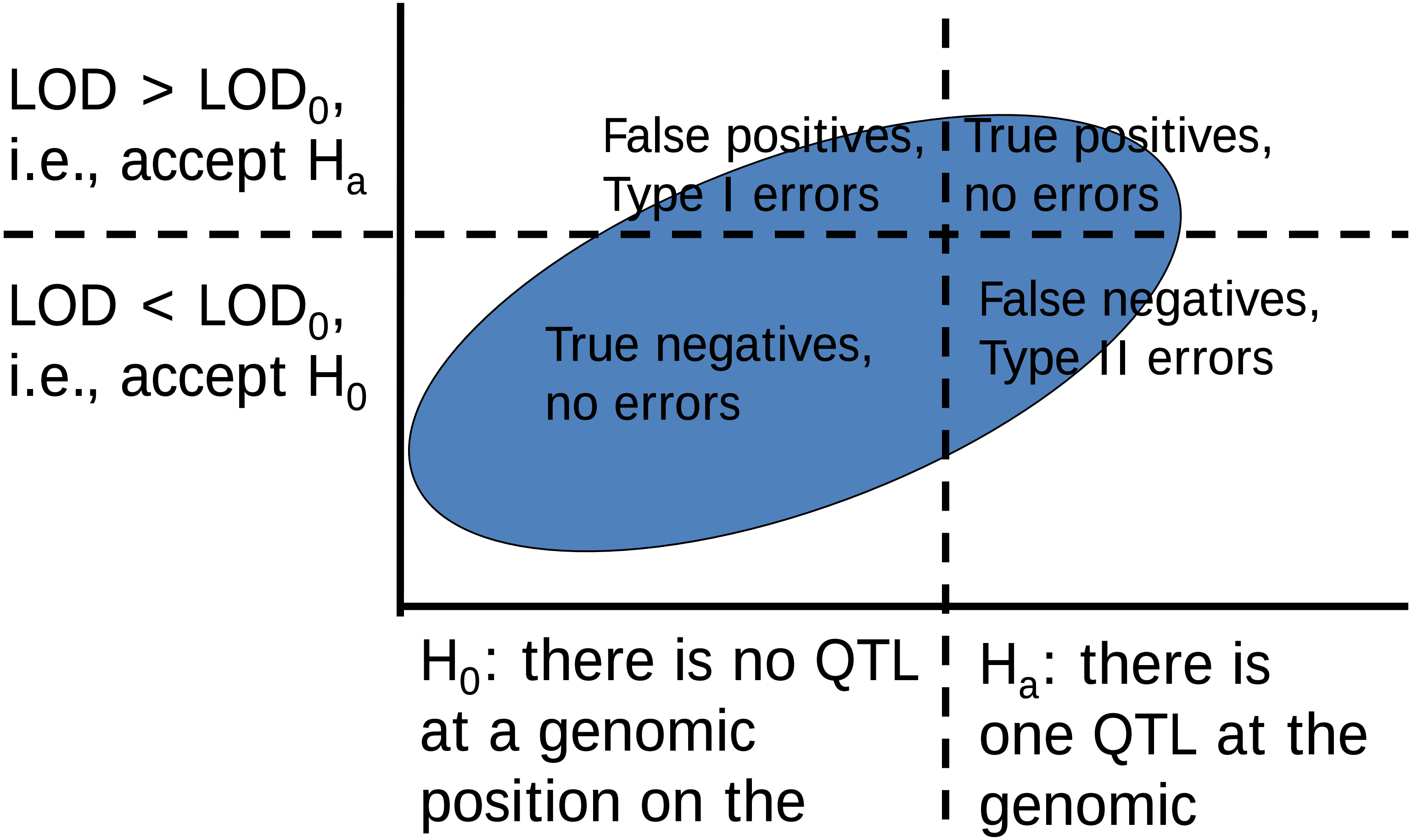
One-week Course on Genetic Analysis and Plant Breeding
21 - 25 January 2013, CIMMYT, Mexico

Frequently Asked Questions in QTL Mapping

Jiankang Wang, CIMMYT China and CAAS
E-mail: jkwang@cgiar.org; wangjiankang@caas.cn
Web: http://www.isbreeding.net

Q1: What is LOD?

Hypothesis test in QTL mapping



LOD > LOD_0,
i.e., accept H_a

False positives,
Type I errors

True positives,
no errors

LOD < LOD_0,
i.e., accept H_0

True negatives,
no errors

False negatives,
Type II errors

H_0: there is no QTL at a genomic position on the trait in interest

H_a: there is one QTL at the genomic position

Likelihood ratio test (LRT)

- Definition of LRT $LRT = -2\ln(\frac{L_0}{L_A})$
- Definition of LOD (likelihood of odd)

$$LOD = \log(\frac{L_A}{L_0}) = \log(L_A) - \log(L_0)$$

- Relationship between LOD and LRT

$$LOD = \frac{LRT}{2\ln(10)} \approx \frac{LRT}{4.61}$$

$$LRT \approx 4.61LOD$$

Q2: How to choose a threshold value of LOD?

