Figure S1  LOD score profiles for the first four traits of the natural bin analysis using the Lasso method, where the 12 chromosomes are separated by the dashed vertical lines and the effective number of tests corrected LOD score threshold for each trait is indicated by the dotted horizontal line. Three bins for KGW have LOD score larger than 16, but the plot is truncated at maximum 16. The LOD score thresholds (drawn from the effective number of tests) for the four traits (YD, TP, GN and KGW) are 2.09, 2.45, 2.23 and 2.48, respectively.
Figure S2  LOD score profiles for the last four traits of the natural bin analysis using the Lasso method, where the 12 chromosomes are separated by the dashed vertical lines and the effective number of tests corrected LOD score threshold for each trait is indicated by the dotted horizontal line. LOD scores larger than 16 are truncated to 16. The LOD score thresholds (drawn from the effective number of tests) for the four traits (GL, GW, HD and OsC1) are 2.56, 2.20, 1.55 and 0.90, respectively.
Figure S3  Functional relationship of the variance of bin genotype indicator variable, $\text{var}(Z_k)$, with the bin size, $\Delta_k$, measured in Morgan in a RIL population generated through repeated selfings. The two coordinates give the values of variance when the bin size equals 0.01 Morgan and 0.25 Morgan, respectively.
Figure S4  LOD score comparison of the Lasso method (color coded in blue) with the composite interval mapping (CIM) method (color coded in red) and interval mapping (IM) method (color coded in green) for the first four traits (YD, TP, GN and KGW). The three horizontal dotted lines are the genome-wide 0.05 Type I error LOD score test critical values drawn from 1000 permuted samples. Natural bins were used in the analysis and the number of bins is 1619. LOD scores for the Lasso method that have reached the upper limit of the y-axis have been truncated, i.e., the actual LOD scores for those bins are higher than this limit.
Figure S5  LOD score comparison of the Lasso method (color coded in blue) with the composite interval mapping (CIM) method (color coded in red) and the interval mapping (IM) method (color coded in green) for the last four traits (GL, GW, HD and OsC1). The three horizontal dotted lines are the genome-wide 0.05 Type I error LOD score test critical values drawn from 1000 permuted samples. Natural bins were used in the analysis and the number of bins is 1619. LOD scores for the Lasso method that have reached the upper limit of the y-axis have been truncated, i.e., the actual LOD scores of those bins are higher than this limit.
Figure S6  LOD score profiles for the first four traits of the artificial bin analysis using the Lasso method, where the 12 chromosomes are separated by the dashed vertical lines and the permutation generated LOD score threshold for each trait is indicated by the dotted horizontal line. Three bins for KGW have LOD score larger than 16, but the plot is truncated at maximum 16. The LOD score thresholds for the four traits (YD, TP, GN and KGW) are 0.83, 0.85, 0.89 and 0.79, respectively.
Figure S7  LOD score profiles for the last four traits of the artificial bin analysis using the Lasso method, where the 12 chromosomes are separated by the dashed vertical lines and the permutation generated LOD score threshold for each trait is indicated by the dotted horizontal line. LOD scores larger than 16 are truncated to 16. The LOD score thresholds for the four traits (GL, GW, HD and OsC1) are 0.84, 0.84, 0.80 and 0.90, respectively.
Figure S8  Comparison of estimated heritability (proportion of phenotypic variance explained by significant bins) between the artificial bin analysis and the natural bin analysis.
Figure S9  Breakpoints and bins in a hypothetical multi-parent MAGIC population initiated with eight parents (color coded). Panel (a): breakpoint patterns, Panel (b): natural bins, Panel (c) artificial bins.
This section provides a proof for equation (4) in the main text and shows the conditions under which the equation holds. This equation is reintroduced below,

\[ y_j = X_j \beta + \sum_{k=1}^{m} Z_{jk} \gamma_k + \varepsilon_j \]  

where

\[ Z_{jk} = \bar{Z}_j(\lambda_k) = \Delta_k^{-1} \int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} Z_j(\lambda)d\lambda \]

is the average \( Z_j \) for all (infinite number of) markers within bin \( k \) and \( \Delta_k \) is the size of the bin, defined as

\[ \Delta_k = \int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} d\lambda \]

The location of the bin is denoted by \( \lambda_k \), which is the middle point position of the bin in the genome. The genetic effect \( \gamma_k \) for bin \( k \) is defined as

\[ \gamma_k = \bar{\gamma}(\lambda_k) \Delta_k = \int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} \gamma(\lambda)d\lambda \]

which is the sum of effects for all (infinite number of) markers in bin \( k \). Equation (1) is an approximation. The exact version of the equation should be

\[ y_j = X_j \beta + \sum_{k=1}^{m} \alpha_{jk} + \varepsilon_j \]

where

\[ \alpha_{jk} = \int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} Z_j(\lambda)\gamma(\lambda)d\lambda \]

Let us rewrite equation (5) using

\[ \alpha_{jk} = \int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} \left[ Z_j(\lambda) - \bar{Z}_j(\lambda_k) + Z_{jk} \right] \left[ \gamma(\lambda) - \bar{\gamma}(\lambda_k) + \Delta_k^{-1} \gamma_k \right] d\lambda \]

