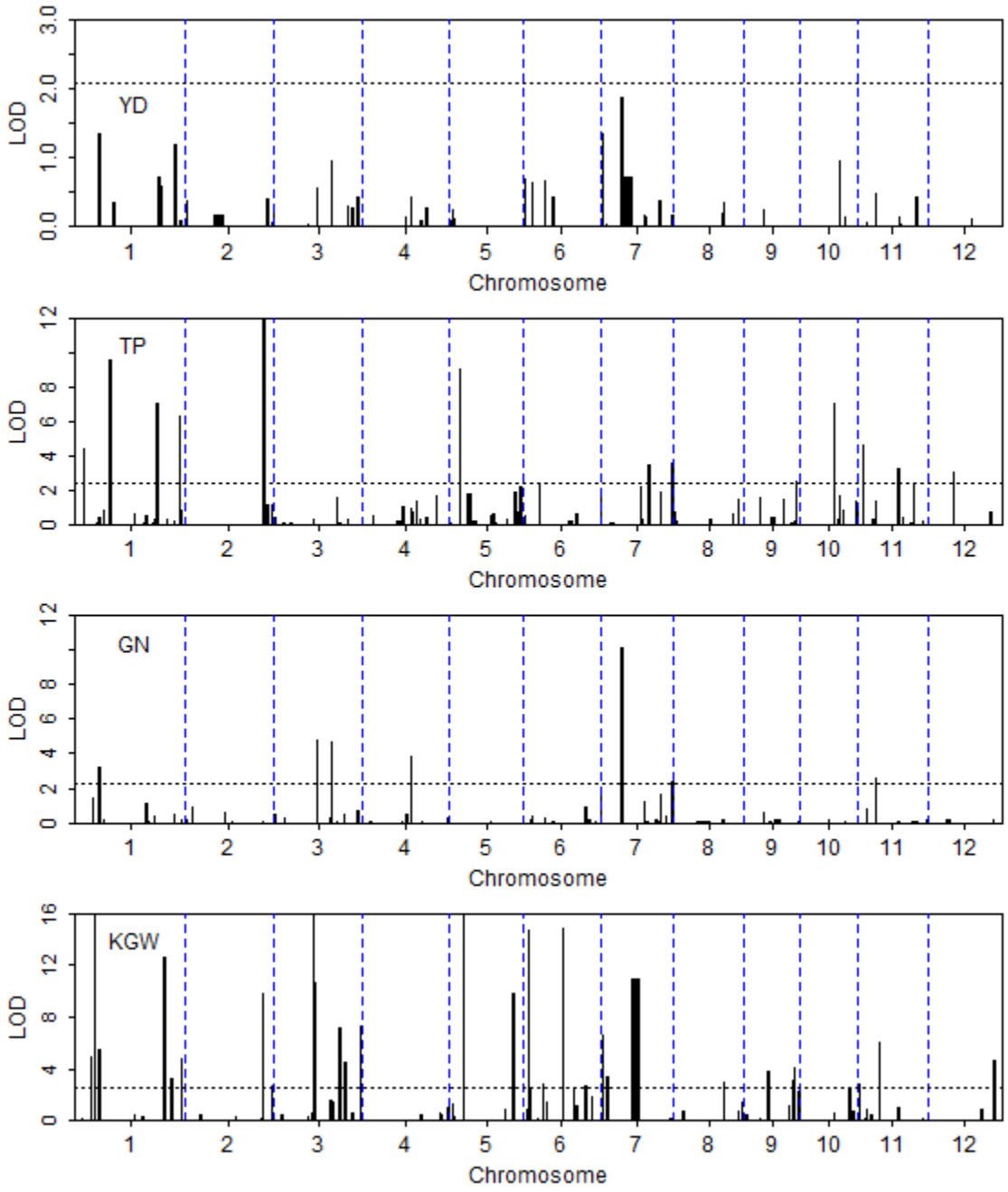

Figure S1 LOD score profiles for the first four traits of the natural bin analysis using the Lasso method, where the 12 chromosomes are separated by the dashed vertical lines and the effective number of tests corrected LOD score threshold for each trait is indicated by the dotted horizontal line. Three 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.

