OXFORD
G3 Genes | Genomes | Genetics

G3, 2022, **12(1),** jkab370
https://doi.org/10.1093/g3journal/jkab370
Advance Access Publication Date: 15 November 2021
Investigation

Optimization of BSA-seq experiment for QTL mapping

Likun Huang,[1] Weiqi Tang,[2] and Weiren Wu [1,*]

[1]Fujian Key Laboratory of Crop Breeding by Design, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China, and
[2]Institute of Oceanography, Marine Biotechnology Center, Minjiang University, Fuzhou, Fujian 350108, China

*Corresponding author: Email: wuwr@fafu.edu.cn

Abstract

Deep sequencing-based bulked segregant analysis (BSA-seq) has become a popular approach for quantitative trait loci (QTL) mapping in recent years. Effective statistical methods for BSA-seq have been developed, but how to design a suitable experiment for BSA-seq remains unclear. In this paper, we show in theory how the major experimental factors (including population size, pool proportion, pool balance, and generation) and the intrinsic factors of a QTL (including heritability and degree of dominance) affect the power of QTL detection and the precision of QTL mapping in BSA-seq. Increasing population size can improve the power and precision, depending on the QTL heritability. The best proportion of each pool in the population is around 0.25. So, 0.25 is generally applicable in BSA-seq. Small pool proportion can greatly reduce the power and precision. Imbalance of pool pair in size also causes decrease of the power and precision. Additive effect is more important than dominance effect for QTL mapping. Increasing the generation of filial population produced by selfing can significantly increase the power and precision, especially from F_2 to F_3. These findings enable researchers to optimize the experimental design for BSA-seq. A web-based program named BSA-seq Design Tool is available at http://124.71.74.135/BSA-seqDesignTool/ and https://github.com/huanglikun/BSA-seqDesignTool.

Keywords: BSA-seq; QTL; power; precision; influencing factor; experimental design

Introduction

Bulked segregant analysis based on deep sequencing (BSA-seq) is an efficient and cost-effective approach for rapid mapping of quantitative trait loci (QTLs). Since it was first reported in yeast (Ehrenreich *et al.* 2010), this approach has been widely applied to many different species especially in plants, such as rice (Arikit *et al.* 2019; Lei *et al.* 2020), wheat (Xin *et al.* 2020), tomato (Ruangrak *et al.* 2019), groundnut (Pandey *et al.* 2017), chickpea (Deokar *et al.* 2019), sunflower (Imerovski *et al.* 2019), squash (Ramos *et al.* 2020), watermelon (Branham *et al.* 2018), cricket (Pascoal *et al.* 2014), and Hessian fly (Navarro-Escalante *et al.* 2020). To facilitate BSA-seq for QTL mapping, a number of different statistical methods have been proposed, such as G′ test (Magwene *et al.* 2011), MULTIPOOL (Edwards and Gifford 2012), EXPLoRA (Duitama *et al.* 2014), Hidden Markov Model (Claesen and Burzykowski 2015), and Nonhomogeneous Hidden Markov Model (Ghavidel *et al.* 2015).

Recently, we proposed a new statistical method named block regression mapping (BRM) for BSA-seq (Huang *et al.* 2020). The method uses the simple and intuitional allele frequency difference (AFD) between two pools as statistic to test putative QTLs. Most importantly, it proves that smoothing by block regression can effectively remove the noise of sequencing (*i.e.*, the random error of resampling) and the expected AFD at a genomic position estimated by block regression is close to the actual AFD at the position even under very low sequencing depth. This means that the variation of AFD is basically determined by the size of the two pools. Based on this fact, the method reasonably resolves the problem of multiple testing correction in the estimation of significance threshold in BSA-seq, and can obtain both the point estimate and the 95% confidence interval (CI) of a QTL's position. In addition, with the expected AFD of a QTL obtained by BRM, the proportion of variance explained by the QTL, or termed the QTL's heritability, can be estimated using a method named Pooled QTL Heritability Estimater (PQHE; Tang *et al.* 2018).

Apart from the statistical method, experimental design is also very important for BSA-seq. An appropriate experimental design can effectively increase the statistical power of QTL detection and the precision of QTL mapping. At present, however, how to make a suitable experimental design for BSA-seq is still a problem to be solved. The size of population used in the BSA-seq experiments so far varies from very small (<100; Deokar *et al.* 2019; Imerovski *et al.* 2019) to very large (>10,000; Yang *et al.* 2013), with most being small (around 200; Das *et al.* 2014; Pandey *et al.* 2017; Branham *et al.* 2018; Luo *et al.* 2018; Arikit *et al.* 2019; Lahari *et al.* 2019; Ramos *et al.* 2020), and some medium (around 500; Xin *et al.* 2020) or large (near or over 1000; Ruangrak *et al.* 2019; Lei *et al.* 2020). The pool size used is also very diverse, varying from very small (only five individuals; Branham *et al.* 2018) to very large (>400; Yang *et al.* 2013), corresponding to a pool proportion (the ratio of pool size to population size) of ∼10% or less in most experiments.

In this paper, based on the principle of the BRM method, we investigate the factors that influence the power (of QTL detection)

Received: September 15, 2021. **Accepted:** October 19, 2021

and the precision (of QTL mapping) in BSA-seq through theoretical derivation and numerical calculation as well as analysis on experimental data from yeast. According to the effects of these influencing factors, it is possible to optimize the experimental design for BSA-seq.

Materials and methods

Experimental design for BSA-seq

There are various kinds of populations derived from a bi-parent cross ($P_1 \times P_2$) for BSA-seq, including temporal populations such as F_k ($k = 2, 3, 4\ldots$) populations, and permanent populations such as recombinant inbred line (RIL) population, doubled haploid (DH) population, and haploid (H) population. F_k generation can be produced from F_{k-1} generation either by selfing or by intercrossing (or random mating). These two different mating ways in F_{k-1} generation will result in different genetic structures of the F_k population when $k \geq 3$. In this paper, we mainly analyze the situation of using the F_k population produced by selfing, and the term of F_k only refers to this type unless otherwise mentioned. Among the permanent populations, H population is usually used in fungi (*e.g.*, yeast), of which the life cycle is dominated by gametophyte generation. A pair of distinct DNA pools, namely, high-trait-value (H) pool *vs.* low-trait-value (L) pool (design A) or selected (S) pool *vs.* random (R) pool (design B), is established from the population and deeply sequenced. By mapping the sequencing reads to a reference genome, a very large number of molecular markers (mainly SNPs) and their counts in each pool can be found. QTL mapping is performed by comparing the two pools based on the marker data (Huang *et al.* 2020). In this paper, we shall focus on the optimization of design A. The principle should be also applicable to design B.

Expectation of the power of QTL detection

Suppose a trait (*y*) is controlled by a QTL with two different alleles, Q from P_1 and *q* from P_2. The trait variation in an F_k or a permanent population can be described as a mixture distribution as below:

$$\begin{aligned} f(y) &= bf_{QQ}(y) + (1-2b)f_{Qq}(y) + bf_{qq}(y) \\ &= b\phi\left[\frac{y-(\mu-a)}{\sigma_e}\right] + (1-2b)\phi\left[\frac{y-(\mu+d)}{\sigma_e}\right] + b\phi\left[\frac{y-(\mu+a)}{\sigma_e}\right] \end{aligned} \tag{1}$$

where *a* and *d* are the additive effect and dominance effect of the QTL, respectively; μ is the population mean; σ_e is the standard deviation of the background (including genetic background and environment) variation; $\phi(\cdot)$ is the probability density function of standard normal distribution; and $b = \left[1 - (1/2)^{k-1}\right]/2$, where $k \to \infty$ in a permanent population.

