An Extensive Comparison of Quantitative Trait Loci Mapping Methods

A. Kleensang, D. Franke, A. Alcaïs, L. Abel, B. Müller-Myhsok, A. Ziegler

Introduction

Over the last two decades complex traits with quantitative intermediate traits gained considerable interest in genetic epidemiology since many common diseases can be measured directly on a quantitative scale. These quantitative intermediate traits have been shown to be more powerful than a dichotomous disease definition [1], and many different approaches that are specifically designed for detecting linkage between genetic markers and a quantitative trait (QT) locus have been proposed quite recently. As stated by Blangero [2], the harvest has come: hundreds of chromosomal regions influencing traits like body weight [3], body stature [4], bone mass density [5, 6] or mild malaria [7] have been identified, and several human genes influencing QTs have already been mapped (see, e.g., [8]).

The choice of the study design and the choice of the statistical approach for mapping a QT are of great importance [9]. Most QT studies for linkage are based on nuclear families with either independent sib-pairs, i.e., one sib-pair per family, or possibly large sibships, i.e., more than one sib-pair per family. Since it is well established that larger sibships are more powerful than small ones for quantitative trait loci (QTL) mapping, though unfortunately only some methods are originally developed or ex-
Comprehensive Comparison of QTL Methods

Among the statistical approaches for linkage mapping, variance component models (VC) received great attention in the past decade because they explicitly allow estimation of major gene effects, polygenic components as well as of environmental effects [14, 15]. This explicit modeling relies on the assumption of multivariate normality, and if this assumption is not met, the test statistics do not keep their nominal levels. For example, Allison and colleagues [16] observed an empirical significance level of about 18% for specific genetic models if the nominal level was 5%. This is especially important for selected samples because they exhibit a deviation from normality as shown by Dolan [17].

For other QT mapping approaches that seem to be more robust with respect to selected samples and/or deviation from normality, however, only very limited information is available. In all but one study, only two different statistical approaches have been compared by Monte-Carlo simulations [18–22], and only two comparisons are available under specific ascertainment schemes [23, 24]. Analytical considerations are rare [18, 20, 25, 26]. Since the available comparisons are based on different sets of assumptions it is difficult to compare or combine them. For some methods including Merlin-QTL [27] and the nonparametric (NPAR) linkage method [28] the situation is even worse because analyses of robustness or power for these methods are completely lacking. Moreover, the effect of violation of multivariate normality assumptions and/or the effect of analyzing selected samples is only known for a limited number of approaches (for a discussion, see next section). Therefore, many authors urgently called for extensive additional studies for gaining a better understanding of QT mapping approaches regarding robustness and power under a wide range of conditions for a variety of study designs and models violating the assumption of normality [10, p 220–221; 16, p 541; 19, p 252; 23, p 872; 24, p 884].

To fill this gap, we compared eight commonly applied QT mapping methods as implemented in six packages (Table 1) that are freely available for non-commercial purposes in a Monte-Carlo simulation study. We studied the robustness and power under a wide range of conditions for a variety of study designs and models violating the assumption of normality [10, p 220–221; 16, p 541; 19, p 252; 23, p 872; 24, p 884].

Table 1. Overview of used quantitative-trait linkage analyses methods, used abbreviations and programs with version numbers