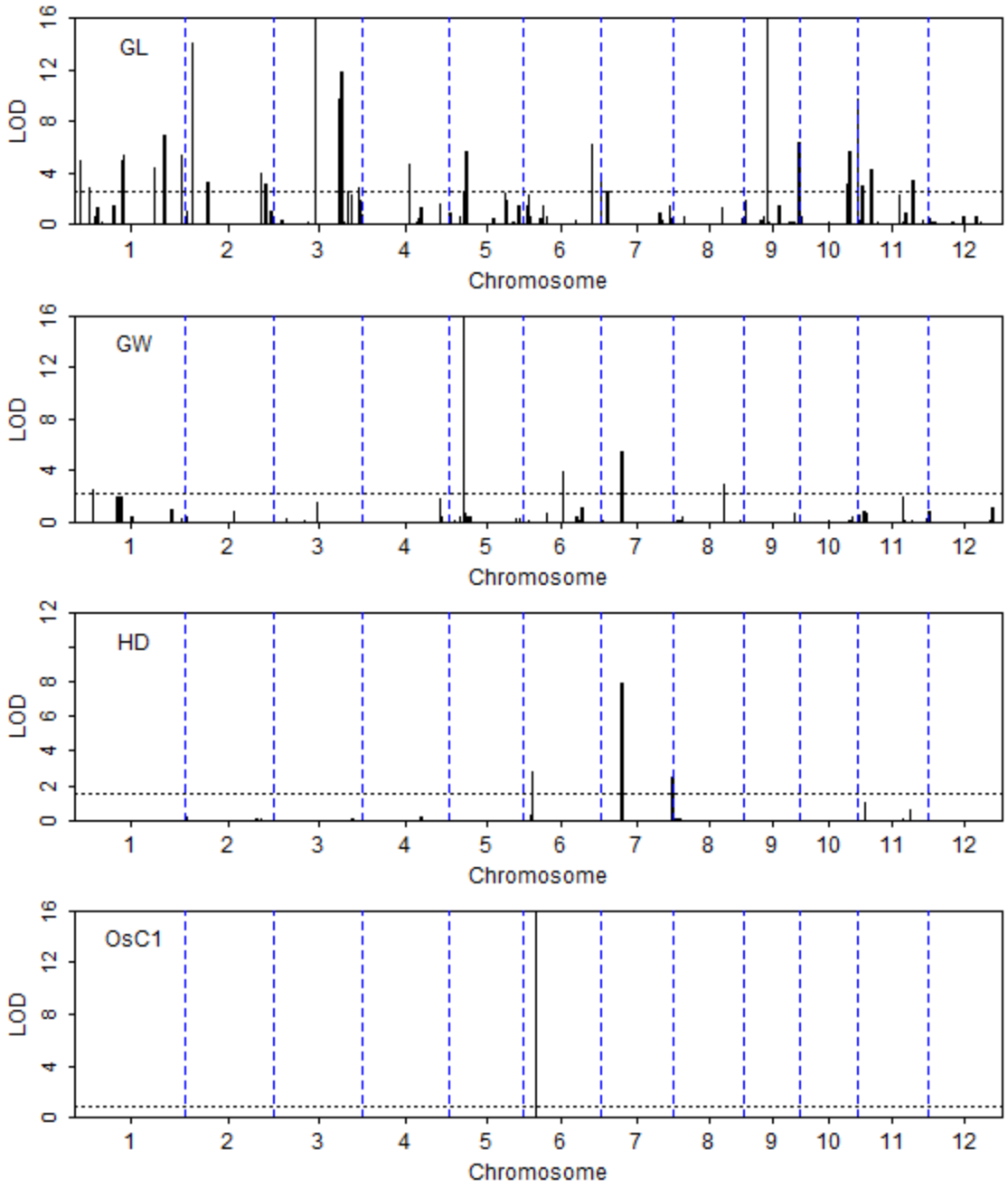

Figure S2 LOD score profiles for the last four traits of the natural bin analysis using the Lasso method, where the 12 chromosomes are separated by the dashed vertical lines and the effective number of tests corrected LOD score threshold for each trait is indicated by the dotted horizontal line. LOD 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.

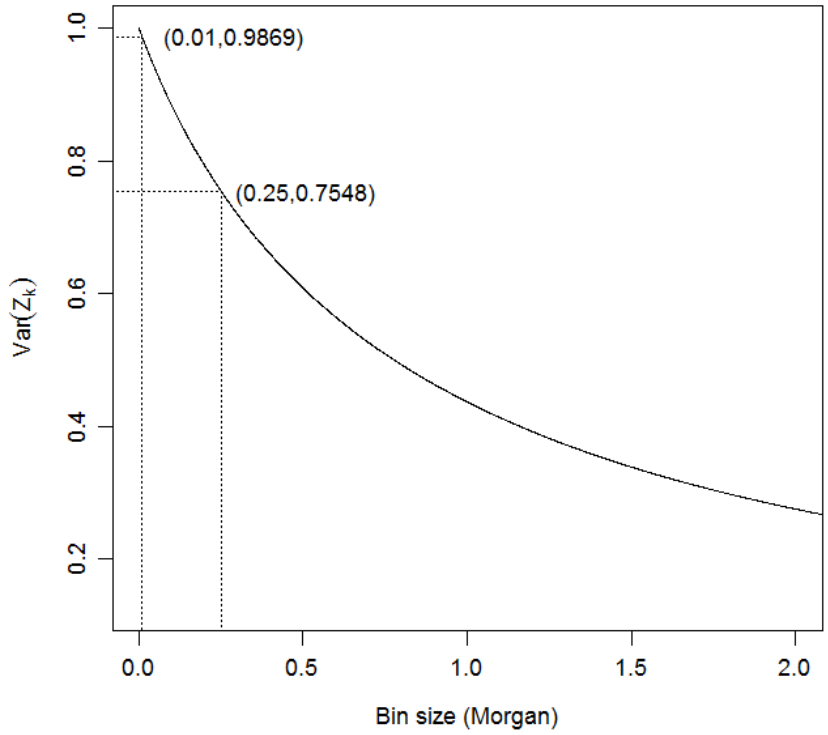

Figure S3 Functional relationship of the variance of bin genotype indicator variable, $\mathrm{var}(Z_k)$, with the bin size, Δ_k, measured in Morgan in a RIL population generated through repeated selfings. The two coordinates give the values of variance when the bin size equals 0.01 Morgan and 0.25 Morgan, respectively.

