

The Power of QTL Mapping with RILs

Shohei Takuno^{1‡}, Ryohei Terauchi², Hideki Innan^{1*}

1 Graduate University for Advanced Studies, Hayama, Kanagawa, Japan, **2** Iwate Biotechnology Research Center 22-174-4, Narita, Kitakami, Iwate, Japan

Abstract

QTL (quantitative trait loci) mapping is commonly used to identify genetic regions responsible to important phenotype variation. A common strategy of QTL mapping is to use recombinant inbred lines (RILs), which are usually established by several generations of inbreeding of an F₁ population (usually up to F₆ or F₇ populations). As this inbreeding process involves a large amount of labor, we are particularly interested in the effect of the number of inbreeding generations on the power of QTL mapping; a part of the labor could be saved if a smaller number of inbreeding provides sufficient power. By using simulations, we investigated the performance of QTL mapping with recombinant inbred lines (RILs). As expected, we found that the power of F₄ population could be almost comparable to that of F₆ and F₇ populations. A potential problem in using F₄ population is that a large proportion of RILs are heterozygotes. We here introduced a new method to partly relax this problem. The performance of this method was verified by simulations with a wide range of parameters including the size of the segregation population, recombination rate, genome size and the density of markers. We found our method works better than the commonly used standard method especially when there are a number of heterozygous markers. Our results imply that in most cases, QTL mapping does not necessarily require RILs at F₆ or F₇ generations; rather, F₄ (or even F₃) populations would be almost as useful as F₆ or F₇ populations. Because the cost to establish a number of RILs for many generations is enormous, this finding will cause a reduction in the cost of QTL mapping, thereby accelerating gene mapping in many species.

Citation: Takuno S, Terauchi R, Innan H (2012) The Power of QTL Mapping with RILs. PLoS ONE 7(10): e46545. doi:10.1371/journal.pone.0046545

Editor: Jinfa Zhang, New Mexico State University, United States of America

Received: February 16, 2012; **Accepted:** September 5, 2012; **Published:** October 9, 2012

Copyright: © 2012 Takuno et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work is supported by the Program for Promotion of Basic Research Activities for Innovative Biosciences (PROBRAIN) from the Bio-oriented Technology Research Advancement Institution in Japan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: innan hideki@soken.ac.jp

‡ Current address: Department of Plant Sciences, University of California Davis, Davis, California, United States of America

Introduction

Mapping quantitative trait loci (QTL) plays crucial roles in a number of research fields in biology. QTL mapping basically relies on detecting correlations between genetic markers and phenotypic traits in a segregating population [1–4]. The development of the interval mapping method [5,6] made it possible to infer the positions of QTL with a limited number of markers. Since then, QTL mapping has been applied to various crop and vegetable species, including an early application to genome-wide QTL analysis of tomato species [7]. With the advent of molecular biology techniques such as sequencing, DNA microarray and primer extension assay [8–10], it became feasible to distribute a large number of markers across the genome and genotype those markers for a large sample of individuals. This revolutionary change in molecular biology further facilitated QTL mapping in many species.

Efficient fine-scale QTL mapping requires a large segregating population (bi-parental mapping population) such as an F₂ population or Recombinant Inbred Lines (or RILs). An F₁ population is first generated by a pair of homozygous parents (usually denoted by P₁ and P₂), and then selfing or sibling mating of the F₁ individual generates an F₂ population. It is common that each of the RILs is further selfed or sib-mated for several more generations, and F₆~F₇ populations are frequently used for QTL analyses.

The advantages of using RILs for a QTL analysis are obvious. First, multiple selfing processes can increase the number of recombination events [11], which results in a finer mapping of QTLs. More importantly, once RILs are established, in which the genotypes of all lines are fixed as homozygotes, these lines can be repeatedly used for investigating QTLs of various phenotypes under different environments. Thus, the establishment of a comprehensive set of RILs will be a substantial contribution to QTL mapping of the species.

In the meantime, QTL mapping is frequently applied to species that do not have substantial resources at the single-lab level. In this case, it is not reasonable to establish comprehensive RILs; rather, it makes more sense to conventionally map a rough location of a QTL with a limited amount of effort. There is an obvious tradeoff between the performance of QTL mapping and the cost required, including the sample size and the number of generations of selfing or sibling mating. The heaviest labor would be to maintain a number of RILs for multiple generations, so that a simple idea is to use a younger generation with a limited number of RILs. As a consequence, as a most aggressive setting, there are a number of QTL mapping studies that conventionally used an F₂ population.

In such a small-scale QTL mapping, it is very useful if we have some ideas about the relationship between the performance (statistical power) and the cost (the number of selfing or sibling mating generations, sample size, and marker density), which will greatly help to optimize the design of the QTL mapping

experiment. This problem has been extensively investigated in simple theoretical models [12–14]. Here, we provide the results of extensive simulations in more realistic situations. We assume that a large number of markers are distributed across the genome, and that they are partially linked. With these results, we discuss how the cost can be reduced by minimizing the reduction of the performance.

Methods

Model and simulation

For simulating QTL mapping process with RILs, we consider a diploid species. It is assumed that the genome consists of L chromosomes with equal lengths and that the genome size is G Mb, which corresponds to R centimorgan (cM). It is also assumed that M markers are evenly distributed across the genome. We set a single QTL in the simulated genome, and ask whether we can find significant phenotype-genotype correlations for markers nearby the QTL. To assess the performance of QTL mapping, we simulate the process of creating a large number of RILs from a single pair of parental lines, P_1 and P_2 , both of which are assumed to be completely homozygote. Their hybrid progeny, F_1 , is created, and then $N \times F_2$ progenies are produced by selfing F_1 . It is assumed that each of the F_2 progenies is successfully inbred by the single-seed-descent method for six generations (*i.e.*, up to F_7). Throughout this process, recombination occurs randomly at rate R , following the four-strand model [15]. It is also assumed that at least one chiasma form in each chromosome in one meiosis event, called obligate chiasma [16–18], but for simplicity, we also assume no crossing-over interference. At each generation from F_2 to F_7 , a simple QTL mapping method (see below) is applied.

In the QTL mapping process, it is assumed that all markers are genotyped for all individuals, and the phenotype of each individual is determined by a simple model, in which there is a particular locus that partially contribute to the quantitative trait of interest [1–5]. Let Q_1 and Q_2 be the two alternative alleles at this QTL inherited from the two parental lines (P_1 and P_2). Then, it is assumed that the phenotype of each diploid individual in the segregating population is determined by the genotype at this locus. There are three possible states, Q_1Q_1 , Q_1Q_2 , Q_2Q_2 , which are denoted by genotypes 1, 2, and 3, respectively. The numbers of individuals with the three genotypes are denoted by n_1 , n_2 , and n_3 , and N is the total number of individuals ($N = n_1 + n_2 + n_3$).

