notation we denote the trait values of the selected sib-pairs by \((X_{1i}, X_{2i}), i = 1, \ldots, n\).

The ERS described above with set size \(k\) is analogous to an upper \((100/k)\%\) TS procedure. The screening size of ERS is a fixed constant \(nk\). The screening size of TS is random, with mean value \(n/Pr(X_{1i} > t_k) = nk\), since \(t_k\) is chosen such that \(Pr(X_{1i} > t_k) = 1/k\). Therefore, on average these two selection procedures have similar screening sizes. We can expect these two selection procedures to have comparable powers in linkage analysis, which will be demonstrated later.

Similar Carey-Williamson’s regression models can be obtained for ERS. The detailed derivations of the following results are similar to those in Carey & Williamson (1991) and are omitted. Like TS we assume the following for ERS:

\[
E(\tilde{X}_2i | X_{1i}, \pi_{mi}) = \hat{\beta}_1 + \hat{\beta}_2 X_{1i} + \hat{\beta}_3 \delta_i + \hat{\beta}_4 \gamma_i. \tag{5}
\]

where \(\pi_{mi}\) is the proportion of alleles IBD shared by the \(i\)th selected sib-pair, and \(\delta_i\) and \(\gamma_i\) are defined through \(\pi_{mi}\) the same as \(\delta_i\) and \(\gamma_i\), through \(\pi_{mi}\). Note that model (5) is analogous to model (3) in Carey & Williamson (1991). We show that the properties of model (3) for TS hold in model (5) for ERS.

In matrix notation, (5) is written as \(E(\tilde{Y}) = \tilde{X}\hat{b}\), where \(\tilde{X} = (\tilde{X}_1, \cdots, \tilde{X}_n)^T\), \(\tilde{Y}\) is the design matrix, \(\hat{b} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)^T\), and the expectation \(E(\tilde{Y})\) is conditional on the same variables as in (5). As \(n\) is large, the LSE of \(\hat{b}\) can be asymptotically approximated by

\[
\hat{b} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y} \approx \left[\frac{1}{n}E(\tilde{X}^T\tilde{X})\right]^{-1}\frac{1}{n}E(\tilde{X}^T\tilde{Y}).
\]

Note that \((\tilde{X}_{1i}, \tilde{X}_{2i}), i = 1, \ldots, n\), are independent and identically distributed. Denote \(\bar{\mu}_i = E(\tilde{X}_{1i}), \bar{\sigma}_i^2 = \text{Var}(\tilde{X}_{1i})\), \(m_{12} = E(\tilde{X}_{1i}\tilde{X}_{2i})\) and \(\bar{\mu}_{2i} = E(\tilde{X}_{2i} | \pi_{mi} = l/2), l = 0, 1, 2\). We obtain

\[
\left[\frac{1}{n}E(\tilde{X}^T\tilde{X})\right]^{-1} = \begin{bmatrix}
(\bar{\mu}_1^2 + \bar{\sigma}_1^2)/\bar{\sigma}_1^2 - \bar{\mu}_1/\bar{\sigma}_1^2 & 0 & 0 & 0 \\
-\bar{\mu}_1/\bar{\sigma}_1^2 & 1/\bar{\sigma}_1^2 & 0 & 0 \\
0 & 0 & 8 & 0 \\
0 & 0 & 0 & 16
\end{bmatrix}
\]

and

\[
\frac{1}{n}E(\tilde{X}^T\tilde{Y}) = \begin{bmatrix}
\hat{\mu}_1 \\
\bar{m}_{12} \\
(\bar{\mu}_{22} - \bar{\mu}_{20})/8 \\
(\bar{\mu}_{22} + \bar{\mu}_{20} - 2\bar{\mu}_{21})/16
\end{bmatrix}.
\]

Thus, the asymptotic approximation of the LSE of \(\hat{\beta}_3\) can be written as

\[
\hat{\beta}_3 \approx \hat{\mu}_{22} - \hat{\mu}_{20}.
\]