Let $x = (y - \mu)/\sigma_e$, $a_0 = a/\sigma_e$, and $d_0 = d/\sigma_e$. Equation (1) can be rewritten as:

$$f(x) = b\phi(x + a_0) + (1-2b)\phi(x - d_0) + b\phi(x - a_0). \tag{2}$$

According to Equation (2), the additive effect heritability (h_a^2) and the dominance effect heritability (h_d^2) of the QTL are (Tang *et al.* 2018):

$$h_a^2 = \frac{2ba_0^2}{1 + 2ba_0^2 + 2b(1-2b)d_0^2} \tag{3}$$

$$h_d^2 = \frac{2b(1-2b)d_0^2}{1 + 2ba_0^2 + 2b(1-2b)d_0^2}. \tag{4}$$

And the total heritability of the QTL is: $h^2 = h_a^2 + h_d^2$. If the QTL does not exist, namely, the null hypothesis ($H_0: a_0 = d_0 = 0$) is true, the mixture distribution $f(x)$ will degenerate into a standard normal distribution $\phi(x)$.

Suppose the proportions of the H pool and the L pool in the population are p_H and p_L, and the corresponding cut points for the H pool and the L pool are x_H and x_L, respectively. According to Equation (2), we find:

$$p_\mathrm{H} = 1 - b\Phi(x_\mathrm{H} + a_0) - (1-2b)\Phi(x_\mathrm{H} - d_0) - b\Phi(x_\mathrm{H} - a_0) \tag{5}$$

$$p_\mathrm{L} = b\Phi(x_\mathrm{L} + a_0) + (1-2b)\Phi(x_\mathrm{L} - d_0) + b\Phi(x_\mathrm{L} - a_0), \tag{6}$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. Equations (5) and (6) indicate that p_H and p_L are determined when x_H and x_L are given and vice versa. According to Equations (2), (5), and (6), the allele frequency (AF, referring to the allele from P_1) of the QTL in the H pool and that in the L pool are expected to be:

$$\mu_{f_\mathrm{H}} = \frac{1 - 2b\Phi(x_\mathrm{H} - a_0) - (1-2b)\Phi(x_\mathrm{H} - d_0)}{2p_\mathrm{H}} \tag{7}$$

$$\mu_{f_\mathrm{L}} = \frac{2b\Phi(x_\mathrm{L} - a_0) + (1-2b)\Phi(x_\mathrm{L} - d_0)}{2p_\mathrm{L}}. \tag{8}$$

Let $p = (p_\mathrm{H} + p_\mathrm{L})/2$ and $\gamma = p_\mathrm{H}/p_\mathrm{L}$, we have $p_\mathrm{H} = 2p\gamma/(1+\gamma)$, and $p_\mathrm{L} = 2p/(1+\gamma)$. Thus, the AFD of the QTL between the two pools is expected to be:

$$\begin{aligned} \mu_{\Delta f} &= \mu_{f_\mathrm{H}} - \mu_{f_\mathrm{L}} \\ &= \frac{(1+\gamma)[1 - 2b\Phi(x_\mathrm{H} - a_0) - (1-2b)\Phi(x_\mathrm{H} - d_0)]}{4p\gamma} \\ &\quad - \frac{(1+\gamma)\left[2b\Phi(x_\mathrm{L} - a_0) + (1-2b)\Phi(x_\mathrm{L} - d_0)\right]}{4p}. \end{aligned} \tag{9}$$

According to the Central Limit Theorem, the sampled AFD will approximately follow a normal distribution with mean $\mu_{\Delta f}$ and variance

$$\begin{aligned} \sigma_{\Delta f}^2 &= \sigma_{f_\mathrm{H}}^2 + \sigma_{f_\mathrm{L}}^2 = \frac{\mu_{f_\mathrm{H}}(1-\mu_{f_\mathrm{H}})}{2^t p_\mathrm{H} n} + \frac{\mu_{f_\mathrm{L}}(1-\mu_{f_\mathrm{L}})}{2^t p_\mathrm{L} n} \\ &= \frac{1+\gamma}{2^{t+1}p\gamma n}\left[\mu_{f_\mathrm{H}}(1-\mu_{f_\mathrm{H}}) + \gamma\mu_{f_\mathrm{L}}(1-\mu_{f_\mathrm{L}})\right], \end{aligned} \tag{10}$$

where *n* is the size of the mapping population; $t = 0$ for permanent populations and $t = 1$ for F_k populations. However, if the QTL does not exist ($H_0: a_0 = d_0 = 0$), then $\mu_{f_\mathrm{H}} = \mu_{f_\mathrm{L}} = 0.5$ and $\mu_{\Delta f} = 0$ according to Equations (5)–(9). In this case, AFD will approximately follow a normal distribution with mean 0 and variance σ_0^2, where, according to Equation (10),

$$\sigma_0^2 = \sigma_{\Delta f}^2(\mu_{\Delta f} = 0) = \frac{(1+\gamma)^2}{2^{t+3}p\gamma n}. \tag{11}$$

So, the threshold of AFD at significance level α (two-tail test) is (Huang *et al.* 2020):

$$T_\pm = \pm u_{\alpha/2}\sigma_0 = \pm\frac{(1+\gamma)u_{\alpha/2}}{\sqrt{2^{t+3}p\gamma n}}. \tag{12}$$

where the ± sign indicates the upper/lower threshold; $u_{\alpha/2}$ is the upper percentile point of $\alpha/2$ in standard normal distribution, which is a constant for a species under a certain overall (genome-wise) significance level. Thus, the statistical power of detecting the QTL is expected to be (Figure 1):

$$\text{Power} = 1 - \Phi\left(\frac{T_+ - \mu_{\Delta f}}{\sigma_{\Delta f}}\right) + \Phi\left(\frac{T_- - \mu_{\Delta f}}{\sigma_{\Delta f}}\right). \tag{13}$$

Expectation of the precision of QTL mapping

The mapping precision of a QTL can be indicated by the size of its CI. A narrower CI means a higher mapping precision. Consider a position M linked with a given QTL. No matter in an F_2 or in an H/DH population, the AF at M in the H pool ($\mu_{f_{MH}}$) and that in the L pool ($\mu_{f_{ML}}$) are expected to be:

$$\mu_{f_{MH}} = (1-\theta)\mu_{f_H} + \theta(1-\mu_{f_H}) = \theta + (1-2\theta)\mu_{f_H} \tag{14}$$

$$\mu_{f_{ML}} = (1-\theta)\mu_{f_L} + \theta(1-\mu_{f_L}) = \theta + (1-2\theta)\mu_{f_L}, \tag{15}$$

where θ is the recombination rate between M and the QTL. Therefore, the AFD at M between the two pools ($\mu_{\Delta f_M}$) is expected to be:

$$\mu_{\Delta f_M} = \mu_{f_{MH}} - \mu_{f_{ML}} = (1-2\theta)(\mu_{f_H} - \mu_{f_L}) = (1-2\theta)\mu_{\Delta f}. \tag{16}$$

Equation (16) indicates that $\mu_{\Delta f_M}$ is a function of $\mu_{\Delta f}$ and θ, which describes the expected AFD curve around a QTL. It can be seen that $\mu_{\Delta f_M}$ varies between 0 (when $\theta = 0.5$, or M is far from the QTL) and $\mu_{\Delta f}$ (when $\theta = 0$, or M is just at the position of the QTL). Therefore, the AFD curve forms a positive peak (when $\mu_{\Delta f} > 0$) or negative peak (when $\mu_{\Delta f} < 0$) with the positive/negative top point being at the position of the QTL.

This AFD curve enables us to estimate the 95% CI (denoted as CI 95) of the QTL (Huang *et al.* 2020). Let $\mu_{\Delta f_M} = \mu_{\Delta f} - 1.65\sigma_{\Delta f}$ (in the case of $\mu_{\Delta f} > 0$) or $\mu_{\Delta f_M} = \mu_{\Delta f} + 1.65\sigma_{\Delta f}$ (in the case of $\mu_{\Delta f} < 0$). Substitute it into Equation (16), we find the recombination rate between the left (or right) border of CI 95 and the QTL:

$$\theta = \frac{0.825\sigma_{\Delta f}}{\mu_{\Delta f}}, \tag{17}$$

where $\mu_{\Delta f}$ and $\sigma_{\Delta f}$ are determined by Equations (9) and (10), respectively. By assuming Kosambi's mapping function and ignoring the influence of phenotypic selection on recombination rate (Lin and Ritland 1996), the corresponding genetic distance D (cM) would approximately be:

$$D = 25\ln\left(\frac{1+2\theta}{1-2\theta}\right). \tag{18}$$

Therefore, the width of CI 95 of the QTL is $2D$ (cM).