Sun, Z., H. Li, L. Zhang, **J. Wang***. 2013. Properties of the test statistic under null hypothesis and the calculation of LOD threshold in quantitative trait loci (QTL) mapping. Acta Agronomica Sinica (accepted)

Two types of error in hypothesis test

- Type I error rate = P {Reject H_0 | True H_0}

- Type II error rate = P {Accept H_0 | False H_0}

Significance level (α) in hypothesis test – The control of Type I error

- Significance level for N times of independent tests: $1\text{-}(1\text{-}\ \alpha)^{N}$
- Bonferroni adjustment: $\approx \alpha\ /\ N$

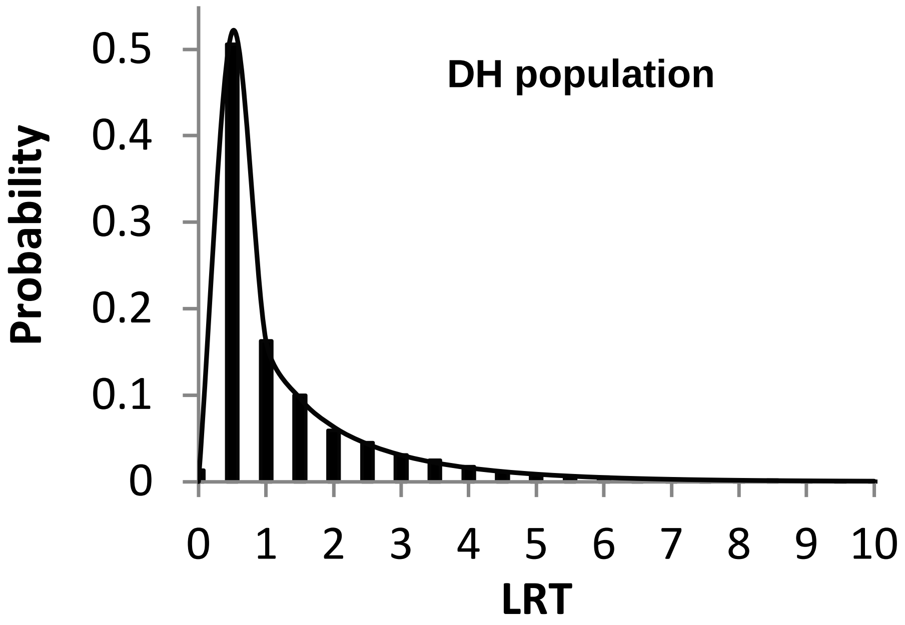
- Problem: Multiple and dependent tests exist in QTL mapping!
- Permutation test in QTL mapping

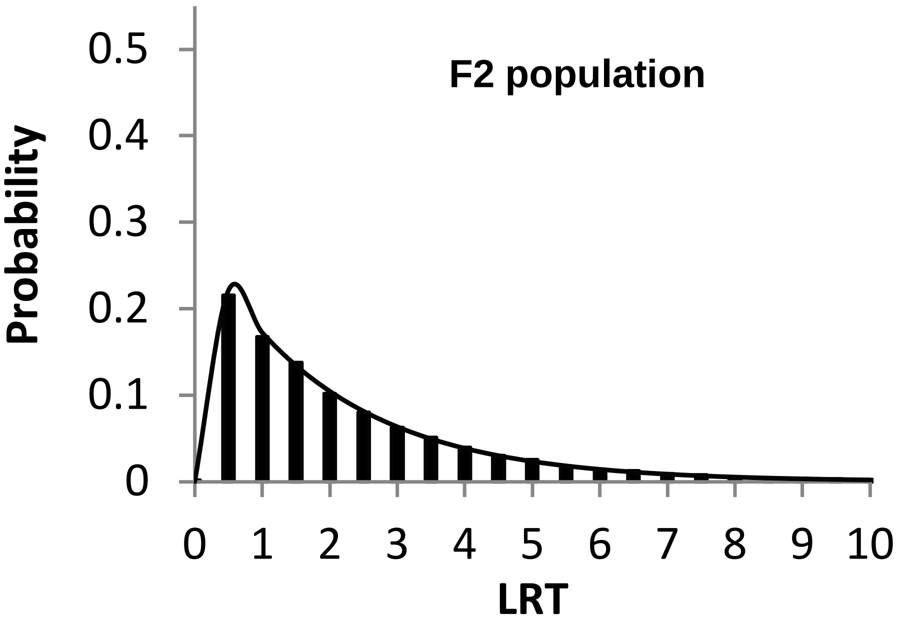
Choice of the threshold of LOD

- For one test： α (e.g., 0.1, 0.05, 0.01)
- N times of independent tests: $1-(1-\alpha)^N$