Let y_{ij} be the quantitative value representing the focal phenotype of the j th individual in the i th genotype ($i = 1, 2, 3$ and $j = 1, 2, \dots, n_i$), then in a simple model with no interaction between genotypes and environment, y_{ij} can be written as

$$y_{1j} = \mu + a + \varepsilon_{1j}, \quad (1)$$

$$y_{2j} = \mu + d + \varepsilon_{2j}, \quad (2)$$

$$y_{3j} = \mu - a + \varepsilon_{3j}, \quad (3)$$

where μ is the mid-parental value, a is the additive genetic effect and d is the dominance effect. Other factors are represented by ε_{ij} , including the environmental variance and the residual genotypic variance due to other unlinked QTLs. ε_{ij} is assumed to follow a normal distribution with mean 0 and variance σ^2 . We assume that this factor ε_{ij} is added at each generation independently. In other words, only μ , a and d are the parameters that determine the

genetic factors that can be inherited through generations, and ε_{ij} is not affected by the phenotype or genotype at the previous generation. Simulations of RILs under this simple model are used for investigating the performance of QTL mapping. Assuming a large number of markers are available across the genome, we simply perform a statistical test of the null hypothesis of no association between the phenotype and each of all markers. We do not need to use the interval mapping method because of the availability of a large number of markers (this condition will be relaxed later). We use two likelihood ratio tests to examine if there is a significant phenotype-genotype correlation.

In the first method (Method I), if B and b represent the two alleles from P_1 and P_2 , respectively, the null model assumes equal average phenotypes of the three genotypes, $\bar{y}_{BB} = \bar{y}_{Bb} = \bar{y}_{bb}$. Alternatively, if the marker and the QTL is completely linked, we expect $E(\bar{y}_{BB}) = \mu + a$, $E(\bar{y}_{Bb}) = \mu + d$ and $E(\bar{y}_{bb}) = \mu - a$. Method I requires the likelihoods of the observation (\bar{y}_{BB} , \bar{y}_{Bb} , \bar{y}_{bb}) under these two extreme cases (null and alternative). It should be noted that this very commonly used method requires estimation of the dominance effect (*e.g.*, [6,19,20]). Alternatively, the second method, which we propose here, is a simplified version (Method II), in which only homozygote individuals with marker genotypes, BB and bb are considered (heterozygotes, Bb , are excluded), and tests the null hypothesis of $\bar{y}_{BB} = \bar{y}_{bb}$. We propose this conventional method because it does have to involve the dominance parameter by excluding heterozygotes from the analysis. Estimation of the dominance parameter has to rely on a relatively small number of heterozygotes, which will likely cause a great deal of uncertainty in the estimate. We suspected that miss-inference of the dominance parameter due to such uncertainty might result in a reduction of the power. Obviously, the situation would be identical when selfing generations increase and all RILs become homozygote in the entire genome. The two methods are described below in detail.

Method I. This method involves computation of the maximum likelihoods of the observation, ($\bar{y}_{BB}, \bar{y}_{Bb}, \bar{y}_{bb}$), under the null and alternative models. The latter involves maximum likelihood estimation of the four unknown parameters, μ , a , d , and $\sigma_{I,Alt}^2$, which are given by

$$\hat{\mu} = \frac{\bar{y}_1 + \bar{y}_3}{2}, \quad (4)$$

$$\hat{a} = \frac{\bar{y}_1 - \bar{y}_3}{2}, \quad (5)$$

$$\hat{d} = \bar{y}_2 - \hat{\mu}, \quad (6)$$

and

$$\hat{\sigma}_{I,Alt}^2 = \frac{1}{N} \left[\sum_j^{n_1} (y_{1j} - \hat{\mu} - \hat{a})^2 + \sum_j^{n_2} (y_{2j} - \hat{\mu} - \hat{d})^2 + \sum_j^{n_3} (y_{3j} - \hat{\mu} + \hat{a})^2 \right]. \quad (7)$$

Then, it can be considered that the maximum likelihood of the observation under the alternative scenario is given these estimates. That is, the log-maximum likelihood is computed by

$$\begin{aligned}
LL_{I,Alt}(\hat{\mu}, \hat{a}, \hat{d}, \sigma_{I,Alt}^2) &= -\frac{1}{2}N \ln(2\pi\sigma_{I,Alt}^2) \\
&- \sum_j^{n_1} \frac{(y_{1j} - \hat{\mu} - \hat{a})^2}{2\sigma_{I,Alt}^2} - \sum_j^{n_2} \frac{(y_{2j} - \hat{\mu} - \hat{a})^2}{2\sigma_{I,Alt}^2} \\
&- \sum_j^{n_3} \frac{(y_{3j} - \hat{\mu} + \hat{a})^2}{2\sigma_{I,Alt}^2} \\
&= -\frac{1}{2}N \ln(2\pi\sigma_{I,Alt}^2) - \frac{1}{2}N
\end{aligned} \quad (8)$$

In the null model, in which only two parameters (μ and $\sigma_{I,Null}^2$) are involved, the maximum log-likelihood of the data is given by

$$LL_{I,Null}(\hat{\mu}, \sigma_{I,Null}^2) = -\frac{1}{2}N \ln(2\pi\sigma_{I,Null}^2) - \frac{1}{2}N, \quad (9)$$

where $\sigma_{I,Null}^2$ is simply given by

$$\sigma_{I,Null}^2 = \frac{1}{N} \sum_j^N (y_j - \hat{\mu})^2. \quad (10)$$

y_j represents the phenotypic value of the j th individual (with no specification of genotype, so that $j = 1, 2, \dots, N$).

Thus, the maximum log-likelihoods under the null and alternative models are computed by equations (9) and (8), respectively, from which the LOD score can be obtained by $(LL_{I,Alt} - LL_{I,Null})/\ln(10)$. For each replication of the simulations, we set a cut-off value of the LOD score by 1,000 replications of a permutation test [21], so that the false positive rate is set at $\alpha = 0.05$ after correcting for multiple testing by multiplying the P -value by the number of markers (*i.e.*, Bonferroni correction). Note that because a permutation test is performed for each data set, the false positive rate is always 5% for any parameter set in all generations. This allows a fair comparison of the performances of different models with different parameters.