For F_3 and F_4 populations, the parameter θ in Equations (14)–(17) should be replaced with θ_3 and θ_4, the apparent recombination rates in F_3 and F_4, respectively, where (Huang *et al.* 2020):

$$\theta_3 = \theta\left[1 + \frac{1}{2}(1-\theta)^2\right] \tag{19}$$

$$\theta_4 = \theta\left[1 + \frac{1}{2}(1-\theta)^2 + \frac{1}{4}(1-\theta)^4\right]. \tag{20}$$

From Equations (19) and (20), the real recombination rate θ can be calculated and thus the width of CI 95 can be calculated from Equation (18).

Numerical analysis

Equations (13) and (18) describe the relationships of various factors (parameters) with the power of QTL detection and the precision of QTL mapping in BSA-seq using the BRM method. To display how the factors affect the power and precision, we used Equations (13) and (18) to analyze yeast H population (representing permanent populations) and rice F_2, F_3, and F_4 populations (representing F_k populations), respectively. In the analyses, the value of $u_{\alpha/2}$ in Equation (12) under the genome-wise significance level of 0.05 for yeast H population and those for rice F_2, F_3, and F_4 populations were 3.93 and 3.65, 3.74, and 3.78, respectively, of which the corresponding nominal significance level (α) was 8.49×10^{-5} and 2.62×10^{-4}, 1.84×10^{-4}, and 1.57×10^{-4}, respectively (Huang *et al.* 2020). For simplicity, equal pool size (namely, $\gamma = 1$ or $p_H = p_L = p$) was assumed except when the effect of pool size imbalance (*i.e.*, $\gamma \neq 1$ or $p_H \neq p_L$) was analyzed.

In addition, to demonstrate the influence of pool proportion in BSA-seq, we also used the BRM method (Huang *et al.* 2020) to analyze a series of simulated BSA-seq experiments based on the experimental data in an H population from yeast (Bloom *et al.* 2015). The data included the genotypes of 28,220 SNPs and phenotypes of trait GCS (end-point Growth on a medium containing Copper Sulfate) in 4276 segregants derived from a cross between a laboratory strain BY and a vineyard strain RM (Huang *et al.* 2020). We randomly extracted 4000 segregants from the total as the mapping population, from which a series of H *vs.* L pool pairs with different pool proportions (including 0.01, 0.05, 0.1, 0.2, 0.25, 0.3, 0.4, and 0.5) were made according to the GCS phenotype data. By dividing the genome into tandem 200-bp blocks, the actual AFD at

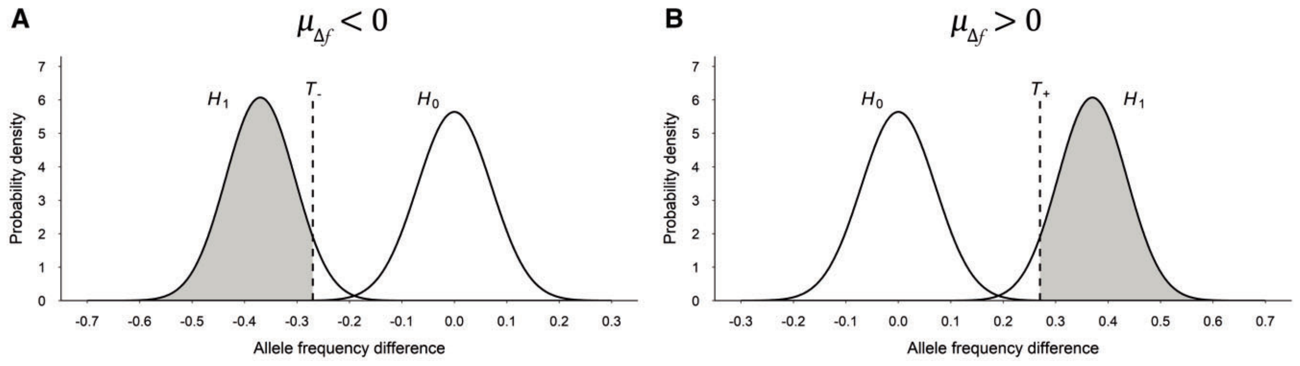

Figure 1 The power of QTL detection in BSA-seq using BRM. The AFD of the QTL follows a normal distribution with mean at AFD = 0 under the null hypothesis H_0 (no QTL) or a normal distribution with mean at AFD = $\mu_{\Delta f}$ under the alternative hypothesis H_1 (there is QTL). The power is equal to the shaded area (probability) under H_1 on the left of the threshold T_- when $\mu_{\Delta f} < 0$ (A) or on the right of the threshold T_+ when $\mu_{\Delta f} > 0$ (B).

every position (block) in the genome was calculated in each pool pair based on the SNP genotype data. The AFD thresholds at the genome-wise significance level of 0.05 were calculated according to the BRM method (Huang *et al.* 2020). It was considered that a QTL existed when the maximum AFD value in a peak region exceeded the threshold.

Results

Effect of population size

Population size (n) and pool proportion (p) are two major experimental factors affecting the power and precision in BSA-seq. When p is fixed, depending on the QTL heritability (h^2), the power and the CI 95 width display a series of S-shape curves (Figure 2, A and B) and L-shape curves (Figure 2, C and D) with n, respectively. So, approximately, the process of power increase along with the n increase could be divided into three stages: slow-fast→slow, while the process of CI 95 width decrease along with the n increase could be divided into two stages: fast→slow. Obviously, increasing n in the last stage is inefficient for power increase and CI 95 width decrease. This suggests that a value of n just before the start point of the last stage would be optimal. However, different QTL may have different optimal n, which is inversely proportional to h^2 (Figure 2). A larger h^2 would have a smaller optimal n.

Effect of pool proportion

Theoretically, p varies between 0 and 0.5. The extreme situation $P = 0.5$ means that the population is divided into two pools of equal size just at the mid-point of the trait. When n is fixed, it is seen that the power (Figure 3, A and B) and the CI 95 width (Figure 3, C and D) are neither a monotonic function of p. There is a peak of power and a valley of CI 95 width, respectively. The highest point of power is mainly located between $P = 0.25$ and $P = 0.3$, while the lowest point of CI 95 width is located around $P = 0.25$, depending on h^2. This suggests that 0.25 is a generally suitable, if not the best, value of p. Nonetheless, the peak top of the power and the valley bottom of the CI 95 width are broad and flat, especially when h^2 is large. Therefore, the suitable value of p can be flexible in a wider range. It is noticeable that small p has very strong unfavorable effects on the power and the CI 95 width. As p decreases toward zero, the power will quickly drop toward zero and the CI 95 width will soar toward infinite, no matter how large h^2 is. Therefore, it is inappropriate to use small p.

Effect of interaction between population size and pool proportion

When the intrinsic factor h^2 is fixed, it is seen that the basic feature of the relationship between the power and p and that between the CI 95 width and p (Figure 3) remain the same under different n (Figure 4). The value of $P = 0.25$ still appears to be generally suitable (Figure 4). However, the effects of p on the power and the CI 95 width are related to n. As n increases, the peak top of the power and the valley bottom of the CI 95 width will become wider and flatter. Therefore, the suitable value of p can be more flexible under larger n. It is noted that the peak top of the power also becomes flatter when n is very small (Figure 4, A and B). But in this case, the power is very low and therefore is meaningless.