Denote the expected frequencies of genotypes MM, MN and NN in the ERS selected sample \((\tilde{X}_{1i}, \tilde{X}_{2i})\) by \((\tilde{f}_{11}, \tilde{f}_{22}), l = 0, 1, 2\), where \(l\) indicates the number of \(N\) alleles in the genotype, e.g., \(\tilde{f}_{10} = Pr(MM|\text{selected sib 1})\). It can be shown that

\[
\hat{\mu}_{22} - \hat{\mu}_{20} = (1 - 2\theta)^2 \sum_{l=0}^{2} \tilde{f}_{l1} \tilde{f}_{l2} g_l, \tag{7}
\]

where \(g_l\) is as defined before. Since the selection in ERS is independent of the trait of the second sib-pair \((X_2)\), we have \(\sum_{l=0}^{2} \tilde{f}_{l1} g_l = 0\). Further, as ERS approximates TS, \(\tilde{f}_{l1} \approx f_{l1}\) for \(l = 0, 1, 2\). Thus, \(\sum_{l=0}^{2} (\tilde{f}_{l1} - f_{l1}) g_l = \sum_{l=0}^{2} f_{l1} g_l \approx \sum_{l=0}^{2} f_{l1} g_l\). Thus, \(\hat{\beta}_3 \approx \hat{\beta}_3^* = (1 - 2\theta)^2 \sum_{l=0}^{2} f_{l1} g_l\), where \(\hat{\beta}_3^*\) was obtained by Carey & Williamson (1991). Thus, the QTL effect can be tested by testing \(H_0 : \hat{\beta}_3 = 0\) against \(H_1 : \hat{\beta}_3 > 0\). The one-sided alternative is used, similar to the TS approach used before.

### ERS for Discordant Sib-Pairs

To obtain DSPs by ERS a two-stage procedure is required. Denote the trait of a sib-pair by \((X_1, X_2)\). First, identify \(k^2\) sib-pairs and divide them into \(k\) groups each of size \(k\), denoted as \((X_{1ij}, X_{2ij})\) for the \(j\)th sib-pair in the \(i\)th group, \(i, j = 1, \ldots, k\). In the \(i\)th group rank all the first sibs, \(X_1(1), \ldots, X_1(k)\), and obtain \(X_{1(i)} \leq \cdots \leq X_{1(k)}\). The corresponding second sibs are denoted by \(X_{2(i)}(1), \ldots, X_{2(k)}(k)\). We retain both pairs \((X_{1(i)}, X_{2(i)})\) and \((X_{1(i)}, X_{2(i)})\) for \(i = 1, \ldots, k\). Then, in the second stage we rank all \(X_{2(i)}(1)\) as \(X_{2(i)}(1) \leq \cdots \leq X_{2(k)}(1)\) and their corresponding \(X_{1(i)}(1), \ldots, X_{1(k)}(1)\), are denoted by \(X_{1(i)}(1), \ldots, X_{1(i)}(1)\). We retain the pair \((X_{1(i)}(1), X_{2(i)}(1))\) and it is a DSP. For the other set of retained pairs \((X_{1(i)}, X_{2(i)})\), \(i = 1, \ldots, k\), a similar procedure is applied and the pair \((X_{1(i)}(1), X_{2(i)}(1))\) will be
selected as a DSP. Thus we retain two DSPs from \( k^2 \) random sib pairs and their genotype data are obtained.

The ERS is repeated \( n \) times to obtain \( 2n \) DSPs. Then the robust statistic (4) can be applied. An example of applying ERS for DSPs is given in Table 1 when \( k = 3 \). This dataset was generated from a bivariate normal distribution with zero means, unit variances, and correlation \( \rho = 0.2 \). Applying ERS we selected two DSPs (1.693, 0.344) and (0.070, 0.106).

### Results: Power Comparison

Simulation studies were conducted to compare the power performance of different selective genotyping. The first simulation was to compare the powers for detecting QTL using SRS, TS, and ERS for regression models of Carey & Williamson (1991). The second simulation was to compare powers of TS and ERS using DSPs. Both simulation procedures were similar to that of Carey & Williamson (1991). For a QTL with two alleles \( M \) and \( N \), there are nine possible unordered genotype pairs for a sib-pair (Table 2). Denote the probability of the \( i \)th sib-pair sharing \( j \) alleles IBD at the QTL conditional on the genotype pair by \( f_{ij} \), \( j = 0, 1, 2 \) and \( i = 1, \ldots, 9 \). The expressions of \( p_i \) and \( f_{ij} \), which were given in Haseman & Elston (1972, Table III), are summarized in Table 2.

