A combinatorial strategy to identify various types of QTLs for quantitative traits using extreme phenotype individuals in an F2 population

Pei Li1, Guo Li1, Ya-Wen Zhang1, Jian-Fang Zuo1, Jin-Yang Liu1,2 and Yuan-Ming Zhang1,*

1Crop Information Center, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, China
2Jiangsu Key Laboratory for Horticultural Crop Genetic Improvement, Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China
*Correspondence: Yuan-Ming Zhang (soyzhang@mail.hzau.edu.cn)

https://doi.org/10.1016/j.xplc.2022.100319

ABSTRACT

Theoretical and applied studies demonstrate the difficulty of detecting extremely over-dominant and small-effect genes for quantitative traits via bulked segregant analysis (BSA) in an F2 population. To address this issue, we proposed an integrated strategy for mapping various types of quantitative trait loci (QTLs) for quantitative traits via a combination of BSA and whole-genome sequencing. In this strategy, the numbers of read counts of marker alleles in two extreme pools were used to predict the numbers of read counts of marker genotypes. These observed and predicted numbers were used to construct a new statistic, \( G_w \), for detecting quantitative trait genes (QTGs), and the method was named dQTG-seq1. This method was significantly better than existing BSA methods. If the goal was to identify extremely over-dominant and small-effect genes, another reserved DNA/RNA sample from each extreme phenotype F2 plant was sequenced, and the observed numbers of marker alleles and genotypes were used to calculate \( G_w \) to detect QTGs; this method was named dQTG-seq2. In simulated and real rice dataset analyses, dQTG-seq2 could identify many more extremely over-dominant and small-effect genes than BSA and QTL mapping methods. dQTG-seq2 may be extended to other heterogeneous mapping populations. The significance threshold of \( G_w \) in this study was determined by permutation experiments. In addition, a handbook for the R software dQTG.seq, which is available at https://cran.r-project.org/web/packages/dQTG.seq/index.html, has been provided in the supplemental materials for the users’ convenience. This study provides a new strategy for identifying all types of QTLs for quantitative traits in an F2 population.

Keywords: F2, bulked segregant analysis, dQTG-seq, extremely over-dominant gene, small-effect gene, rice

Li P., Li G., Zhang Y.-W., Zuo J.-F., Liu J.-Y., and Zhang Y.-M. (2022). A combinatorial strategy to identify various types of QTLs for quantitative traits using extreme phenotype individuals in an F2 population. Plant Comm. 3, 100319.

INTRODUCTION

In the past several decades, many methods have been established to mine elite genes for quantitative traits in animals, plants, and human beings; these include quantitative trait locus (QTL) mapping (Lander and Botstein, 1989), genome-wide association studies (Zhang et al., 2005; Yu et al., 2006), and bulked segregant analysis (BSA) (Giovannoni et al., 1991; Michelmore et al., 1991; Zhang et al., 1994). In real data analysis, it is difficult to identify small-effect and linked genes (Krommann and Mitchell-Olids, 2005; Mackay et al., 2009; Wang et al., 2016; Wen et al., 2019). In theory, it is also difficult to detect extremely over-dominant genes in BSA (Takagi et al., 2013a; Schneberger, 2014; Haase et al., 2015; Wang et al., 2019). As we know, quantitative traits are controlled by both a few major genes and a series of polygenes, whereas over-dominant genes can be used to explain heterosis, which is a natural phenomenon. Thus, it is necessary to investigate methods for detecting extremely over-dominant and small-effect genes in BSA (Cockram and Mackay, 2018).

Giovannoni et al. (1991) and Michelmore et al. (1991) established the BSA method in the early 1990s to quickly detect the association of a molecular marker with a trait of interest at a relatively low cost. To date, many BSA methods have been
Plant Communications  A Combinatorial Strategy to Detect all Types of QTLs in the F2
decided, and they have been summarized in Li and Xu, 2021. These methods are mainly used to detect genes for qualitative traits, mutants, and quantitative traits. In the early stage, BSA methods were widely used to identify dominant genes for qualitative traits and then increase molecular marker density in each F2 individual to fine-map target genes such as the dominant white locus in chicken (Ruyter-Spira et al., 1997). Recently, microsatellite markers were replaced by single nucleotide polymorphism (SNP) markers in the references of Li et al. (2018), Feng et al. (2019), and Liu et al. (2019). BSA methods were then widely used to identify mutant genes for quantitative traits. Schneeberger et al. (2009) first integrated BSA with whole-genome resequencing in a large pool of mutant F2 plants to rapidly identify recessive/dominant mutant genes based on the frequency of mutant alleles using SHOREmap software. When a smaller number of mutant plants was available, Cuperus et al. (2010) used wild-type SNP enrichment in a pool of mutant F2 plants to identify mutant genes. Austin et al. (2011) used the frequency of wild-type SNP alleles in a pool of mutant F2 plants to locate recessive mutant genes by the next generation mapping method, and Lindner et al. (2012) used the causative SNP/non-causative SNP ratio in a pool of mutant BC2 individuals to clone lethal mutations. For species with large and incomplete reference genomes, CloudMap (Minevich et al., 2012), SNPtrack (Leshchiner et al., 2012), Fishyskeleton (Bowen et al., 2012), MegaMapper (Obholzer et al., 2012), and MMAPP (Hill et al., 2013) are available. When a mutant is dominant, the reverse mapping and F3 screen strategies of Smith et al. (2016) are available. Abe et al. (2012), Fekih et al. (2013), and Takagi et al. (2013b) used the SNP index in a pool of mutant F2 (M3) plants from a cross between a mutant and its wild type to identify mutant genes with MutMap, MutMap+, and MutMap-Gap, respectively. To simultaneously identify multiple causal mutations, SIMM (Yan et al., 2016) and BSA-seq (Wang et al., 2021) are available. However, these recessive/dominant mutant traits are different from the dominant effects of quantitative traits.

Finally, BSA methods are widely used to identify quantitative trait genes (QTGs). As we know, quantitative traits are controlled by a few major genes and a series of polygenes with small genetic effects, making polygene identification difficult (Kroymann and Mitchell-Olds, 2005; Mackay et al., 2009). To identify these polygenes, SHOREmap (Schneeberger et al., 2009), MMAPP (Hill et al., 2013), MutMap (Abe et al., 2012), MutMap+ (Fekih et al., 2013), and MutMap-Gap (Takagi et al., 2013b) are available. Ehrenreich et al. (2010) used allelic frequency to identify yeast polygenes via BSA-seq. The Z-statistic is also used to detect allelic frequency differences at the target locus in high and low pools (Huang et al., 2012; Benowitz et al., 2019). If several markers in one window (or region) are used to calculate the average Z-statistic, this is the Z-statistic (Haase et al., 2015). Similarly, a smoothed G′ of the standard G statistic was proposed in Magwene et al. (2011). In addition, a dynamic Bayesian network via MULTI POOL (Edwards and Gifford, 2012) is used to capture the association among nearby loci in a pooled sequencing experiment; a hidden Markov model (HMM) is used to calculate the most probable states via EXPLoRator (ExploRator), indicating genomic regions that may be likely to contain trait-related genes (Duitama et al., 2014; Claesen and Burzykowski, 2015); a nonhomogeneous HMM with a transition matrix is used to detect the association of markers with genes (Ghavidel et al., 2015). All the approaches above make use of high and low pools. To increase the power of QTG detection, Wang et al. (2019) added an additional, middle pool in GradedPool-Seq to identify the rice heterotic QTL GW3p6. If no secondary F2 populations are available, a primary F2 population from a cross between any two homozygous parents may be used. In this case, the SNP index is also used to detect QTGs, and the procedure is termed QTL-seq (Takagi et al., 2013a). Clearly, the results may be affected by genetic background, although the method may be effective in some instances. To address this issue, Zhang et al. (2019) proposed a new design and a smooth-LOD (logarithm of odds) statistic. In analyses of real data, however, it is difficult to detect extremely over-dominant and small-effect genes for quantitative traits (Ehrenreich et al., 2010; Takagi et al., 2013a; Schneeberger, 2014; Haase et al., 2015; Wang et al., 2019).

