QTL Mapping with F_2 Populations

Outlines

- Genetic and statistical models in F_2
- Additive and dominance mapping in F_2
- Simulation study
- QTL mapping in an actual F_2 population
- Conclusion

One-locus model in F_2

- **One-locus model:** $G = \mu + aw + dv$

  where μ is mean of the two homozygous genotypes *QQ* and *qq*, *a* is the additive effect, *d* is the dominance effect . *w* and *v* are the indicators for genotypes at the QTL, valued at 1 and 0 for *QQ*, 0 and 1 for *Qq*, and -1 and 0 for *qq*, respectively

The expected genotypic value of an individual with known marker types

$$E(G \mid x_1, x_2, y_1, y_2) = \mu + a \times E(w \mid x_1, x_2, y_1, y_2) + d \times E(v \mid x_1, x_2, y_1, y_2)$$

where x_1 and y_1 are the indicators for the left marker, x_2 and y_2 are the indicators for the right marker

Probability of the three QTL genotypes under given marker types

Left marker	Right marker	QQ (w=1, v=0) ($m+a$)	Qq (w=0, v=1) ($m+d$)	qq (w=-1, v=0) ($m-a$)
AA	BB	$\frac{1}{4}(1-r_1)^2(1-r_2)^2$	$\frac{1}{2}r_1(1-r_1)r_2(1-r_2)$	$\frac{1}{4}r_1^2r_2^2$
AA	Bb	$\frac{1}{2}(1-r_1)^2r_2(1-r_2)$	$\frac{1}{2}r_1(1-r_1)(1-r_2)^2+\frac{1}{2}r_1(1-r_1)r_2^2$	$\frac{1}{2}r_1^2r_2(1-r_2)$
AA	bb	$\frac{1}{4}(1-r_1)^2r_2^2$	$\frac{1}{2}r_1(1-r_1)r_2(1-r_2)$	$\frac{1}{4}r_1^2(1-r_2)^2$

Estimation of marker class mean

Marker class	n	Frequency	Indicator for marker x_1	x_2	y_1	y_2	$E(w \mid x_1, x_2, y_1, y_2)$	$E(v \mid x_1, x_2, y_1, y_2)$	Genetic mean of the class
AABB	n_1	$\frac{1}{4}(1-r)^2$	1	1	0	0	f_1	g_1	$\mu + f_1 a + g_1 d$
AABb	n_2	$\frac{1}{2}r(1-r)$	1	0	0	1	f_2	g_2	$\mu + f_2 a + g_2 d$
AAbb	n_3	$\frac{1}{4}r^2$	1	-1	0	0	f_3	g_3	$\mu + f_3 a + g_3 d$

Where

$$1 - 2r_1 r_2 /(1-r) \hat{=} f_1 \qquad 2r_1(1-r_1)r_2(1-r_2)/(1-r)^2 \hat{=} g_1$$

$$[(1-2r_1)r_2(1-r_2)]/(r-r^2) \hat{=} f_2 \qquad r_1(1-r_1)(1-2r_2+2r_2^2)/(r-r^2) \hat{=} g_2$$

$$(r_2 - r_1)/r \hat{=} f_3 \qquad 2r_1(1-r_1)r_2(1-r_2)/r^2 \hat{=} g_3$$

Relationship between marker class mean and marker effect
(including marker interactions)

$$\begin{bmatrix} \mu + f_1 a + g_1 d \\ \mu + f_2 a + g_2 d \\ \mu + f_3 a + g_3 d \\ \mu + f_4 a + g_4 d \\ \mu + g_5 d \\ \mu - f_4 a + g_4 d \\ \mu - f_3 a + g_3 d \\ \mu - f_2 a + g_2 d \\ \mu - f_1 a + g_1 d \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & 1 & 0 & 0 & 0 & -1 & 0 \\ 1 & -1 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} \mu + (d)\mu_d \\ (a)A_1 \\ (a)A_2 \\ (d)D_1 \\ (d)D_2 \\ (d)AA_{12} \\ AD_{12} \\ DA_{12} \\ (d)DD_{12} \end{bmatrix}$$