- Empirical LOD threshold for an overall significance level of 0.05: 2.0 – 3.0

Distribution of LRT under H_0 at each scanning position

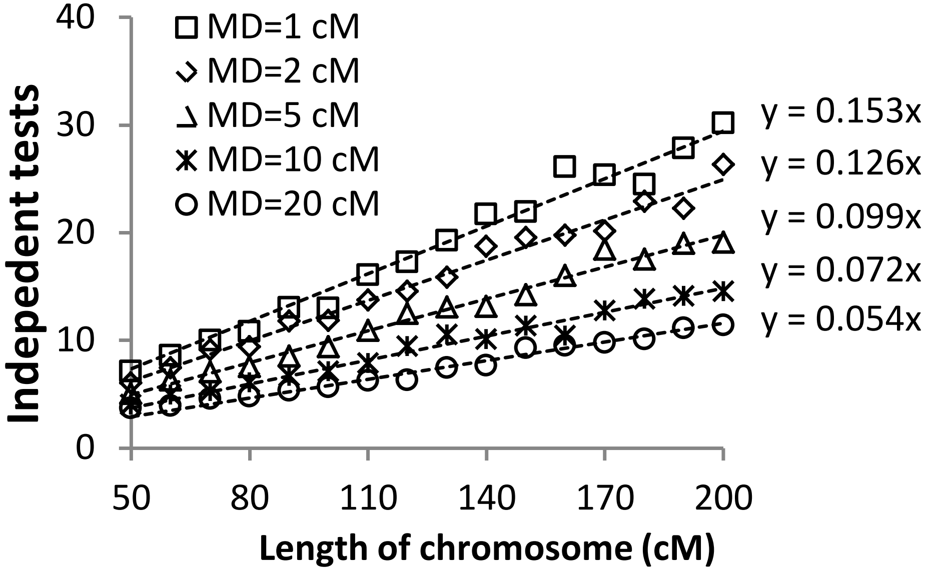




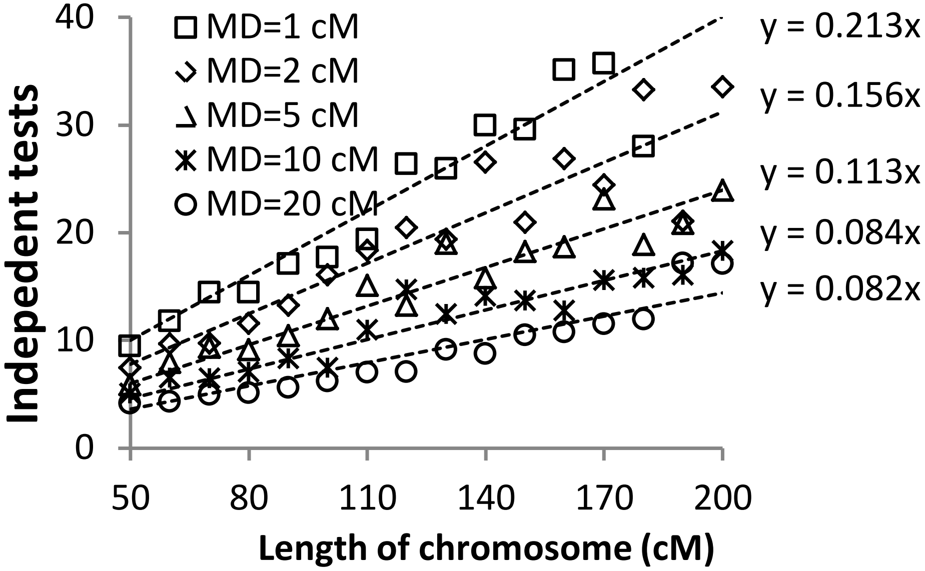
- In DH populations, LRT ~ χ^2(df=1)
- In F2 populations, LRT ~ χ^2(df=2)
- D.F. is equal to the number of independent genetic effects to be estimated