A dominant mutant of a single-gene qualitative trait is different from the dominant effect of a QTL. In theory, the former is a qualitative trait, and the latter is a quantitative trait; they have different genetic structures in the two extreme phenotype pools. The structure of high and low pools is QQ + Qq via qq for a dominant mutant and QQ + qq via Qq for a fully over-dominant (zero additive and non-zero dominance) QTL for a quantitative trait (Haase et al., 2015). Thus, the allelic frequency of a homozygous genotype pool is indistinguishable from the allelic frequency of a heterozygous genotype pool in the fully over-dominant QTL model, and the fully over-dominant QTL becomes undetectable (Supplemental File 1). In the same way, it is difficult to detect extremely over-dominant QTGs (Schneeberger, 2014; Wang et al., 2019).

To address the issues above, we propose a combinatorial strategy for identifying various types of QTGs for quantitative traits in the F2 population by a combination of BSA and whole-genome sequencing. Specifically, equally mixed DNA sample of all plants in each extreme pool was sequenced. On the basis of the observed numbers of read counts of marker alleles in the two extreme pools, we propose a method for predicting the numbers of read counts of marker genotypes. The observed allele and predicted genotype numbers were used to construct a new statistic, Gw, to identify marker-trait associations. If the goal was to detect extremely over-dominant and small-effect genes, another reserved DNA/RNA sample from each extreme phenotype individual was sequenced one plant at a time, and the observed numbers of marker alleles and genotypes were used to calculate Gw to identify marker-trait associations. The former strategy is a new BSA method named dQTG-seq1, and the latter is a new individual segregant analysis named dQTG-seq2, where “d” indicates the dominant effect of the QTL (see materials and methods). Real and simulated datasets were used to analyze the sensitivity and specificity of the new strategy, and we also investigated the best sampling plan for the new strategy.

RESULTS
A combinatorial strategy for mapping all types of QTGs for quantitative traits in an F2 population by a combination of BSA and whole-genome sequencing

The new strategy in this study was proposed to quickly identify QTGs for quantitative traits in an F2 population, especially for
extremely over-dominant and small-effect genes (Figure 1). There were six types of F2 designs available. They included an F2 from a cross between any two homozygous parents (F2-I), an immortalized F2 (IMF2) derived from a DH (doubled haploid) or RIL (recombinant inbred line) population (F2-II) (Gardiner et al., 1993; Hua et al., 2003), a secondary F2 from a cross between a mutant and its wild type (F2-III) (Abe et al., 2012), a secondary F2 from a cross between two near-isogenic lines (F2-IV), a secondary F2 from the selfing of residual heterozygous lines of a targeted QTL (F2-V) (Yamanaka et al., 2005), and the improved secondary F2 of Zhang et al. (2019) from the selfing of residual heterozygous lines of a targeted QTL (F2-VI). When one of these F2 populations was constructed, the new strategy was available. F2-I to F2-V have been described in previous studies; F2-VI here is slightly different from that in Zhang et al. (2019). The improved F2 design involves the following steps: (1) Develop F1 and F2 populations from the cross between any two inbred lines. (2) Identify QTL for the quantitative trait of interest using trait-phenotype and marker-genotype datasets from the F2 population. (3) If the additive (a) and dominant (d) effects of the target QTL have the same sign, P2 should be crossed with F1 to construct BC1(P2)F1. If they have opposite signs, P1 should be crossed with F1 to construct BC1(P1)F1. This is because the large difference in the two genotypic values in BC1(Ps)F1 (s = 1, 2) can increase the power of QTL detection in BSA. To select BC1(Ps)F1 individuals with a heterozygous genotype at the target QTL and homozygous genotypes at non-target QTLs, molecular markers tightly linked to each detected QTL in the F2 population are available. The selected BC1(Ps)F1 individuals are self-pollinated to produce a secondary F2 population (Zhang et al., 2019). To ensure a large population, multiple target BCsF1 individuals should be selected. (4) The above-mentioned secondary F2 individuals are measured for the trait of interest, and the plants with minimum and maximum phenotypes are selected separately within each family. This is the F2-VI design used in the present study. In each aforementioned F2 design, an equally mixed DNA/RNA sample of all individuals in each extreme pool is separately subjected to whole-genome sequencing with deep genome coverage. The sequenced datasets for each mixed sample are mapped to the reference genome, and genomic variants across the genome are identified. On the basis of the observed numbers of read counts of marker alleles \( n_Q \) and \( n_q \) in low \((i = 1)\) and high \((i = 2)\) pools, we propose a method to predict the numbers of read counts of marker genotypes \( QQ \) and \( Qq \) in the two extreme pools (see materials and methods). Thus, the observed allele and predicted genotype numbers are used to calculate the new statistic \( G_W \)

\[
G_W = \frac{G_1}{G_1 + G_2} \times G_1 + \frac{G_2}{G_1 + G_2} \times G_2
\]

where \( G_1 \) and \( G_2 \) are the standard G statistics of Magwene et al. (2011), \( G_1 = \sum_{i=1}^{2} \sum_{j=1}^{2} n_{Qij} \ln[n_{Qij} / E(n_{Qij})] \), using allelic
Plant Communications

A Combinatorial Strategy to Detect all Types of QTLs in the F2

read numbers \( n_{Qij} \) and their expectations \( E(n_{Qij}) \), and \( G_2 = \sum_{i=1}^{n} \sum_{j=1}^{n} n_{Qij} \ln(n_{Qij}/E(n_{Qij})) \) using genotypic read numbers \( n_{Qij} \) and their expectations \( E(n_{Qij}) \). The \( G_2 \) statistic was used to identify QTLs for quantitative traits, and the strategy was named dQTG-seq1.

Although dQTG-seq1 has higher power for QTL detection than existing BSA methods, it may fail to detect extremely and fully over-dominant and small-effect genes, for which all existing BSA methods including dQTG-seq1 do not work well. To address this issue, additional reserved DNA/RNA sample from each extreme phenotype individual is separately subjected to whole-genome sequencing one plant at a time. In this case, the observed numbers of marker alleles and genotypes in the two pools can be used to calculate \( G_2 \) to identify QTGs in a strategy we have named dQTG-seq2.

To validate the new strategy, other BSA and QTL mapping approaches, such as the \( \Delta \text{SNP} \) index (Takagi et al., 2013a, 2013b), \( G' \) (Magwene et al., 2011), composite interval mapping (CIM) (Zeng, 1993), and inclusive CIM (ICIM) (Li et al., 2007), were also used in this study.

Monte Carlo simulation studies

dQTG-seq1 has higher power than other BSA methods

To demonstrate that dQTG-seq1 has higher power than other BSA methods, one small-effect (5%) QTL was simulated under four kinds of genetic models: additive \((a = 0)\) and over-dominant \((d = 2a)\), completely dominant \((a = d)\), and over-dominant \((d = 2a)\) and \(d = 3a\) (Supplemental Table 1). All samples in simulated dataset I were analyzed by the dQTG-seq1, ED (Euclidean distance), \( G' \), and \( \Delta \text{SNP} \) methods. dQTG-seq1 had significantly higher power than the other BSA methods; its power was 10.0%, 8.5%, 4.7%, and 4.1% higher than those of the other BSA methods; therefore be used to detect multiple QTLs on a chromosome.

\( \Delta \text{SNP} \) index

To investigate whether the dQTG-seq1 and dQTG-seq2 methods can identify extremely and fully over-dominant and small-effect genes in the F2 population, one small-effect (5%) QTL was simulated under extremely \((d/a = 4)\) and fully over-dominant \((a = 0 \text{ and } d = 0)\) over-dominant models in a secondary F2 (F2-III to F2-VI) population. Each simulation sample in simulated dataset II was analyzed by the dQTG-seq1, dQTG-seq2, ED, \( G' \), \( \Delta \text{SNP} \), and ICIM methods, and the power, false positive rate (FPR), false negative rate (FNR), and accuracy of QTL position estimates in QTL detection were calculated. Under the extremely over-dominant model, the powers of QTL detection using dQTG-seq2 (71.3%) and ICIM (70.0%) were significantly higher than those using the dQTG-seq1 (14.3%), ED (12.7%), \( G' \) (13.7%), and \( \Delta \text{SNP} \) (13.3%) methods. Similar trends were observed in the FPR, FNR, and accuracy of QTL position estimates (Figure 2B; Supplemental Table 2). Under the fully over-dominant model, the powers of QTL detection using dQTG-seq2 (74.8%) and ICIM (73.0%) increased slightly, whereas the powers of the other methods decreased significantly (<5%). The trends in FPR, FNR, and accuracy of QTL position estimates were similar to those in the extremely over-dominant model (Figure 2B; Supplemental Table 2).