Relationship between marker effects and QTL effects

$$\begin{bmatrix} \mu+(d)\mu_d \\ (a)A_1 \\ (a)A_2 \\ (d)D_1 \\ (d)D_2 \\ (d)AA_{12} \\ AD_{12} \\ DA_{12} \\ (d)DD_{12} \end{bmatrix} = \begin{bmatrix} \mu+\frac{1}{2}(g_1+g_3)d \\ f_2 a \\ \frac{1}{2}(f_1-f_3)a \\ (-\frac{1}{2}g_1-\frac{1}{2}g_3+g_4)d \\ (-\frac{1}{2}g_1+g_2-\frac{1}{2}g_3)d \\ \frac{1}{2}(g_1-g_3)d \\ 0 \\ 0 \\ (\frac{1}{2}g_1-g_2+\frac{1}{2}g_3-g_4+g_5)d \end{bmatrix}$$

The additive QTL effect "*a*" only causes additive marker effects. But the dominance QTL effect "*d*" causes additive by additive, and dominance by dominance marker interactions as well as dominance marker effects

The linear model of genotypic values on markers in F_2

- **The expectations of *w* and *v* (indicators of QTL genotype) under each marker class can be proved as:**

$$E(w \mid x_1, x_2, y_1, y_2) = \lambda_1' x_1 + \lambda_2' x_2$$

$$E(v \mid x_1, x_2, y_1, y_2) = \delta + \rho_1' y_1 + \rho_2' y_2 + \lambda\lambda_{12}' x_1 x_2 + \rho\rho_{12}' y_1 y_2$$

The linear model of genotypic values on markers in F_2

- **So the genotypic value of an F_2 individual with known marker class is**

$$E(G \mid x_1, x_2, y_1, y_2) = \beta + (a)A_1 x_1 + (d)D_1 y_1 + (a)A_2 x_2 + (d)D_2 y_2$$

$$+ (d)AA_{12} x_1 x_2 + (d)DD_{12} y_1 y_2$$

- **Two multiplication variables between each flanking markers have to be considered**
  - **to completely absorb the effects of QTL with dominance**
  - **to reduce the residual error variance**
  - **to increase the detection precision**

Properties of the linear model in F_2

- **The additive and dominance effects of the flanked QTL are completely absorbed by the six variables in the model above.**
- **Interactions between marker variables may be declared as interaction between QTL by mistake when using ANOVA.**
- **But from our analysis, interactions between marker variables can be caused simply by dominance effects of QTL .**

Multiple QTL model in F_2

- **For multiple QTL, assume there are *m* QTL located on *m* intervals defined by *m*+1 markers on one chromosome, then the genotypic value of an F_2 individual is defined as:**

$$G = \mu + \sum_{j=1}^{m} [a_j w_j + d_j v_j]$$

where w_j and v_j are the indicators for genotypes at the *j* th QTL.

The linear model in F_2 under multiple QTL

- **The genotypic value of an F_2 individual with known marker types can be re-organized as:**

$$E(G) = \beta + \sum_{j=1}^{m+1} \lambda_j x_j + \sum_{j=1}^{m+1} \rho_j y_j$$

$$+ \sum_{j=1}^{m} \lambda\lambda_{j,j+1} x_j x_{j+1} + \sum_{j=1}^{m} \rho\rho_{j,j+1} y_j y_{j+1}$$

The linear model for QTL mapping in F_2

- **The inclusive linear model containing all markers simultaneously containing all markers and phenotyping errors is**

$$P = E(G) + \varepsilon = \beta + \sum_{j=1}^{m+1} \lambda_j x_j + \sum_{j=1}^{m+1} \rho_j y_j$$

$$+ \sum_{j=1}^{m} \lambda\lambda_{j,j+1} x_j x_{j+1} + \sum_{j=1}^{m} \rho\rho_{j,j+1} y_j y_{j+1} + \varepsilon$$

where *P* is the phenotypic value of a trait in interest, ε is the random environmental error.