<table>
<thead>
<tr>
<th>Used abbreviation</th>
<th>Program</th>
<th>Version</th>
<th>Quantitative-trait linkage analysis methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gh.HE.Trad</td>
<td>Genehunter</td>
<td>2.1_r4</td>
<td>Haseman-Elston method</td>
</tr>
<tr>
<td>Npar</td>
<td></td>
<td></td>
<td>Wilcoxon signed rank test</td>
</tr>
<tr>
<td>Gh.VC</td>
<td></td>
<td></td>
<td>Variance-component models</td>
</tr>
<tr>
<td>Linkage</td>
<td></td>
<td></td>
<td>Linkage</td>
</tr>
<tr>
<td>Merlin-Regess</td>
<td></td>
<td>5.1</td>
<td>Merlin-Regess</td>
</tr>
<tr>
<td>Merlin.K&amp;C</td>
<td></td>
<td>0.10.2</td>
<td>Merlin-QTL</td>
</tr>
<tr>
<td>Merlin.W&amp;H</td>
<td></td>
<td></td>
<td>Maximum-likelihood-binomial method for quantitative phenotypes</td>
</tr>
<tr>
<td>Mlbqt.N</td>
<td>Mlbgh</td>
<td>1.0</td>
<td>Haseman-Elston method</td>
</tr>
<tr>
<td>Mlbqt.Cat</td>
<td>S.A.G.E.</td>
<td>4.3</td>
<td>Revised Haseman-Elston method</td>
</tr>
<tr>
<td>Sage.HE</td>
<td></td>
<td></td>
<td>Variance-component models</td>
</tr>
<tr>
<td>rHE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solar.VC</td>
<td>Solar</td>
<td>1.7.4</td>
<td></td>
</tr>
</tbody>
</table>

Comprehensive Comparison of QTL Methods

Hum Hered 2010;69:202–211

203
the alternative hypothesis of linkage. To investigate the behavior of the test statistics under appropriately small significance levels, we generated 100,000 replicates under H₀ and 1,000 replicates under H₁ per model. The high number of replicates under H₀ also allowed the calculation of significance thresholds from empirical distributions for investigating empirical power as recommended by Yu et al. [22].

This article is organized as follows: in the next section we give a brief introduction to common QTL mapping methods followed by the material and methods section with a detailed description of this simulation study and compared QTL mapping. The results are explained as a whole and for each method separately, followed by a thorough discussion.

**Common Quantitative Trait Linkage Mapping Methods**

In this section we sketch common methods for mapping QT by linkage. For a detailed overview the interested reader is referred to the literature [11, 29]. The Haseman-Elston method [HE] is the root of modern model-free linkage approaches for QT [30]. The basic idea relies on the principle of similarity which traces back to Penrose [31]: genetic similarity should imply phenotypic similarity. The most commonly used measure of genetic similarity among a sib-pair is the number of alleles shared identical by descent (IBD), and Haseman and Elston used the squared trait difference to measure phenotypic similarity. The standard statistical test for linkage is a simple linear regression of the squared trait differences on the proportion of alleles shared IBD. Because of its simplicity and robustness against deviations from the assumption of normality, the HE is still commonly applied in practice.

In the last two decades the classical HE has been extended in several ways (for an overview, see, e.g., special issue Hum Hered 2003:55 and [11]). A major point of improvement is based on the fact that the squared trait difference does not utilize all phenotypic information available [32–34]. Wright suggested the combination of the trait difference and the trait sum which retain the total information [34]. He specifically proposed to use the difference between the mean-corrected squared trait sum and the squared trait difference; an expression which is equivalent to four times the mean corrected cross-product. The resulting approach is termed Haseman-Elston method revisited (rHE) [35]. rHE has greater power than HE if the residual sib correlation ρ is weak, but lower power if ρ is high [36, 37]. An interesting aspect of rHE is that it shows correct type I error fractions even under a model violating normality assumptions, strong residual correlation and an ESP ascertainment strategy [19]. Both HE and rHE were originally developed for independent sib-pairs, and the methods therefore tend to be liberal if they are applied to large sibships without appropriate corrections [38].