Method II. This method is a simplified version of Method I, in which marker-heterozygous individuals are excluded so that it does not involve the process of estimating the dominance parameter. In the alternative model of Method II, μ and a can be estimated from the average phenotypes, \bar{y}_1 and \bar{y}_3 :

$$\hat{\mu} = \frac{\bar{y}_1 + \bar{y}_3}{2}, \quad (11)$$

$$\hat{a} = \frac{\bar{y}_1 - \bar{y}_3}{2}, \quad (12)$$

and $\sigma_{II,Alt}^2$ is given by

$$\sigma_{II,Alt}^2 = \frac{1}{n_1 + n_3} \left[\sum_j^{n_1} (y_{1j} - \hat{\mu} - \hat{a})^2 + \sum_j^{n_3} (y_{3j} - \hat{\mu} + \hat{a})^2 \right]. \quad (13)$$

This process is basically identical to that for Method I. Then, the maximum likelihood of the observation under the alternative scenario is given with these given estimates:

$$\begin{aligned}
LL_{II,Alt}(\hat{\mu}, \hat{a}, \sigma_{II,Alt}^2) &= -\frac{1}{2}(n_1 + n_3) \ln(2\pi\sigma_{II,Alt}^2) \\
&- \sum_j^{n_1} \frac{(y_{1j} - \hat{\mu} - \hat{a})^2}{2\sigma_{II,Alt}^2} - \sum_j^{n_3} \frac{(y_{3j} - \hat{\mu} + \hat{a})^2}{2\sigma_{II,Alt}^2} \\
&= -\frac{1}{2}(n_1 + n_3) \ln(2\pi\sigma_{II,Alt}^2) - \frac{1}{2}(n_1 + n_3)
\end{aligned} \quad (14)$$

In the null model, where only two parameters (μ and $\sigma_{II,Null}$) are involved as well as Method I, the maximum log-likelihood of the data is given by

$$\begin{aligned}
LL_{II,Null}(\hat{\mu}, \sigma_{II,Null}^2) &= \\
&- \frac{1}{2}(n_1 + n_3) \ln(2\pi\sigma_{II,Null}^2) - \frac{1}{2}(n_1 + n_3), \quad (15)
\end{aligned}$$

where $\sigma_{II,Null}^2$ is simply given by

$$\sigma_{II,Null}^2 = \frac{1}{n_1 + n_3} \left[\sum_j^{n_1} (y_{1j} - \hat{\mu})^2 + \sum_j^{n_3} (y_{3j} - \hat{\mu})^2 \right]. \quad (16)$$

Then, from equations (14) and (15), the LOD score can be computed as $(LL_{II,Alt} - LL_{II,Null})/\ln(10)$.

Results

Simulation results

We designed simulations to quantitatively evaluate the effect of the number of generations on the performance of QTL mapping. Throughout this article, we fix $\mu = 0$ and $a = 1$. We assume a simple model, in which the simulated genome consists of $L = 12$ chromosomes with equal length $G = 30$ Mb, so that the genome size (360 Mb) is similar to that of rice, a species to which QTL mapping is frequently applied. In total $M = 1,200$ codominant DNA markers are evenly distributed on the genome, such that the interval length is 300 kb (100 markers per chromosome). The recombination rate is assumed to be 4 cM/Mb, which is roughly consistent with estimates of rice [22]. Some of these simulation conditions will be relaxed later.

We are interested in the power of QTL mapping to detect a particular QTL that has a significant genetic contribution. It is assumed that this QTL locates at the center of one chromosome. This location is also the middle of two adjacent markers; therefore, the distance to the closest marker is 150 kb. Although the model does not set other specific QTLs, their effect is incorporated in the environmental factor, ε in equations (1–3). For each of these parameter settings, we performed 10,000 independent replications of simulations from F_1 to F_7 , and at each generation (except for F_1) the LOD scores were computed for all markers.

A typical pattern of the results is shown in Figure 1, in which $N = 200$, $\sigma^2 = 2$, and no dominance ($d = 0$) were assumed. The expected heritability in the F_2 population is given by

$$Q_{PIF_2} = \frac{V_A + V_H}{V_A + V_H + V_E}. \quad (17)$$

Therefore, with this parameter set, we expect that the expected heritability is 20% (note that the heritability changes in the following F_3 , F_4 , ... generations). It was found that on the chromosome with the QTL (left panel in Figure 1A), both