Number of independent tests

DH population, genome-wide Type I error = 0.05



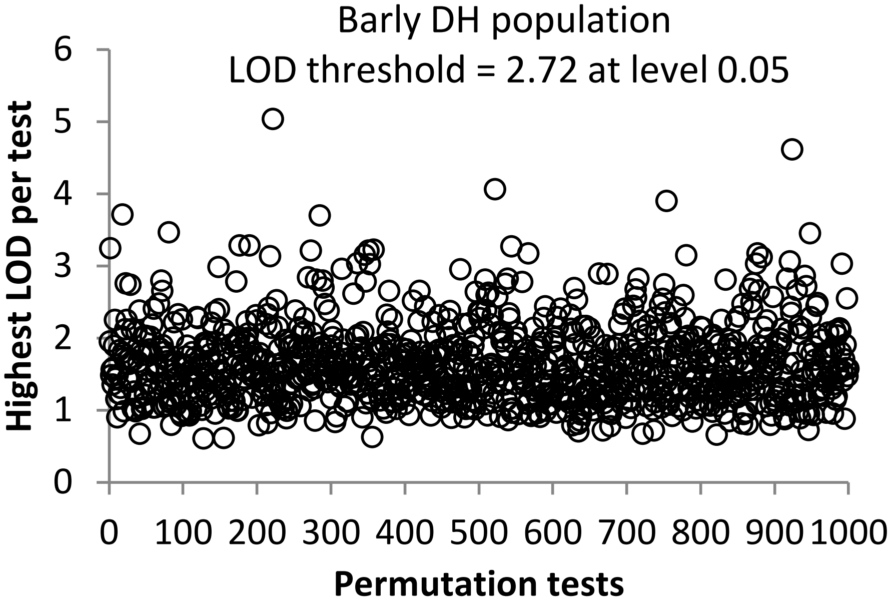
DH population, genome-wide Type I error = 0.01

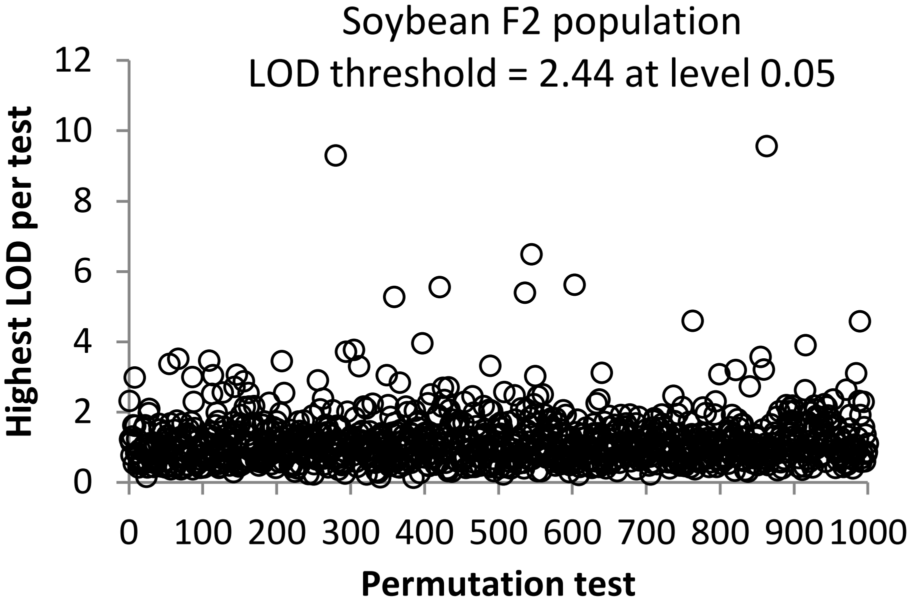


LOD threshold, assuming marker density is 1 cM

Genome size	Genome-wide α=0.05			Genome-wide α=0.01		
	DH	RIL	F2	DH	RIL	F2
50	1.61	1.84	2.40	2.37	2.56	3.18
75	1.77	2.01	2.57	2.53	2.73	3.36
100	1.88	2.12	2.70	2.65	2.84	3.49
150	2.04	2.28	2.87	2.81	3.01	3.66
200	2.16	2.40	3.00	2.93	3.13	3.79
250	2.24	2.49	3.10	3.02	3.22	3.88
300	2.32	2.56	3.17	3.10	3.29	3.96
500	2.52	2.77	3.40	3.31	3.50	4.18
1000	2.80	3.05	3.70	3.59	3.79	4.49
1500	2.97	3.22	3.87	3.76	3.95	4.66
2000	3.09	3.33	4.00	3.88	4.07	4.79
3000	3.25	3.50	4.17	4.04	4.24	4.96
4000	3.37	3.62	4.30	4.16	4.36	5.09