To assess the applicability of dQTG-seq1 and dQTG-seq2 to a primary F2 population from a cross between any two homozygous parents (F2-I and F2-II), five QTLs were simulated under extremely \((d/a = 4)\) and fully over-dominant \((a = 0 \text{ and } d = 0)\) models in simulated dataset III, and their sizes were set to 2.5%, 5.0%, 5.0%, 10.0%, and 10.0%. Under the extremely over-dominant model, the average powers of dQTG-seq2 (49.0%) and ICIM (30.1%) were significantly higher than those of the dQTG-seq1 (15.8%), ED (9.3%), \( G' \) (15.9%), and \( \Delta \text{SNP} \) (13.3%) methods, and the FPR, FNR, and accuracy of QTL position estimates from dQTG-seq2 and ICIM were better than those from the other methods (Figure 2C; Supplemental Table 3). Under the fully over-dominant model, the average powers of dQTG-seq2 (50.2%) and ICIM (28.4%) showed slight changes, whereas the average powers of the dQTG-seq1 (7.7%), ED (4.0%), \( G' \) (6.8%), and \( \Delta \text{SNP} \) (6.6%) methods decreased. The trends in FPR, FNR, and accuracy of QTL position estimates were similar to those in the extremely over-dominant model (Figure 2D; Supplemental Table 3). Interestingly, we found that QTL4 and QTL5 with a size of 10% had lower powers than QTL2 and QTL3 with a size of 5%. This may be because the effects of QTL4 and QTL5 were in opposite directions.

In conclusion, from the second and third simulation experiments, dQTG-seq2 can detect extremely and fully over-dominant and small-effect genes, but the other BSA methods, such as dQTG-seq1, ED, \( G' \), and \( \Delta \text{SNP} \), do not work well.

To determine whether dQTG-seq1 and dQTG-seq2 can be used to detect multiple QTLs on a chromosome, two 10% QTLs on one chromosome were simulated in an F2 population; the genetic model was \( a = 0.5a \), and all the samples in simulated dataset IV were analyzed by dQTG-seq1 and dQTG-seq2. The two QTLs could be detected (Supplemental Table 4), and the new methods can therefore be used to detect multiple QTLs on a chromosome.

Analyses of real data from maize and rice

dQTG-seq1, as well as the \( G' \) and \( \Delta \text{SNP} \) methods, can identify a QTG for maize plant height in a secondary F2 population

To validate dQTG-seq1 in a secondary F2 population (F2-III to F2-VI), the maize dataset from Zhang et al. (2019) was reanalyzed. In Zhang et al. (2019), the numbers of read counts of marker alleles obtained from whole-genome sequencing were used to detect a QTN for plant height. The target QTL on chromosome 7 in the F2 population was also detected in a secondary F2 population, and its candidate gene around this QTN was confirmed by a molecular biology experiment (Zhang et al., 2019). In the present study, we predicted the numbers of read counts of marker genotypes, and the observed allele and predicted genotype numbers were used to detect the QTN. In Figure 3, the QTN in Zhang et al. (2019) was also found in the secondary F2 population by the dQTG-seq1, \( G' \), and \( \Delta \text{SNP} \) methods, indicating the feasibility of dQTG-seq1 for detection of target QTN in a secondary F2 population.

dQTG-seq1 identifies more known genes for rice yield in a primary F2 population than the previous BSA (\( \Delta \text{SNP} \) and \( G' \)) methods

To validate dQTG-seq1 in a primary IMF2 population (F2-I and F2-II) from a cross between any two homozygous parents, the
Figure 2. Comparison of statistical power for QTL detection across various methods under different genetic models.

(A) Comparison of dQTG-seq1 with existing BSA (ED, G, and ΔSNP) methods under various degrees of dominance (d/a = 0, 1, 2, and 3) models of one simulated QTL in a secondary F2 population.

(B) Comparison of dQTG-seq2 with dQTG-seq1, ED, G, ΔSNP, and ICIM methods under extremely and fully over-dominant models of one simulated QTL in a secondary F2 population.

(C and D) Comparison of dQTG-seq2 with dQTG-seq1, ED, G, ΔSNP, and ICIM methods under extremely (C) and fully (D) over-dominant models of five simulated QTLs in a primary F2 population. *, **, and *** indicate the 0.05, 0.01, and 0.001 significance probability levels using a two-sample percentage test.
In Zhou et al. (2012), 11 QTLs were identified as significantly associated with rice yield. Around these QTLs, there were eight previously reported genes (Supplemental Table 5). In this study, 20% extremely low-yield plants and 20% extremely high-yield plants were grouped into low and high pools. In each pool, we calculated the numbers of marker alleles and predicted the numbers of marker genotypes. These numbers were used to calculate \( G_{M} \). As a result, 18 QTLs were found to be significantly associated with rice yield (Supplemental Table 5). Around these QTLs, 11 previously reported genes were found to be truly associated with rice yield (Supplemental Table 6). Thus, dQTG-seq1 and existing BSA methods have shortcomings in the detection of small-effect (5%) QTL detection in 500 F2 individuals. Here, we provide some simulation and theoretical analyses to support the thresholds of \( |d/a| > 2 \) and \( |\Delta \text{SNP index}| \leq 0.15 \). To confirm this viewpoint, we calculated the \( |d/a| \) and \( |\Delta \text{SNP index}| \) of the 11 known genes identified by dQTG-seq1. The value of \( |d/a| \) ranged from 0.00 to 1.73, and its average was 0.72 ± 0.55; \( |\Delta \text{SNP index}| \) ranged from 0.13 to 0.22, and its average was 0.19 ± 0.026 (Supplemental Table 6). The low \( |d/a| \) and large \( |\Delta \text{SNP index}| \) phenomena were also observed in the G’ and \( \Delta \text{SNP} \) methods. Thus, dQTG-seq1 and existing BSA methods have shortcomings in the detection of large \( |d/a| \) and small \( |\Delta \text{SNP index}| \) genes, although dQTG-seq1 has higher power than existing BSA methods.

Here, we provide some simulation and theoretical analyses to support the thresholds of \( |d/a| > 2 \) and \( |\Delta \text{SNP index}| \leq 0.15 \). First, in Figure 2A, \( |d/a| = 2.0 \) was viewed as a critical value based on the stability of power changes. That is, the power of small-effect (5%) QTL detection in 500 F2 individuals is approximately 30%. Second, we carried out one theoretical analysis to explain why the critical value of \( |\Delta \text{SNP index}| \) is 0.15. In situations where QTL size ranged from 1% to 5% and sampling fraction was 5%, 10%, 15%, and 20% in the F2 population, we calculated allelic frequency differences between the two extreme pools. We found that most allelic frequency differences were less than 0.15 (Supplemental Figure 1), in which it was difficult to detect these loci. The above two critical values were consistent with those in analysis of real data (Supplemental Table 6).
above-mentioned rice dataset from Zhou et al. (2012) was reanalyzed. All the IMF2 individuals were sorted according to their phenotypic values, and 20% of the extreme high and extreme low individuals were selected to form the high and low pools. From the 20% extreme low and high pools, we could obtain 5%, 10%, and 15% extreme low and high pools. In these four datasets, dQTG-seq2 was used to detect QTLs for rice yield. As a result, 33 QTLs were found to be significantly associated with rice yield (Figure 4F; Supplemental Table 5). Among these QTLs, 26 previously reported genes were found to be truly associated with rice yield, \( |d/a| < 2.0 \), small effects, and \( |d/a| > 2.0 \) are indicated by black, red, and pink, respectively. If the dominance ratio of a small-effect gene is larger than 2.0, the gene name is in pink, and its corresponding solid line is in red. The dataset was derived from Zhou et al. (2012).