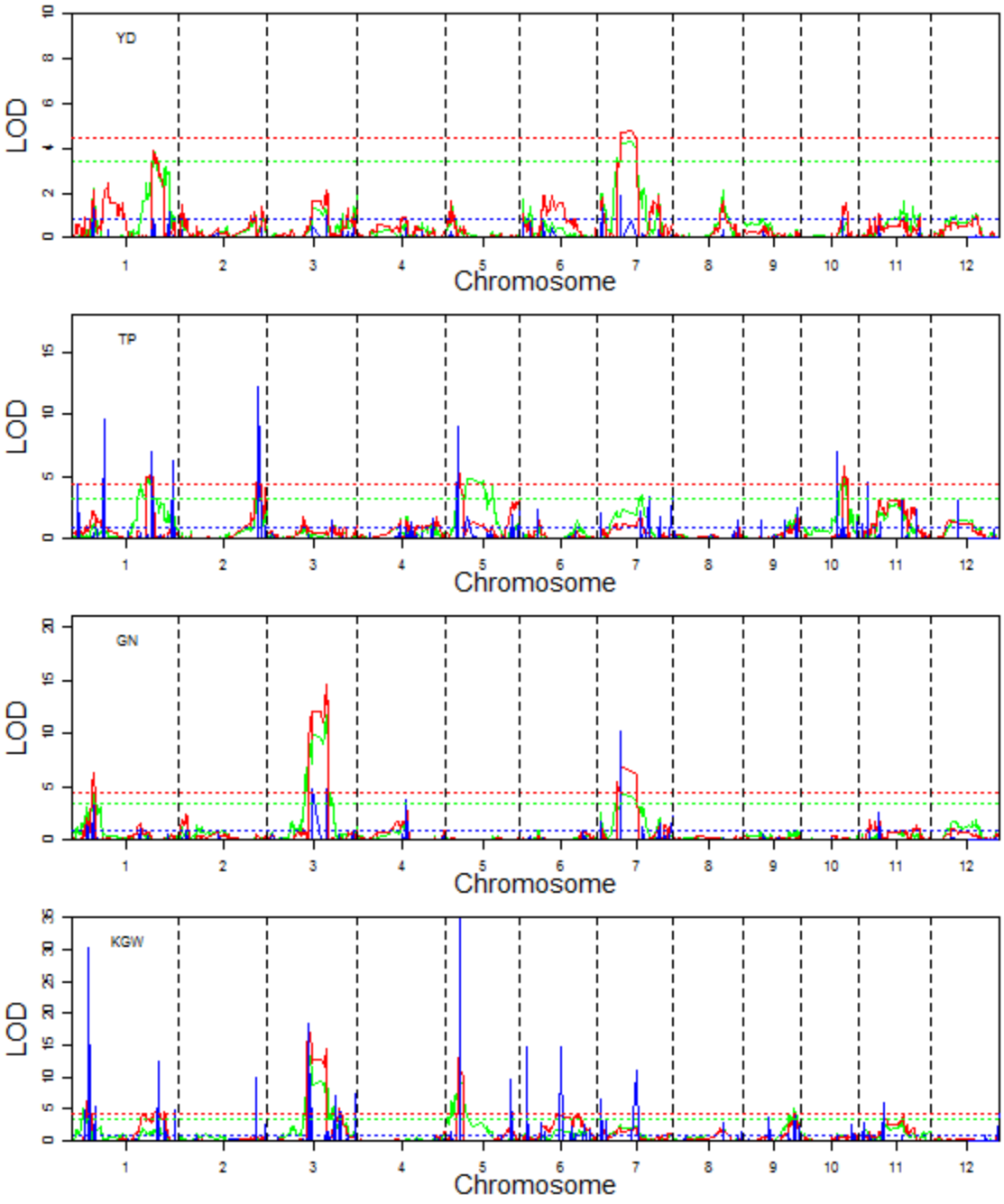

Figure S4 LOD score comparison of the Lasso method (color coded in blue) with the composite interval mapping (CIM) method (color coded in red) and interval mapping (IM) method (color coded in green) for the first four traits (YD, TP, GN and KGW). The three horizontal dotted lines are the genome-wide 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.

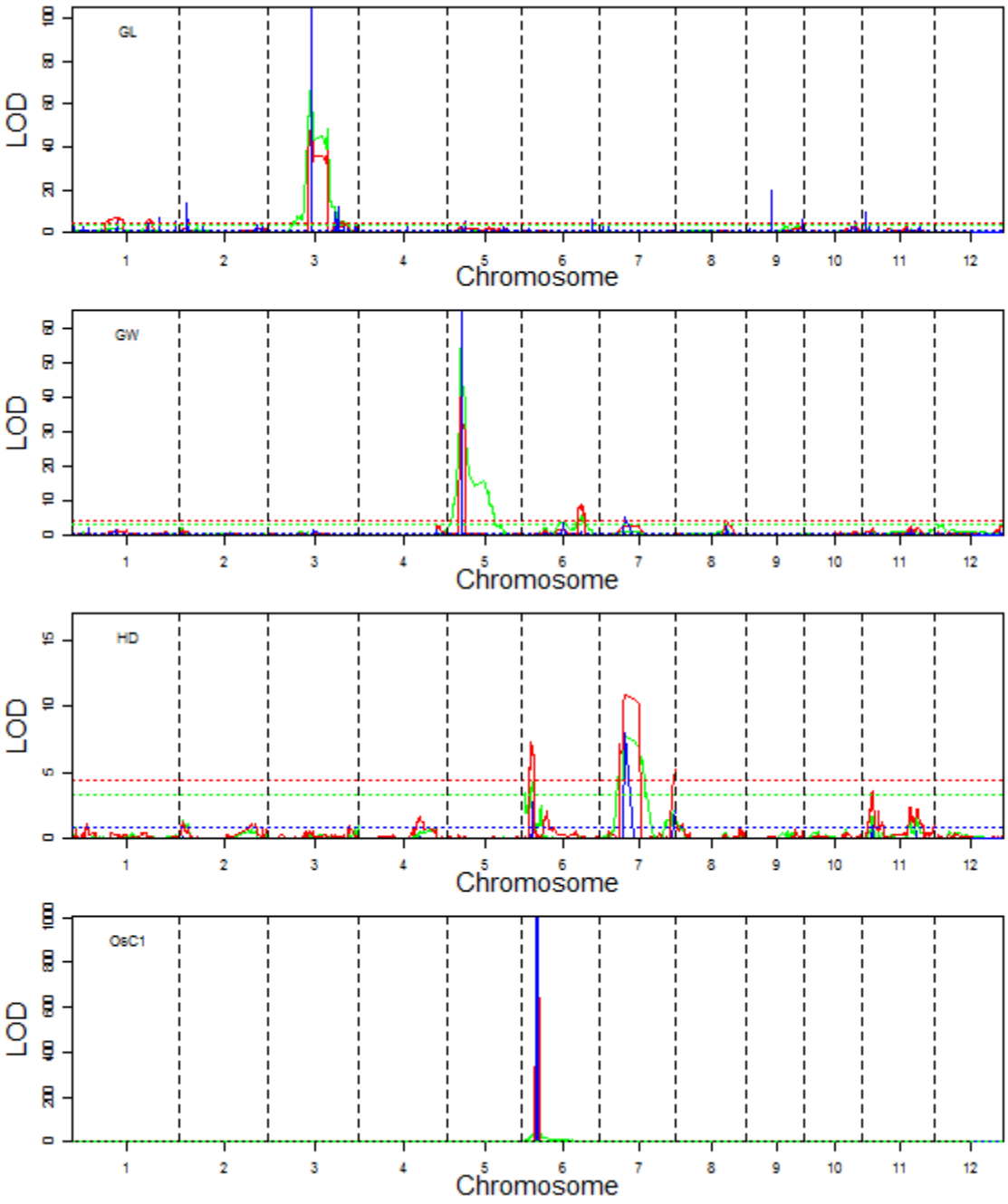

Figure S5 LOD score comparison of the Lasso method (color coded in blue) with the composite interval mapping (CIM) method (color coded in red) and the interval mapping (IM) method (color coded in green) for the last four traits (GL, GW, HD and OsC1). The three horizontal dotted lines are the genome-wide 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.