Sham et al. [21] further improved HE and rHE and implemented their method (Merlin-Regress) in the Merlin software package [27]. They reversed the original HE method and regressed the IBD sharing on the squared trait sums and squared trait differences in a multivariate model. In contrast to the original HE, Merlin-Regress automatically allows inclusion of all relative pairs in a pedigree. It requires, however, population-based estimates of the mean, the variance and the heritability of the trait. The correct specification of these parameters is however challenging, especially in case of selected families or non-normally distributed traits. If they are correctly specified, Merlin-Regress reveals correct type I error fractions for normally as well as for non-normally distributed traits, even for varying sibship sizes as shown by Sham et al. in an extensive simulation study [21]. Moreover, type I error fractions were robust to ESP sampling. Therefore, Merlin-Regress seems to be applicable to both unselected and selected samples. Its power is comparable to VC [21], and therefore both Merlin-Regress and VC are more powerful than the original HE for normally distributed traits (see, e.g., [39]). Misspecification of exactly one of the parameters that need to be specified in Merlin-Regress results in a loss of power but not in a violation of the type I error fractions if the trait is normally distributed and if the samples are randomly ascertained [21]. However, it remains unclear whether this robustness of Merlin-Regress can be generalized to the more complex situation of selected samples and/or non-normal traits.

A second approach implemented in the software package Merlin is an extension to the framework of allele sharing statistics introduced by Whittemore and Halpern [40] and Kong and Cox [41]. As pointed out by Ferreira [29], this framework is most appropriate for the analysis of binary traits, but it has been adapted to QTs [27]. Instead of using the number of alleles shared IBD for scoring – which is commonly used in the affected sib-pair approach, the squared scores of all the founder alleles present in a specific inheritance vector are summed over all possible inheritance vectors. The score for each founder allele in a specific inheritance vector is calculated as the sum of the mean deviate for all individuals who carry that founder allele in the pedigree. The scores can be transformed in z-scores in the usual way, and this results in the Whittemore and Halpern type test statistic.
shown in a simulation study [18]. Mlbqt considers the whole sibship and relies on the idea of binomial distributions of parental alleles among offspring. A latent, i.e., unobserved binary variable \( Y \) is introduced to capture the linkage information between the QT and the genetic marker. For each observed phenotypic value, \( Y \) represents values 1 or 0 for affected and unaffacted subjects respectively, and the probability for the latent variable \( Y \) to be equal to 1 increases as the observed QT increases. Linkage is then investigated by a single parameter likelihood ratio test for all possible sets of \( Y \) values within sibships weighted by their probabilities. Mlbqt can be applied for normally as well as non-normally distributed traits. In the latter case, the distributional assumption is replaced by the empirical distribution as a step-function defined from the cumulative frequencies, e.g., by use of empirical deciles. This also allows the application of Mlbqt to selected samples. As shown in a simulation study [18], Mlbqt generally has higher power than the classical HE, and it reveals correct type I error fractions for normally distributed traits.

Similar to Mlbqt for deciles, the model-free nonparametric linkage method for quantitative phenotypes (Npar) should also be robust to deviations from normality [28]. The authors proposed their method especially for situations where the phenotypes are non-normally distributed. The method is based upon a Wilcoxon signed rank test and is therefore applicable to any phenotypic distributions [28].

Apart from the model-free methods described before, the linkage program package can be used to run a fully parameterized LOD-Score analysis for QT mapping [42]. However, because for a complex disease the detailed inheritance model – with its allele numbers, their frequencies, genotypic means and variance – can hardly be specified, model-free mapping methods are almost always used.

**Material and Methods**

**Simulations**

**Phenotypic Models**

The phenotype simulation is based on a variance analytic model which is described in detail, e.g., by Falconer and Mackay [43]. The phenotypic value \( x_{ik} \) of an individual \( k \) within family \( i \) is additively decomposed into an overall mean, a major gene effect \( g_k \) being determined by the genotype of a diallelic quantitative trait locus together with its specified inheritance model, an environmental effect \( e_k \) simulated as family effect which assigns each member of the pedigree the same random value and, finally, an error term \( \epsilon_{ik} \):

\[
x_{ik} = \mu + g_k + G_i + \epsilon_{ik}
\]