Among the above-mentioned 27 known genes, 25 were validated by one-way ANOVA by calculating the average values of individuals with each genotype (Supplemental Figure 2).

Suitable sampling plans via Monte Carlo simulation studies

To obtain suitable sampling plans for the dQTG-seq1 and dQTG-seq2 methods, we conducted a series of Monte Carlo simulation studies. In these studies, sample sizes were set to 250, 500, 1 000, and 2 000; QTL sizes \( (r^2, \%) \) were set to 2%, 5%, and 10%; 30, 50, and 80 extreme individuals were sampled in each extreme pool. There were 1 000 replicates. All samples in simulated dataset V were analyzed by dQTG-seq1 and dQTG-seq2, and the results are shown in Figure 5 and Supplemental Table 7. If sampling plans in which the power of QTL detection is >40% were suggested to be suitable to users, some suitable sampling plans were obtained. That is, 80 extreme individuals in 2 000 F2 plants were selected for each pool to identify 2%
QTLs via dQTG-seq2, whereas more than 2,000 F₂ individuals were needed for dQTG-seq1. Thirty extreme individuals in at least 250 F₂ individuals were selected for each pool to identify 5%–10% QTLs via dQTG-seq1 and dQTG-seq2. In summary, selection of 80 extreme individuals for each pool from a large F₂ population was suitable for identifying small-effect (2%) genes, whereas selection of 30 extreme individuals for each pool from a general F₂ population was suitable for identifying large-effect genes.

**DISCUSSION**

Significant progress in BSA has been made in this study. First, a new BSA strategy was proposed to quickly detect QTLs for...
A Combinatorial Strategy to Detect all Types of QTLs in the F2

Table 1. Mathematical differences among dQTG-seq1, dQTG-seq2, and other BSA methods.

| Methods          | No. of marker alleles and genotypes | Observed number of genotypes | Observed number of genotypes | Statistic                                      |
|------------------|------------------------------------|-----------------------------|-----------------------------|------------------------------------------------|
| dQTG-seq1        | ✔                                  | ✔                           |                             | \( G_w = c_1 G_1 + c_2 G_2 \)                  |
| dQTG-seq2        | ✔                                  | ✔                           |                             | \( G_w = c_1 G_1 + c_2 G_2 \)                  |
| ED               | ✔                                  | ✔                           |                             | \( ED = |p_{AL} - p_{AH}| + |p_{AL} - p_{AH}|^2 \)                                    |
| \( G' \)         | ✔                                  | ✔                           |                             | \( G' = \sum_{j=1}^{k} k_j G \)                  |
| \( \Delta SNP \) | ✔                                  |                             |                             | \( \Delta SNP index = |p_{AL} - p_{AH}| \)                                    |

Table 1. Mathematical differences among dQTG-seq1, dQTG-seq2, and other BSA methods.

\[ G_1 = G = \sum_{j=1}^{n} n_{Qij} \ln \left( \frac{n_{Qij}}{E(n_{Qij})} \right), \]
\[ n_{Qij} \text{ are the number of read counts of marker alleles, and } E(n_{Qij}) \text{ is the expectation of } n_{Qij}; \]
\[ G_2 = \sum_{j=1}^{n} \sum_{k=1}^{c_{ij}} n_{Qij} \ln \left( \frac{n_{Qij}}{E(n_{Qij})} \right) c_{ij} = \frac{G_1}{G_1 + G_2}; \]
\[ G_2 = \frac{G_1}{G_1 + G_2}; \]
\[ n_{Qij} \text{ are the number of read counts of marker genotypes, and } E(n_{Qij}) \text{ is the expectation of } n_{Qij}; \]
\[ p_{AH} \text{ and } p_{AH} \text{ are the frequencies of marker allele A (a) in extreme low and high pools, respectively.} \]

quantitative traits in primary or secondary F2 population. Specifically, equally mixed DNA/RNA samples of all the plants in each extreme pool are sequenced one pool at a time. The sequencing results are used to identify QTLs via dQTG-seq1. If the goal is to identify large \(|d/a|\) and/or small \(\Delta SNP\) index genes, reserved DNA/RNA sample from each extreme plant is sequenced one plant at a time. The observed numbers of marker alleles and genotypes are used to detect QTLs via dQTG-seq2. In analyses of simulated and real data, dQTG-seq2 can identify extremely and fully over-dominant and small-effect QTLs, whereas dQTG-seq1 is feasible for detecting low \(|d/a|\) and large-effect loci. The two methods had significantly higher powers of QTL detection than existing BSA methods. This is because the new methods utilize both allelic and genotypic frequency differences between high and low pools (Table 1). As shown in Figure 6A–6E, \(\Delta SNP\) index decreases as \(|d/a|\) increases, especially in the fully over-dominant model in which \(\Delta SNP\) index \(= 0\), but differences in marker genotypic frequencies exist, demonstrating the effectiveness of detecting a gene or locus using dQTG-seq2. Then, in BSA, we predicted the numbers of read counts of marker genotypes from the numbers of read counts of marker alleles, and the observed allele and predicted genotype numbers were used to construct the new statistic \(G_w\). We found that the predicted values were acceptable when \(|d/a| < 2.0\) and \(\Delta SNP\) index \(> 0.15\), even in a primary F2 population (Supplemental Tables 1 and 5; Supplemental Figure 1). If \(|d/a| \geq 2.0\), \(\Delta SNP\) index \(< 0.15\), \(a = 0\), and \(d \neq 0\), the predicted values were inaccurate, the power of QTL detection was low, and the FPR was high (>5%)(Supplemental Tables 1 and 5; Supplemental Figure 1), indicating the difficulty of detecting extremely and fully over-dominant and small-effect genes or loci via dQTG-seq1. We therefore proposed sequencing each extreme individual in the two pools (Figure 1) so that the numbers of marker alleles and genotypes could be obtained, and extremely and fully over-dominant and small-effect genes could be detected by dQTG-seq2 with high power and accuracy (Figures 2 and 4; Supplemental Table 6) and a low false-positive rate (Figure 2; Supplemental Table 2). In the analysis of real data for rice yield, dQTG-seq2 identified eight over-dominant known genes, four of which were detected by CIM and ICIM; the other BSA methods found no known genes. More importantly, dQTG-seq2 identified eight small-effect known genes, whereas the other methods detected no known small-effect genes. Finally, dQTG-seq2 had some slight advantages over CIM and ICIM in simulation and real data analysis, such as the low cost of marker genotyping and slightly higher power and accuracy. As shown in Figure 4, dQTG-seq2 identified more known genes than CIM and ICIM, suggesting that CIM and ICIM can be replaced by dQTG-seq2 in QTL mapping.

In quantitative genetics, over-dominance and complete dominance are defined as \(|d/a| > 1.2\) and \(0.8 < |d/a| < 1.2\), respectively (Stuber et al., 1987). In this study, extreme over-dominance was defined as \(|d/a| \geq 4.0\), and full over-dominance was defined as \(d \neq 0\) and \(a = 0\). Small- and large-effect QTLs (genes) were distinguished between \(r^2 < 5\%\) and \(r^2 \geq 5\%\) based on the results in Supplemental Figure 1 and the analysis of real rice data. The definitions in this study are helpful for understanding the new methods.

The new methods can be used to identify QTLs in the aforementioned six types of F2 population. As we know, a secondary F2 population is better than a primary F2 population for fine-mapping of QTLs. This is also true for BSA. If no QTLs in a secondary F2 (F2-III to F2-VI) or no ideal QTLs in a primary F2 (F2-I and F2-II) are detected by dQTG-seq1, we recommend sequencing another reserved DNA/RNA sample from each extreme plant, and the observed numbers of marker alleles and genotypes can be used to identify QTLs via dQTG-seq2. We should point out that F2-VI is an improved F2 design based on Zhang et al. (2019), and the designs differ in whether two parents may be used to backcross the F1. In addition, dQTG-seq2 is theoretically applicable to any heterogeneous mapping population, such as BC1F5, F6 (\(s \geq 3\)), genetic mating design, or association mapping population, because the numbers of marker alleles and genotypes are observed, and their theoretical numbers are calculated from a contingency table. Thus, dQTG-seq2 is feasible. Detailed results will be reported soon.