This is because \( Z_{jk} = \bar{Z}_j(\lambda_k) \) and \( \gamma_k = \bar{\gamma}(\lambda_k) \Delta_k \) as defined in equations (2) and (4), respectively. Expanding the product of equation (7) yields
\[ \alpha_{jk} = \int_{\lambda_{k}-\Delta_k}^{\lambda_{k}+\Delta_k} Z_{jk} \gamma_k \Delta_k^{-1} d\lambda + \int_{\lambda_{k}-\Delta_k}^{\lambda_{k}+\Delta_k} \left[ Z_j(\lambda) - \bar{Z}_j(\lambda_k) \right] [\gamma(\lambda) - \bar{\gamma}(\lambda_k)] d\lambda \]

\[ = Z_{jk} \gamma_k \Delta_k^{-1} \int_{\lambda_{k}-\Delta_k}^{\lambda_{k}+\Delta_k} d\lambda + \int_{\lambda_{k}-\Delta_k}^{\lambda_{k}+\Delta_k} \left[ Z_j(\lambda) - \bar{Z}_j(\lambda_k) \right] [\gamma(\lambda) - \bar{\gamma}(\lambda_k)] d\lambda \quad (8) \]

\[ = Z_{jk} \gamma_k \int_{\lambda_{k}-\Delta_k}^{\lambda_{k}+\Delta_k} \left[ Z_j(\lambda) - \bar{Z}_j(\lambda_k) \right] [\gamma(\lambda) - \bar{\gamma}(\lambda_k)] d\lambda \]

When we derived the above equation, we used the following equivalents to simplify the derivation,

\[ \int_{\lambda_{k}-\Delta_k}^{\lambda_{k}+\Delta_k} \left[ Z_j(\lambda) - \bar{Z}_j(\lambda_k) \right] d\lambda = \int_{\lambda_{k}-\Delta_k}^{\lambda_{k}+\Delta_k} [\gamma(\lambda) - \bar{\gamma}(\lambda_k)] d\lambda = 0 \quad (9) \]

The second term of equation (8) will disappear if either \( Z_j(\lambda) = \bar{Z}_j(\lambda_k) \) or \( \gamma(\lambda) = \bar{\gamma}(\lambda_k) \)

for \( \lambda \in \Delta_k \). For the natural bins, all markers within a bin have an identical genotype and thus the first condition applies. The second condition means that all loci within the same bin have the same genetic effect. This condition is out of our control. Further examination of equation (8), we realized that this integration of the product can be interpreted as the covariance between \( Z_j \) and \( \gamma \),

\[ \text{cov}(Z_j, \gamma) = \int_{\lambda_{k}-\Delta_k}^{\lambda_{k}+\Delta_k} \left[ Z_j(\lambda) - \bar{Z}_j(\lambda_k) \right] [\gamma(\lambda) - \bar{\gamma}(\lambda_k)] d\lambda \quad (10) \]

There is no reason to believe that the genetic effect profile (a population parameter) is correlated to the segregation pattern of markers of an individual within the population. Therefore, we may safely ignore this covariance and replace \( \alpha_{jk} \) by \( Z_{jk} \gamma_k \).
File S2

Variance of Lasso Estimate of Bin Effect

The GlmNet/R program (Friedman et al. 2010) does not provide the standard error for the Lasso estimate $\hat{\gamma}_k$ (Tibshirani 1996). Here we present an approximate formula to calculate the standard error. Let $\text{var}(\hat{\gamma}_k)$ be the variance of the estimated bin effect. The standard error simply takes the square root of this variance. The Lasso estimate of a bin effect is a kind of shrinkage estimate. Therefore, $\hat{\gamma}_k$ and $\text{var}(\hat{\gamma}_k)$ are considered as the “posterior mean” and “posterior variance” of $\gamma_k$, respectively. We may find the “prior variance” retrospectively using $\hat{\gamma}_k$ and the data. If $\gamma_k$ were estimated from a single bin model, the variance of that estimate would be

$$ \text{var}(\gamma_k) = (Z_k^T Z_k)^{-1} \hat{\sigma}^2 $$  \hspace{1cm} (11)

where

$$ \hat{\sigma}^2 = \frac{1}{n} (y - X \hat{\beta} - Z \hat{\gamma})^T (y - X \hat{\beta} - Z \hat{\gamma}) $$ \hspace{1cm} (12)

is the estimated residual error variance. Let $\sigma_k^2$ be the “prior variance” of $\gamma_k$. Combining the prior variance $\sigma_k^2$ and the variance from the data $\text{var}(\hat{\gamma}_k)$, we can generate the posterior variance expressed by

$$ \text{var}(\hat{\gamma}_k) = \left( \frac{1}{\sigma_k^2} + \frac{1}{\text{var}(\hat{\gamma}_k)} \right)^{-1} = \frac{\sigma_k^2 \hat{\sigma}^2}{\hat{\sigma}^2 + \sigma_k^2 Z_k^t Z_k} $$ \hspace{1cm} (13)

When the prior variance is estimated from the data, it can be interpreted as the expectation of $\gamma_k^2$,

$$ \sigma_k^2 = E(\gamma_k^2) = \hat{\gamma}_k^2 + \text{var}(\hat{\gamma}_k) $$ \hspace{1cm} (14)

Therefore,

$$ \sigma_k^2 = \hat{\gamma}_k^2 + \frac{\sigma_k^2 \hat{\sigma}^2}{\hat{\sigma}^2 + \sigma_k^2 Z_k^t Z_k} $$ \hspace{1cm} (15)