Property of the linear model for QTL mapping in F_2

- **The QTL effects will be completely absorbed by the six variables of the two closest markers. Model 1 is suitable for QTL mapping in F_2 populations, as it completely explains both additive and dominance variations.**

ICIM (Inclusive Composite Interval Mapping) in F_2

- **Then assume there are *n* individuals in an F_2 population. Two steps are included in ICIM**
  - **Step1: stepwise regression was used to estimate the parameters in model 1. Coefficients of those variables not retained by stepwise regression were set at 0.**
  - **Step2: traditional interval mapping was conducted on adjusted phenotypic values, i.e.,**

$$\Delta P_i = P_i - \sum_{j \neq k, k+1} [\hat{\lambda}_j x_{ij} + \hat{\rho}_j y_{ij}]$$

$$- \sum_{j \neq k} [\lambda\hat{\lambda}_{j,j+1} x_{ij} x_{i,j+1} + \rho\hat{\rho}_{j,j+1} y_{ij} y_{i,j+1}]$$

Hypothesis test of QTL mapping in F_2

- **The two hypotheses used to test the existence of QTL at the scanning position are:**

  vs. $H_0 : \mu_1 = \mu_2 = \mu_3$

  H_A : at least two of μ_1, μ_1 and μ_3 are not equal

- **The logarithm likelihood under H_A is**

$$L_A = \sum_{j=1}^{9} \sum_{i \in S_j} \log[\sum_{k=1}^{3} \pi_{jk} f(\Delta P_i; \mu_k, \sigma^2)]$$

where S_j denotes individuals belonging to the j^{th} marker class (j=1, 2, …, 9), π_{jk} (k=1, 2, 3) is the proportion of the k^{th} QTL genotype in the j^{th} class, and $f(\cdot; \mu_k, \sigma^2)$ is the density function of the normal distribution $N(\mu_k, \sigma^2)$.

EM algorithm of QTL mapping in F_2

- **Use EM algorithm to get the estimation of** μ_1, μ_2 and μ_3

- **So the genetic effects in** $G = \mu + aw + dv$ **were therefore estimated by**

$$\hat{\mu} = \tfrac{1}{2}(\hat{\mu}_1 + \hat{\mu}_3) \quad \hat{a} = \tfrac{1}{2}(\hat{\mu}_1 - \hat{\mu}_3) \quad \hat{d} = \hat{\mu}_2 - \hat{\mu}$$

EM algorithm of QTL mapping in F_2

- **Parameters under H_0 were calculated as:**

$$\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n}\Delta P_i \qquad \hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(\Delta P_i - \hat{\mu}_0)^2$$

- **From which the maximum likelihood under H_0, and the LOD score between H_A and H_0 can be calculated.**

QTL distribution models in simulation

- **Six QTL with different levels of dominance and a genome consisting of 8 chromosomes.**
- **Each chromosome is 140 cM long, with 15 evenly distributed co-dominant markers.**
- **Three QTL distribution models.**

QTL distribution models in simulation

QTL	*a*	*d*	Model I		Model II		Model III		Genotypic variation explained (%)	Phenotypic variation explained (%)
			Chr	cM	Chr	cM	Chr	cM		
QTL1	1	0	1	25	1	25	1	25	11.43	8.00
QTL2	0	1	2	55	1	55	2	55	5.71	4.00
QTL3	1	1	3	25	2	25	3	25	17.14	12.00
QTL4	1	-1	4	55	2	55	1	55	17.14	12.00
QTL5	1	1.5	5	25	3	25	2	25	24.29	17.00
QTL6	1	-1.5	6	55	3	55	3	55	24.29	17.00

QTL distribution models in simulation

- **F_2 populations were simulated by the genetics and breeding simulation tool of QuLine.**
- **QTL mapping using ICIM was implemented by the software QTL IciMapping.**

Theoretical marker effects in the genetic model used in simulation

- **The expected additive, dominance, additive by additive, and dominance by dominance effects of the two flanking markers associated with each QTL is shown in the following table.**
- **It indicated that the dominance of a QTL could complicate the coefficients of the two markers flanking a QTL, and cause the interactions between markers.**