Sometimes, F2 individuals have a large experimental error. First, the F2:3 design (Zhang and Xu, 2004) can be used to overcome this problem. That is, DNA/RNA is extracted from each F2 plant,
and extreme individuals are determined based on the average phenotypes of F_{2:3} families. This design is frequently used in maize and cotton genetic analysis. IMF_{2} derived from DH lines or RILs is also used to overcome this problem (Hua et al., 2003). This is because replicated experiments can be implemented in IMF_{2}, especially from RILs with more recombinants.

The new methods are different from previous BSA and QTL mapping methods. To reduce the time required to construct the mapping population and the cost of marker genotyping (Darvasi and Soller, 1994), BSA is proposed at an early stage to detect the association of a marker with a trait (Giovannoni et al., 1991; Michelmore et al., 1991). This method is also frequently used to locate a single-gene trait in Mendelian genetics (Ruyter-Spira et al., 1997). Specifically, some closely linked markers around one marker with significant differences in allele frequency between the two extreme pools are used to fine-map the gene via linkage analysis. With the development of sequencing technology and reductions in sequencing costs, the integration of BSA with sequencing technology provides new opportunities for quickly detecting QTGs. To date, a series of methods have been proposed (Magwene et al., 2011; Abe et al., 2012; Takagi et al., 2013a, 2013b; Yan et al., 2016). In these methods, the numbers of read counts of marker alleles in low and high pools in primary or secondary F_{2} populations are used to construct their statistics. In this study, the numbers of (read counts of) marker alleles and genotypes in the two extreme pools are used to construct the new statistic $G_{w}$. If the trait is controlled by a fully over-dominant gene, in theory there are no differences in allelic frequencies of markers around the trait gene between the two extreme pools (Supplemental File 1), highlighting the inability of existing BSA methods to locate fully over-dominant genes (Schneeberger, 2014; Wang et al., 2019). Clearly, the present study solves this problem, which has been pending for some time.

As described in model 1, $G_{w}$ consists of $G_{1}$ and $G_{2}$, which are based on the numbers of read counts of marker alleles and genotypes, respectively. If $G_{1} = 0$, $G_{2}$ is available; if $G_{2} = 0$, $G_{w}$ is $G'$ of Magwene et al. (2011). Thus, $G_{w}$ is an extension of $G'$. Although the distributions of $G_{1}$ and $G_{2}$ are $\chi^{2}$ (Magwene et al., 2011), $G_{1}$ and $G_{2}$ are correlated; thus, the distribution of $G_{w}$ is unknown. This makes it difficult to determine the critical value of $G_{w}$ for significant QTGs. In the present study, 1 000 permutation experiments were used to determine the critical value. The results are presented in Supplemental Figure 3. They show that the critical value of $G_{w}$ is 6.52–7.24 at the 0.10 probability level and 7.60–8.52 at the 0.05 probability level using the dQTG-seq2 method. We suggest that the 0.10 threshold be used with dQTG-seq1 and the 0.05 threshold be used with dQTG-seq2. This is because the predicted numbers of read counts of marker alleles and genotypes in the two extreme pools are used to construct the new statistic $G_{w}$. If the trait is controlled by a fully over-dominant gene, in theory there are no differences in allelic frequencies of markers around the trait gene between the two extreme pools (Supplemental File 1), highlighting the inability of existing BSA methods to locate fully over-dominant genes (Schneeberger, 2014; Wang et al., 2019). Clearly, the present study solves this problem, which has been pending for some time.
genotypes have residual error. This may explain why different critical values in Figures 4A (6.52) and 4F (7.60) are used. In addition, we have listed additional $G_w$ thresholds under various probability levels in Supplemental Table 8 and Supplemental Figure 4.

The new method is also applicable in other situations. If structural variants are obtained from sequencing, the markers with structural variants can also be used in the new methods. Although the new methods have many advantages over the existing methods, they have some limitations. First, in the additive–dominant model, the numbers of marker genotypes in the two extreme pools need to be predicted when dQTG-seq1 is used, and its predicted accuracy decreases with increasing dominance ratio; in particular, when a QTG is extremely over-dominant ($|d/a| > 4.0$), dQTG-seq1 does not work. To address this issue, each extreme plant is sequenced, and dQTG-seq2 is used. Second, the new and existing BSA methods are affected by QTL size, sample size, sampling fraction, and sequencing depth. The best values for these factors can improve the accuracy of these methods (Zou et al., 2016) (Supplemental Table 9). In this study, we also determined the best sampling plan for the new methods (Figure 5).

Although marker segregation distortion is a common phenomenon in bi-parental populations, only three markers around 27 known genes (Supplemental Table 6) had segregation distortion, indicating that marker segregation distortion had little importance in the analysis of real rice data. In our simulation study, marker segregation distortion affected the detection of genes (Supplemental Table 10), and its methodological extension needs to be further addressed in the near future.

MATERIALS AND METHODS

Materials
The maize dataset was derived from Zhang et al. (2019), in which the $F_1$ plants between HZS (an elite inbred line) and 1462 (a temperate inbred line) were self-pollinated and backcrossed to HZS to produce $F_2$ and BC$_1F_2$ seeds, respectively. The 20% tallest (580) and 20% shortest (367) plants were selected from each BC$_1F_2$ family and bulked in equal ratios into high and low pools. The two pools were subjected to whole-genome sequencing with up to >280× coverage, and 197 021 high-quality SNPs were generated.

The genotypic and phenotypic observations from a rice IMF$_2$ population were downloaded from Zhou et al. (2012) (http://www.pnas.org/content/suppl/2012/09/07/121411109.DCSupplemental). In the IMF$_2$ population, there were 278 individuals and 1619 bin markers, and yield per plant in 1998 was re-analyzed. All the IMF$_2$ individuals were sorted according to their phenotypic values, and 20% of the extreme high and low individuals were selected to form the high and low pools. Using dQTG-seq1, the observed numbers of marker alleles were used to predict the numbers of marker genotypes. The observed and predicted numbers in extreme pools were used to calculate $G_w$. Using dQTG-seq2, the observed numbers of marker alleles and genotypes in extreme pools were used to calculate $G_w$.


g_1 = \frac{0.25(1 - r)^2}{P_L} + \frac{0.5r(1 - r)\Phi(x_L - d) + 0.25r^2\Phi(x_L + a)}{2} \times n_{G11} + n_{G12} \\
g_2 = \frac{0.25(1 - r)^2}{P_L} + \frac{0.5r(1 - r)\Phi(x_L - d) + 0.25r^2\Phi(x_L + a)}{2} \times n_{G21} + n_{G22} \\
g_3 = \frac{0.5r(1 - r)\Phi(x_L - d) + 0.25r^2\Phi(x_L + a)}{2} \times n_{G31} + n_{G32} \\
g_4 = \frac{0.25(1 - r)^2}{P_L} + \frac{0.5r(1 - r)\Phi(x_L - d) + 0.25r^2\Phi(x_L + a)}{2} \times n_{G41} + n_{G42} \\
g_5 = \frac{0.5r(1 - r)\Phi(x_L - d) + 0.25r^2\Phi(x_L + a)}{2} \times n_{G51} + n_{G52}