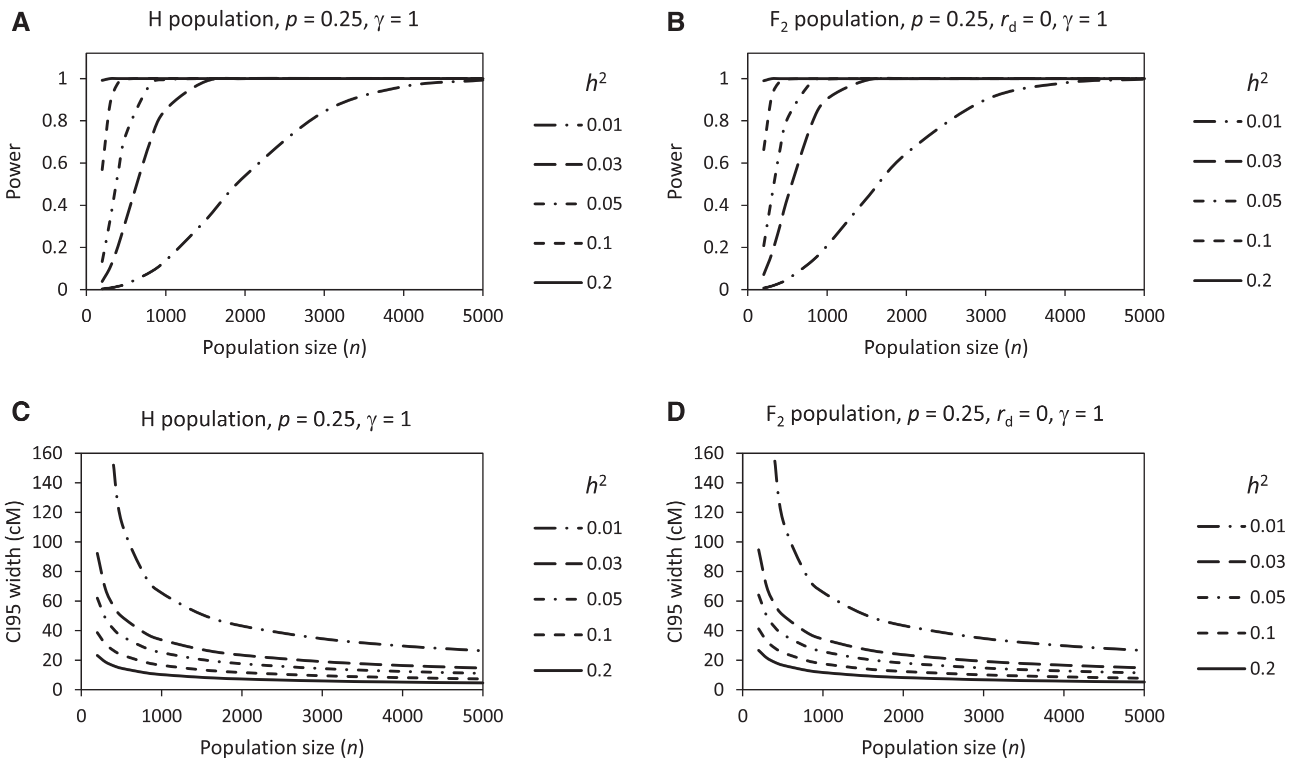

Figure 2 Relationships of power and CI 95 width with population size depending on QTL heritability in yeast H population (A, C) and rice F_2 population (B, D).

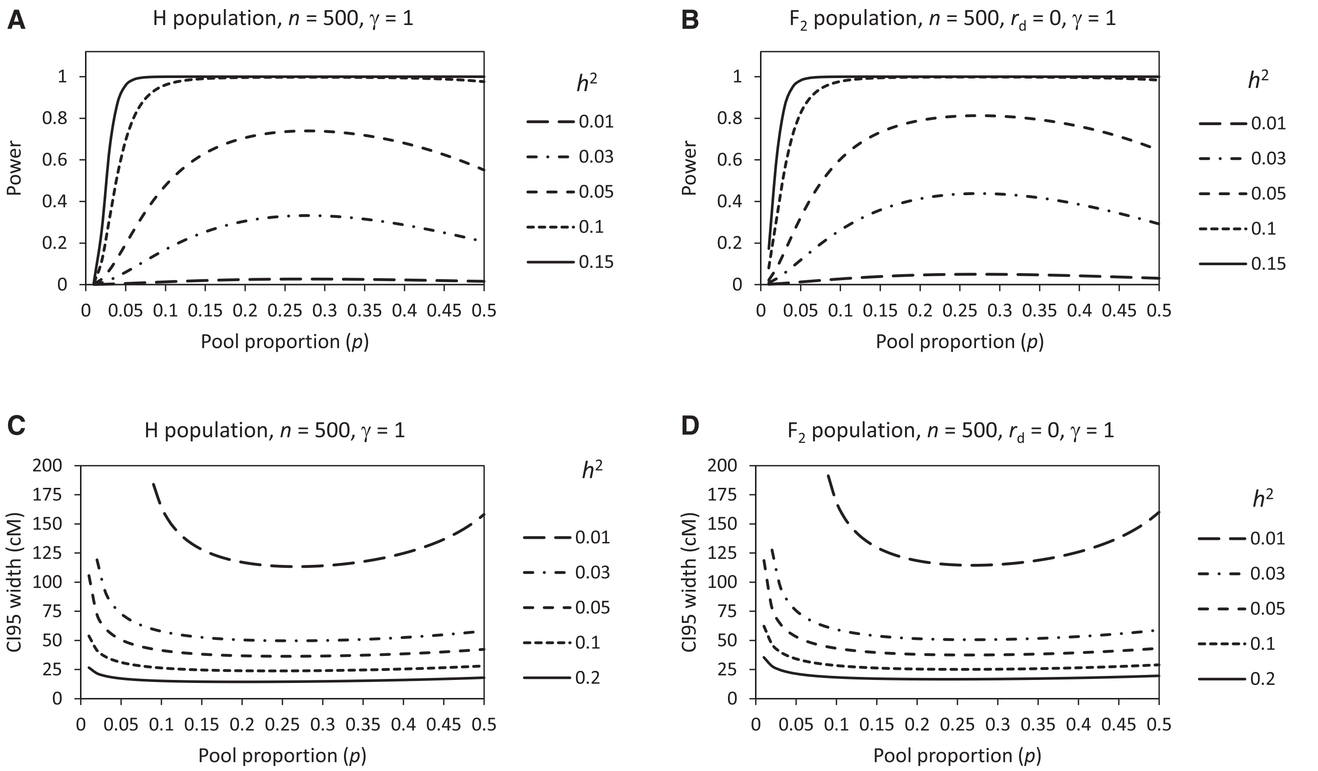

Figure 3 Relationships of power and CI 95 width with pool proportion depending on QTL heritability in yeast H population (A, C) and rice F_2 population (B, D).