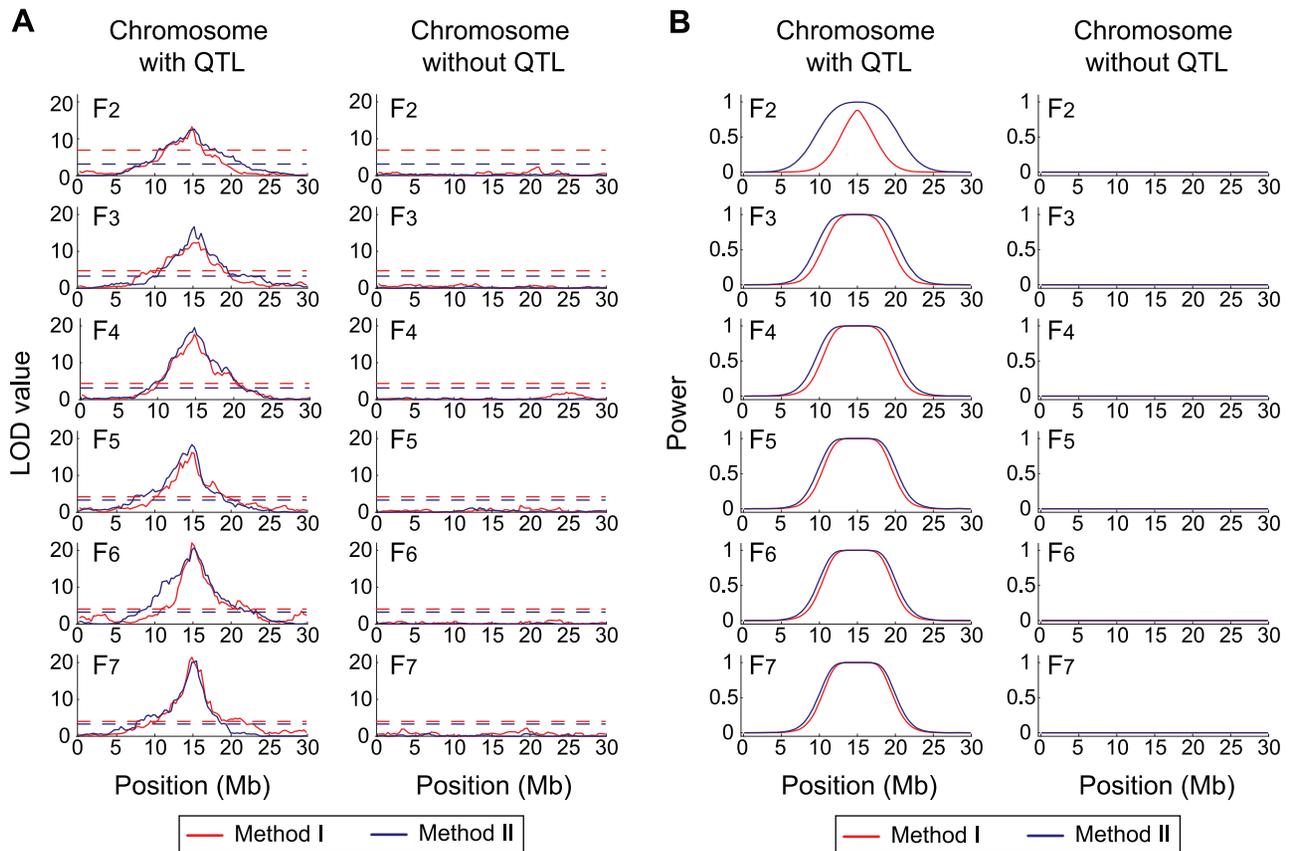


Figure 1. Evaluating the power of QTL mapping by simulations. **a** The distributions of the LOD values at markers along chromosomes (left: the chromosome with the QTL, right: a chromosome representing the other chromosomes without the QTL). The QTL is located at the middle of the chromosome (left panel). The red and blue lines show the LOD scores of Method I and Method II, respectively. The result is from a single replication of the simulation with $N = 200$, $d = 0$, and $\sigma^2 = 2$. The 5% cutoff values for the two methods are shown by broken lines. **b** The distributions of the power of the two methods, which were obtained by 10,000 replications.
 doi:10.1371/journal.pone.0046545.g001

Methods I and II provide the highest LOD score around the QTL, creating a sharp peak, whereas the LOD scores on all other chromosomes are low (plot for one representative chromosome is shown in the right panel in Figure 1a). We confirmed that similar patterns hold for all simulated parameter sets unless σ^2 is very large.

We found that there are at least two notable observations in Figure 1. (i) The distributions of LOD scores do not change much through generations, suggesting that significant power of detecting QTL may be expected even in early generations. If so, QTL mapping does not necessarily require many generations of inbreeding, so that a huge amount of time and cost could be saved. (ii) The performance of Method II exceeds that of Method I in many cases, especially at early generations. Method II is a simplified method that does not use heterozygous markers, whereas Method I uses all samples. It is suggested that the simpler method without considering the dominance effect (Method II) may be more efficient even with an obvious drawback of reducing sample size. These two observations have significant implications that $F_3 \sim F_4$ populations could have reasonable power for QTL mapping and that Method II would perform better at such early generations.

In order to quantitatively evaluate these hypotheses, we investigated the power of QTL mapping. The right panel of Figure 1 summarizes the results of 10,000 replications of the

simulations with the same parameters as those used for the left panel. The power was computed for each SNPs, which is defined as the proportion of the replications, in which the LOD score is significant at the 5% level ($\alpha < 0.05$, after correcting for multiple testing). The spatial distributions of the power support our two hypotheses; the performance of Method II (blue line) overall exceeds that of Method I (red line) and the power at F_4 is almost comparable to that at F_7 .

Further simulations with wide ranges of parameters were carried out to confirm if this holds. The results are summarized in Figure 2. In this figure, we mainly focus on how the environmental variance (σ^2) affects the power in two sample sizes, $N = 200$ and 1,000. We also considered two cases: no dominance ($d = 0$) and complete dominance ($d = 1$). We used a wide range of $\sigma^2 = \{2.4, 5, 9.5, 19.5, 49.5, 99.5\}$ (the corresponding heritability at the F_2 generation are $Q_{PIF_2} = \{20\%, 10\%, 5\%, 2.5\%, 1\%, 0.5\%\}$), and partial results are shown in Figure 2 such that the power at F_7 distributes roughly from 0.1 to 1. The power is here defined as the proportion of simulation replications in which the LOD scores of both of the two closest markers to the QTL are significant at the 5% level (after correcting for multiple testing). As the power is overall much higher when $N = 1,000$, we found that the QTL can be detected with probability ~ 1 when σ^2 is smaller than 9.5 (that is, larger heritability; Figure 2B), while the QTL with $\sigma^2 = 9.5$

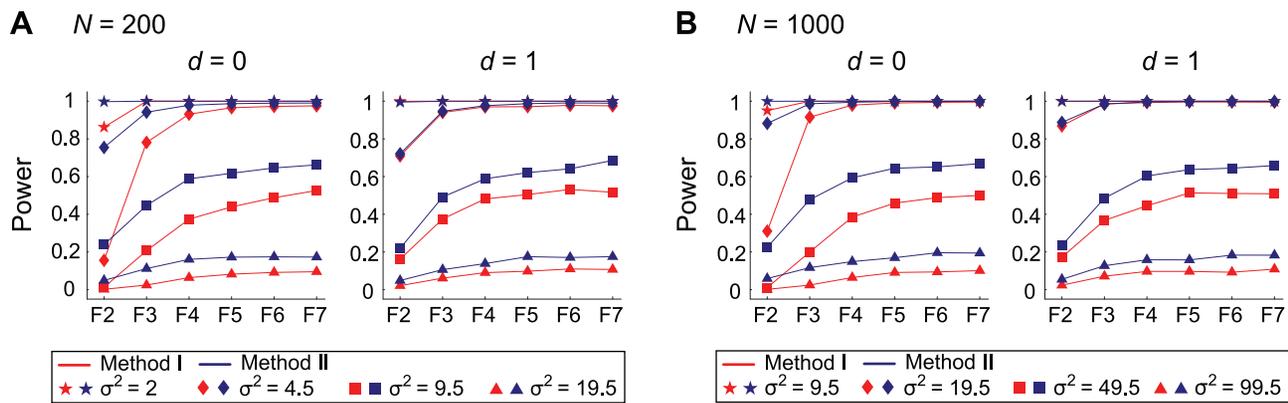


Figure 2. The power of QTL mapping as a function of the number of inbreeding generations. **a** $N=200$ and **b** $N=1000$ are assumed. The red and blue lines are for the results of Method I and Method II, respectively. See text for details.
doi:10.1371/journal.pone.0046545.g002

would be detected with probability roughly 0.5 when $N=200$ (Figure 2A).

These simulations supported that our two hypotheses hold with these wide ranges of parameters. For all the parameter sets, the performance of Method II exceeds that of Method I especially at early generations and the power of Method II at F_4 is almost comparable to that of F_7 . These seem to be true regardless of the degree of dominance. It should be noted that as mentioned earlier, the power is measured by a permutation applied to each data set, so that the false positive rate is always controlled to be 5% for all parameter sets. Therefore, the comparison of power is statistically fair.

In Figure 3, we investigated the effects of other parameters including the recombination rate, genome size, and marker density. It is found that overall the effects of these parameters are small. In Figure 3A, the power is shown for the recombination rate is changes from $R=1$ to $R=8$, while all other parameters remained the same as those used for Figures 1 and 2A. The panel in the broken square is identical to Figure 2A. In Figure 3B, the effect of genome size is investigated. Because our initial setting may be applied to species with small genomes such as Arabidopsis and rice, the genome size is increased up to 4 Gb, which is almost as large as maize and wheat. In Figure 3C, the marker density is reduced to up to 10 times. We found that the overall patterns are similar to one another, although the power becomes relatively weak when marker density is low (the leftmost panel in Figure 3C, and also see the leftmost panel of Figure 3B). There also seems to be a weak negative correlation between the power and the recombination rate (Figure 2A). Thus, our conclusion could be robust to these parameters.