Plant Communications 3, 100319, May 9 2022 © 2022 The Author(s).
where \( r \) is the recombination fraction between the current marker and putative QTG; \( P_L \) and \( P_H \) are the sampling fractions in low and high pools, respectively; \( P_L(x_i) \) and \( P_H(x_i) \) are the probabilities of genotypes in low and high pools, respectively; \( n_Q \) and \( n_G \) are the numbers of read counts of marker genotypes in low and high pools, respectively; \( a \) and \( d \) are the additive and dominant effects standardized by \( e \); respectively; \( x_L = (x_i - a)/e \) and \( x_H = (x_i - a)/e \) are the standardized cut points in low and high pools, respectively; \( \mu \) and \( \sigma^2 \) are mean and residual variance, respectively; and \( \Phi(\cdot) \) is the probability function of the standard normal distribution (Lander and Botstein, 1989);

\[ P_L = \frac{0.25(1 - r)^2 \Phi(x_L - a) + 0.5 r(1 - r) \Phi(x_L - d) + 0.25 r^2 \Phi(x_L) + 0.5(1 - r) r \Phi(x_L - a) + 0.5(1 - 2r + 2r^2) \Phi(x_L - d) + 0.5(1 - r) r \Phi(x_L + a)}{P_L} \]

\[ P_H = \frac{0.25(1 - r)^2 [1 - \Phi(x_H - a)] + 0.5 r(1 - r) [1 - \Phi(x_H - d)] + 0.25 r^2 [1 - \Phi(x_H + a)] + 0.5 r(1 - r) [1 - \Phi(x_H - a)] + 0.25 r^2 [1 - \Phi(x_H + a)] + 0.5 r(1 - r) [1 - \Phi(x_H - d)] + 0.25 r^2 [1 - \Phi(x_H + d)]}{P_H} \]

\[ f_{ML} = \frac{0.5(1 - r) \Phi(x_L - a) + 0.5(1 - 2r + 2r^2) \Phi(x_L - d) + 0.5(1 - r) r \Phi(x_L + a)}{P_L} \]

\[ f_{MLH} = \frac{0.5(1 - r) r \Phi(x_H - a) + 0.5(1 - 2r + 2r^2) \Phi(x_H - d) + 0.5(1 - r) [1 - \Phi(x_H + a)]}{P_H} \]

\[ |f_{ML} - f_{MLH}| = (1 - 2r)|f_{ML} - f_{MLH}| \]

where \( f_{ML} \) (\( f_{MLH} \)) and \( f_{MLH} \) (\( f_{ML} \)) are the frequencies of read counts of marker allele \( M \) (\( Q \)) at the currently detected marker (or QTG) in low and high pools, respectively. The five parameters above are solved using the R package BB. The initial value for \( f_{ML} - f_{MLH} \) is set as the maximum difference of marker allele read frequencies in the current chromosome between high and low pools; the initial values for \( x_L \) and \( x_H \) were determined based on \( P_L \) and \( P_H \); \( a = 0 \), and \( d = 0 \); the initial values for \( a \) and \( d \) are set to zero; the initial value for \( r \) is set to the recombinant fraction between the putative QT and the marker with the maximum allele frequency difference on the chromosome based on the genetic relationship between physical and genetic distances (Supplemental Table 1) and the Haldane function (Haldane, 1919). If it does not converge in the estimation of five parameters, let \( G_2 \) be zero.

**Construction of a new statistic, \( G_w \)**

A new statistic, \( G_w \), is a weighted mean of \( G_1 \) and \( G_2 \) and is indicated by

\[ G_w = \frac{G_1}{G_1 + G_2} \times G_1 + \frac{G_2}{G_1 + G_2} \times G_2 \]

where \( G_1 = \sum_{i=1}^{2} \sum_{j=1}^{3} n_Q \ln \left( \frac{n_Q}{E(n_Q)} \right) \) and \( G_2 = \sum_{i=1}^{2} \sum_{j=1}^{3} n_G \ln \left( \frac{n_G}{E(n_G)} \right) \) are standard G statistics (Magwene et al., 2011); \( E(n_Q) \) and \( E(n_G) \) are the theoretical numbers of read counts of marker alleles and genotypes, which are calculated from the \( 2 \times 2 \) and \( 2 \times 3 \) contingency tables, respectively. The Smooth \( G_w \) value is a weighted average of the original statistic (\( G_w \)) values for all the SNPs in a window (Magwene et al., 2011) and is indicated by

\[ \text{Smooth} \ G_w = \sum_{j \in \text{a window}} k_j G_w(j) \]

Although \( G_1 \) and \( G_2 \) obey the \( \chi^2 \) distribution, they are correlated. Thus, the distribution of \( G_w \) is unclear. In this case, permutation experiments are used to determine the critical value of \( G_w \) for significant QTGs (Fisher, 1935; Pitman, 1937).

**Monte Carlo simulation studies**

The purposes of a series of Monte Carlo simulation studies were to confirm whether the new methods could identify fully and extremely over-dominant and small-effect genes, to explore the conditions under which dQTL-seq was feasible, and to obtain the best sampling plans for the new methods.

In the first simulation experiment, one small-effect (5%) QTL with various degrees of dominance (0.0, 1.0, 2.0, and 3.0) was simulated to imitate a secondary F₂ population with a sample size of 500. This QTL was placed at the position of the 250th marker on a chromosome 100 cM in length covered by 500 evenly distributed markers. Their phenotypes were
simulated from the population mean (100), the QTL effect, and residual variance (10). The sampling fraction was 20% for the high (low) pool. This was simulated dataset I. Each sample was analyzed by the dQTG-seq1, G’, aSNP, and ED methods. The critical values of significant QTLs at the 0.10 and 0.05 probability levels were determined by 10,000 permutation experiments for ED, dQTG-seq1, and dQTG-seq2 and by using the R package QTLeqPCR (Mansfeld and Grumet, 2018) for the G’ and aSNP methods. The number of replications was 100 for ICIM and 1,000 for the other methods, and their results were used to evaluate the statistical power for this QTL, while the detected QTLs within 1 cM of the simulated QTL were viewed as true.

In the second simulation experiment, one small-effect (5%) QTL was simulated under extremely (\(d/a = 4\)) and fully (\(a = 0\) and \(d = 0\)) over-dominant models to imitate a secondary \(F_2\) population with a sample size of 500. The others were the same as those in simulated dataset I. This was simulated dataset II. In the third simulation experiment, five QTLs were simulated under extremely (\(d/a = 4\)) and fully (\(a = 0\) and \(d = 0\)) over-dominant models to imitate a primary \(F_2\) population with a sample size of 500. Their positions and effects are shown in Supplemental Table 3. The others were the same as those in simulated dataset I. This was simulated dataset III.

In the fourth simulation experiment, two QTLs with \(d/a = 0.5\) were simulated on one chromosome. The sample size in the \(F_2\) was 500, each QTL size was 10%, and QTL positions were placed at 25.00 cM and 45.00 cM. The others were the same as those in simulated dataset I. This was simulated dataset IV. In the fifth simulation experiment, we investigated the effect of sample size, QTL size, and number of plants in each extreme pool on the statistical power of dQTG-seq1 and dQTG-seq2. QTL sizes were set to 2%, 5%, and 10%; sample sizes in the \(F_2\) were set to 250, 500, 1,000, and 2,000; and the number of plants in each extreme pool was set to \(30, 50, 80\). The others were the same as those in simulated dataset I. The simulated datasets of the 36 parameter combinations were analyzed by dQTG-seq1 and dQTG-seq2 to obtain suitable sampling plans for the two new methods. This was simulated dataset V. In the sixth simulation experiment, to investigate the effect of marker segregation distortion on QTL detection, one QTL with \(d/a = 1.0\) was simulated in the \(F_2\). The sample size was 500. The QTL position overlapped with a segregation distortion locus (SDL) at 50.00 cM, the sizes of both QTL and SDL were 10%, and the genetic models of the SDL were set to additive (\(a \neq 0\) and \(d = 0\)), completely dominant (\(d/a = 1.0\)), and fully over-dominant (\(a = 0\) and \(d \neq 0\)). The others were the same as those in simulated dataset I. This was simulated dataset VI.

The above simulated datasets are found in Supplemental Data 1.

SUPPLEMENTAL INFORMATION
Supplemental Information is available at Plant Communications Online.

FUNDING
This work was supported by the National Natural Science Foundation of China (32070557 and 31871242), the Fundamental Research Funds for the Central Universities (2662020ZKPY017), and the Huazhong Agricultural University Scientific and Technological Self-Innovation Foundation (2014RC020).

AUTHOR CONTRIBUTIONS
Y.-M.Z. conceived and designed the study. P.L. wrote the code and conducted Monte Carlo simulation studies. P.L., G.L., Y.-W.Z., J.-F.Z., J.-Y.L., and Y.-M.Z. performed the data analyses. P.L. and Y.-M.Z. developed the method, wrote the draft, and revised the manuscript. All authors reviewed the manuscript.