The above equation has only one unknown quantity, $\sigma_k^2$, which has two solutions with the positive one being

$$ \hat{\sigma}^2 = \frac{\hat{\gamma}_k^2 Z_k^t Z_k + \sqrt{\hat{\gamma}_k^2 Z_k^t Z_k})^2 + 4\sigma_k^2 \hat{\gamma}_k^2 Z_k^t Z_k}}{2Z_k^t Z_k} $$ \hspace{1cm} (16)

Therefore, the posterior variance $\text{var}(\hat{\gamma}_k)$ is obtained by substituting $\sigma_k^2$ appearing in equation (13) by the solution given in equation (16),

$$ \text{var}(\hat{\gamma}_k) = \left( \frac{1}{\hat{\sigma}^2} + \frac{1}{\text{var}(\hat{\gamma}_k)} \right)^{-1} = \frac{\sigma_k^2 \hat{\sigma}^2}{\hat{\sigma}^2 + \sigma_k^2 Z_k^t Z_k} $$ \hspace{1cm} (17)

When $m \gg n$ and $n$ is very large, the Lasso estimate of $\sigma^2$ is often near zero (perfect fit), leading to $\sigma_k^2 \approx \hat{\gamma}_k^2$. The corresponding Wald test under $\sigma_k^2 \approx \hat{\gamma}_k^2$ is then
\[
W_k = \hat{\gamma}_k^2 \frac{\hat{\sigma}_k^2 + \hat{\gamma}_k^2 Z_k^T \hat{\gamma}_k^2 Z_k^T}{\hat{\sigma}_k^2} \approx \hat{\gamma}_k^2 \frac{\hat{\gamma}_k^2 Z_k^T \hat{\gamma}_k^2 Z_k^T}{\hat{\gamma}_k^2 \hat{\gamma}_k^2 Z_k^T \hat{\gamma}_k^2 Z_k^T + 1}
\]

which was given by Hu et al. (2012) without providing a proof.

The posterior variance \( \text{var}(\hat{\gamma}_k) \) provides a convenient way to draw the Wald test statistic. The retrospectively inferred “prior variance” \( \hat{\sigma}_k^2 \) also gives us a chance to evaluate the “degree of confidence” for each bin effect and “the effective number of tests”, which were proposed by MacKay (1992) and Tipping (2001). The degree of confidence (also called degree of freedom) for bin \( k \) is defined by

\[
d_k = 1 - \frac{\text{var}(\hat{\gamma}_k)}{\hat{\sigma}_k^2} = \frac{\hat{\sigma}_k^2 - \text{var}(\hat{\gamma}_k)}{\hat{\sigma}_k^2} = \frac{\hat{\sigma}_k^2 Z_k^T Z_k}{\hat{\sigma}_k^2 + \hat{\sigma}_k^2 Z_k^T Z_k}
\]

Note that \( 0 \leq d_k \leq 1 \), with 1 indicating perfect confidence and 0 indicating no confidence. In fact, \( d_k \) is the reduction of the posterior variance relative to the prior variance. In shrinkage analysis, not every single test is counted as a test. If a bin effect has a confidence \( d_k = 0.5 \), this test only counts as a “half test”. For the entire genome, we have \( m \) bins and thus \( m \) Wald tests, but the “effective number of tests” is only

\[
m_e = \sum_{k=1}^{m} d_k = \sum_{k=1}^{m} \frac{\hat{\sigma}_k^2 Z_k^T Z_k}{\hat{\sigma}_k^2 + \hat{\sigma}_k^2 Z_k^T Z_k}
\]

Recall that when \( m \geq n \) and \( n \) is very large, the Lasso estimate of \( \sigma^2 \) is often near zero and \( \sigma_k^2 \approx \hat{\gamma}_k^2 \). However, most bin effects will be estimated at exactly zero. In this case, the effective number of tests simply equals the number of non-zero effects. The effective number of tests implies a different method to correct multiple tests, if it is necessary to consider such a correction. Assume that the nominal \( p \)-value criterion is 0.05. The genome-wide \( p \)-value criterion after Bonferroni correction for multiple tests should be \( 0.05 / m_e \), rather than \( 0.05 / m \). The conventional Bonferroni correction using the number of bins to adjust for the \( p \)-value is too conservative (over-correction) and the effective number of tests will correct the over-correction. The effective number of test share a similar nature as the QTL intensity in the Bayesian QTL mapping (Sillanpää and Arjas 1998; Sillanpää and Arjas 1999).