The expected marker effects in simulation

QTL	$(d)\mu_d$	$(a)A_1$	$(a)A_2$	$(d)D_1$	$(d)D_2$	$(d)AA_{12}$	$(d)DD_{12}$	Interaction variation (%)
QTL1	0.000	0.498	0.498	0.000	0.000	0.000	0.000	0.0
QTL2	0.253	0.000	0.000	0.248	0.248	-0.248	0.243	21.8
QTL3	0.253	0.498	0.498	0.248	0.248	-0.248	0.243	5.7
QTL4	-0.253	0.498	0.498	-0.248	-0.248	0.248	-0.243	5.7
QTL5	0.379	0.498	0.499	0.371	0.371	-0.371	0.364	9.6
QTL6	-0.379	0.498	0.498	-0.371	-0.371	0.371	-0.364	9.6

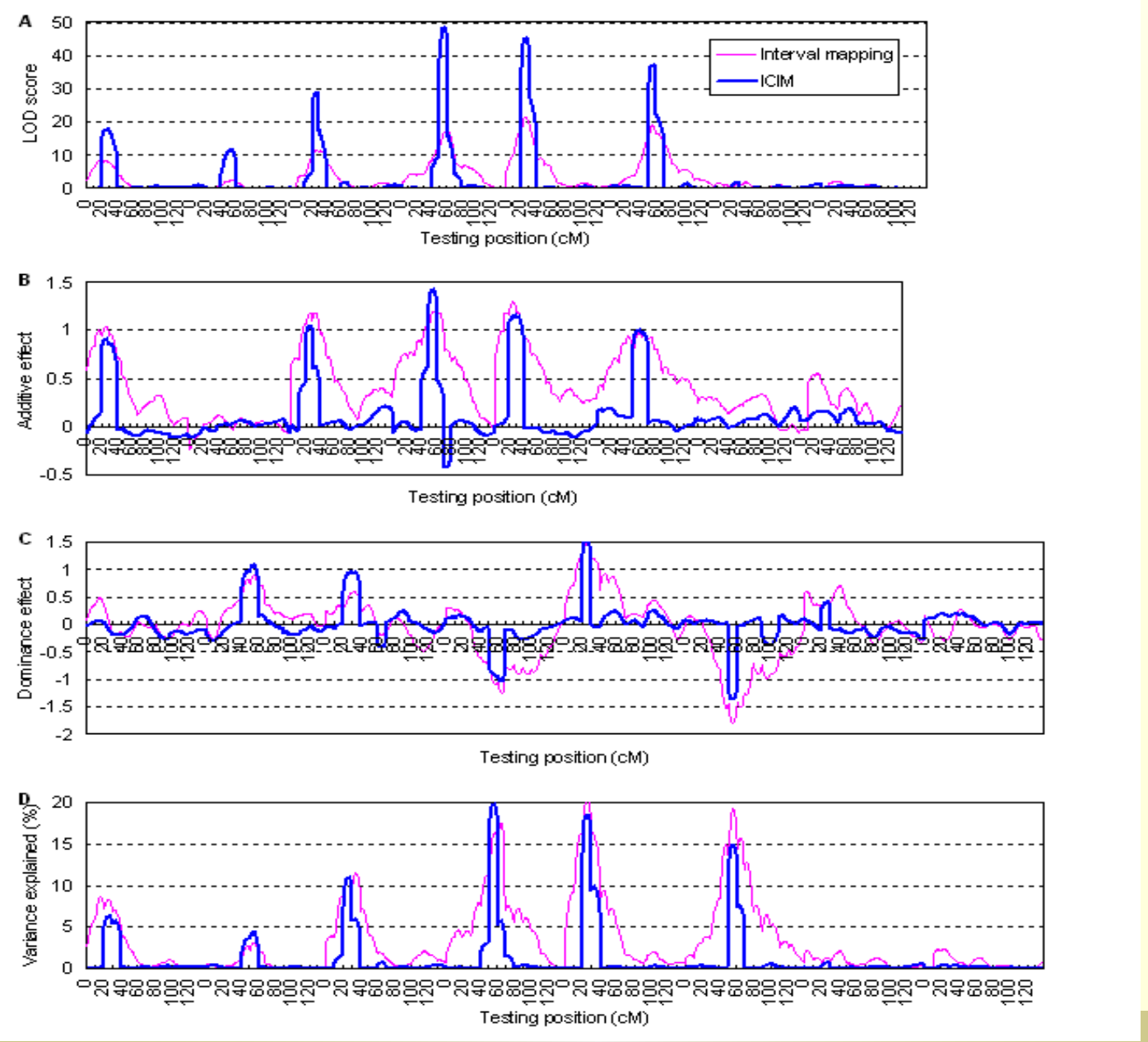
QTL mapping in simulated F_2 populations

- There is no clear classification on the phenotype, and it is impossible to deduce the number of QTL without the assistance of molecular markers.
- Set the LOD threshold at 3.0. The two probabilities for entering and removing variables were set at 0.01 and 0.02. The scanning step is 1 cM.
- The mapping results is shown in the next page.