First we describe the procedure for simulating the data for a SRS of sib-pairs. We fixed the sample size \( n \) and the allele frequency \( p \). The counts of nine different genotypes of \( n \) sib-pairs were generated using the multinomial distribution \( (n_1, \ldots, n_9) \sim Mul(n; p_1, \ldots, p_9) \). For each sib-pair with the \( i \)th genotype \( (i = 1, \ldots, 9) \), the IBD status at the QTL was generated using Haseman & Elston (1972, Table III), are summarized in Table 2.
values \((X_1, X_2)\) were generated using the bivariate normal distribution with mean \((d_1, d_2)\) and covariance matrix given by \(\sigma^2_i \omega^i \) with \(\sigma^2 = 1\). For each sib-pair the IBD status at the marker locus \((\pi_m)\) was generated for a given recombination fraction \(\theta\), using the joint distributions of IBDs at the QTL and marker loci, which were given in Haseman & Elston (1972, Table V). Finally, \(\delta\) and \(\gamma\) were computed for each sib-pair from \(\pi_m\).

To generate data with TS or ERS, SRS were generated first according to the above procedure. Then TS and ERS were applied to obtain the selected samples. The pre-specified threshold value for TS was found by using simple random sib-pairs simulated independently.

To compare the powers of the three sampling approaches for detecting linkage under various genetic models, we chose minor allele frequencies \(p = 0.1\) and \(0.3\), recombination fraction \(\theta = 0.05\) and \(0.10\), heritability \(H = 0.05, 0.20, 0.35\) and \(0.50\). Sample size \(n = 240\) and correlation \(\omega = 0.2\) were fixed in all simulation studies. The significance level used in the simulation was \(\alpha = 0.05\). For TS selection was based on the top \(10\%\) portion of the trait distribution, while for the corresponding ERS the set size was taken as \(k = 10\). The Carey and Williamson regression models were calculated for testing the null hypothesis of no linkage.

The empirical powers presented in Table 3 for recessive (REC), additive (ADD) and dominant (DOM) models are based on 1000 replications.

### Table 3 Linkage analysis of Carey and Williamson regression models (correlation \(\omega = 0.2\), \(n = 240\) sib-pairs, and 1000 replications)