Table S3 lists the effective numbers of tests for all the eight traits under the natural bin analysis. Taking the yield (YD) trait for example, the effective number of tests is 26.56, which leads to a critical genome-wide \( p \)-value of \( 0.05 / 26.56 = 0.00188 \) after the Bonferroni correction for multiple tests. Whether a bin is associated or not with YD, we should use 0.00188 as the criterion, rather than the 0.05 criterion. Alternatively, the Chi-square one criterion should be \( \chi^2 (1 - 0.00188) = \chi^2 (0.99812) = 9.63 \), which is converted into a LOD score criterion of \( 9.63/4.61 = 2.09 \). In other words, a bin is declared as significant if its Wald test statistic is larger than 9.63 or its LOD score is greater than 2.09. When this criterion (effective number of tests) is used, no bins are associated with YD. If the original Bonferroni correction were conducted, the \( p \)-value criterion would be \( 0.05/1619 = 3.0883 \times 10^{-5} \), which is even more stringent. The LOD score tests and the thresholds using this effective number of tests criterion are illustrated
in Supplemental Figure S1 for traits YD, TP, GN, KGW and Figure S2 for traits GL, GW, HD and OsC1. Among the eight traits, the first six traits have relatively large effective numbers of tests. The last two traits have small effective numbers of tests, especially, the single-gene-controlled trait OsC1 has $m_e = 1.2$, which is virtually a single gene test. This is a known single-gene-controlled trait and, theoretically, the effective number of tests should be exactly one. The known gene is located in bin 868 and this bin completely co-segregates with the color trait, whose LOD score and confidence are $LOD_{868} = 46737$ and $d_{868} = 0.999995$, respectively. However, out of 210 lines, 209 lines co-segregate with the bin 867 (the one next to the bin containing the color gene). This bin (bin 867) has a LOD score 0.0609, a $p$-value 0.5964 and a degree of confidence 0.212056. As a result, the effective number of tests for OsC1 is $0.999995 + 0.212056 = 1.2121$. Interestingly, a single mismatch causes a LOD score drop from 46737 (perfect match) to 0.6 (one mismatch). This kind of result would not be possible if one bin were fit to the model at a time.

**References**


Variance of Bin Genotype Indicator Variable

1. Backcross (BC) population

Let \( A_1A_1 \) and \( A_2A_2 \) be the genotypes of two inbred lines \( (P_1 \) and \( P_2 \) ) used to initiate the cross of interest. The \( F_1 \) hybrid has a genotype of \( A_1A_2 \). A backcross population generated by \( F_1 \times P_1 \) is denoted by mating type \( A_1A_2 \times A_1A_1 \). The BC population contains two possible genotypes, \( A_1A_1 \) and \( A_2A_2 \). Let us denote the genotype indicator variable of individual \( j \) at locus \( \tau \) by \( Z_j(\tau) \), which is numerically coded by

\[
Z_j(\tau) = \begin{cases} 
1 & \text{for } A_1A_1 \\
0 & \text{for } A_1A_2 
\end{cases}
\]

Under Mendelian segregation, the two genotypes have an equal probability. The variance of \( Z_j(\tau) \) across all individuals within this BC population is

\[
\text{var}[Z_j(\tau)] = \mathbb{E}[Z_j^2(\tau)] - \mathbb{E}^2[Z_j(\tau)] = \frac{1}{2} - \left( \frac{1}{2} \right)^2 = \frac{1}{4}
\]

The covariance of \( Z_j \) between loci \( \tau \) and \( \omega \) across all individuals within the BC family is

\[
\text{cov}[Z_j(\tau), Z_j(\omega)] = (1 - 2r_{\tau\omega}) \sqrt{\text{var}[Z_j(\tau)] \text{var}[Z_j(\omega)]}
\]

where

\[
r_{\tau\omega} = \frac{1}{2} (1 - e^{-2|\tau - \omega|})
\]

is the recombination fraction between loci \( \tau \) and \( \omega \) (Haldane 1919) and \( 1 - 2r_{\tau\omega} \) is the correlation between the two loci. The distance between the two loci \( |\tau - \omega| \) is measured in Morgan. Because the variance of \( Z_j(\tau) \) is a constant \((1/4)\), equation (23) can be reformulated as

\[
\text{cov}[Z_j(\tau), Z_j(\omega)] = \frac{1}{4} (1 - 2r_{\tau\omega}) = \frac{1}{4} e^{-2|\tau - \omega|}
\]

where \( \tau \in \Delta_k \) and \( \omega \in \Delta_k \), and \( \Delta_k \) is the size of the \( k \)th bin in the genome. Let us define the average of \( Z_j(\tau) \) for all loci within this bin (with bin size \( \Delta_k \) ) for individual \( j \) (only this individual, not across all individuals with the family) by

\[
\bar{Z}_j(\Delta_k) = \frac{1}{\Delta_k} \int_{\lambda_k - \frac{1}{2}\Delta_k}^{\lambda_k + \frac{1}{2}\Delta_k} Z_j(\tau) d\tau
\]

This is an observed data point and determined by the breakpoints within bin \( k \) for this particular individual. The upper and lower limits of the above integral may be redefined by subtracting \( \lambda_k - \frac{1}{2}\Delta_k \) from them, leading to a simple expression,
\[ \bar{Z}_j(\Delta_k) = \frac{1}{\Delta_k} \int_0^{\Delta_k} Z_j(\tau) d\tau \] (27)

It should be noted that the genome location \( \tau \) here has been redefined as a number relative to this bin. The variance of \( \bar{Z}_j(\Delta_k) \) across all individuals within the BC population is defined as

\[
\text{var}[\bar{Z}_j(\Delta_k)] = \frac{1}{\Delta_k^2} \int_0^{\Delta_k} \int_0^{\Delta_k} \text{cov}[Z_j(\tau), Z_j(\omega)] d\omega d\tau \]

\[ = \frac{1}{4\Delta_k^2} \int_0^{\Delta_k} \int_0^{\Delta_k} e^{-2|\tau-\omega|} d\omega d\tau \] (28)

\[ = \frac{1}{2\Delta_k^2} \int_0^{\Delta_k} \int_0^{\Delta_k} e^{-2(\tau-\omega)} d\omega d\tau \]