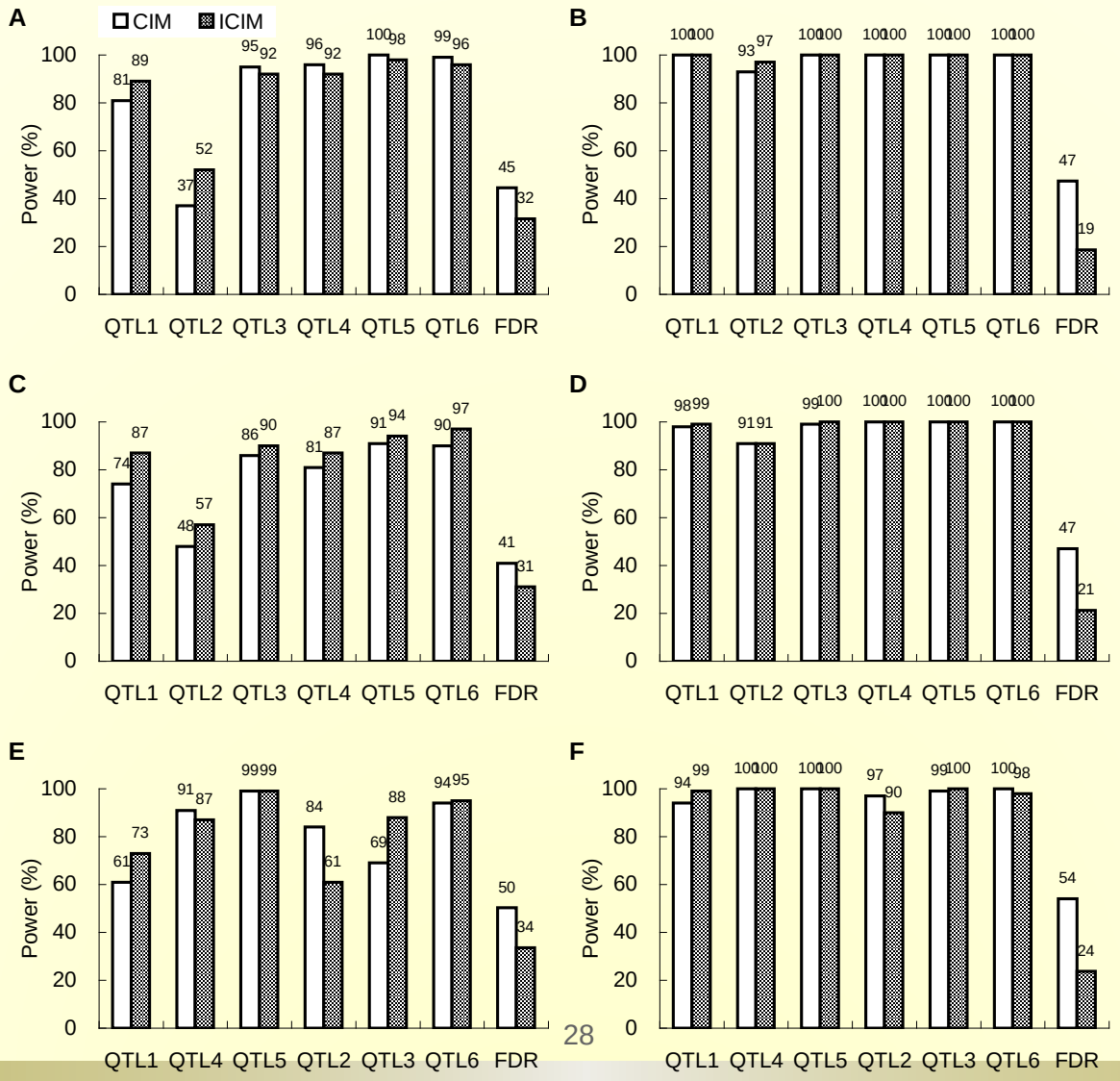
Mapping results of simulated populations