Datasets were simulated under dominant, additive and recessive genetic models with respective frequencies of the 'high allele' of 0.05, 0.2 and 0.3. All models were simulated with a mean of 0, a variance of 1 and heritability in a narrow sense of 0.8. The total variance of 1 is decomposed into the variance of the major gene effect of 0.2, variance of the environmental effect of 0.3, and variance of the error term of 0.5. The major gene and the environmental effects are simulated from a normal distribution. To assess the effect of a violation of normality, the error term was chosen either from a normal distribution or a three parametric lognormal distribution. To obtain a standard three parametric lognormally distributed variable we took the exponent of a standard normal variable, subtracted the mean of the standard lognormal distribution and, finally divided this by the standard deviation of the standard lognormally distributed variable.

**Family Structures**

Two different types of nuclear families were chosen as family structures. Within every dataset, we simulated either 300 families with two offspring (independent sibships) or a more realistic situation of 100 families with two, three, four and five sib-pairs per family and their respective proportions of 40, 30, 18 and 12% (dependent sibships) according to Speer et al. [44].

**Genetic Marker**

One genetic marker with ten alleles with equal frequencies was simulated. Under \( H_0 \) the genetic marker was simulated with a \( \theta = 0.5 \) between genetic marker and QT; under \( H_1 \) we assumed complete linkage, i.e., \( \theta = 0 \).

**Selection of Families**

To take into account different ascertainment schemes, the following constellations were simulated: We chose families without selection (random selection), families with at least one sibling in the highest quartile (single proband selection according to a SPSP design) and families with at least two siblings which were either both in the highest quartile, or both in the lowest quartile, or of which one was in the highest and one in the lowest quartile (double proband selection according to an ESP design).

**Datasets**

For the purpose of this article, we developed the software package SIBSIM [45] which is tailored for rapid simulation of QTs in nuclear families as well as in extended pedigrees. Specifically, we simulated 100,000 datasets under \( H_0 \) and 1,000 datasets under \( H_1 \). Entirely new data sets were generated for every model, selection scheme, family structure and phenotype distribution. Genotypes and phenotypes were simulated for all family members, thus no missing data were simulated. Altogether, we simulated 36 different models under both \( H_0 \) and \( H_1 \) (3 genetic models \( \times 2 \) family structures \( \times 3 \) selection schemes \( \times 2 \) distributions).
Implementations of QT Linkage Methods in Software Packages

Eight different QT methods for analyzing linkage implemented in six software packages were chosen that are freely available for non-commercial use. The methods, their corresponding programs, version numbers and abbreviations are displayed in table 1. The analysis options are described below in detail. LOD scores were converted to p values as described elsewhere [11, p 159]. To assess possible implementation differences between different programs, the Haseeman-Elston method and the variance component models were calculated with two different programs.

Genehunter
Genehunter (version 2.1_r4) [46] was used to calculate the HE method (Gh.HE.Trad), the Npar and the VC (Gh.VC). For all analyses we used the options ‘all pairs unweighted’ and ‘no dominance variance’. Because parental genotypes were complete, we ignored the EM results.

Linkage
The linkage package (version 5.1) [42] was used to perform model-based quantitative-trait linkage analyses (Linkage). The true values for genotype means of the major gene effect, their corresponding frequencies of the diallelic QTL as well as population-based parameters for variances of the phenotypes (for all models var = 1) were employed for the analyses. The genotype means for the three different genetic models are given in table 2, where the allele A1 corresponds to the ‘high allele’.

Merlin
The program Merlin-Regress which is a part of the Merlin program package (version 0.10.2) [21] was used to calculate the method proposed by Sham et al. (Merlin-Regress). The true population-based values for phenotypic mean, variance and heritability were applied for the analyses (for all models mean = 0, variance = 1 and heritability = 0.5). To assess a possible influence of one-parameter model misspecification, all datasets were additionally analyzed under a wide range of parameter values for mean (range: ±5), variance (range: 0.1–10) and heritability (range: 0.05–0.95). Merlin was also used for calculating Merlin-QTL: We evaluated both the Whittemore and Halpern (Merlin.W&H) and the Kong and Cox (Merlin.K&C) test statistics.