The double integration has a closed form expression, and so does the variance. The closed form variance is

\[
\text{var}[\bar{Z}_j(\Delta_k)] = \frac{1}{8\Delta_k^2} (e^{-2\Delta_k} + 2\Delta_k - 1) \] (29)

which is a function of the bin size \( \Delta_k \). One can verify that

\[
\lim_{\Delta_k \to 0} \text{var}[\bar{Z}_j(\Delta_k)] = \lim_{\Delta_k \to 0} \frac{1}{8\Delta_k^2} (e^{-2\Delta_k} + 2\Delta_k - 1) = \frac{1}{4} \] (30)

\[
\lim_{\Delta_k \to \infty} \text{var}[\bar{Z}_j(\Delta_k)] = \lim_{\Delta_k \to \infty} \frac{1}{8\Delta_k^2} (e^{-2\Delta_k} + 2\Delta_k - 1) = 0 \]

The situation where \( \Delta_k \to 0 \) is equivalent to single marker and thus the variance is 1/4.

2. Double haploid (DH) population

A double haploid population is generated by duplicating the gametes of the \( F_1 \) individuals derived from the cross of two inbred lines (\( A_1A_1 \) and \( A_2A_2 \)). As a result, the DH population contains two possible genotypes, \( A_1A_1 \) and \( A_2A_2 \). The genotype indicator variable of individual \( j \) at locus \( \tau \) is coded by

\[
Z_j(\tau) = \begin{cases} 
+1 & \text{for } A_1A_1 \\
-1 & \text{for } A_2A_2 
\end{cases} \] (31)

Under Mendelian segregation, the two genotypes have an equal probability. The variance of \( Z_j(\tau) \) across all individuals within this DH population is

\[ \text{var}[Z_j(\tau)] = E[Z_j^2(\tau)] - E^2[Z_j(\tau)] = 0 \] (32)

The covariance of genotype indicator variables between two different loci within the same bin is

\[ \text{cov}[Z_j(\tau), Z_j(\omega)] = 1 - 2r_{\tau\omega} = e^{-2|\tau-\omega|} \] (33)

Using the same approach as that described in the BC population, we obtained
\[
\text{var}\left[ Z_j(\Delta_k) \right] = \frac{2}{\Delta_k} \int_0^{\Delta_k} e^{-2(\tau-\omega)} d\omega d\tau = \frac{1}{2\Delta_k^2} (e^{-2\Delta_k} + 2\Delta_k - 1) \quad (34)
\]

The limits of the variance are \( \lim_{\Delta_k \to 0} \text{var}\left[ Z_j(\Delta_k) \right] = 1 \) and \( \lim_{\Delta_k \to \infty} \text{var}\left[ Z_j(\Delta_k) \right] = 0 \).

3. F\textsubscript{2} population

A \( F\textsubscript{2} \) population is generated by selfing the \( F\textsubscript{1} \) individuals derived from the cross of two inbred lines, \( A_1A_1 \) and \( A_2A_2 \). As a result, the \( F\textsubscript{2} \) family contains three possible genotypes, \( A_1A_1, A_1A_2 \) and \( A_2A_2 \). Let us denote the genotype indicator variable of individual \( j \) at locus \( \tau \) by \( Z_j(\tau) \), which is coded by

\[
Z_j(\tau) = \begin{cases} 
+1 & \text{for } A_1A_1 \text{ with probability } 1/4 \\
0 & \text{for } A_1A_2 \text{ with probability } 1/2 \\
-1 & \text{for } A_2A_2 \text{ with probability } 1/4 
\end{cases}
\quad (35)
\]

The variance of \( Z_j(\tau) \) across all individuals within this \( F\textsubscript{2} \) population is

\[
\text{var}\left[ Z_j(\tau) \right] = \text{E}\left[ Z_j^2(\tau) \right] - \text{E}^2\left[ Z_j(\tau) \right] = \frac{1}{2} \quad (36)
\]

The covariance between the genotype indicator variables of two different loci is

\[
\text{cov}\left[ Z_j(\tau), Z_j(\omega) \right] = \frac{1}{2}(1 - 2r_{\omega\tau}) = \frac{1}{2} e^{-2|\tau-\omega|} \quad (37)
\]

Using the same approach as that described before, we obtained

\[
\text{var}\left[ Z_j(\Delta_k) \right] = \frac{1}{\Delta_k} \int_0^{\Delta_k} e^{-2(\tau-\omega)} d\omega d\tau = \frac{1}{4\Delta_k^2} (e^{-2\Delta_k} + 2\Delta_k - 1) \quad (38)
\]

The limits of the variance are \( \lim_{\Delta_k \to 0} \text{var}\left[ Z_j(\Delta_k) \right] = 1/2 \) and \( \lim_{\Delta_k \to \infty} \text{var}\left[ Z_j(\Delta_k) \right] = 0 \).