QTL	LOD score	PVE (%)	True Position (cM)	Est. Position (cM)	True add. effect	Est. add. effect	True dom. effect	Est. dom. effect
QTL distribution model I								
QTL1	16.52	6.67	25	28	1	0.88	0	-0.11
QTL2	7.67	3.27	55	53	0	0.03	1	0.85
QTL3	25.11	11.28	25	24	1	0.86	1	1.08
QTL4	35.46	16.43	55	57	1	0.74	-1	-1.58
QTL5	37.12	16.74	25	26	1	1.05	1.5	1.38
QTL6	28.44	13.16	55	55	1	0.84	-1.5	-1.22

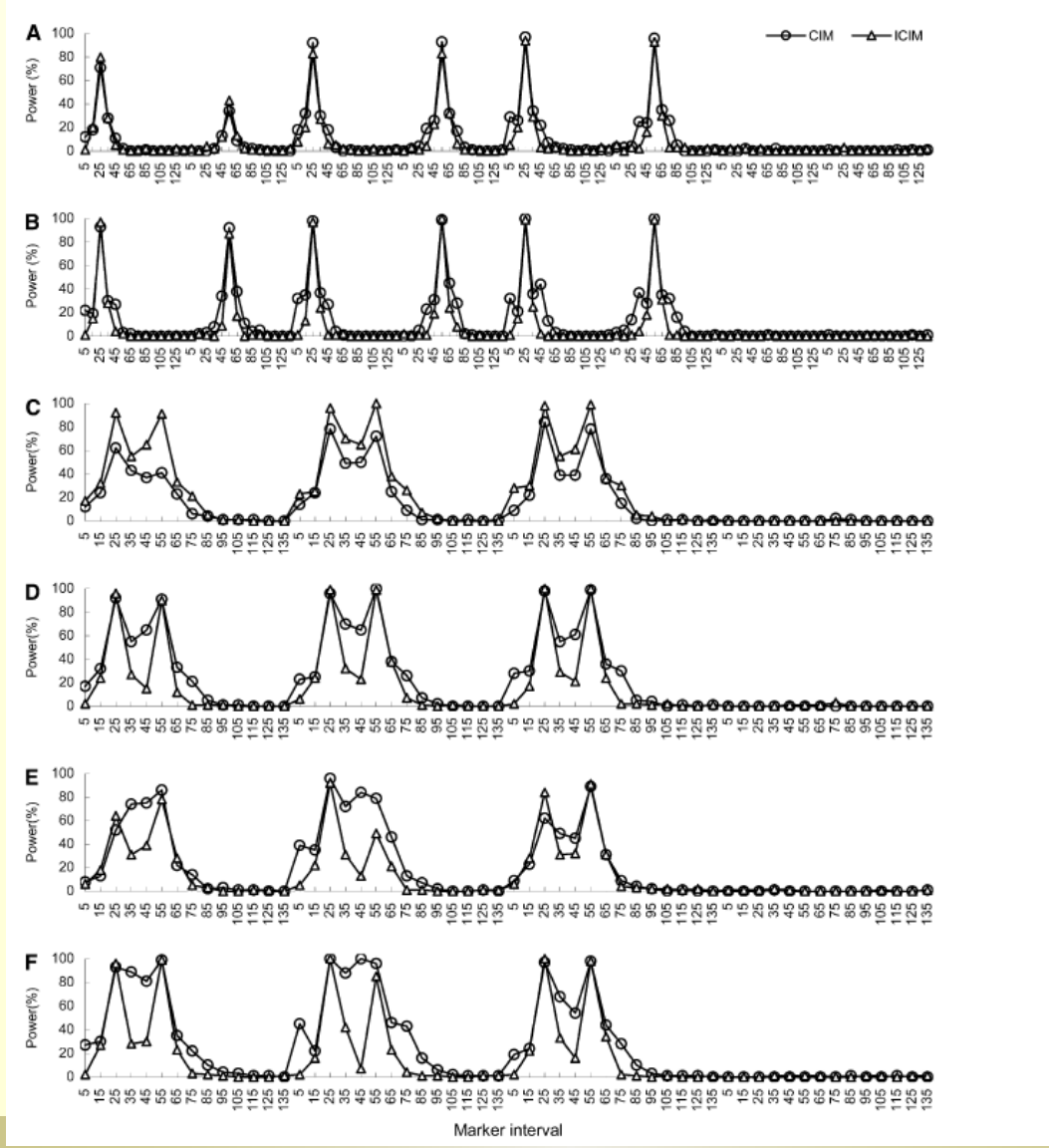
Comparison of ICIM with IM



Power analysis of CIM and ICIM from 100 simulations



Power analysis of CIM and ICIM for marker intervals



An actual rice F_2 population

- **180 individuals**