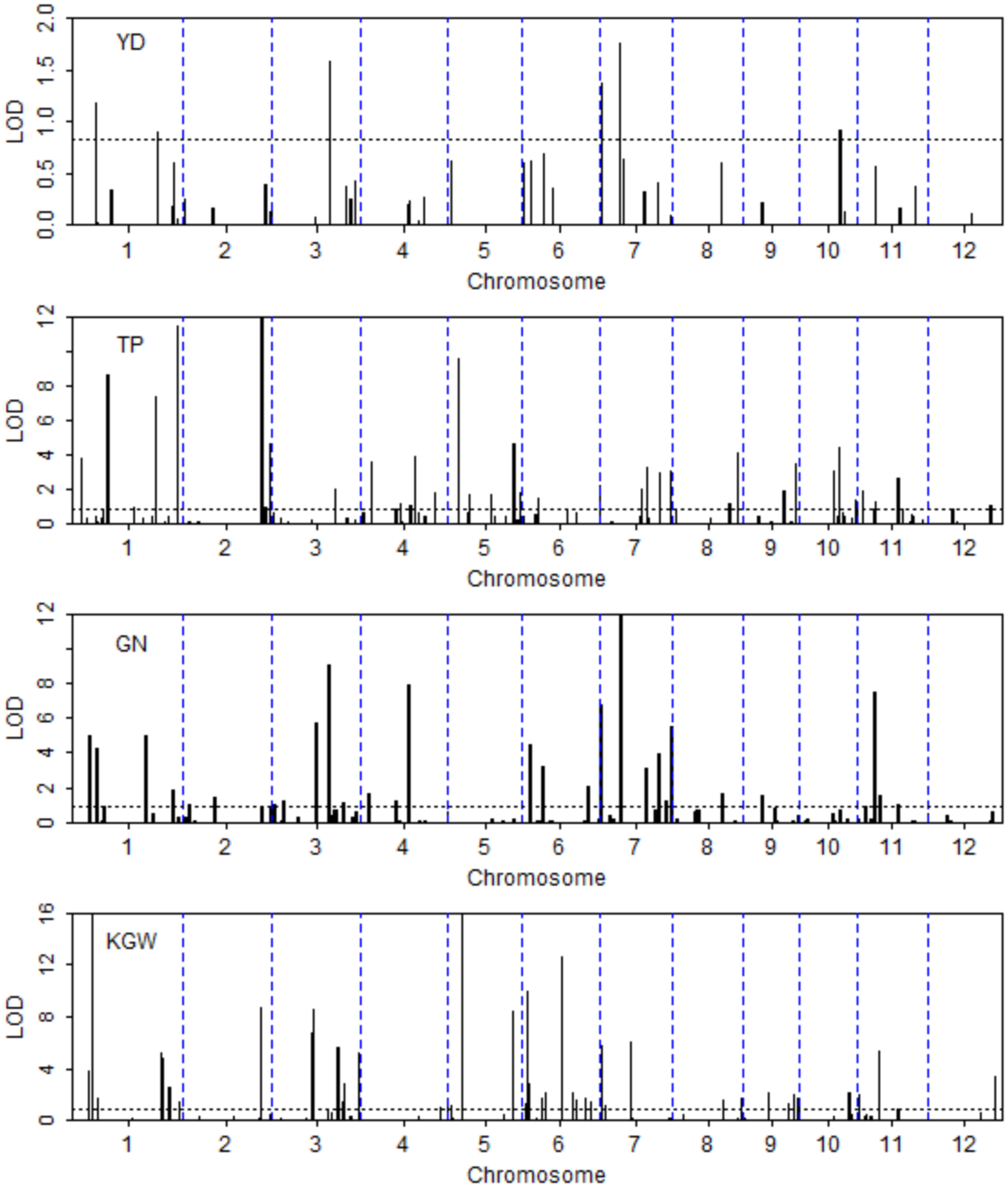

Figure S6 LOD score profiles for the first four traits of the artificial bin analysis using the Lasso method, where the 12 chromosomes are separated by the dashed vertical lines and the permutation generated LOD score threshold for each trait is indicated by the dotted horizontal line. Three bins for 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.