These results are for the cases of relatively normal settings with additive phenotype effect at the focal QTL. However, there are cases where this does not obviously hold. One example is overdominance. Suppose the phenotypic value of heterozygote individuals at the focal QTL are expected to be larger than those of homozygotes. Such a situation can be realized by setting $d > 1$, so that the expected phenotype value for Q_1Q_2 heterozygotes exceeds that of Q_1Q_1 homozygotes (Q_2Q_2 homozygotes always have smallest values. See equations (1–3)). To investigate the power of the two methods under this setting, we repeated the same power simulations by assuming $d=1.5$ and 2. (we don't need to mention $\mu=0$, $a=1$ and $\sigma^2=2\sim 99.5$ if they are identical to those above.) With these settings, because the phenotype of heterozygotes are very informative to identify the QTL, the overall

performance of Method I is quite good (Fig. 4A). This is remarkable especially in earlier generations, but the situation becomes similar to those with the QTL with the additive phenotype effect as the number of generation increases because almost all individuals become homozygotes. This pattern is remarkable in the extreme case, symmetric overdominance, where $a=0$ is given so that the expected phenotype values of Q_1Q_1 and Q_2Q_2 homozygotes are identical and the phenotype of heterozygotes exceeds homozygotes by d (Fig. 4B). In earlier generations, Method I works fairly well, but the power is almost zero in F_6 and F_7 because almost all individuals are homozygotes, either Q_1Q_1 and Q_2Q_2 , between which there is no difference in phenotype.

Linked QTLs

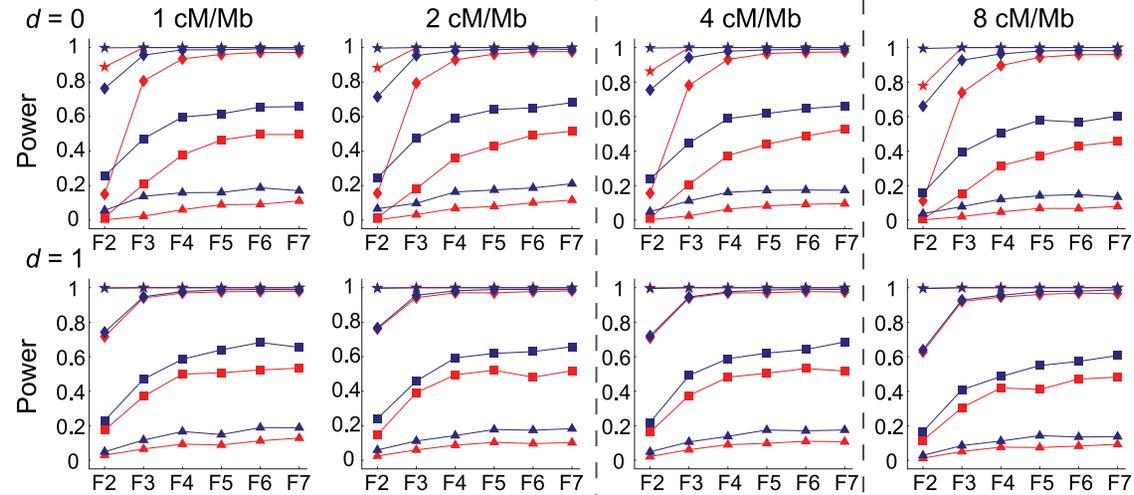
We also consider a more complicated model, in which there are QTLs that are linked to the focal QTL. It should be noted that our basic model described above takes into account the effect of multiple QTLs, whose effects are included in the third term of the right-hand side of equations (1–3). The assumption was that those QTLs are not linked to the focal QTL. We here investigate the effect of linked QTLs to the focal QTL.

We use a simple two-locus model. The alleles from P_1 at the two loci are denoted by $Q_{I,1}$ and $Q_{II,1}$ and those from P_2 are denoted by $Q_{I,2}$ and $Q_{II,2}$. μ , a and σ^2 were set such that their Q_{PIF_2} are 20% and 10%, respectively, in the codominance case. Other parameters follow those used in the earlier simulations for Figure 2A. These two QTL are linked, and four different distances between them were considered ($\{30,21,12,3\}$ Mb). No epistasis between QTLs was assumed.

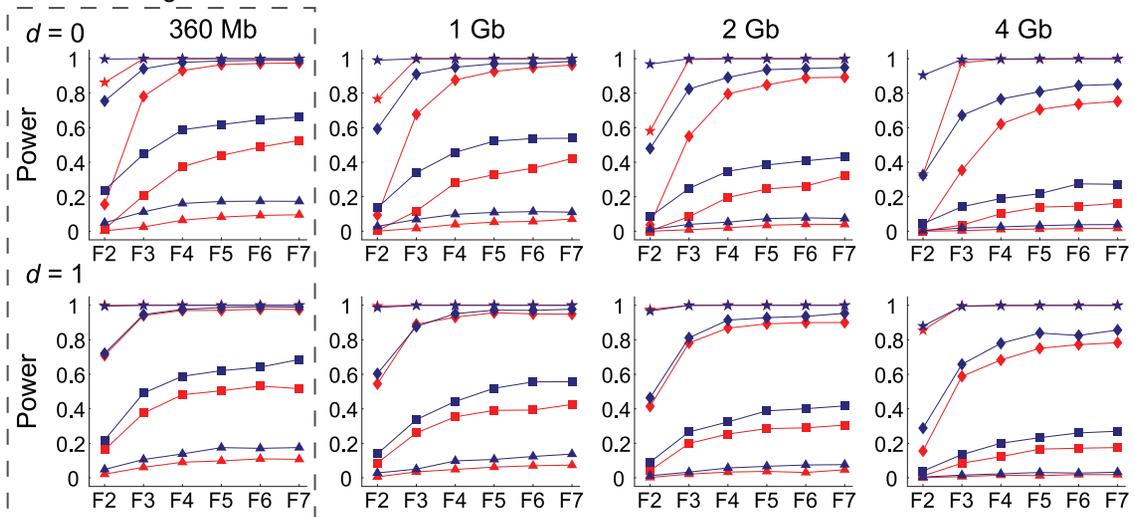
We first consider the cases of coupling phenotype effect, that is, both of the two alleles from P_1 (*i.e.*, $Q_{I,1}$ and $Q_{II,1}$) have positive effects on the phenotype. The results are summarized in Figure 5A, which shows the power to detect each QTL in the codominance and dominance cases. The overall patterns are quit similar to each other. When the distance is short (12 and 3 Mb), we observe very high power because the two QTLs behave almost as a single QTL with relative contribution $\sim 30\%$. As the distance increases, the power decreases because of recombination. If the distance is significantly long (*i.e.*, $\gg 30$ Mb), the two QTLs behave almost independently, so that the power to detect them should become comparable to those shown in Figure 2A. The performance of Method II is better than Method I in all cases.