- **The cross was made in Chengdu, China, in July 2002 between the *indica* rice variety and Nipponbare.**
- **137 SSR markers.**
- **The whole genome was of 2046.2 cM, and the average marker distance was 17.1 cM.**
- **A number of agronomic traits were investigated in the field.**

QTL mapping in the actual F_2 population

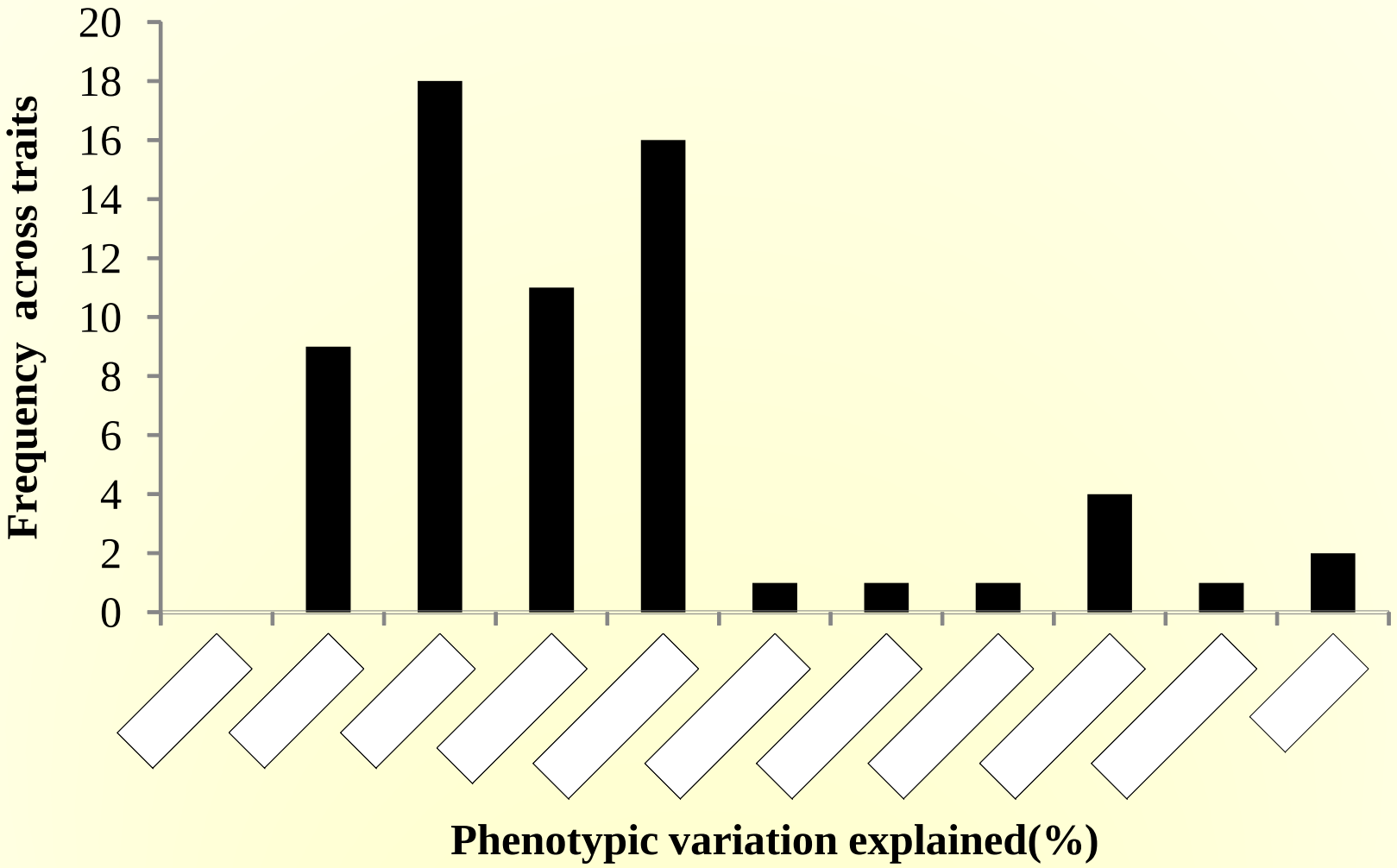
- **In the real population, few F_2 individuals are shorter than PA64s, indicating most, if not all, reduced height alleles are harbored by PA64s.**
- **Set the LOD threshold at 3.0. The two probabilities for entering and removing variables were set at 0.01 and 0.02. Scanning step is 1 cM.**