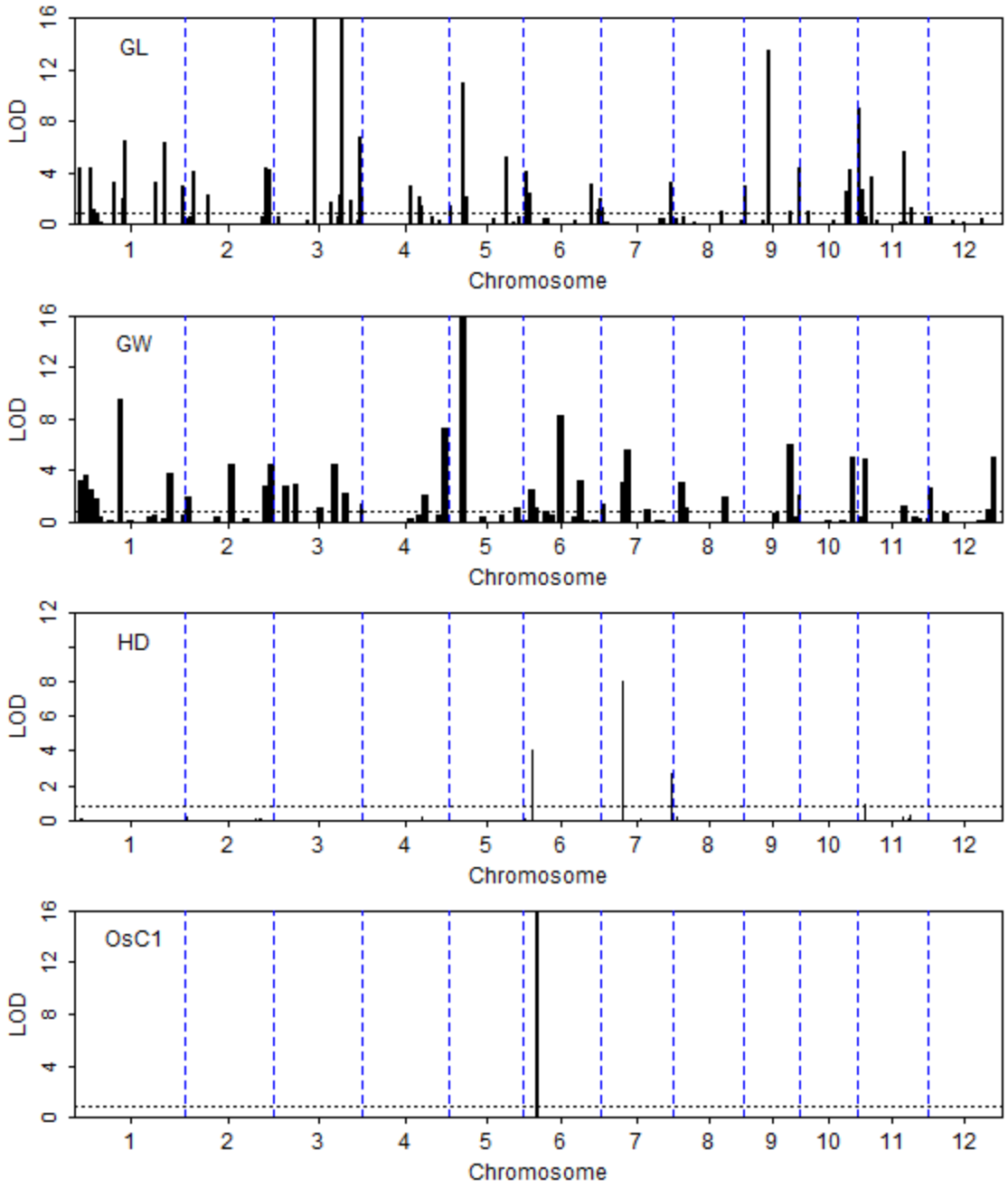

Figure S7 LOD score profiles for the last four traits of the artificial bin analysis using the Lasso method, where the 12 chromosomes are separated by the dashed vertical lines and the permutation generated LOD score threshold for each trait is indicated by the dotted horizontal line. LOD scores larger 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.

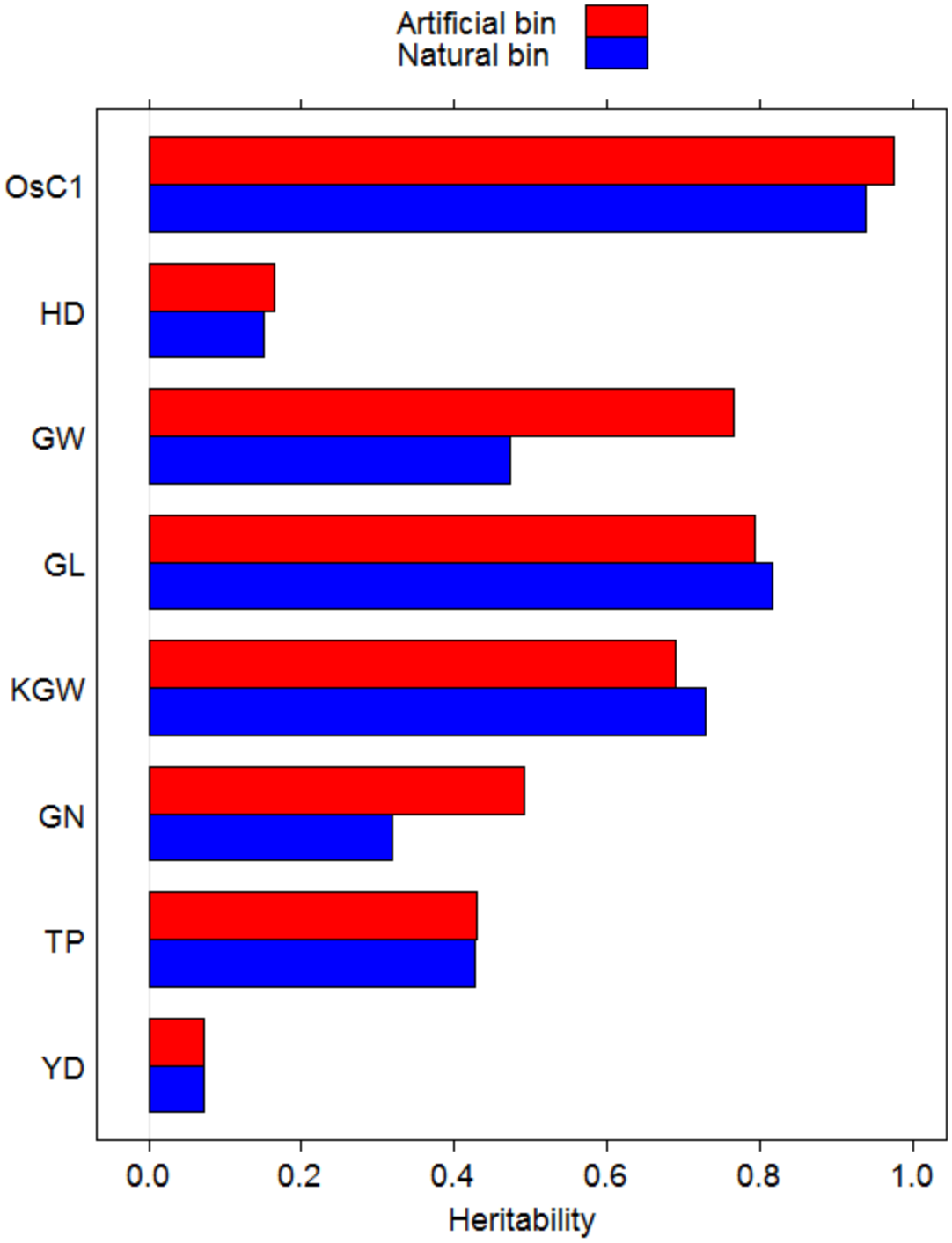

Figure S8 Comparison of estimated heritability (proportion of phenotypic variance explained by significant bins) between the artificial bin analysis and the natural bin analysis.