<table>
<thead>
<tr>
<th>(p)</th>
<th>(\theta)</th>
<th>(H)</th>
<th>(REC) SRS</th>
<th>(REC) TS</th>
<th>(REC) ERS</th>
<th>(ADD) SRS</th>
<th>(ADD) TS</th>
<th>(ADD) ERS</th>
<th>(DOM) SRS</th>
<th>(DOM) TS</th>
<th>(DOM) ERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td>0.35</td>
<td>0.27</td>
<td>0.06</td>
<td>0.17</td>
<td>0.14</td>
<td>0.09</td>
<td>0.16</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.15</td>
<td>0.88</td>
<td>0.83</td>
<td>0.21</td>
<td>0.89</td>
<td>0.81</td>
<td>0.26</td>
<td>0.87</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>0.22</td>
<td>0.94</td>
<td>0.92</td>
<td>0.44</td>
<td>1.00</td>
<td>1.00</td>
<td>0.46</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.22</td>
<td>0.98</td>
<td>0.96</td>
<td>0.62</td>
<td>1.00</td>
<td>1.00</td>
<td>0.67</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>0.05</td>
<td>0.06</td>
<td>0.26</td>
<td>0.20</td>
<td>0.08</td>
<td>0.11</td>
<td>0.12</td>
<td>0.06</td>
<td>0.11</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.13</td>
<td>0.71</td>
<td>0.66</td>
<td>0.16</td>
<td>0.71</td>
<td>0.62</td>
<td>0.13</td>
<td>0.66</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>0.15</td>
<td>0.80</td>
<td>0.78</td>
<td>0.32</td>
<td>0.98</td>
<td>0.96</td>
<td>0.38</td>
<td>0.98</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.22</td>
<td>0.88</td>
<td>0.86</td>
<td>0.47</td>
<td>1.00</td>
<td>1.00</td>
<td>0.51</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.05</td>
<td>0.05</td>
<td>0.07</td>
<td>0.20</td>
<td>0.17</td>
<td>0.10</td>
<td>0.12</td>
<td>0.10</td>
<td>0.08</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.19</td>
<td>0.98</td>
<td>0.92</td>
<td>0.21</td>
<td>0.69</td>
<td>0.61</td>
<td>0.23</td>
<td>0.50</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>0.37</td>
<td>1.00</td>
<td>1.00</td>
<td>0.52</td>
<td>0.98</td>
<td>0.97</td>
<td>0.50</td>
<td>0.84</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.57</td>
<td>1.00</td>
<td>1.00</td>
<td>0.79</td>
<td>1.00</td>
<td>1.00</td>
<td>0.74</td>
<td>0.96</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>0.05</td>
<td>0.07</td>
<td>0.14</td>
<td>0.14</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td>0.08</td>
<td>0.09</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.13</td>
<td>0.89</td>
<td>0.77</td>
<td>0.20</td>
<td>0.53</td>
<td>0.45</td>
<td>0.16</td>
<td>0.35</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>0.30</td>
<td>1.00</td>
<td>0.99</td>
<td>0.36</td>
<td>0.90</td>
<td>0.85</td>
<td>0.43</td>
<td>0.65</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.39</td>
<td>1.00</td>
<td>1.00</td>
<td>0.59</td>
<td>1.00</td>
<td>0.99</td>
<td>0.62</td>
<td>0.83</td>
<td>0.81</td>
<td></td>
</tr>
</tbody>
</table>
the DEXA measurement of fat). Once a sib-pair is selected based on the ranks of BMI, their DEXA measurement of fat and genotypes are also obtained. Given the set size \( k \), the screening using ERS is independent of the correlation of two sibs while the screening using TS depends on the correlation. Thus it is difficult to compare the screening sizes of both approaches. In our simulation, in each replication we first simulated \( n = 240 \) DSPs by TS as described before and recorded its screening size. The 10% and 90% cutoff points were used. Then the set size \( k \) of ERS was calculated so that both approaches had the same screening size. Finally, \( n = 240 \) DSPs using ERS were obtained. The simulation of IBD data and the choices of parameter values in the second simulation were similar to the first simulation. The powers of linkage analysis using TS and ERS are plotted in Figures 1-3 for three genetic models (recessive, additive and dominant) based on 100 replications. In each Figure different combinations of allele frequency \( p \) and recombination fraction \( \theta \) were used. This shows that, with the same screening size and sample size, TS is slightly more powerful than ERS in most situations. As in the first simulation study, the small loss of power was due to the fact that ERS is an approximation of TS.

**Discussion**

Our study indicates that the extreme rank selection (ERS) procedure for selecting sib-pairs in genetic linkage analysis can be used as an alternative approach to the truncation selection (TS) method. Both approaches can be used to select extremely discordant sib-pairs and moderately discordant sib-pairs. When the cut-off point, e.g. the top 10%, for the TS is known the TS consists of independent samples. When the cut-off point is estimated by a pilot study and some of the samples in the top 10% in the pilot study may also be included in the final analysis, the correlation of these samples is negligible. Likewise the ERS consists of independent selected samples, as only a single extreme sample is selected from every \( k \) samples. For selecting discordant sib-pairs two oppositely discordant sib-pairs are selected in every \( k^2 \) samples; their correlation is also negligible.

The two methods, however, are different in design and analysis. When applying TS the cutoff points for truncation need to be estimated and the screening size is random. This may cause difficulty with a study of cost-effectiveness at the design stage. For example, in Xu et al. (1999) more than 200,000 samples were finally screened to obtain the discordant sib-pairs. To plan such

---

**Figure 1** Powers of linkage analysis using TS and ERS under the recessive model.
Figure 2 Pairs of linkage analysis using TS and ERS under the additive model.