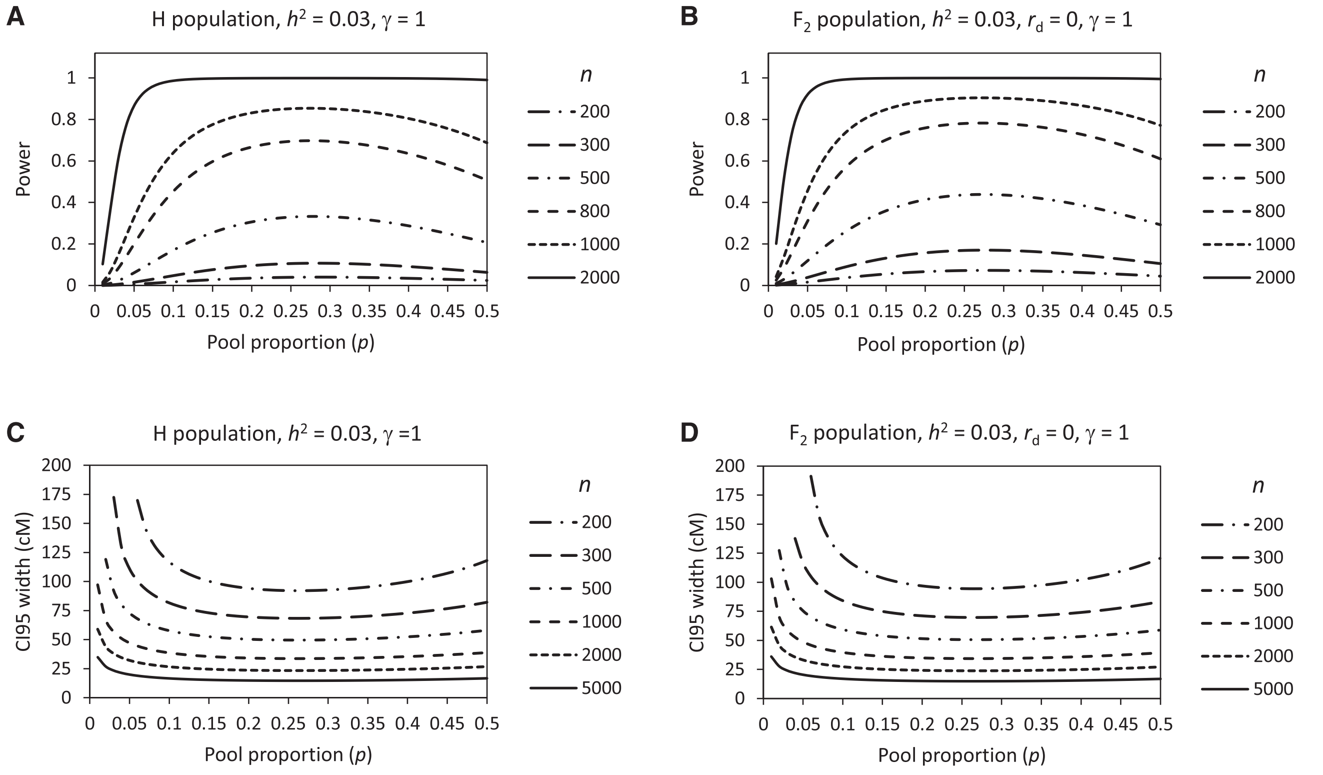

Figure 4 Relationships of power and CI 95 width with pool proportion depending on population size in yeast H population (A, C) and rice F_2 population (B, D).

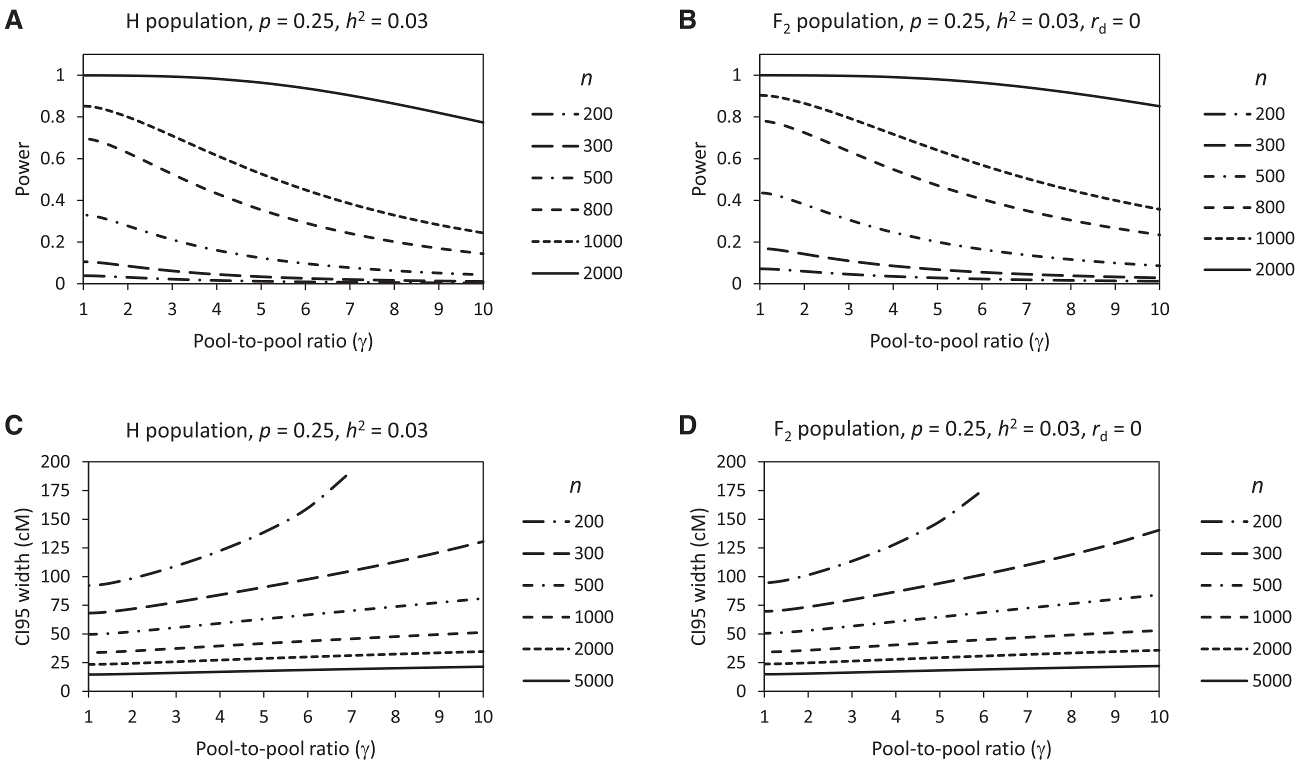

Figure 5 Relationships of power and CI 95 width with pool-to-pool ratio depending on population size in yeast H population (A, C) and rice F_2 population (B, D). Only the case of $\gamma \geq 1$ is shown because $p_H/p_L < 1$ is equivalent to $p_L/p_H > 1$.

Effect of pool imbalance

In the above analyses, it is assumed that the two pools are balanced in size, namely, $\gamma = 1$ or $p_H = p_L = p$. By fixing p and h^2, it is found that imbalance of pool size ($\gamma \neq 1$) can reduce the power (Figure 5, A and B) and increase the CI 95 width (Figure 5, C and D). The more the γ deviates from 1, the stronger the effect of pool imbalance, but increasing n can attenuate the effect of pool imbalance to some extent. The result suggests that the optimal design is to use two pools of equal size.

Effects of degree of dominance and generation

In F_k populations, the dominance effect of a QTL may exist and therefore affect the result of BSA-seq. In the above analyses, it is assumed that there is no dominance effect (namely, the degree of dominance $r_d = 0$) in the F_2 population. However, if $r_d > 0$ but h^2 is fixed, the power will be reduced (Figure 6A) and the CI 95 width will be increased (Figure 6B). The larger the r_d is, the smaller the power and the larger the CI 95 width will be. This suggests that additive effect is more beneficial than dominance effect to QTL mapping in BSA-seq.

Unlike permanent populations, F_k populations have different genetic structures in different generations. This can affect the result of BSA-seq. When other conditions are the same, the power increases (Figure 6C) and the CI 95 width decreases (Figure 6D) as the generation increases. The increment of power and the decrement of CI 95 width are particularly significant from F_2 to F_3. However, the power increase is attenuated or even disappeared when the population is large, while the CI 95 decrease remains significant (the relative decrease from F_2 to F_3 is always ~50%) with little influence by the population size.

Simulation of BSA-seq based on experimental data from yeast

To demonstrate the effect of pool proportion in practical populations, we used a set of experimental data from yeast (Bloom *et al.* 2015) to simulate BSA-seq using different pool proportions. The results were consistent with the theoretical expectation (Figure 7). A total of 15 putative QTLs were detected, with 5, 11, 11, 14, 14, 15, 15, and 13 QTLs detected under the pool proportion of 0.01, 0.05, 0.1, 0.2, 0.25, 0.3, 0.4, and 0.5, respectively. The number of detected QTLs was greatly reduced when p was small, but basically remained stable when $P \geq 0.2$, varying only at two QTLs with small peaks (QTL1 and 11), which were marginally significant (just reaching or slightly exceeding the threshold) at the maximum.