4. Recombinant inbred lines (RIL)

There are two ways to generate recombinant inbred lines: (1) repeated selfings starting from the \( F\textsubscript{1} \) individual for multiple generations until all genes are fixed and (2) repeated brother-sister matings for multiple generations until all genes are fixed. The RIL generated from selfing is called RIL1 and that generated from brother-sister mating is called RIL2. The two genotypes in the RIL population are \( A_1A_1 \) and \( A_2A_2 \). The genotype indicator variable of individual \( j \) at locus \( \tau \) is coded by

\[
Z_j(\tau) = \begin{cases} 
+1 & \text{for } A_1A_1 \\
-1 & \text{for } A_2A_2 
\end{cases}
\quad (39)
\]

Under Mendelian segregation, the two genotypes have an equal probability. The variance of \( Z_j(\tau) \) across all individuals within the RIL population is

\[
\text{var}\left[ Z_j(\tau) \right] = \text{E}\left[ Z_j^2(\tau) \right] - \text{E}^2\left[ Z_j(\tau) \right] = 1 \quad (40)
\]
which is the same as the DH population. The covariance of the genotype indicator variables between two different loci within the same bin is

\[
\text{cov}[Z_j(\tau), Z_j(\omega)] = 1 - 2R_{\omega\omega}
\]  

(41)

where

\[
R_{\omega\omega} = \frac{ar_{\omega\omega}}{1 + br_{\omega\omega}}
\]

(42)

The constant numbers, \(a\) and \(b\), depend on the type of RIL with \(a = b = 2\) for RIL1 and \(a = 4\) and \(b = 6\) for RIL2 (Haldane and Waddington 1931). Substituting \(R_{\omega\omega}\) in equation (41) by equation (42) and replacing \(r_{\omega\omega}\) by the Haldane map function (Haldane 1919) yields

\[
\text{cov}[Z_j(\tau), Z_j(\omega)] = 1 - 2R_{\omega\omega} = \frac{2 + b - 2a + (2a - b)e^{-2|\tau - \omega|}}{2 + b - be^{-2|\tau - \omega|}}
\]

(43)

which is a function of the distance between the two loci measured in Morgan. The variance of \(Z_j(\Delta_k)\) is defined by

\[
\text{var}[Z_j(\Delta_k)] = \frac{2}{\Delta_k^2} \int_0^{\Delta_k} \int_0^{\Delta_k} \frac{2 + b - 2a + (2a - b)e^{-2|\tau - \omega|}}{2 + b - be^{-2|\tau - \omega|}} \, d\omega \, d\tau
\]

(44)

A closed form expression is

\[
\text{var}[Z_j(\Delta_k)] = \frac{1}{b(b + 2)\Delta_k^2} \left[ \Theta_0(e^{-2\Delta_k}) + \Theta_1(\Delta_k^2) + \Theta_2(\Delta_k) + \Theta_3 \right]
\]

(45)

where

\[
\Theta_0 = 2a\xi_2 \left( \frac{2}{b + 2} \right) + 2\ln(2)a\ln \left( \frac{b}{b + 2} \right)
\]

\[
\Theta_1(\Delta_k) = -4\ln(2)a\Delta_k
\]

\[
\Theta_2(\Delta_k^2) = -(2a - b - 2)b\Delta_k^2
\]

\[
\Theta_3(e^{-2\Delta_k}) = -2a\xi_2 \left( 1 - \frac{b}{b + 2} e^{-2\Delta_k} \right) - 2a\ln \left( \frac{b}{b + 2} e^{-2\Delta_k} \right) \ln \left( b + 2 - be^{-2\Delta_k} \right)
\]

and \(\xi_2(x)\) is the dilogarithm function defined as

\[
\xi_2(x) = \sum_{i=1}^{\infty} \frac{x^i}{i^2}
\]

(47)

The limits of the variance are \(\lim_{\Delta_k \to 0} \text{var}[Z_j(\Delta_k)] = 1\) and \(\lim_{\Delta_k \to \pi} \text{var}[Z_j(\Delta_k)] = 0\).

For the RIL1 population where \(a = b = 2\), the variance is simplified into

\[
\text{var}[Z_j(\Delta_k)] = \frac{1}{2\Delta_k^2} \left\{ \xi_2 \left( \frac{1}{2} e^{-2\Delta_k} \right) + 2\ln(2)\Delta_k + \frac{1}{12} \left[ 6(\ln(2))^2 - \pi^2 \right] \right\}
\]

(48)

Simplification of the variance for RIL2 does not help too much except that \(\Theta_2(\Delta_k^2) = 0\).

Figure S3 shows the function graphically (for RIL1) in the range of \(0 \leq \Delta_k \leq 2\). For example, when \(\Delta_k = 0.01\) Morgan (1 cM), the variance is \(\text{var}(Z_k) = 0.9868\). When
\[ \Delta_k = 0.25 \text{ Morgan (25 cM)}, \text{ the variance is } \text{var}(Z_k) = 0.7548, \text{ where the new notation appearing in the figure } \text{var}(Z_k) = \text{var}\left[ \tilde{Z}_j(\Delta_k) \right], \text{ which is too complicated to draw in the figure.} \]

All derivations were conducted using Mathematica (Wolfram 1999). The final forms of the equations were then subject to manual simplification.

References

Files S4-S7


**File S4**: This is an excel spreadsheet data file with eight sheets, one for each trait. The file gives the estimated effect of each natural bin (a total of 1619 natural bins) along with the estimation error, Wald test, LOD score, p-value (drawn from permutation) and the degree of freedom (confidence). Any bins with p-value < 0.05 can be declared significant at the genome-wide Type I error of 0.05. The method used for the data analysis is the Lasso method.

**File S5**: This is an excel spreadsheet data file with eight sheets, one for each trait. The file gives the estimated effect of each natural bin (a total of 1619 natural bins) along with the F-test (equivalent to Wald test), LOD score and p-value (drawn from permutation). Any bins with p-value < 0.05 can be declared significant at the genome-wide Type I error of 0.05. The method used for the data analysis is the interval mapping (IM) procedure.