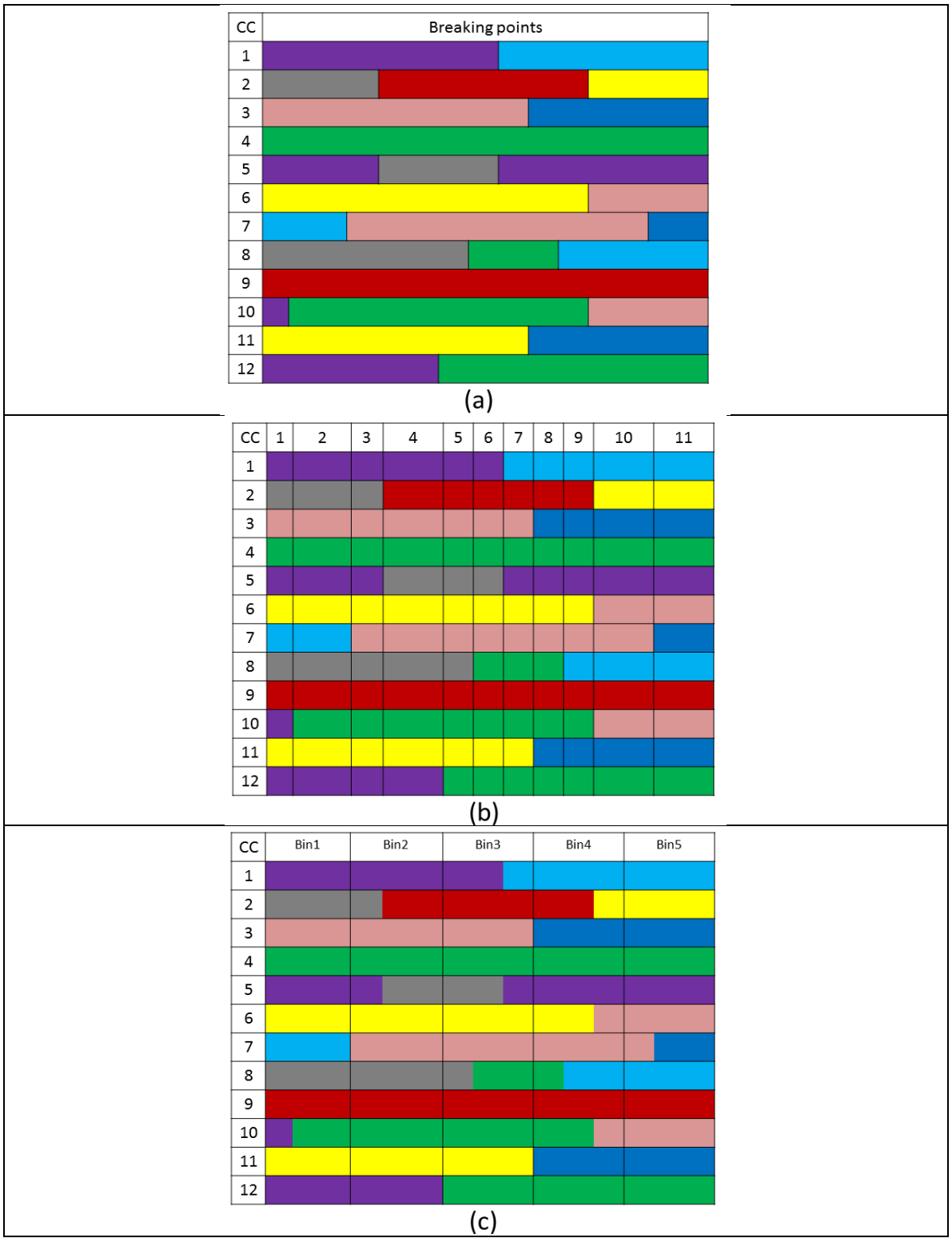

Figure S9 Breakpoints and bins in a hypothetical multi-parent MAGIC population initiated with eight parents (color coded). Panel (a): breakpoint patterns, Panel (b): natural bins, Panel (c) artificial bins.

File S1

Proof of Equation (4) of the Main Text

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,

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

where

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

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

$$\Delta_k = \int_{\lambda_k - \frac{1}{2}\Delta_k}^{\lambda_k + \frac{1}{2}\Delta_k} d\lambda \tag{3}$$

The location of the bin is denoted by λ_k, which is the middle point position of the bin in the genome. The genetic effect γ_k for bin *k* is defined as

$$\gamma_k = \bar{\gamma}(\lambda_k)\Delta_k = \int_{\lambda_k - \frac{1}{2}\Delta_k}^{\lambda_k + \frac{1}{2}\Delta_k} \gamma(\lambda) d\lambda \tag{4}$$

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

$$y_j = X_j\beta + \sum_{k=1}^{m} \alpha_{jk} + \varepsilon_j \tag{5}$$

where

$$\alpha_{jk} = \int_{\lambda_k - \frac{1}{2}\Delta_k}^{\lambda_k + \frac{1}{2}\Delta_k} Z_j(\lambda)\gamma(\lambda) d\lambda \tag{6}$$

Let us rewrite equation (6) using

$$\alpha_{jk} = \int_{\lambda_k - \frac{1}{2}\Delta_k}^{\lambda_k + \frac{1}{2}\Delta_k} \left[Z_j(\lambda) - \bar{Z}_j(\lambda_k) + Z_{jk} \right] \left[\gamma(\lambda) - \bar{\gamma}(\lambda_k) + \Delta_k^{-1}\gamma_k \right] d\lambda \tag{7}$$

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

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

$$= Z_{jk}\gamma_k\Delta_k^{-1}\int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} d\lambda + \int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} \left[Z_j(\lambda)-\bar{Z}_j(\lambda_k)\right]\left[\gamma(\lambda)-\bar{\gamma}(\lambda_k)\right]d\lambda \quad (8)$$

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

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

$$\int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} \left[Z_j(\lambda)-\bar{Z}_j(\lambda_k)\right]d\lambda = \int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} \left[\gamma(\lambda)-\bar{\gamma}(\lambda_k)\right]d\lambda = 0 \quad (9)$$

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 γ,

$$\text{cov}(Z_j,\gamma) = \int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} \left[Z_j(\lambda)-\bar{Z}_j(\lambda_k)\right]\left[\gamma(\lambda)-\bar{\gamma}(\lambda_k)\right]d\lambda \quad (10)$$

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 α_{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 γ_k, respectively. We may find the "prior variance" retrospectively using $\hat{\gamma}_k$ and the data. If γ_k were estimated from a single bin model, the variance of that estimate would be

$$\text{var}(\tilde{\gamma}_k) = (Z_k^T Z_k)^{-1}\hat{\sigma}^2 \tag{11}$$

where

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

is the estimated residual error variance. Let σ_k^2 be the "prior variance" of γ_k. Combining the prior variance σ_k^2 and the variance from the data $\text{var}(\tilde{\gamma}_k)$, we can generate the posterior variance expressed by

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

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

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

Therefore,

$$\sigma_k^2 = \hat{\gamma}_k^2 + \frac{\sigma_k^2\hat{\sigma}^2}{\hat{\sigma}^2 + \sigma_k^2 Z_k^T Z_k} \tag{15}$$

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

$$\hat{\sigma}_k^2 = \frac{\hat{\gamma}_k^2 Z_k^T Z_k + \sqrt{(\hat{\gamma}_k^2 Z_k^T Z_k)^2 + 4\hat{\sigma}^2\hat{\gamma}_k^2 Z_k^T Z_k}}{2Z_k^T Z_k} \tag{16}$$

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

$$\text{var}(\hat{\gamma}_k) = \left(\frac{1}{\hat{\sigma}_k^2} + \frac{1}{\text{var}(\tilde{\gamma}_k)}\right)^{-1} = \frac{\hat{\sigma}_k^2\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\sigma}_k^2 Z_k^T Z_k} \tag{17}$$