Discussion

We have shown above the effects of several factors, including experimental factors (population size, pool proportion, pool balance, and generation) and intrinsic factors (QTL heritability and degree of dominance), on the power of QTL detection and the precision (CI 95 size) of QTL mapping in BSA-seq in two different types of populations, H population (yeast) and F_2 population (rice). The factors display similar relationships with the power and CI 95 size in the two types of populations (Figures. 2–5), suggesting that the laws of the relationships revealed in this study are universal in BSA-seq. The intrinsic factors are mainly determined by the characteristics of QTLs, while the experimental factors can be managed in experimental design. Hence, according to the effects of various experimental factors, it is able to optimize the experimental design for BSA-seq.

Generally speaking, the experimental design for BSA-seq mainly involves three aspects, namely, population type (what

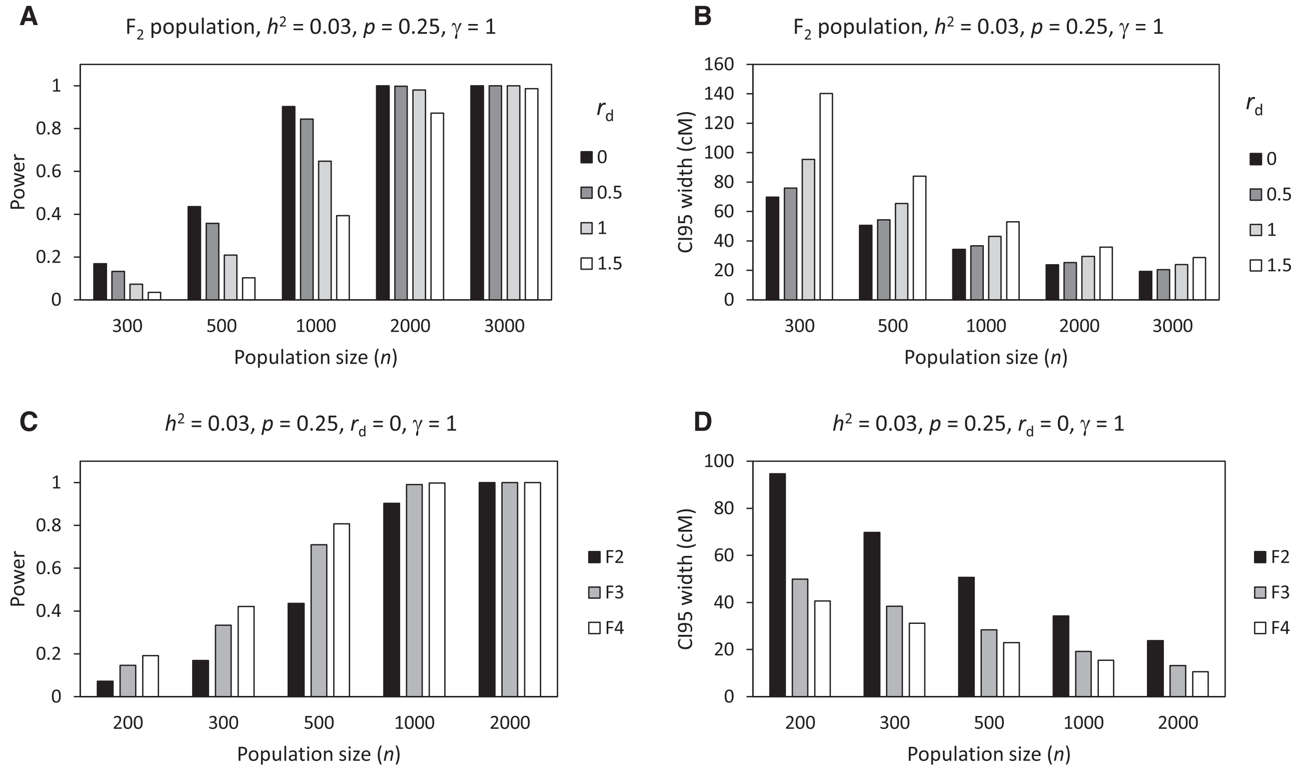

Figure 6 Relationships of power and CI 95 width with degree of dominance (A, B) and generation (C, D) depending on population size in rice.

kind of population), population size, and pool proportion. Population type is not an issue for fungi because only H population is applicable. For plants, however, there are multiple choices, such as F_k population, RIL population, and DH population. Among them, the most convenient type is probably F_2 population. However, we have seen that higher F_k generations are better for the power and precision (Figure 6, C and D). This is understandable because the frequency of homozygous genotypes of a QTL will increase and more recombination events between the QTL and flanking markers will occur as the generation increases. The former will likely increase the AFD between the two pools and thus increase the power because additive effect generally contributes more to genetic variation than dominance effect; the latter will increase the resolution of mapping and thus reduce the CI 95 size. As the improvements of power and precision are the most significant from F_2 to F_3 (Figure 6, C and D), it is recommended using F_3 instead of F_2 in practice if time permits. Nonetheless, it is necessary to point out that for the type of F_k population produced by random mating instead of selfing, the higher generations ($k \geq 3$) all have the same population structure as F_2 in terms of a single locus. In this case, the power does not increase with generation. Instead, the power may decrease with generation because the significance threshold of AFD increases with generation (Huang *et al.* 2020).

According to the principle of BSA (Michelmore *et al.* 1991), the smaller the pool proportion is, the greater the difference between the two pools will be. However, studies of conventional BSA method by theoretical analysis based on the infinitesimal model (Gallais *et al.* 2007) and computer simulation (Navabi *et al.* 2009) and simulation study of BSA-seq based on G′ test (Magwene *et al.* 2011) all show that the power of QTL detection reach the maximum value not under small pool proportions but under larger ones. This was verified in our study (Figures 3, A and B 4, A and B). The reason is that the power is determined not only by the AFD between the two pools but also by the variance of AFD. Decreasing pool proportion can increase AFD and its variance simultaneously, which will make the power increase and decrease, respectively. When the effect of AFD increase is smaller than the effect of AFD variance increase due to the decrease of pool proportion, the power will decrease. Apart from the effect on power, we also analyzed the effect of pool proportion on the CI 95 size (Figures 3, C and D 4, C and D), revealing that pool proportion affects the power and the CI 95 size correspondingly. So, considering the power and the CI 95 size simultaneously, we found that a pool proportion of 0.25 is generally suitable for BSA-seq. Nonetheless, pool proportion can be flexible in a wide range when the population size is large.

Intuitively, the two pools are usually set to be equal or balanced in size in BSA-seq. However, the benefit of pool balance has not been studied. In some studies, the two pools are very different in size. For example, in a BSA-seq experiment for mapping QTLs underlying the high ethanol tolerance in yeast, the two pools consisted of 32 and 237 segregants, respectively (Pais *et al.* 2013). In this study, we proved that imbalance of pool size is harmful, especially when the difference between the two pools is large, which can reduce the power and increase the CI95 size significantly (Figure 5). Hence, pool balance is important.