LOD threshold from permutation test



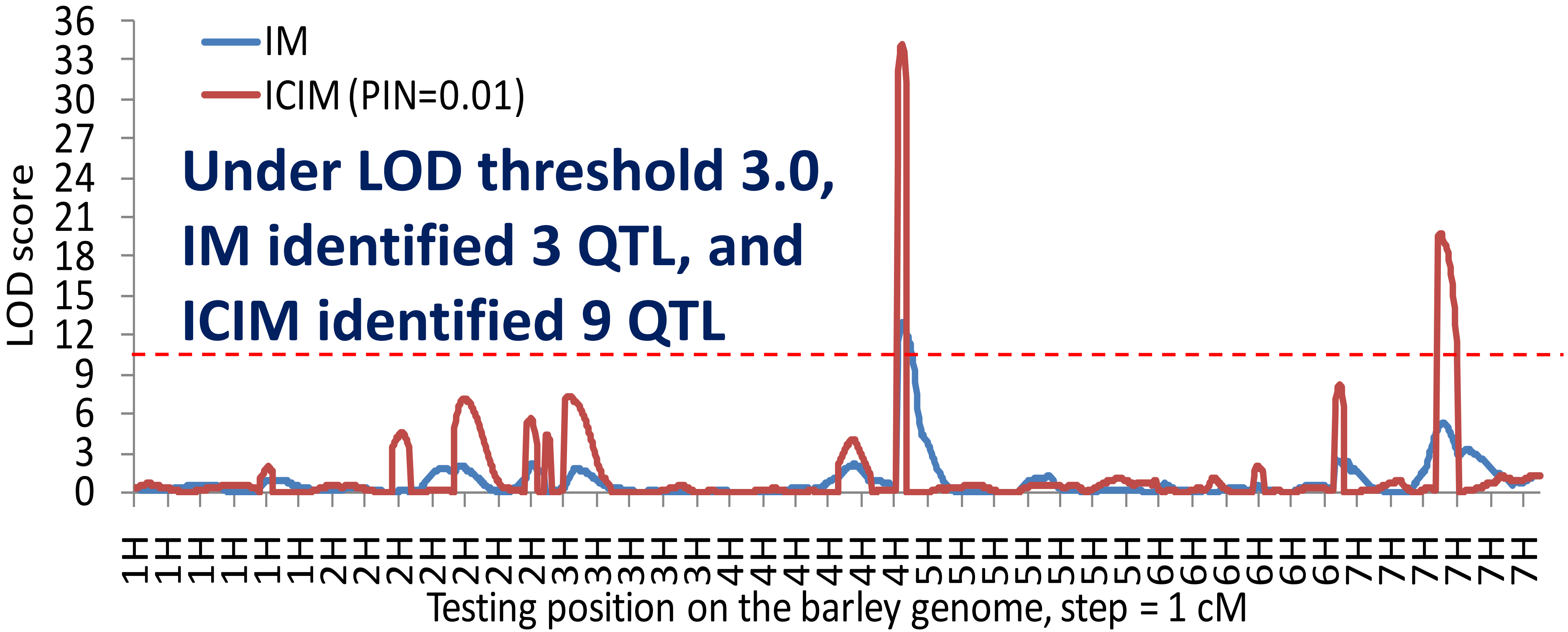


Q3: Which method to use?

Power of a statistical test

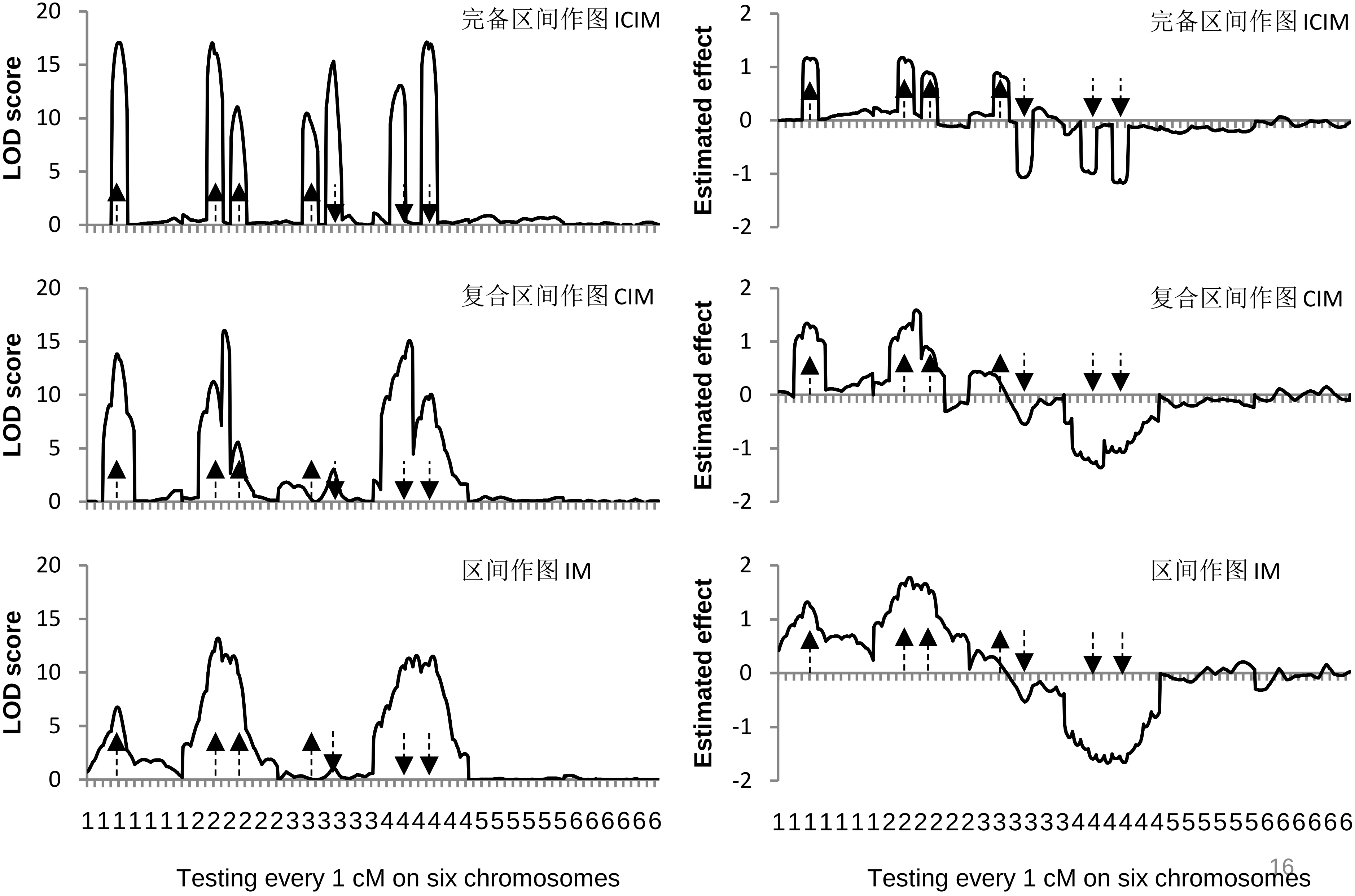
- The power of a statistical test is the probability that the test will reject a false null hypothesis (i.e., it will not make a Type II error).
- Power= 1.0 – Type II error