When $m \gg n$ and n is very large, the Lasso estimate of σ^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

$$W_k = \hat{\gamma}_k^2 \frac{\hat{\sigma}^2 + \hat{\sigma}_k^2 Z_k^T Z_k}{\hat{\sigma}_k^2 \hat{\sigma}^2} \approx \hat{\gamma}_k^2 \frac{\hat{\sigma}^2 + \hat{\gamma}_k^2 Z_k^T Z_k}{\hat{\gamma}_k^2 \hat{\sigma}^2} \approx \frac{\hat{\gamma}_k^2}{\hat{\sigma}^2} Z_k^T Z_k + 1 \qquad (18)$$

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

$$d_k = 1 - \frac{\text{var}(\hat{\gamma}_k)}{\hat{\sigma}_k^2} = \frac{\hat{\sigma}_k^2 - \text{var}(\hat{\gamma}_k)}{\hat{\sigma}_k^2} = \frac{\hat{\sigma}_k^2 Z_k^T Z_k}{\hat{\sigma}^2 + \hat{\sigma}_k^2 Z_k^T Z_k} \qquad (19)$$

Note that $0 \le d_k \le 1$, 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

$$m_e = \sum_{k=1}^{m} d_k = \sum_{k=1}^{m} \frac{\hat{\sigma}_k^2 Z_k^T Z_k}{\hat{\sigma}^2 + \hat{\sigma}_k^2 Z_k^T Z_k} \qquad (20)$$

Recall that when $m \square\ n$ and *n* is very large, the Lasso estimate of σ^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 $\chi_1^{-2}(1-0.00188) = \chi_1^{-2}(0.99812) = 9.63$, which is converted into to 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

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1): 1-22.

Hu Z, Wang Z, Xu S. 2012. An infinitesimal model for quantitative trait genomic value prediction. *PLoS One* **7**: e41336.

MacKay DJC. 1992. Bayesian interpolation. *Neural Computation* **4**: 415-447.

Sillanpää MJ, Arjas E. 1998. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**(3): 1373-1388.

Sillanpää MJ, Arjas E. 1999. Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **151**: 1605-1619.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* **58**: 267-288.

Tipping ME. 2001. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **1**: 211-244.

File S3

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_1A_2. Let us denote the genotype indicator variable of individual *j* at locus τ by $Z_j(\tau)$, which is numerically coded by

$$Z_j(\tau) = \begin{cases} 1 & \text{for } A_1A_1 \\ 0 & \text{for } A_1A_2 \end{cases} \tag{21}$$

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

$$\text{var}\left[Z_j(\tau)\right] = \text{E}\left[Z_j^2(\tau)\right] - \text{E}^2\left[Z_j(\tau)\right] = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4} \tag{22}$$

The covariance of Z_j between loci τ and ω across all individuals within the BC family is

$$\text{cov}\left[Z_j(\tau), Z_j(\omega)\right] = (1-2r_{\tau\omega})\sqrt{\text{var}\left[Z_j(\tau)\right]\text{var}\left[Z_j(\omega)\right]} \tag{23}$$

where

$$r_{\tau\omega} = \frac{1}{2}(1-e^{-2|\tau-\omega|}) \tag{24}$$

is the recombination fraction between loci τ and ω (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

$$\text{cov}\left[Z_j(\tau), Z_j(\omega)\right] = \frac{1}{4}(1-2r_{\tau\omega}) = \frac{1}{4}e^{-2|\tau-\omega|} \tag{25}$$

where $\tau \in \Delta_k$ and $\omega \in \Delta_k$, and Δ_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 Δ_k) for individual *j* (only this individual, not across all individuals with the family) by

$$\bar{Z}_j(\Delta_k) = \frac{1}{\Delta_k}\int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} Z_j(\tau)d\tau \tag{26}$$

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,

$$\bar{Z}_j(\Delta_k) = \frac{1}{\Delta_k}\int_0^{\Delta_k} Z_j(\tau)d\tau \tag{27}$$

It should be noted that the genome location τ 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

$$\begin{aligned}\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] &= \frac{1}{\Delta_k^2}\int_0^{\Delta_k}\int_0^{\Delta_k}\mathrm{cov}\left[Z_j(\tau), Z_j(\omega)\right]d\omega d\tau \\ &= \frac{1}{4\Delta_k^2}\int_0^{\Delta_k}\int_0^{\Delta_k} e^{-2|\tau-\omega|}d\omega d\tau \\ &= \frac{1}{2\Delta_k^2}\int_0^{\Delta_k}\int_0^{\tau} e^{-2(\tau-\omega)}d\omega d\tau\end{aligned} \tag{28}$$

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

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{1}{8\Delta_k^2}(e^{-2\Delta_k} + 2\Delta_k - 1) \tag{29}$$

which is a function of the bin size Δ_k. One can verify that

$$\begin{cases}\lim\limits_{\Delta_k\to 0}\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \lim\limits_{\Delta_k\to 0}\dfrac{1}{8\Delta_k^2}(e^{-2\Delta_k} + 2\Delta_k - 1) = \dfrac{1}{4} \\ \lim\limits_{\Delta_k\to \infty}\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \lim\limits_{\Delta_k\to \infty}\dfrac{1}{8\Delta_k^2}(e^{-2\Delta_k} + 2\Delta_k - 1) = 0\end{cases} \tag{30}$$

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 τ is coded by

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

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

$$\mathrm{var}\left[Z_j(\tau)\right] = \mathrm{E}\left[Z_j^2(\tau)\right] - \mathrm{E}^2\left[Z_j(\tau)\right] = 1 \tag{32}$$

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

$$\mathrm{cov}\left[Z_j(\tau), Z_j(\omega)\right] = 1 - 2r_{\tau\omega} = e^{-2|\tau-\omega|} \tag{33}$$

Using the same approach as that described in the BC population, we obtained

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{2}{\Delta_k^2}\int_0^{\Delta_k}\int_0^{\tau} e^{-2(\tau-\omega)}d\omega d\tau = \frac{1}{2\Delta_k^2}(e^{-2\Delta_k} + 2\Delta_k - 1) \tag{34}$$

The limits of the variance are $\lim_{\Delta_k \to 0} \mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 1$ and $\lim_{\Delta_k \to \infty} \mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 0$.