We have shown that increasing population size can increase the power and reduce the CI95 size under a constant pool proportion, but the improvement of power and CI95 size due to population size increase is very small when the population is sufficiently large (Figure 2). In practice, population size is also restricted by the cost of phenotyping. So, it is not that the larger the population, the better. However, determining the suitable population size is not easy because it depends on the heritability of each QTL. We suggest that the suitable population size can be

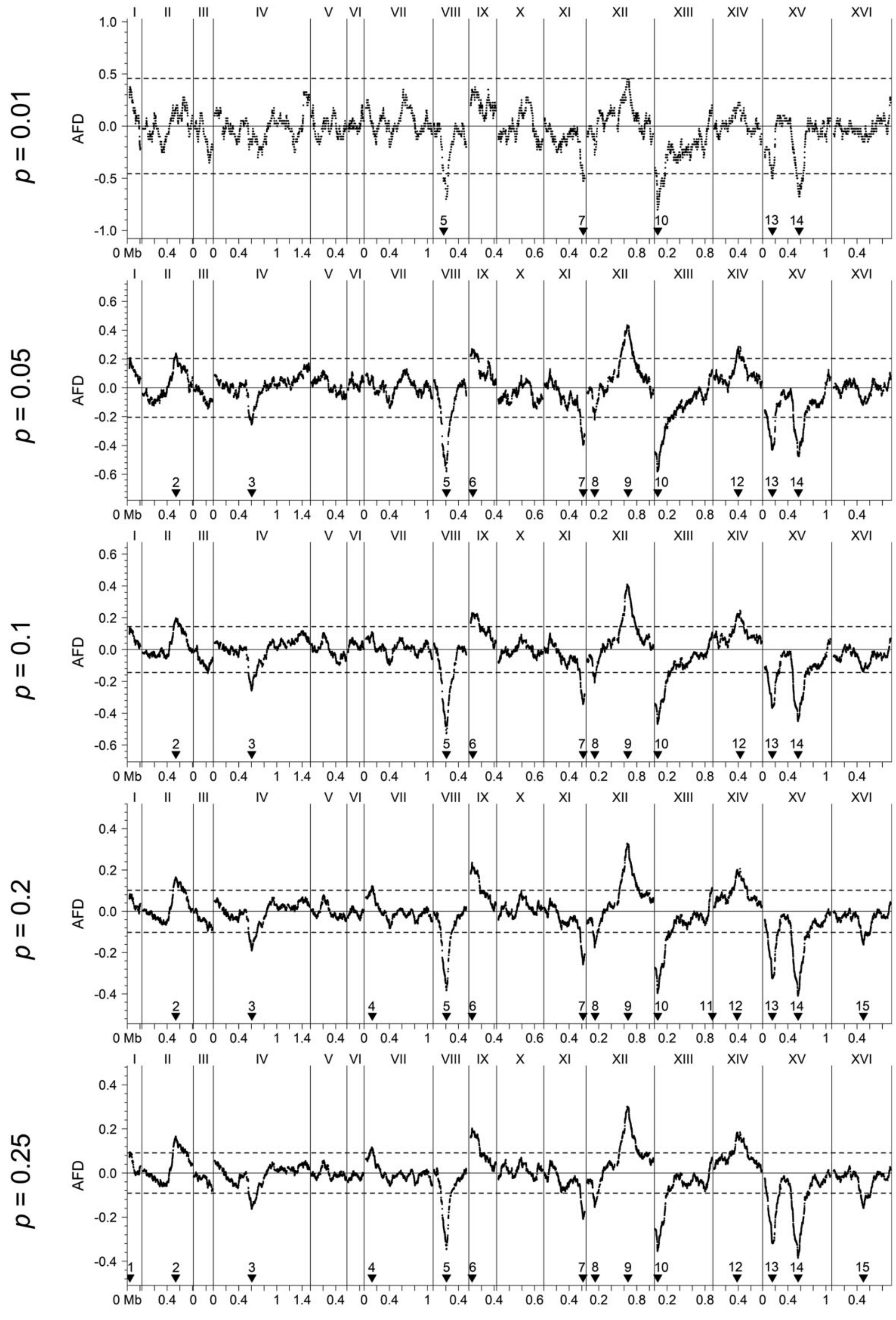

Figure 7 Actual AFD profiles under different pool proportions in a real yeast H population consisting of 4000 strains. The horizontal dashed lines are thresholds at the genome-wise significance level of 0.05. The black filled reversed triangles indicate the positions of detected QTLs, which are numbered from left to right.

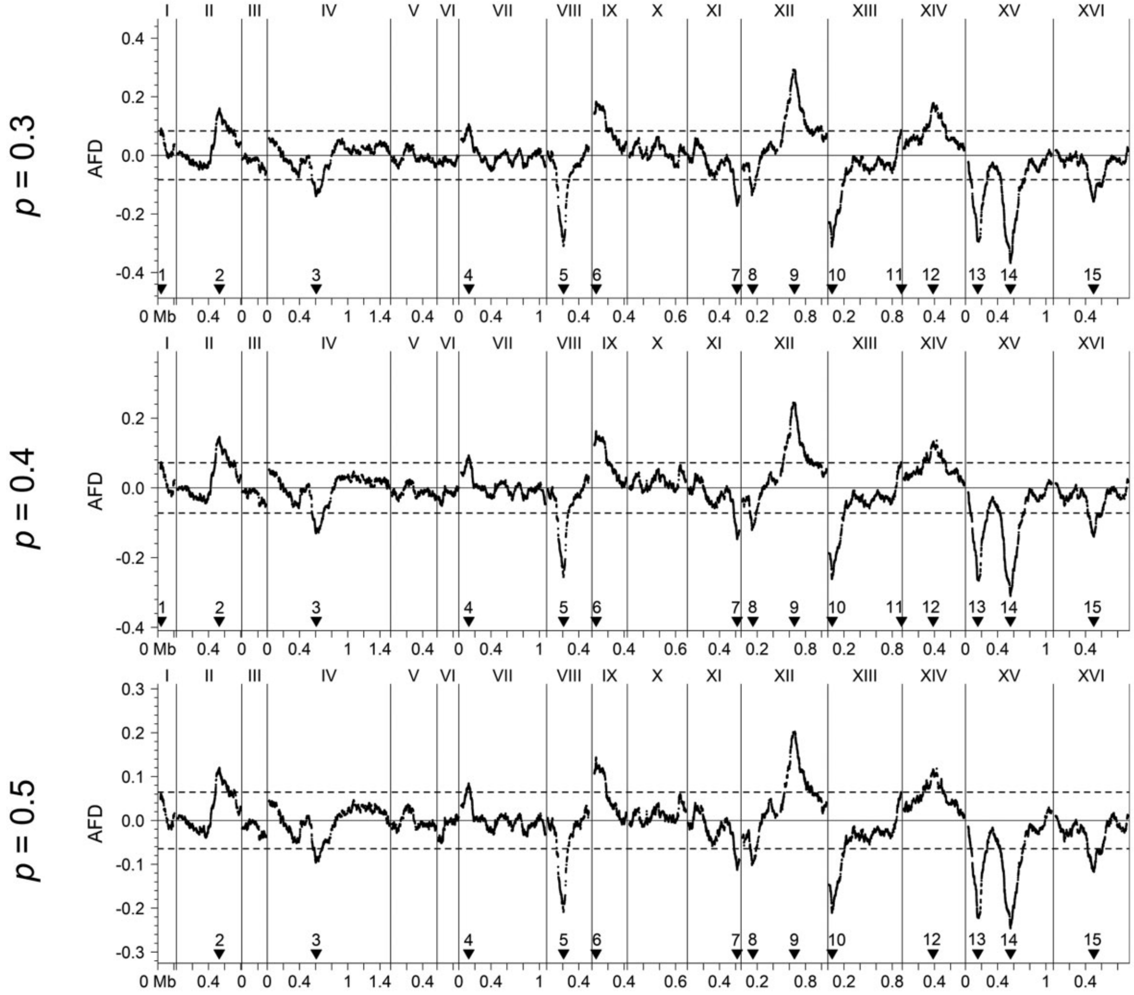

Figure 7 Continued.

chosen in light of a typical minor QTL (*e.g.*, $h^2 = 0.03$). It can be seen from Figure 2 that in both yeast H population and rice F_2 population, 1500 can be considered to be a suitable population size for a QTL with $h^2 = 0.03$, at which the power basically reaches 100% and the CI95 size has been within the slow decrease stage. In practice, researchers may want to know what the minimum population size is needed to detect a QTL with a given or higher heritability at a required power, or what the power is expected for a QTL with a given heritability when the population size is fixed. To meet these needs, we developed a web-based tool named BSA-seq Design Tool, which can be visited at http://124.71.74.135/BSA-seqDesignTool/ or downloaded from https://github.com/huanglikun/BSA-seqDesignTool. The tool will facilitate researchers to optimize their experimental designs of BSA-seq for QTL mapping.