**File S6**: This is an excel spreadsheet data file with eight sheets, one for each trait. The file gives the LOD score of each natural bin (a total of 1619 natural bins) along with the p-value (drawn from permutation). Any bins with p-value < 0.05 can be declared significant at the genome-wide Type I error of 0.05. The method used for the data analysis is the composite interval mapping (CIM) procedure.

**File S7**: This is an excel spreadsheet data file with eight sheets, one for each trait. The file gives the estimated effect of each artificial bin (number of bins varies across traits) along with the estimation error, Wald test, LOD score and the p-value (drawn from permutation). Any bins with p-value < 0.05 can be declared significant at the genome-wide Type I error of 0.05. The method used for the data analysis is the Lasso method.
<table>
<thead>
<tr>
<th>Trait</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield (YD)</td>
<td>13.42065</td>
<td>15.77650</td>
<td>18.62915</td>
<td>27.17197</td>
</tr>
<tr>
<td>Tiller number (TP)</td>
<td>13.19609</td>
<td>14.64264</td>
<td>19.48538</td>
<td>24.24969</td>
</tr>
<tr>
<td>Grain number (GN)</td>
<td>13.87269</td>
<td>15.80410</td>
<td>19.45575</td>
<td>25.80274</td>
</tr>
<tr>
<td>K grain weight (KGW)</td>
<td>14.05583</td>
<td>15.39363</td>
<td>18.51520</td>
<td>23.72287</td>
</tr>
<tr>
<td>Grain length (GL)</td>
<td>13.96338</td>
<td>15.56000</td>
<td>18.97977</td>
<td>27.83338</td>
</tr>
<tr>
<td>Grain width (GW)</td>
<td>14.04309</td>
<td>15.42061</td>
<td>18.30397</td>
<td>23.41500</td>
</tr>
<tr>
<td>Heading date (HD)</td>
<td>13.99390</td>
<td>15.58231</td>
<td>19.25128</td>
<td>27.55085</td>
</tr>
<tr>
<td>Apicule color (OsC1)</td>
<td>13.99859</td>
<td>15.43835</td>
<td>20.25753</td>
<td>25.67522</td>
</tr>
<tr>
<td>Mean threshold</td>
<td>13.81803</td>
<td>15.45227</td>
<td>19.10975</td>
<td>25.67771</td>
</tr>
<tr>
<td>Theoretical threshold</td>
<td>2.7055</td>
<td>3.8414</td>
<td>6.6349</td>
<td>∞</td>
</tr>
</tbody>
</table>

The $x\%$ percentile represents $\alpha = 1 - x\%$ Type I error rate. For example, the Chi-square threshold under 95% percentile gives the threshold used to control $\alpha = 1 - 95\% = 0.05$ genome-wide Type I error. The Chi-square threshold divided by $2 \ln(10) \approx 4.61$ gives the LOD score threshold.
**Table S2**  Threshold values of the LRT test statistic (one degree of freedom) under various genome-wide Type I error rates obtained from 1000 permuted samples for the composite interval mapping (CIM) procedure.

<table>
<thead>
<tr>
<th>Trait</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield (YD)</td>
<td>17.63537</td>
<td>20.31503</td>
<td>24.36029</td>
<td>27.97551</td>
</tr>
<tr>
<td>Tiller number (TP)</td>
<td>17.54152</td>
<td>18.97397</td>
<td>22.94701</td>
<td>30.24515</td>
</tr>
<tr>
<td>Grain number (GN)</td>
<td>16.59045</td>
<td>19.05533</td>
<td>23.75548</td>
<td>32.35316</td>
</tr>
<tr>
<td>K grain weight (KGW)</td>
<td>17.42943</td>
<td>19.34218</td>
<td>23.52755</td>
<td>30.84495</td>
</tr>
<tr>
<td>Grain length (GL)</td>
<td>17.79907</td>
<td>19.47785</td>
<td>23.77822</td>
<td>28.63580</td>
</tr>
<tr>
<td>Grain width (GW)</td>
<td>17.40540</td>
<td>19.40530</td>
<td>24.15972</td>
<td>28.06567</td>
</tr>
<tr>
<td>Heading date (HD)</td>
<td>17.38699</td>
<td>19.07095</td>
<td>23.6893</td>
<td>30.38192</td>
</tr>
<tr>
<td>Apicule color (OsC1)</td>
<td>17.93634</td>
<td>20.54178</td>
<td>27.30653</td>
<td>2773.310</td>
</tr>
<tr>
<td><strong>Mean threshold</strong></td>
<td>17.46557</td>
<td>19.52280</td>
<td>23.74545</td>
<td>29.78602</td>
</tr>
<tr>
<td><strong>Theoretical threshold</strong></td>
<td>2.7055</td>
<td>3.8414</td>
<td>6.6349</td>
<td>∞</td>
</tr>
</tbody>
</table>

The $x\%$ percentile represents $\alpha = 1 - x\%$ Type I error rate. For example, the Chi-square threshold under 95% percentile gives the threshold used to control $\alpha = 1 - 95\% = 0.05$ genome-wide Type I error. The Chi-square threshold divided by $2 \ln(10) \approx 4.61$ gives the LOD score threshold.
Table S3  The effective numbers of tests for eight traits of rice in the natural bin analysis.