3. F_2 population

A F_2 population is generated by selfing the F_1 individuals derived from the cross of two inbred lines, A_1A_1 and A_2A_2. As a result, The F_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 τ by $Z_j(\tau)$, which is coded by

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

The variance of $Z_j(\tau)$ across all individuals within this F_2 population is

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

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

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

Using the same approach as that described before, we obtained

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{1}{\Delta_k^2}\int_0^{\Delta_k}\int_0^{\tau} e^{-2(\tau-\omega)}d\omega d\tau = \frac{1}{4\Delta_k^2}(e^{-2\Delta_k} + 2\Delta_k - 1) \tag{38}$$

The limits of the variance are $\lim_{\Delta_k \to 0} \mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 1/2$ and $\lim_{\Delta_k \to \infty} \mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 0$.

4. Recombinant inbred lines (RIL)

There are two ways to generate recombinant inbred lines: (1) repeated selfings starting from the F_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 τ is coded by

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

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

$$\mathrm{var}\left[Z_j(\tau)\right] = \mathrm{E}\left[Z_j^2(\tau)\right] - \mathrm{E}^2\left[Z_j(\tau)\right] = 1 \tag{40}$$

which is the same as the DH population. The covariance of the genotype indicator variables between two different loci within the same bin is

$$\mathrm{cov}\left[Z_j(\tau), Z_j(\omega)\right] = 1 - 2R_{\tau\omega} \tag{41}$$

where

$$R_{\tau\omega} = \frac{ar_{\tau\omega}}{1 + br_{\tau\omega}} \tag{42}$$

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 $R_{\tau\omega}$ in equation (41) by equation (42) and replacing $r_{\tau\omega}$ by the Haldane map function (Haldane 1919) yields

$$\mathrm{cov}\left[Z_j(\tau), Z_j(\omega)\right] = 1 - 2R_{\tau\omega} = \frac{2 + b - 2a + (2a - b)e^{-2|\tau - \omega|}}{2 + b - be^{-2|\tau - \omega|}} \tag{43}$$

which is a function of the distance between the two loci measured in Morgan. The variance of $\bar{Z}_j(\Delta_k)$ is defined by

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{2}{\Delta_k^2}\int_0^{\Delta_k}\int_0^{\tau} \frac{2 + b - 2a + (2a - b)e^{-2|\tau - \omega|}}{2 + b - be^{-2|\tau - \omega|}} d\omega d\tau \tag{44}$$

A closed form expression is

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{1}{b(b+2)\Delta_k^2}\left[\Theta_3(e^{-2\Delta_k}) + \Theta_2(\Delta_k^2) + \Theta_1(\Delta_k) + \Theta_0\right] \tag{45}$$

where

$$\begin{aligned}
\Theta_0 &= 2a\xi_2\left(\frac{2}{b+2}\right) + 2\ln(2)a\ln\left(\frac{b}{b+2}\right) \\
\Theta_1(\Delta_k) &= -4\ln(2)a\Delta_k \\
\Theta_2(\Delta_k^2) &= -(2a - b - 2)b\Delta_k^2 \\
\Theta_3(e^{-2\Delta_k}) &= -2a\xi_2\left(1 - \frac{b}{b+2}e^{-2\Delta_k}\right) - 2a\ln\left(\frac{b}{b+2}e^{-2\Delta_k}\right)\ln\left(b + 2 - be^{-2\Delta_k}\right)
\end{aligned} \tag{46}$$

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

$$\xi_2(x) = \sum_{i=1}^{\infty}\frac{x^i}{i^2} \tag{47}$$

The limits of the variance are $\lim_{\Delta_k \to 0}\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 1$ and $\lim_{\Delta_k \to \infty}\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 0$.

For the RIL1 population where $a = b = 2$, the variance is simplified into

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{1}{2\Delta_k^2}\left\{\xi_2\left(\frac{1}{2}e^{-2\Delta_k}\right) + 2\ln(2)\Delta_k + \frac{1}{12}\left[6\left(\ln(2)\right)^2 - \pi^2\right]\right\} \tag{48}$$

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 $0 \le \Delta_k \le 2$. For example, when $\Delta_k = 0.01$ Morgan (1 cM), the variance is $\mathrm{var}(Z_k) = 0.9868$. When

$\Delta_k = 0.25$ Morgan (25 cM), the variance is $\text{var}(Z_k) = 0.7548$, where the new notation appearing in the figure $\text{var}(Z_k)$ is $\text{var}\left[\bar{Z}_j(\Delta_k)\right]$, 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

Haldane JB, Waddington CH. 1931. Inbreeding and linkage. *Genetics* **16**(4): 357-374.

Haldane JBS. 1919. The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**: 299-309.

Wolfram S. 1999. *The Mathematica Book*. Wlfram Media/Cambridge University Press, Oxfordshire, UK.

Files S4-S7

Available for download at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.155309/-/DC1

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 S1 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 interval mapping (IM) procedure.

Trait	90%	95%	99%	100%
Yield (YD)	13.42065	15.77650	18.62915	27.17197
Tiller number (TP)	13.19609	14.64264	19.48538	24.24969
Grain number (GN)	13.87269	15.80410	19.45575	25.80274
K grain weight (KGW)	14.05583	15.39363	18.51520	23.72287
Grain length (GL)	13.96338	15.56000	18.97977	27.83338
Grain width (GW)	14.04309	15.42061	18.30397	23.41500
Heading date (HD)	13.99390	15.58231	19.25128	27.55085
Apicule color (OsC1)	13.99859	15.43835	20.25753	25.67522
Mean threshold	13.81803	15.45227	19.10975	25.67771
Theoretical threshold	2.7055	3.8414	6.6349	∞

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.

Trait	90%	95%	99%	100%
Yield (YD)	17.63537	20.31503	24.36029	27.97551
Tiller number (TP)	17.54152	18.97397	22.94701	30.24515
Grain number (GN)	16.59045	19.05533	23.75548	32.35316
K grain weight (KGW)	17.42943	19.34218	23.52755	30.84495
Grain length (GL)	17.79907	19.47785	23.77882	28.63580
Grain width (GW)	17.40540	19.40530	24.15972	28.06567
Heading date (HD)	17.38699	19.07095	23.6893	30.38192
Apicule color (OsC1)	17.93634	20.54178	2730.653	2773.310
Mean threshold	17.46557	19.52280	23.74545	29.78602
Theoretical threshold	2.7055	3.8414	6.6349	∞

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.

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

m_e : Effective number of tests.

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

Chi-square criterion: The $(1-p)\times 100$ percentile of the χ_1^2 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.

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

[a]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.

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

[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.