QTL mapping from IM and ICIM



Under LOD threshold 3.0, IM identified 3 QTL, and ICIM identified 9 QTL

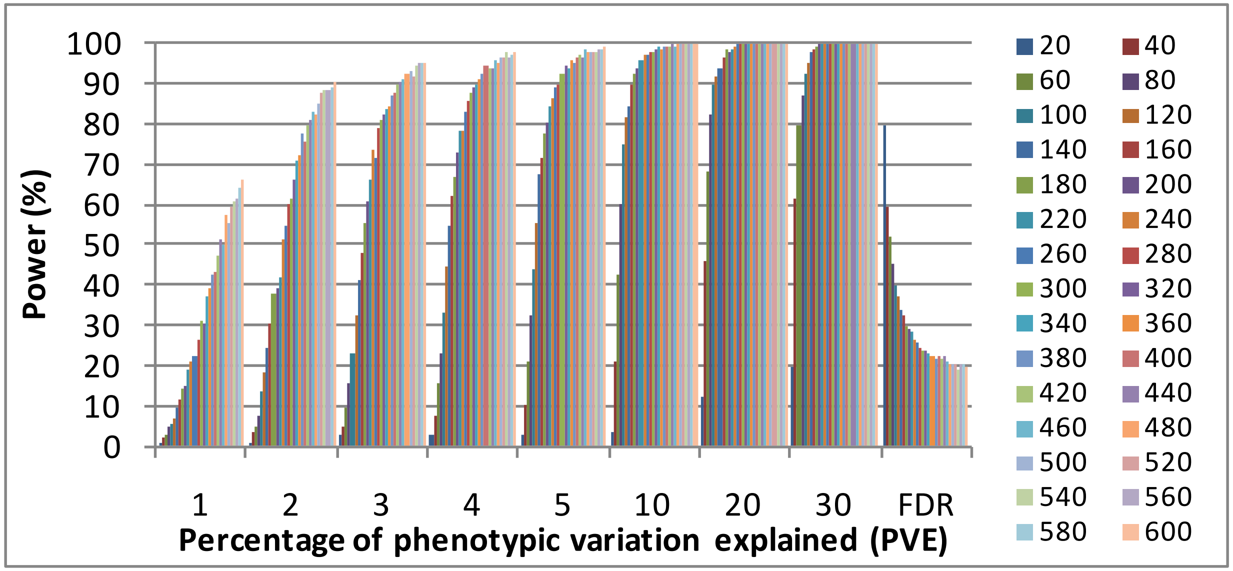
QTL mapping in a simulated population

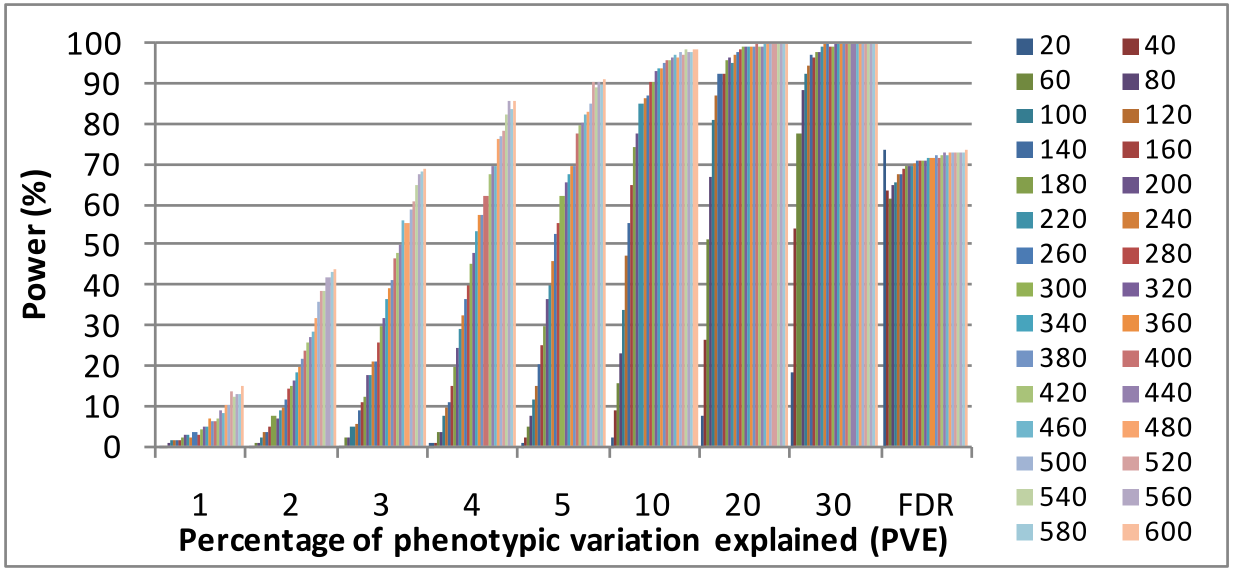


Power from 3 simulated populations, each of 200 RILs

Pop	Chr.	Pos.	Effect	LOD	PVE (%)	CI=10 cM
1	2	25.1	0.19	2.56	3.48	False QTL
	5	51.1	0.29	6.05	8.14	IQ5
	6	60.0	0.30	6.72	8.86	IQ6
	7	40.0	0.20	2.94	3.71	False QTL
	7	70.0	0.42	11.87	16.64	IQ7
2	2	30.5	0.27	5.35	7.78	IQ2
	5	45.0	0.27	5.25	7.94	False QTL
	6	59.1	0.26	4.94	7.50	IQ6
	7	59.4	0.38	9.84	15.61	False QTL
3	2	30.0	0.21	2.50	3.96	IQ2
	6	55.4	0.29	4.47	7.81	IQ6
	7	70.0	0.28	4.42	7.14	IQ7
	7	90.0	0.25	3.39	5.41	False QTL