MLbgh
The program MLbgh (version 1.0) [18, 47] is a modification of Genehunter with the implementation of MLbqt [18]. The analyses were performed both under the assumption of a standard normal distribution for the phenotypes (MLbqt.N) and by defining a step function to divide the phenotypes in empirical deciles (MLbqt.Cat).

S.A.G.E.
The program package S.A.G.E. (version 4.3) was used to calculate the HE (Sage.HE) and the rHE. The analyses were performed using the standard parameters, and for rHE the population-based true mean value was set to the correct value of 0.

Solar
Solar (version 1.7.4) [14] was used to calculate VC (Solar.VC). The analyses were performed under the given default parameters.

Table 2. Genotype means for the major gene effect as used for the model-based analyses with the linkage program

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dominant</td>
<td>-0.147</td>
<td>1.361</td>
<td>1.361</td>
</tr>
<tr>
<td>Additive</td>
<td>-0.316</td>
<td>0.474</td>
<td>1.265</td>
</tr>
<tr>
<td>Recessive</td>
<td>-0.141</td>
<td>-0.141</td>
<td>1.422</td>
</tr>
</tbody>
</table>

The allele A1 corresponds to the ‘high allele’.

Empirical Type I Error and Empirical Power

For every model and method we investigated the robustness by estimating the empirical type I error at nominal type I error levels of 5, 1 and 0.1% under H0. For power comparisons, the power was calculated as empirical power at an empirical type I error of 0.05.

Results

Type I Error

The empirical type I error fractions at a nominal type I error of 1% are displayed in table 3, and corresponding tables for the 0.1 and 5% nominal significance levels are given in the online suppl. tables 5, 6 (for all online suppl. material, see www.karger.com/doi/10.1159/000289596). Under normality assumptions, no selection and independent sibships, VC were too liberal while Merlin.W&H and linkage were conservative for almost all scenarios considered. The other methods showed the correct type I error fraction. A violation of one or more of those assumptions affected the type I error level for many of the methods. This is described below in detail for each method.

We note that the relative increase of a violation of the type I error fraction increases with decreasing nominal type I error. Specifically, at a nominal type I error of 5%, independent randomly ascertained sib-pairs, a dominant genetic model and a normally distributed error term, VC showed an empirical type I error of 6.14%, i.e., a relative inflation of ~23%. This increased to a relative inflation of ~70% at the nominal 0.1% test level (empirical type I error level 0.170%).

Empirical Power

The empirical power estimates at an empirical type I error fraction of 5% are shown in table 4. VC and Merlin-Regress showed high empirical power for all models considered if the normality assumption was met.
### Table 3. Empirical type I error fractions at a nominal type I error fraction of 1%

<table>
<thead>
<tr>
<th>Methods</th>
<th>Sibships</th>
<th>Normality assumptions</th>
<th>Violation of normality assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>dom</td>
<td>add</td>
</tr>
<tr>
<td>Gh.HE.Trad</td>
<td>Independent</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Sage.HE</td>
<td>1.0</td>
<td>1.0</td>
<td>0.9</td>
</tr>
<tr>
<td>rHE</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Merlin-Regression</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Solar.VC</td>
<td>1.4</td>
<td>1.4</td>
<td>0.7</td>
</tr>
<tr>
<td>Npar</td>
<td>1.0</td>
<td>1.0</td>
<td>0.9</td>
</tr>
<tr>
<td>Merlin.K&amp;C</td>
<td>1.0</td>
<td>1.0</td>
<td>0.9</td>
</tr>
<tr>
<td>Merlin.W&amp;H</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Mlbqt.Cat</td>
<td>1.1</td>
<td>1.1</td>
<td>1.0</td>
</tr>
<tr>
<td>Linkage</td>
<td>0.4</td>
<td>0.4</td>
<td>0.2</td>
</tr>
</tbody>
</table>