Figure 3 Pairs of linkage analysis using TS and ERS under the dominant model.

A study and estimate the cost the power and sample size cannot be obtained unless the percentiles of the trait distribution in the population are known or estimated using a pilot study. In contrast, using ERS the screening size is fixed and there is no need to estimate the cutoff points. Furthermore, using ERS the selection of samples is based on the ranks of the trait values. Thus, the actual trait values are not always required for all samples screened. This approach is useful when a surrogate phenotype is available and/or when traits can be ranked easily without quantifications, e.g. weight, height, the size of an infant’s head, etc.
Acknowledgments

The authors would like to thank two anonymous referees for their thoughtful comments and helpful suggestions that greatly improved our presentation. The research of Zehua Chen was supported by National University of Singapore grant R-155-000-043-112. The research of Zhaohai Li was partially supported by grant EY014478 of The National Eye Institute.

References


Received: 24 August 2005
Accepted: 07 November 2005
Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

<table>
<thead>
<tr>
<th>Instruction to printer</th>
<th>Textual mark</th>
<th>Marginal mark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leave unchanged</td>
<td>... under matter to remain</td>
<td>Stet</td>
</tr>
<tr>
<td>Insert in text the matter indicated in the margin</td>
<td>\under matter to remain</td>
<td>New matter followed by</td>
</tr>
<tr>
<td>Delete</td>
<td>\through matter to be deleted</td>
<td>\under matter to be deleted</td>
</tr>
<tr>
<td>Delete and close up</td>
<td>\through letter or \ through word</td>
<td>\ durch letter oder durch word</td>
</tr>
<tr>
<td>Substitute character or substitute part of one or more word(s)</td>
<td>\under matter to be changed</td>
<td>\ durch character oder \ durch word</td>
</tr>
<tr>
<td>Change to italics</td>
<td>\under matter to be changed</td>
<td>\ durch character oder \ durch word</td>
</tr>
<tr>
<td>Change to capitals</td>
<td>\under matter to be changed</td>
<td>\ durch character oder \ durch word</td>
</tr>
<tr>
<td>Change to small capitals</td>
<td>\under matter to be changed</td>
<td>\ durch character oder \ durch word</td>
</tr>
<tr>
<td>Change to bold type</td>
<td>\under matter to be changed</td>
<td>\ durch character oder \ durch word</td>
</tr>
<tr>
<td>Change to bold italic</td>
<td>\under matter to be changed</td>
<td>\ durch character oder \ durch word</td>
</tr>
<tr>
<td>Change to lower case</td>
<td>Encircle matter to be changed</td>
<td>\ durch character oder \ durch word</td>
</tr>
<tr>
<td>Change italic to upright type</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Insert ‘superior’ character</td>
<td>\ through character or \ where required</td>
<td>\ durch character oder \ durch word</td>
</tr>
<tr>
<td>Insert ‘inferior’ character</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Insert full stop</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Insert comma</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Insert single quotation marks</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Insert double quotation marks</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Insert hyphen</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Start new paragraph</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>No new paragraph</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Transpose</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Close up</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Insert space between letters</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Insert space between words</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Reduce space between letters</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Reduce space between words</td>
<td>(As above)</td>
<td>(As above)</td>
</tr>
<tr>
<td>Linking letters</td>
<td>linking \ letters</td>
<td>linking \ letters</td>
</tr>
<tr>
<td>Insert ‘superior’ character</td>
<td>\ between letters affected</td>
<td>\ between letters affected</td>
</tr>
<tr>
<td>Insert ‘inferior’ character</td>
<td>\ between words affected</td>
<td>\ between words affected</td>
</tr>
<tr>
<td>Insert full stop</td>
<td>\ between letters affected</td>
<td>\ between letters affected</td>
</tr>
<tr>
<td>Insert comma</td>
<td>\ between words affected</td>
<td>\ between words affected</td>
</tr>
<tr>
<td>Insert single quotation marks</td>
<td>\ between letters affected</td>
<td>\ between letters affected</td>
</tr>
<tr>
<td>Insert double quotation marks</td>
<td>\ between words affected</td>
<td>\ between words affected</td>
</tr>
</tbody>
</table>

If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.