Figure 5B shows the power when the phenotype effects of the two QTLs are decoupling or repulsion, that is, $Q_{I,1}$ and $Q_{II,2}$ have

A Effect of recombination rate



B Effect of genome size



C Effect of marker density

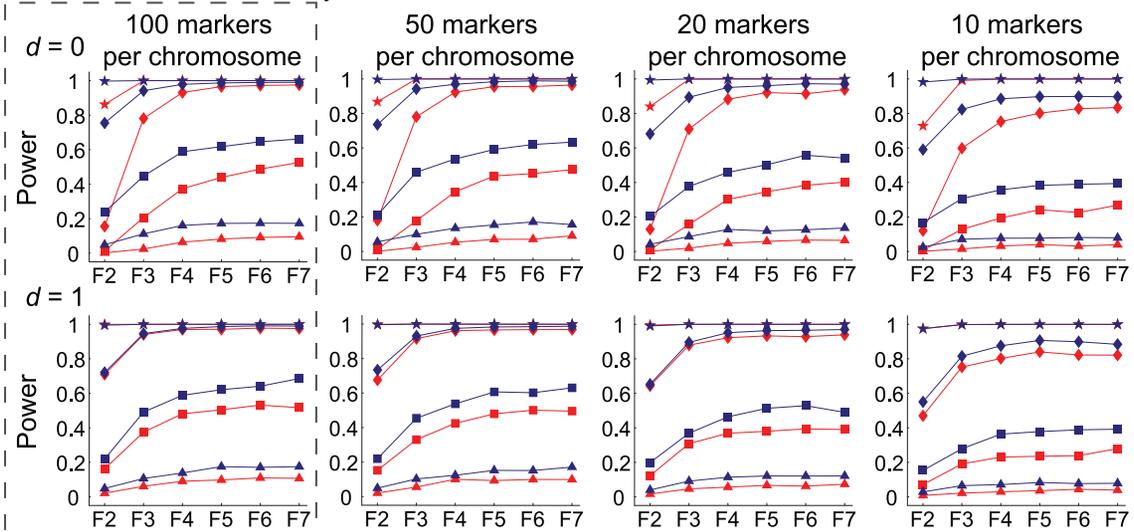


Figure 3. The effects of recombination rate (a), genome size (b) and marker density (c) on the power of QTL mapping. The panels in broken squares are identical to Fig. 2a.
doi:10.1371/journal.pone.0046545.g003

positive effects on the phenotype. We first consider the codominance case. Because alleles with positive and negative effects are initially coupled, the power is much more lower than in the case of coupling (Figure 5A). Recombination between the two QTLs creates coupling haplotypes, $Q_{J,1}Q_{J,1}$ and $Q_{J,2}Q_{J,2}$, thereby increasing the power. Indeed, the power increases with increasing the number of generations and the distance between the QTLs. The performance of Method II is overall better than Method I.

The pattern is more complicated in the dominance case. With few recombinations (*i.e.*, in younger generations with short distance), heterozygote individuals have the largest phenotypic values, so that they are very informative. This is why we observe higher performance of Method I. When the distance is 3 Mb, the power of Method I in F_2 is almost one because of the striking difference between homozygotes and heterozygotes. As more recombination events occur, the advantage of Method I is getting smaller, and the pattern becomes similar to the codominance case.

Thus, when there are multiple QTLs especially with dominance effect and/or epistatic interaction, the relationship between the phenotype parameters (σ^2) and the power is complicated. In such a case, it is quite common that we observe a single peak of high LOD scores encompassing the two QTLs. In a practical case, the problem would be that it is very difficult to know whether a single peak of the LOD score involves only one QTL or multiple QTLs. To distinguish these cases, further breeding should be required. For example, see refs. [23,24].

Discussion

QTL mapping plays significant roles to identify genetic regions responsible to important phenotype variation. One of the common strategies of QTL mapping uses a large number of RILs, which are established for at least several generations of inbreeding (typically up to F_6 or F_7). We here used simulations to quantitatively evaluate the performance of QTL mapping using RILs. Under the simple model with one focal QTL, it was found

that the performance of QTL mapping with F_4 population could be almost comparable to that with F_6 or F_7 populations (Figures 2 and 3). It was also found that Method II has more power than Method I especially at earlier generations. Method II is a simplified version of Method I, and it does not involve the process to estimate the dominance parameter, d . An obvious drawback of Method II is a reduction of sample size because it discards marker-heterozygote samples. For example, roughly 25% and 12.5% of RILs are excluded at F_3 and F_4 , respectively. Nevertheless, the performance of Method II exceeds that of Method I, suggesting that the uncertainty of d might reduce the power of Method I. Thus, our results imply that QTL mapping does not necessarily require RILs at F_6 or F_7 generations; rather, F_4 (or even F_3) populations would be almost as useful as F_6 or F_7 populations. Although we quit the simulations at F_7 , it is expected that the results for further generations can be intuitively understood; Because the power is almost saturated at $F_6 \sim F_7$ for many parameter sets, the power for $F_{>7}$ cannot be much larger than that of F_7 . Only when the power is still increasing at F_7 , more power is expected for $F_{>7}$, but it would eventually saturate in a few generations. Soller and Beckmann [12] suggested relatively little gain of the power by increasing the number of inbreeding generations when heritability is large, based on their theoretical analysis under a two-locus model (*i.e.*, QTL vs. marker). Our simulations support their implication in more practical situations with a number of markers for a wide range of σ^2 . While we only simulated RILs with selfing, these conclusions should hold for RILs with sibling mating, which was confirmed by a limited amount of additional simulations. We found that the only effect of sibling mating is that the decrease of heterozygous loci is slightly retarded (data not shown).