### Table 4. Empirical power at an empirical type I error fraction of 5%

<table>
<thead>
<tr>
<th>Methods</th>
<th>Sibships</th>
<th>Normality assumptions</th>
<th>Violation of normality assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>dom</td>
<td>add</td>
</tr>
<tr>
<td>Gh.HE.Trad</td>
<td>Independent</td>
<td>34</td>
<td>38</td>
</tr>
<tr>
<td>Sage.HE</td>
<td>34</td>
<td>38</td>
<td>32</td>
</tr>
<tr>
<td>rHE</td>
<td>24</td>
<td>28</td>
<td>26</td>
</tr>
<tr>
<td>Merlin-regress</td>
<td>42</td>
<td>49</td>
<td>40</td>
</tr>
<tr>
<td>Solar.VC</td>
<td>44</td>
<td>47</td>
<td>42</td>
</tr>
<tr>
<td>Npar</td>
<td>21</td>
<td>27</td>
<td>23</td>
</tr>
<tr>
<td>Merlin.K&amp;C</td>
<td>13</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>Merlin.W&amp;H</td>
<td>15</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>Mlbqt.Cat</td>
<td>21</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td>Linkage</td>
<td>69</td>
<td>41</td>
<td>58</td>
</tr>
</tbody>
</table>

Downloaded by: 54.70.40.11 - 1/27/2018 3:27:08 PM
ever, for non-normal models, Npar, Merlin.W&H and Merlin.K&C generally performed better. Merlin-Regress and VC were superior over HE for all considered models. As expected, the highest power was observed for the dominant and the recessive model for the gold standard of comparison, the fully parameterized LOD score analysis. For Npar and the allele sharing approaches Merlin and Mlbqt the empirical power increased under deviation from normality but decreased for the other methods with some unsystematic exceptions. The results will be explained below in more detail for each method separately.

**HE and rHE**

Dependent sibships led to only slightly inflated type I error levels for HE in the Gh implementation; a remarkable inflation was observed for the GEE approach taken by S.A.G.E. for HE and rHE. The empirical type I error was almost not affected by selection for both methods. Violation of normality assumptions led to conservative type I levels for HE with the exception of Sage.HE under random or double selection. For rHe the type I error was almost not affected with the exception of dependent sibships where the type I error tended to inflate.

The empirical power was in general inferior compared to Merlin-Regress and VC for all considered models with some exceptions for rHE under violation of normality assumptions.

**Merlin-Regress**

The empirical type I error is not affected by selection but became clearly too liberal for dependent sibships and slightly too conservative under violation of normality assumptions.

Merlin-Regress showed in general a very good power comparable to VC. Under violation of normality assumptions the power was in general higher than with VC.

**VC**

Gh.VC and Solar.VC implementations showed no clear differences in empirical type I error and power. To increase the readability of the tables the results for the Gh.VC are not shown. VC did not show the correct type I error for any model, not even for normality assumptions and independent sibships. VC is unacceptably too liberal, especially under violation of normality assumptions. The effect of single or double selection heavily influenced the type I error in both directions. The effect was much stronger for independent sibships where in general the VC was very conservative under single selection and unacceptably too liberal for double selection.

In general, VC showed a very good power, comparable to Merlin-Regress. Under violation of normality assumptions the power was generally lower than for Merlin-Regress.

**Npar**

For independent sibships Npar fitted the nominal type I error for all models very well. The empirical type I error was not affected by selection or violation of normality assumptions but tended to become slightly too liberal for dependent sibships.

Under normality assumptions the power was in general lower than for Merlin-Regress, VC, HE and rHe. Under violation of normality assumptions the situation changed: While without selection the power was very high and comparable to the power of VC and Merlin-Regress Npar outperformed the other methods under single selection and double selection with the exception of independent sibships.