<table>
<thead>
<tr>
<th>Trait</th>
<th>( m_e )</th>
<th>p-value criterion</th>
<th>Chi-square criterion</th>
<th>LOD criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield (YD)</td>
<td>26.56</td>
<td>0.00188</td>
<td>9.63</td>
<td>2.09</td>
</tr>
<tr>
<td>Tiller number (TP)</td>
<td>65.45</td>
<td>0.00076</td>
<td>11.29</td>
<td>2.45</td>
</tr>
<tr>
<td>Number of grains (GN)</td>
<td>37.55</td>
<td>0.00133</td>
<td>10.28</td>
<td>2.23</td>
</tr>
<tr>
<td>1000 grain weight (KGW)</td>
<td>69.50</td>
<td>0.00072</td>
<td>11.43</td>
<td>2.48</td>
</tr>
<tr>
<td>Grain length (GL)</td>
<td>85.81</td>
<td>0.00058</td>
<td>11.80</td>
<td>2.56</td>
</tr>
<tr>
<td>Grain width (GW)</td>
<td>34.82</td>
<td>0.00143</td>
<td>10.14</td>
<td>2.20</td>
</tr>
<tr>
<td>Heading date (HD)</td>
<td>6.79</td>
<td>0.00736</td>
<td>7.14</td>
<td>1.55</td>
</tr>
<tr>
<td>Apicule color (OsC1)</td>
<td>1.21</td>
<td>0.04125</td>
<td>4.15</td>
<td>0.90</td>
</tr>
</tbody>
</table>

\( m_e \): Effective number of tests.

p-value: The genome-wide p-value criterion used to declare significance (\( p \)-value = 0.05 / \( m_e \)).

Chi-square criterion: The \((1 - p) \times 100\) percentile of the \( \chi^2 \) distribution.

LOD criterion: The LOD score criterion is the Chi-square criterion divided by 4.61.
Table S4  Numbers of natural bins associated with traits detected under genome-wide Type I error of 0.05 after Bonferroni correction using the effective number of tests.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Number of significant bins</th>
<th>Phenotypic variance</th>
<th>Genetic variance</th>
<th>Heritability&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield (YD)</td>
<td>0</td>
<td>19.8324</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Tiller number (TP)</td>
<td>13</td>
<td>1.4845</td>
<td>0.2858</td>
<td>0.1925</td>
</tr>
<tr>
<td>Number of grains (GN)</td>
<td>7</td>
<td>374.4867</td>
<td>76.1498</td>
<td>0.2033</td>
</tr>
<tr>
<td>1000 grain weight (KGW)</td>
<td>33</td>
<td>6.4193</td>
<td>3.7104</td>
<td>0.5780</td>
</tr>
<tr>
<td>Grain length (GL)</td>
<td>27</td>
<td>0.3095</td>
<td>0.2060</td>
<td>0.6654</td>
</tr>
<tr>
<td>Grain width (GW)</td>
<td>5</td>
<td>0.0479</td>
<td>0.0205</td>
<td>0.4285</td>
</tr>
<tr>
<td>Heading date (HD)</td>
<td>3</td>
<td>63.7318</td>
<td>9.3167</td>
<td>0.1461</td>
</tr>
<tr>
<td>Apicule color (OsC1)</td>
<td>1</td>
<td>0.2467</td>
<td>0.2316</td>
<td>0.9388</td>
</tr>
</tbody>
</table>

<sup>a</sup>Heritability: Defined as the ratio of the genetic variance to the phenotypic variance.
Table S5  Numbers of artificial bins associated with traits detected under genome-wide Type I error of 0.05 and proportions of the phenotypic variance explained by the associated bins.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Number of bins(^a)</th>
<th>Associated bins(^b)</th>
<th>Phenotypic variance</th>
<th>Genetic variance</th>
<th>Heritability(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield (YD)</td>
<td>3729</td>
<td>6</td>
<td>19.8324</td>
<td>1.4276</td>
<td>0.0719</td>
</tr>
<tr>
<td>Tiller number (TP)</td>
<td>1869</td>
<td>39</td>
<td>1.4845</td>
<td>0.6393</td>
<td>0.4306</td>
</tr>
<tr>
<td>Grain number (GN)</td>
<td>501</td>
<td>32</td>
<td>374.4867</td>
<td>184.4399</td>
<td>0.4925</td>
</tr>
<tr>
<td>K grain weight (KGW)</td>
<td>7451</td>
<td>49</td>
<td>6.4193</td>
<td>4.4327</td>
<td>0.6905</td>
</tr>
<tr>
<td>Grain length (GL)</td>
<td>750</td>
<td>49</td>
<td>0.3095</td>
<td>0.2459</td>
<td>0.7944</td>
</tr>
<tr>
<td>Grain width (GW)</td>
<td>191</td>
<td>40</td>
<td>0.0479</td>
<td>0.0367</td>
<td>0.7656</td>
</tr>
<tr>
<td>Heading date (HD)</td>
<td>1247</td>
<td>4</td>
<td>63.7318</td>
<td>10.5615</td>
<td>0.1657</td>
</tr>
<tr>
<td>Apicule color (OsC1)</td>
<td>1869</td>
<td>1</td>
<td>0.2467</td>
<td>0.2409</td>
<td>0.9765</td>
</tr>
</tbody>
</table>

\(^a\) Number of bins: This is the optimal number of bins with the smallest cross-validation generated mean squared error.

\(^b\) Associated bins: This is the number bins significantly associated with the trait.

\(^c\) Heritability: Defined as the ratio of the genetic variance to the phenotypic variance.