Further simulations under various conditions were performed (Figure 3) to investigate the effects of the parameters that were fixed in the basic simulation for Figures 2 and 3. The investigated parameters are the recombination rate, genome size, and marker density, while the sample size was fixed to be 200. It was found that these factors have relatively minor effects on the results,

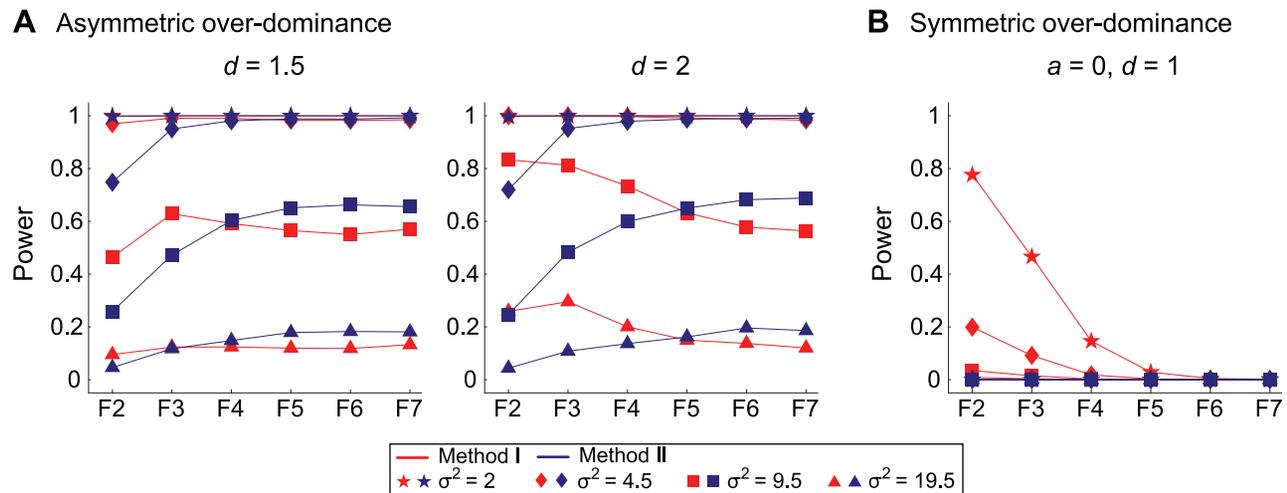


Figure 4. The power of QTL mapping with over-dominance. (a) Asymmetric over-dominance. (b) Symmetric over-dominance. See text for details.
doi:10.1371/journal.pone.0046545.g004

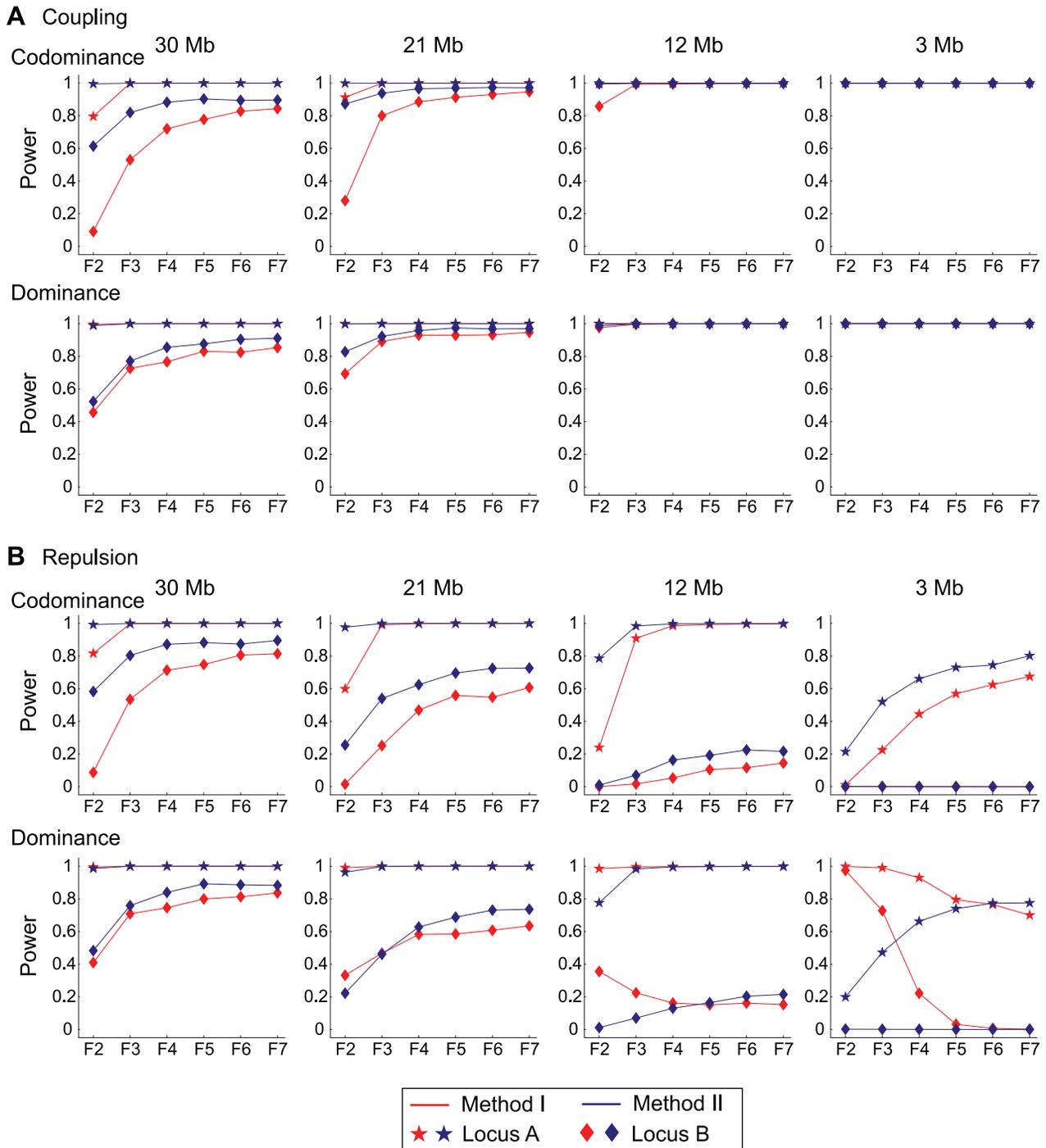


Figure 5. The power to detect two linked QTLs under the two-locus model. (a) Results when the two QTLs have a coupling effect. **(b)** Results when the two QTLs have a decoupling effect. The power is shown for each QTL. See text for details. doi:10.1371/journal.pone.0046545.g005

indicating that our conclusions should hold under wide ranges of the parameters. It was surprising that the power did not decrease much when we have only 10 markers on a 30 Mb (120 cM) of chromosome. An implication is that in order to reduce the cost, a reasonable level of power could be expected when there are roughly every 10 cM.

In contrast, it seems that the effect of the sample size is much larger than those of the factors explored in Figure 3. As shown in

Figure 2, QTLs with much larger σ^2 can be detected when $N=1000$ in comparison with the case of $N=200$. Increasing sample size is costly, may be as much as extending inbreeding generations, but our results imply that the former may be more efficient than the latter. We would suggest that increasing the sample size is one of the best strategies to improve the performance rather than continuing inbreeding for many generations. Because the cost to establish a number of RILs for many generations is

enormous, it is important to understand the relationship between the cost and output. Our results provide several ideas to obtain better performance with a limited cost, there by accelerating gene mapping in many species.