In general, Merlin.W&H was too conservative for all models, while Merlin.K&C fit the nominal type I error very well. The empirical type I error was not affected by ascertainment, violation of normality assumptions or dependent sibships.

Both methods generally showed a low power under normality assumptions. The situation became better under violation of normality assumptions, especially under single or double selection where Merlin.W&H and Merlin.K&C showed a better power than VC and Merlin-Regress, comparable to Mlbqt.

**Mlbqt.N and Mlbqt.Cat**

Mlbqt.N and Mlbqt.Cat showed nearly identical results. To increase the readability of the tables the results for the Mlbqt.N are not shown. Mlbqt fitted the nominal type I errors for all models very well. The empirical type I error was therefore not influenced by either selection, violation of normality assumptions or dependent sibships.

Mlbqt showed a low power under normality assumptions even under selection. Under violation of normality assumptions, with random selection, the power was intermediate. However, the method performed much better under single or double selection resulting in an empirical power being higher than for VC and Merlin-Regress, even comparable to Merlin.K&C.
Linkage

With some rare exceptions Linkage showed extremely too conservative type I error levels for all models, even under normality assumptions, independent sibships and random selection. Under single or double selection the type I error increased in general. The impact of dependent sibships is unclear.

As expected, the fully parameterized LOD score analysis showed the highest power for the dominant and the recessive model. For the additive model the power was comparable or a little bit lower than for Merlin-Regress and VC.

Effect of Model Misspecification for Merlin-Regress

The effects of a one-parameter model misspecification are shown in online supplementary figures 1–3. A misspecification of one parameter only resulted in a loss of power but still showed correct type I error fractions: a misspecification of the phenotypic mean showed the highest effect on empirical power while an underestimation of the phenotypic variance had only a tiny effect. The effect of a heritability misspecification was small for realistic misspecifications but increased at the lower and upper border of the possible scale of values.

Discussion

In this Monte-Carlo simulation study we investigated the type I error and power of different quantitative trait linkage approaches which are, to the best of our knowledge, currently used to map QTs in nuclear families. We specifically investigated dependent sibships, three common ascertainment schemes for three major gene models and the impact of non-normally distributed residuals. There are two important differences to previous simulation studies: (a) while others studied the effect of deviation from normality by transforming the phenotype [16, 19, 21, 23, 24], we used a log-normally distributed error term. This allowed preserving the major effect and the familial component, while at the same time the residual was strongly skewed and kurtotic. (b) Most simulation studies used 10,000 replicates for a specific model under $H_0$. However, confidence intervals of $p$ values are $\sim$1% wide at the nominal 5% test level for this low number of replicates. Thus, lower nominal type I error levels cannot be investigated although $p$ values $\sim$0.001 are of great interest in linkage studies. We therefore generated 100,000 replicates for each scenario under $H_0$, resulting in a $\sim$0.002 length of the 95% confidence interval of the $p$ value. It must be pointed out that these results can be generalized with certainty only to the conditions considered in this study.

Some results are well known and could be confirmed by this study; some are new and will be discussed in this section.

Our results indicate that the VC approach results in substantially inflated type I errors even for normally distributed residuals, realistic sample sizes and randomly drawn nuclear families. There are two reasons that may explain this unexpected finding. First, Ferreira [29] noted that the expected covariance matrix includes six free parameters. However, when applied to sib-pair data, these are not identifiable, and additional restrictions to retain identifiability are required. Moreover, the standard errors of variances are fourth order moments. And it is well known that statistics relying on fourth-order moments are unstable for relatively small sample sizes that characterize most real data (see, e.g., [48]).

Some recommendations are given in the literature to improve the robustness of the VC. For example a simple transformation of the phenotype by a function to obtain a like-normally distributed phenotype that is hopefully multivariate normally distributed. However, there is no guarantee that a transformation exists that will result in a multivariate normal distribution. For example, the non-normally distributed phenotypes in our study could not be transformed to multivariate normality by any function. A second recommendation is to utilize robust estimators [49], but the utilization of robust estimators could result in a power decrease. A third option for selected samples is to incorporate the selection probabilities into the calculation of the likelihood, but this may reduce the power too.