QTL distribution

Trait	R^2 of additive (%)	R^2 of additive and dominance (%)	Absolute degree of dominance (|d/a|)				Total
			<=0.25	(0.25, 0.75]	(0.75, 1.25]	>1.25	
PH	25.84	51.56	2	1	1	5	9
HD	16.12	41.37	1	1	1	3	6
PL	25.58	61.26	5	3	1	8	17
FL	20.86	40.00	0	2	0	3	5
SPK	25.64	27.09	1	1	1	1	4
TKW	20.11	20.11	2	0	2	1	5
DP	19.45	24.87	1	1	0	1	3
GL	30.69	41.96	1	1	0	0	2
GW	26.63	26.63	2	2	0	0	4
RLW	37.63	45.70	1	3	1	1	6
		Total	16	15	7	23	61

Notes: PH=plant height, HD=heading date, PL= panicle length, FL=flag leaf length, SPK=spikelets per panicle, TKW=thousand kernel weight, DP=density of panicle, GL=grain length, GW=grain width, RLW=ratio of grain length and width.

PVE distribution



QTL for height and maturity

Trait	QTL	Chr	Distance to left marker	Add	Dom	LOD	PVE(%)
Plant height (Ph)	QPh1-1	1	12	-0.57	-7.98	8.04	12.03
	QPh1-2	1	19.5	-8.59	0.59	15.54	25.57
	QPh3-1	3	16.9	4.35	-4.86	6.51	13.30
	QPh3-2	3	11.4	-4.69	-1.00	5.04	6.84
	QPh4	4	13.7	-3.56	-2.09	4.61	5.53
	QPh5	5	13	-0.44	-4.48	3.13	3.86
	QPh6	6	6.2	-0.79	-5.05	3.17	4.96
	QPh7	7	7	0.26	6.48	5.27	7.56
	QPh12	12	2.4	-1.66	3.93	3.98	5.44
Heading date (Hd)	QHd1	1	22.1	1.74	-0.30	3.65	7.27
	QHd3	3	19.9	0.88	-3.70	6.04	21.09
	QHd4	4	0.2	-0.77	1.85	3.58	5.24
	QHd8	8	5.7	-1.41	-1.46	4.79	8.20
	QHd10	10	0.3	-1.78	-0.80	4.85	7.21
	QHd11	11	6.2	0.15	-3.03	5.71	11.70

Conclusions

- **Two multiplication variables between each flanking markers have to be considered.**
- **The inclusive linear model in F_2 is**

$$P = E(G) + \varepsilon = \beta + \sum_{j=1}^{m+1} \lambda_j x_j + \sum_{j=1}^{m+1} \rho_j y_j$$

$$+ \sum_{j=1}^{m} \lambda\lambda_{j,j+1} x_j x_{j+1} + \sum_{j=1}^{m} \rho\rho_{j,j+1} y_j y_{j+1} + \varepsilon$$

to absorb the effects of QTL with dominance. And then use ICIM algorithm for mapping.