In summary, we demonstrated that our idea of ignoring heterozygotes (incorporated in Method II) works quite well in a relatively simple situations. The major difference between the two methods is that Method I has an additional parameter (d) that has to be estimated from data. Our demonstration might indicate that simple methods with no estimation process work well. In this sense, one might think that a linear regression analysis might also work well [13,14]. However, although this analysis does not involve estimation of the dominance parameter, it assumes a certain level of dominance (most commonly no dominance). Therefore, when the true dominance parameter is different from the assumption, the power might be reduced. In other words, it still involves uncertainty of the dominance parameter. As expected, we confirmed that the performance of the linear regression analysis did not exceed that of Method II for all parameter range (data not shown). Our Method II provides a general framework in evaluating likelihood ignoring heterozygote. This can be readily incorporated in the interval mapping method [1–6], or recently developed more computationally sophisticated QTL mapping algorithms, such as, Bayesian shrinking method e.g., [25,26] and penalized maximum likelihood e.g., [27].

We mainly obtained these conclusions under a simple model with one focal QTL, but they can be applied to broad cases because the model does not necessarily assumes that there is only one QTL in the genome. We simply focused on a single QTL with its phenotype effect specified by parameter σ^2 (the effects of other QTLs are included in the environmental factors, ε , in equations 1–3). Therefore, as long as the focal QTL is not linked to other QTLs, our conclusions should hold. We confirmed this by additional simulations in a model allowing multiple QTLs with various quantitative effects, although too obvious theoretically.

It should be noted that there are some cases where the performance of Method I exceeds that of Method II, as demonstrated in Figures 4 and 5. The consensus of these cases is that the phenotype of heterozygotes is informative. One is the case of overdominance, where the performance of Method I is much better in earlier generations because there are a number of heterozygotes. The situation is similar when there are two linked QTLs that have decoupling phenotype effects with complete

dominance. Also in this case, the phenotype value of double heterozygotes is the highest, Method I performs well particularly in earlier generations. We should keep in our mind that our major conclusions may not hold in these cases (may not be very common though).

Conclusions

QTL mapping plays significant roles to identify genetic regions responsible to important phenotype variation. One of the common strategies of QTL mapping uses a large number of RILs, which are established for at least several generations of inbreeding (typically up to F_6 or F_7). We here used simulations to quantitatively evaluate the performance of QTL mapping using RILs. It was found that the performance of QTL mapping with F_4 population could be almost comparable to that with F_6 or F_7 populations (Figs. 2 and 3). It was also found that Method II has more power than Method I especially at earlier generations. Method II is a simplified version of Method I, and it does not involve the process to estimate the dominance parameter, d . An obvious drawback of Method II is a reduction of sample size because it discards marker-heterozygote samples. For example, roughly 25% and 12.5% of RILs are excluded at F_3 and F_4 , respectively. Nevertheless, the performance of Method II exceeds that of Method I, suggesting that the uncertainty of \hat{d} might reduce the power of Method I. Thus, our results imply that in most cases, QTL mapping may not necessarily require RILs at F_6 or F_7 generations; rather, F_4 (or even F_3) populations would be almost as useful as F_6 or F_7 populations. Because the cost to establish a number of RILs for many generations is enormous, this finding will cause a reduction in the cost of QTL mapping, thereby accelerating gene mapping in many species.

Acknowledgments

The authors thank S. Shiokai and the anonymous reviewers for various comments.

Author Contributions

Conceived and designed the experiments: HI RT. Performed the experiments: ST. Analyzed the data: ST. Contributed reagents/materials/analysis tools: ST. Wrote the paper: ST HI.

References

- Weir BS, Eisen EJ, Goodman MM, Namkoong G (1987) Proceedings of the Second International Conference on Quantitative Genetics. Sunderland, MA: Sinauer Associates.
- Tanksley SD (1993) Mapping polygenes. *Annu Rev Genet* 27: 205–233.
- Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics, Ed 4. Harlow, Essex, UK: Longmans Green.
- Lynch M, Walsh JB (1998) Genetics and Analysis of Quantitative Traits. Sunderland, MA: Sinauer Associates.
- Lander ES, Botstein D (1986) Mapping complex genetic traits in human: New methods using a complete RFLP linkage map. *Cold Spring Harbor Symp on Quant Biol* 51: 49–62.
- Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121: 185–199.
- Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, et al. (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* 335: 721–726.
- Singer T, Fan Y, Chang HS, Zhu T, Hazen SP, et al. (2006) A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genet* 2: e144.
- Paterson AH (2006) Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nat Rev Genet* 7: 174–184.
- Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. *Heredity* 101: 5–18.
- Jansen RC (2003) Quantitative trait loci in inbred lines. In: Balding DJ, Bishop M, Cannings C, editors. *Handbook of Statistical Genetics*. Chichester, UK: John Wiley & Sons. 589–618.
- Soller M, Beckmann JS (1990) Marker-based mapping of quantitative trait loci using replicated progenies. *Theor Appl Genet* 80: 205–208.
- Hu Z, Xu S (2008) A simple method for calculating the statistical power for detecting a QTL located in a marker interval. *Heredity* 101: 48–52.
- Kao CH, Zeng MH (2010) An investigation of the power for separating closely linked QTL in experimental populations. *Genet Res (Camb)* 92: 283–294.
- Emerson S (1969) Linkage and recombination at the chromosome level. In: Caspari EW, Ravin AW, editors. *Genetic Organization*. New York & London: Academic Press. 267–360.
- Haldane J (1931) The cytological basis of genetical interference. *Cytologia* 3: 54–65.
- Mather K (1937) The determination of position in crossing-over. II. The chromosome length-chiasma frequency relation. *Cytologia Fujii Jubilee Vol*: 514–526.
- Henderson SA (1963) Chiasma distribution at diplotene in a locust. *Heredity* 18: 173–190.
- Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69: 315–324.
- Hayashi T, Ukai Y (1994) Detection of additive and dominance effects of QTLs in interval mapping of F2 RFLP data. *Theor Appl Genet* 87: 1021–1027.

21. Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. *Genetics* 138: 963–971.
22. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436: 793–800.
23. Zhang L, Li H, Li Z, Wang J (2008) Interactions between markers can be caused by the dominance effect of quantitative trait loci. *Genetics* 180: 1177–1190.
24. Li H, Hearne S, Bänziger M, Li Z, Wang J (2010) Statistical properties of qtl linkage mapping in biparental genetic populations. *Heredity* 105: 257–267.
25. Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819–1829.
26. Xu S (2003) Estimating polygenic effects using markers of the entire genome. *Genetics* 163: 789–801.
27. Zhang YM, Xu S (2005) A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity* 95: 96–104.