In contrast to VC, Merlin-Regress is very robust and the type I error fractions are not influenced by selected samples without the need of robust estimators or other pull-ups with the only disadvantage of a liberal type I error in the situation of dependent sibships, as demonstrated in this publication. If, as concluded by several authors [21, 22] and our results, VC and Merlin-Regress show nearly the same high empirical power, Merlin-Regress should be preferred to VC for mapping QTs in nuclear families with realistic sibship sizes as used in this study. Moreover, we demonstrated that the type I error of Merlin-Regress is not influenced by a one-parameter model misspecification and that only a minimal loss of power resulted even in the complex situations of selected data and non-normally distributed traits. However, as pointed out by one reviewer, model misspecification in a real scenario will result from a combination of various parameters and is much more complicated than the one-param-
eter model used in this study. In addition Huang and colleagues identified at the GAW 15 a liberal type I error of Merlin-Regress if applied to extended pedigrees [50]. Further studies are needed to address these points in detail before a clear recommendation for Merlin-Regress can be given.

The next well studied approaches are HE and rHE. The remarkably lower power compared to Merlin-Regress and the strong influence of non-normally distributed residuals on the empirical type I error should make it unnecessary to use these methods anymore.

The allele sharing approach Merlin.W&H was in generally too conservative for all models. An explanation is given by Kong & Cox [41] where the authors showed that for approaches like the W&H, a less than perfectly informative genetic marker can result in an unacceptable conservative type I error. The W&H approach should therefore not be used. The other two allele sharing approaches Merlin.K&C and Mlbqt were very robust but the empirical power is very low compared to other methods with some exceptions under non-normality and single or double ascertainment. Since this is the first comparison under violation of normality assumptions these results could not be generalized. The low power in comparison to the other methods prevents a clear recommendation for them. Mlbqt.Cat could still be a method of choice for some study designs if the phenotype is for example ordinal distributed and an empirical distribution could be defined.

The results for the Npar method were somewhat surprising. The method is robust like the allele sharing approaches with a small increase in type I-error for dependent sibships. Under normality assumptions the power was generally lower than for Merlin-Regress, VC, HE and rHE as expected for a nonparametric method. But surprisingly the situation changed under non-normal distribution. Npar showed a high power and outperformed the other methods under single selection and double selection with the exception of independent sibships. The remarkable power difference between the additive versus the dominant or recessive models under violation of normality can be explained by the chosen function for centering the number of alleles IBD in the test statistic as implemented in Gh. Kruglyak & Lander recommend the implemented function especially to test for a QT with an additive effect [28, 51]. For Npar and the allele sharing approaches the empirical power increased under deviation from normality but decreased for the other methods with some unsystematic exceptions. The reason could be the high third moment introduced by the log-normal distributed error term. The resulting distribution for each of different genotypes of the diallelic QTL is highly right skewed but with the same expectation values like for the normal distributed error term. The allele sharing approaches and the nonparametric method seem to utilize this information in a better way than the other approaches.

Finally, the extremely conservative type I error levels for Linkage are well known. For example, Rao and colleagues showed for a large number of empirical studies that the observed type I error was considerably less than the nominal type I error [52]. The high power of linkage, as shown in this study, can almost not be archived in real data studies because the correct mode of inheritance can hardly be specified for a complex disease. Therefore, model-free mapping methods should almost always the method of choice.

Electronic-Database Information

URLs for data presented herein are as follows:


Linkage, ftp://linkage.rockefeller.edu/software/linkage

Merlin, http://www.sph.umich.edu/csg/abecasis/Merlin/download/


References


Comprehensive Comparison of QTL Methods

Hum Hered 2010;69:202–211

211