ACKNOWLEDGMENTS
We thank Prof. Lin Li (Huazhong Agricultural University) for providing the maize dataset, Prof. Jim M. Dunwell (School of Agriculture, Policy and Development, University of Reading) and Mr. Hanwen Zhang (Hence Education Ltd., Vancouver, Canada; hywenzhang@henceedu.com) for improving the language in the manuscript, and senior engineer Ms. Ju Huang for help with the data analysis. No conflicts of interest by the authors are declared.

Received: July 27, 2021
Revised: March 7, 2022
Accepted: March 22, 2022
Published: May 09, 2022

REFERENCES
Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., Matsumura, H., Yoshida, K., Mitsuoka, C., Tamiru, M., et al. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. Nat. Biotechnol. 30:174–178. https://doi.org/10.1038/nbt.2095. PMID: 22267009.
Austin, R.S., Vidaurre, D., Stamatiou, G., Breit, R., Provart, N.J., Bonetta, D., Zhang, J., Fung, P., Gong, Y., Wang, P.W., et al. (2011). Next-generation mapping of Arabidopsis genes. Plant J. 67:715–725. https://doi.org/10.1111/j.1365-313X.2011.04619.x.
Benowitz, K.M., Coleman, J.M., and Matzkin, L.M. (2019). Assessing the architecture of Drosophila mojavensis locomotor evolution with bulk segregant analysis. G3 9:1767–1775. https://doi.org/10.1534/g3.119.400036.
Bowen, M.E., Henke, K., Siegfried, K.R., Warman, M.L., and Harris, M.P. (2012). Efficient mapping and cloning of mutations in zebrafish by low-coverage whole-genome sequencing. Genetics 190:1017–1024. https://doi.org/10.1534/genetics.111.136069.
Claesen, J., and 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. https://doi.org/10.1515/sagmb-2014-0007.
Cockram, J., and Mackay, I. (2018). Genetic mapping populations for conducting high-resolution trait mapping in plants. Adv. Biochem. Eng. Biotechnol. 164:109–138. https://doi.org/10.1007/10_2017_48.
Cuperus, J.T., Montgomery, T.A., Fahlgren, N., Burke, R.T., Townsend, T., Sullivan, C.M., and Carrington, J.C. (2010). Identification of MIR390a precursor processing-defective mutants in Arabidopsis by direct genome sequencing. Proc. Natl. Acad. Sci. U S A 107:466–471. https://doi.org/10.1073/pnas.0913203107.
Darvasi, A., and Soller, M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor. Appl. Genet. 85:353–359. https://doi.org/10.1007/BF00228881.
Darvasi, A., and Soller, M. (1994). Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. Genetics 138:1365–1373. https://doi.org/10.1093/genetics/138.4.1365.
Duitama, J., Sánchez-Rodríguez, A., Goovaerts, A., Pulido-Tamayo, S., Hubmann, G., Foulqué-Moreno, M.R., Thevelein, J.M., Verstrepen, K.J., and Marchal, K. (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. https://doi.org/10.1186/1471-2164-15-207.
Edwards, M.D., and Gifford, D.K. (2012). High-resolution genetic mapping with pooled sequencing. BMC Bioinf. 13:S8. https://doi.org/10.1186/1471-2105-13-S6-S8.
Ehrenreich, I.M., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, J.A., Gresham, D., Caudy, A.A., Kruglyak, L., et al. (2010). Dissection of genetically complex traits with extremely large pools of...
A Combinatorial Strategy to Detect all Types of QTLs in the F₂

yeast segregants. Nature 464:1039–1042. https://doi.org/10.1038/nature08923.

Fekih, R., Takagi, H., Tamiru, M., Abe, A., Natsume, S., Yagashii, H., Sharma, S., Kanzaki, H., Matsumura, H., and Saitho, H. (2013). MutMap+: genetic mapping and mutant identification without crossing in rice. PLoS One 8:e68529. https://doi.org/10.1371/journal.pone.0068529.

Feng, X., Cheng, H., Zuo, D., Zhang, Y., Wang, Q., Liu, K., Ashraf, J., Yang, Q., Li, S., Chen, X., et al. (2019). Fine mapping and identification of the fuzzless gene GaFd in DPL972 (Gossypium arboreum). Theor. Appl. Genet. 132:2169–2179. https://doi.org/10.1007/s00122-019-03330-3.

Fisher, R.A. (1935). The Design of Experiments (Edinburgh: Oliver & Boyd).

Gardiner, J.M., Coe, E.H., Meliahancock, S., Hoisington, D.A., and Chao, S. (1993). Development of a core RFLP map in maize using an immortalized F₂ population. Genetics 134:917–930. https://doi.org/10.1093/genetics/134.3.917.

Ghavideh, F.Z., Claesen, J., and Burzykowski, T. (2015). A nonhomogeneous hidden Markov model for gene mapping based on next-generation sequencing data. J. Comput. Biol. 22:178–188. https://doi.org/10.1089/cmb.2014.0258.

Giovannoni, J.J., Wing, R.A., Ganal, M.W., and Tanksley, S.D. (1991). Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. Nucleic Acids Res. 19:6553–6558. https://doi.org/10.1093/nar/19.23.6553.

Haase, N.J., Beissinger, T., Hirsch, C.N., Vaillancourt, B., Deshpande, S., Barry, K., Buell, C.R., Kaeppler, S.M., and deLeon, N. (2015). Shared genomic regions between derivatives of a large segregating population of maize identified using bulked segregant analysis with next-generation sequencing data. J. Comput. Biol. 22:178–188. https://doi.org/10.1089/cmb.2014.0258.

Haldane, J.B.S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. J. Genet. 8:299–309.

Hill, J.T., Demarest, B.L., Bisgrove, B.W., Gorsi, B., Su, Y.C., and Yost, H.J. (2013). MMAPP: mutation mapping analysis pipeline for pooled RNA-seq. Genome Res. 23:687–697. https://doi.org/10.1101/gr.149636.112.

Hua, J., Xing, Y., Wu, W., Xu, C., Sun, X., Yu, S., and Zhang, G. (2003). Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. Proc. Natl. Acad. Sci. U S A 100:2574–2579. https://doi.org/10.1073/pnas.0437907100.

Huang, W., Richards, S., Carbone, M.A., Zhu, D., Anholt, R.R., Ayroles, J.F., Duncan, L., Jordan, K.W., Lawrence, F., Magwire, M.M., et al. (2012). Epistasis dominates the genetic architecture of Drosophila quantitative traits. Proc. Natl. Acad. Sci. U S A 109:15553–15559. https://doi.org/10.1073/pnas.1213423109.

Kroymann, J., and Mitchell-Olds, T. (2005). Epistasis and balanced polymorphism influencing complex trait variation. Nature 435:95–98. https://doi.org/10.1038/nature03480. PMID: 15875023.

Lander, E.S., and Botstein, D. (1989). Mapping Mendelian factor underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199. https://doi.org/10.1093/genetics/121.1.185.

Lebowitz, R.J., Soller, M., and Beckmann, J.S. (1987). Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor. Appl. Genet. 73:556–562. https://doi.org/10.1007/BF00289194.

Leshchiner, I., Alexa, K., Kelsey, P., Adzhubei, I., Austin-Tse, C.A., Cooney, J.D., Anderson, H., King, M.J., Stottrmann, R.W., Garnaas, M.K., et al. (2012). Mutation mapping and identification by whole-genome sequencing. Genome Res. 22:1541–1548. https://doi.org/10.1101/gr.135541.111.

Li, Z., and Xu, Y. (2021). Bulk segregation analysis in NGS era: a review for its teenage years. Plant J 109:1355–1374. https://doi.org/10.1111/tpj.15646.

Li, H., Ye, G., and Wang, J. (2007). A modified algorithm for the improvement of composite interval mapping. Genetics 175:361–374. https://doi.org/10.1534/genetics.106.066811.

Liu, P., Du, C., Zhang, Y., Yin, S., Zhang, E., Fang, H., Lin, D., Xu, C., and Yang, Z. (2018). Combined bulked segregant sequencing and traditional linkage analysis for identification of candidate gene for purple leaf sheath in maize. PLoS One 13:e0190670. https://doi.org/10.1371/journal.pone.0190670.