Data availability

A web-based program named BSA-seq Design Tool is available at http://124.71.74.135/BSA-seqDesignTool/ and https://github.com/huanglikun/BSA-seqDesignTool.

Funding

This study was supported by the Sci-Tech Innovation Special Fund of Fujian Agriculture and Forestry University (grant number CXZX2017248).

Conflict of interest

The authors declare that there is no conflicts of interest.

Literature cited

Arikit S, Wanchana S, Khanthong S, Saensuk C, Thianthavon T, *et al.* 2019. QTL-seq identifies cooked grain elongation QTLs near soluble starch synthase and starch branching enzymes in rice (*Oryza sativa* L.). Sci Rep. 9:1–10. doi:10.1038/s41598-019-44856-2.

Bloom JS, Kotenko I, Sadhu MJ, Treusch S, Albert FW, *et al.* 2015. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nat Commun. 6:1–6. doi:10.1038/ncomms9712.

Branham SE, Patrick W, Lambel S, Massey L, Ma M, *et al.* 2018. QTL-seq and marker development for resistance to *Fusarium* *oxysporum* f. sp. niveum race 1 in cultivated watermelon. Mol Breed. 38:139. doi:10.1007/s11032-018-0896-9.

Claesen J, Burzykowski T. 2015. A hidden Markov-model for gene mapping based on whole-genome next generation sequencing data. Stat Appl Genet Mol Biol. 14:21–34. doi:10.1515/sagmb-2014-0007.

Das S, Upadhyaya HD, Bajaj D, Kujur A, Badoni S, *et al.* 2014. Deploying QTL-seq for rapid delineation of a potential candidate gene underlying major trait-associated QTL in chickpea. DNA Res. 22:193–203. doi:10.1093/dnares/dsv004.

Deokar A, Sagi M, Daba K, Tar'an B. 2019. QTL sequencing strategy to map genomic regions associated with resistance to ascochyta blight in chickpea. Plant Biotechnol J. 17:275–288. doi: 10.1111/pbi.12964.

Duitama J, Sánchez-Rodríguez A, Goovaerts A, Pulido-Tamayo S, Hubmann G, *et al.* 2014. Improved linkage analysis of quantitative trait loci using bulk segregants unveils a novel determinant of high ethanol tolerance in yeast. BMC Genom. 15:207. doi: 10.1186/1471-2164-15-207

Edwards MD, Gifford DK. 2012. High-resolution genetic mapping with pooled sequencing. BMC Bioinform. 13(Suppl. 6):S8.doi: 10.1186/1471-2105-13-s6-s8.

Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, *et al.* 2010. Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature. 464:1039–1042. doi:10.1038/nature08923.

Gallais A, Moreau L, Charcosset A. 2007. Detection of marker-QTL associations by studying change in marker frequencies with selection. Theor Appl Genet. 114:669–681. doi:10.1007/s00122-006-0467-z.

Ghavidel FZ, Claesen J, Burzykowski T. 2015. A nonhomogeneous hidden Markov model for gene mapping based on next-generation sequencing data. J Comput Biol. 22:178–188. doi: 10.1089/cmb.2014.0258.

Huang L, Tang W, Bu S, Wu W. 2020. BRM: a statistical method for QTL mapping based on bulked segregant analysis by deep sequencing. Bioinformatics. 36:2150–2156. doi:10.1093/bioinformatics/btz861.

Imerovski I, Dedić B, Cvejić S, Miladinović D, Jocić S, *et al.* 2019. BSA-seq mapping reveals major QTL for broomrape resistance in four sunflower lines. Mol Breed. 39:41. doi:10.1007/s11032-019-0948-9

Lahari Z, Ribeiro A, Talukdar P, Martin B, Heidari Z, *et al.* 2019. QTL-seq reveals a major root-knot nematode resistance locus on chromosome 11 in rice (*Oryza sativa* L.). Euphytica. 215:1–13. doi: 10.1007/s10681-019-2427-0.

Lei L, Zheng H, Bi Y, Yang L, Liu H, *et al.* 2020. Identification of a major QTL and candidate gene analysis of salt tolerance at the bud burst stage in rice (*Oryza sativa* L.) using QTL-seq and RNA-seq. Rice. 13:14.doi:10.1186/s12284-020-00416-1.

Lin JZ, Ritland K. 1996. The effects of selective genotyping on estimates of proportion of recombination between linked quantitative trait loci. Theor Appl Genet. 93:1261–1266. doi:10.1007/BF00223458.

Luo X, Liu J, Zhao J, Dai L, Chen Y, *et al.* 2018. Rapid mapping of candidate genes for cold tolerance in *Oryza rufipogon* Griff. by QTL-seq of seedlings. J Integr Agric. 17:265–275. doi:10.1016/S2095-3119(17)61712-X.

Magwene PM, Willis JH, Kelly JK. 2011. The statistics of bulk segregant analysis using next generation sequencing. PLoS Comput Biol. 7:1–9. doi:10.1371/journal.pcbi.1002255.

Michelmore RW, Paran I, Kesseli RV. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci U S A. 88: 9828–9832. doi:10.1073/pnas.88.21.9828.

Navabi A, Mather DE, Bernier J, Spaner DM, Atlin GN. 2009. QTL detection with bidirectional and unidirectional selective genotyping: marker-based and trait-based analyses. Theor Appl Genet. 118:347–358. doi:10.1007/s00122-008-0904-2.

Navarro-Escalante L, Zhao C, Shukle R, Stuart J. 2020. BSA-seq discovery and functional analysis of candidate Hessian fly (*Mayetiola destructor*) avirulence genes. Front Plant Sci. 11:956. doi: 10.3389/fpls.2020.00956.

Pandey MK, Khan AW, Singh VK, Vishwakarma MK, Shasidhar Y, *et al.* 2017. QTL-seq approach identified genomic regions and diagnostic markers for rust and late leaf spot resistance in groundnut (*Arachis hypogaea* L.). Plant Biotechnol J. 15:927–941. doi: 10.1111/pbi.12686.

Pais TM, Foulquie-Moreno MR, Hubmann G, Duitama J, Swinnen S, *et al.* 2013. Comparative polygenic analysis of maximal ethanol accumulation capacity and tolerance to high ethanol levels of cell proliferation in yeast. PLoS Genet. 9:e1003548.doi:10.1371/journal.pgen.1003548.

Pascoal S, Cezard T, Eik-Nes A, Gharbi K, Majewska J, *et al.* 2014. Rapid convergent evolution in wild crickets. Curr Biol. 24: 1369–1374. doi:10.1016/j.cub.2014.04.053.

Ramos A, Fu Y, Michael V, Meru G. 2020. QTL-seq for identification of loci associated with resistance to Phytophthora crown rot in squash. Sci Rep. 10:1–8. doi:10.1038/s41598-020–62228-z.

Ruangrak E, Du Y, Htwe NMPS, Pimorat P, Gao J. 2019. Identification of early tomato fruit ripening loci by QTL-seq. J Agric Sci. 11: 51.doi:10.5539/jas.v11n2p51.

Tang W, Huang L, Bu S, Zhang X, Wu W. 2018. Estimation of QTL heritability based on pooled sequencing data. Bioinformatics. 34: 978–984. doi:10.1093/bioinformatics/btx703.

Xin F, Zhu T, Wei S, Han Y, Zhao Y, *et al.* 2020. QTL Mapping of kernel traits and validation of a major QTL for kernel length-width ratio using SNP and bulked segregant analysis in wheat. Sci Rep. 10: 1–12. doi:10.1038/s41598-019-56979-7.

Yang Z, Huang D, Tang W, Zheng Y, Liang K, *et al.* 2013. Mapping of quantitative trait loci underlying cold tolerance in rice seedlings via high-throughput sequencing of pooled extremes. PLoS One. 8: e68433. doi:10.1371/journal.pone.0068433.

Communicating editor: P. Ingvarsson