Lindner, H., Raissig, M.T., Sailer, C., Shimosak-Asano, H., Bruggmann, R., and Grossniklaus, U. (2012). SNP-Ratio Mapping (SRM): identifying lethal alleles and mutations in complex genetic backgrounds by next-generation sequencing. Genetics 191:1381–1386. https://doi.org/10.1534/genetics.111.141341.

Liu, G., Zhao, T., You, X., Jiang, J., Li, J., and Xu, X. (2019). Molecular mapping of the Cl-10 gene by combining SNP/InDel-index and linkage analysis in tomato (Solanum lycopersicum). BMC Plant Biol. 19:15. https://doi.org/10.1186/s12870-018-1616-7.

Mackay, T.F., Stone, E.A., and Ayroles, J.F. (2009). The genetics of quantitative traits: challenges and prospects. Nat. Rev. Genet. 10:565–577. https://doi.org/10.1038/nrg2612.

Magwene, P.M., Willis, J.H., and Kelly, J.K. (2011). The statistics of bulk segregant analysis using next generation sequencing. PLoS Comput. Biol. 7:e1002255. https://doi.org/10.1371/journal.pcbi.1002255.

Mansfeld, B.N., and Grumet, R. (2018). QTLseqr: an R package for bulk segregant analysis with next-generation sequencing. Plant Genome 11:180006. https://doi.org/10.3835/plantgenome2018.01.0006.

Michelmore, R.W., Paran, I., and Kesseli, R.V. (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. https://doi.org/10.1073/pnas.88.21.9828.

Minevich, G., Park, D.S., Blankenberg, D., Poole, R.J., and Hobert, O. (2012). CloudMap: a cloud-based pipeline for analysis of mutant genome sequences. Genetics 192:1249–1269. https://doi.org/10.1534/genetics.112.144204.

Obholzer, N., Swinburne, I.A., Schwab, E., Nechiporuk, A.V., Nicolson, T., and Megason, S.G. (2012). Rapid positional cloning of zebratfish mutations by linkage and homzygosy mapping using whole-genome sequencing. Development 139:4280–4290. https://doi.org/10.1242/dev.083931.

Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any populations. J. Roy. Stat. Soc. B4:119–130. https://doi.org/10.2307/2984124.

Ruyter-Spira, C.P., Gu, Z.L., VanderPoel, J.J., and Groenen, M.A.M. (1997). Bulked segregant analysis using microsatellites: mapping of the dominant white locus in the chicken. Poult. Sci. 76:386–391. https://doi.org/10.1093/ps/76.2.386.

Schneeberger, K. (2014). Using next-generation sequencing to isolate mutant genes from forward genetic screens. Nat. Rev. Genet. 15:662–676. https://doi.org/10.1038/nrg3745.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jørgensen, J.E., Weigel, D., and Andersen, S.U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat. Methods 6:550–551. https://doi.org/10.1038/nmeth0809-550.
A Combinatorial Strategy to Detect all Types of QTLs in the $F_2$

Smith, H.E., Fabritius, A.S., Jaramillo-Lambert, A., and Golden, A. (2016). Mapping challenging mutations by whole-genome sequencing. G3 6:1297–1304. https://doi.org/10.1534/g3.116.028316.

Stubber, C.W., Edwards, M.D., and Wendel, J.F. (1987). Molecular marker-facilitated investigations of quantitative trait loci in maize. II. factors influencing yield and its component traits. Crop Sci. 27:639–648. https://doi.org/10.1094/CS-27-0639.

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., et al. (2013a). QTLseq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J. 74:174–183. https://doi.org/10.1111/tpj.12105.

Takagi, H., Uemura, A., Yaegashi, H., Tamiru, M., Abe, A., Mitsuoka, C., Utsushi, H., Natsume, S., Kanzaki, H., Matsumura, H., et al. (2013b). MutMap-Gap: whole-genome resequencing of mutant $F_2$ progeny bulk combined with de novo assembly of gap regions identifies the rice blast resistance gene $Pii$. New Phytol. 200:276–283. https://doi.org/10.1111/nph.12369.

Wang, S.B., Wen, Y.J., Ren, W.L., Ni, Y.L., Zhang, J., Feng, J.Y., and Zhang, Y.M. (2016). Mapping small-effect and linked quantitative trait loci for complex traits in backcross or DH populations via a multi-locus GWAS methodology. Sci. Rep. 6:29951. https://doi.org/10.1038/srep29951.

Wang, C., Tang, S., Zhan, Q., Hou, Q., Zhao, Y., Zhao, Q., Zhou, C., Lyu, D., Cui, L., et al. (2019). Dissecting a heterotic gene through GradedPool-Seq mapping informs a rice-improvement strategy. Nat. Commun. 10:2982. https://doi.org/10.1038/s41467-019-11076-y.

Wang, L., Lu, Z., Regulski, M., Jiao, Y., Chen, J., Ware, D., and Xin, Z. (2021). BSAseq: an interactive and integrated web-based workflow for identification of causal mutations in bulked $F_2$ populations. Bioinformatics 37:382–387. https://doi.org/10.1093/bioinformatics/btaa709.

Wen, Y.J., Zhang, Y.W., Zhang, J., Feng, J.Y., Dunwell, J.M., and Zhang, Y.M. (2019). An efficient multi-locus mixed model framework for the detection of small and linked QTLs in $F_2$. Brief. Bioinform. 20:1913–1924. https://doi.org/10.1093/bib/bby058.

Yamanaka, N., Watanabe, S., Toda, K., Hayashi, M., Fuchigami, H., Takahashi, R., and Harada, K. (2005). Fine mapping of the $FT1$ locus for soybean flowering time using a residual heterozygous line derived from a recombinant inbred line. Theor. Appl. Genet. 110:634–639. https://doi.org/10.1007/s00122-004-1886-3.

Yan, W., Chen, Z., Lu, J., Xu, C., Xie, G., Li, Y., Deng, X.W., He, H., and Tang, X. (2016). Simultaneous identification of multiple causal mutations in rice. Front. Plant Sci. 7:2055. https://doi.org/10.3389/fpls.2016.02055.

Yu, J., Pressoir, G., Briggs, W.H., Vroh, B.I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., and Holland, J.B. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208. https://doi.org/10.1038/ng1702.

Zeng, Z.B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. U S A 90:10972–10976. https://doi.org/10.1073/pnas.90.23.10972.

Zhang, Y.M., and Xu, S. (2004). Mapping quantitative trait loci in $F_2$ incorporating phenotypes of $F_3$ progeny. Genetics 166:1981–1993. https://doi.org/10.1534/genetics.166.1.1981.

Zhang, Q., Shen, B.Z., Dai, X.K., Mei, M.H., Maroof, M.A., and Li, Z.B. (1994). Using bulked extremes and recessive class to map genes for photoperiod-sensitive genic male sterility in rice. Proc. Natl. Acad. Sci. U S A 91:8675–8679. https://doi.org/10.1073/pnas.91.18.8675.

Zhang, Y.M., Mao, Y., Xie, C., Smith, H., Luo, L., and Xu, S. (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize ($Zea$ mays L.). Genetics 166:2267–2275. https://doi.org/10.1534/genetics.104.03217.

Zhang, H., Wang, X., Pan, Q., Li, P., Liu, Y., Lu, X., Zhong, W., Li, M., Han, L., and Li, J. (2019). QTG-Seq accelerates QTL fine mapping through QTL partitioning and whole-genome sequencing of bulked segregant samples. Mol. Plant 12:426–437. https://doi.org/10.1016/j.molp.2018.12.018.

Zhou, G., Chen, Y., Yao, W., Zhang, C., Xie, W., Hua, J., Xing, Y., Xiao, J., and Zhang, Q. (2012). Genetic composition of yield heterosis in an elite rice hybrid. Proc. Natl. Acad. Sci. U S A 109:15847–15852. https://doi.org/10.1073/pnas.1214141109.

Zou, C., Wang, P., and Xu, Y. (2016). Bulked sample analysis in genetics, genomics and crop improvement. Plant Biotechnol. J. 14:1941–1955. https://doi.org/10.1111/pbi.12559.