Power comparison: ICIM vs. IM



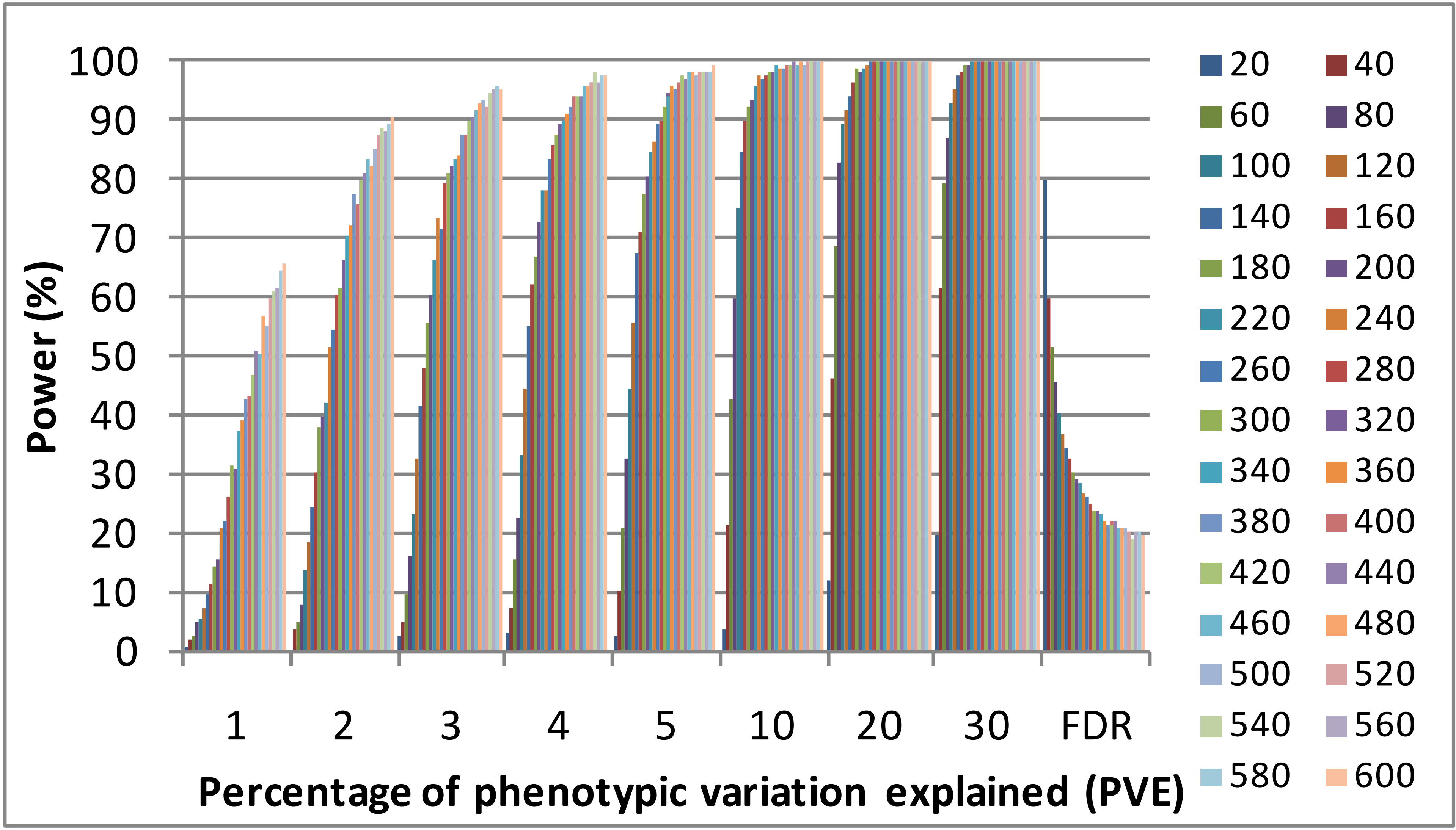


Simulation can help to determine the population size

	Probability			
PVE (%)	0.9	0.8	0.7	0.6
1			>600	540
2	600	420	340	280
3	430	280	230	200
4	340	250	190	160
5	280	200	160	130
10	160	120	100	80
20	100	80	60	50
30	100	60	50	40

FDR: false discovery rate

(False QTL out of all positives)



Q4: What are the ways that can improve mapping efficiency?

Possible ways

- Large population
- Precision phenotyping and genotyping
- Efficient method
- High marker density?!
  - For association mapping, yes.
  - For linkage mapping, probably no.
- Two-stage mapping strategy?!

The length of empirical 95% confidence intervals of QTL

PVE (%)	PS=200				PS=400			
	MD=5 cM	MD=10 cM	MD=20 cM	MD=40 cM	MD=5 cM	MD=10 cM	MD=20 cM	MD=40 cM
1	93.30	104.55	103.10	120.03	47.00	61.94	76.83	88.24
2	54.14	62.37	73.66	86.08	37.55	32.38	38.34	47.59
3	52.65	47.12	50.02	48.18	25.64	21.95	18.78	33.36
4	38.89	46.14	41.94	56.72	25.01	18.46	22.74	36.57
5	25.99	37.83	44.73	59.74	16.35	16.39	22.54	36.30
10	10.31	8.35	11.29	46.45	3.72	4.78	7.92	22.74
20	8.55	10.19	14.78	26.97	4.90	6.04	8.70	15.56
30	5.33	8.23	11.56	18.62	3.18	4.78	6.62	12.62

Q5: How to calculate the contribution of individual QTL?

Contribution of a QTL on phenotypic variation

- PVE = Phenotypic variation explained (%)
- $PVE_g = V_g/V_p * 100\%$
  - BC, DH and RIL， $V_g=a^2$ (a is the additive effect)
  - F2, $V_g=a^2/2+d^2/4$ (d is the dominance effect)

Does high effect mean high PVE?

- In DH or RIL, when there is segregation distortion,
  - V_g=(1-q)*a^2+q*a^2-[(1-2q)*a]^2=4q(1-q)a^2
  - V_g depends on effect and allele frequency
  - When p=q=0.5, V_g is maximized; otherwise, smaller than that of non-distortion
- It is possible that one higher-effect QTL has lower PVE

Non-additive PVE

- For two random variables X and Y
  - $E(X+Y) = E(X) + E(Y)$
  - $V(X+Y) = V(X) + V(Y) + 2Cov(X, Y)$

- When QTL are unlinked, PVE of multiple QTL is the sum of individual PVE
- When QTL are linked, PVE of multiple QTL is not equal to the sum of individual PVE

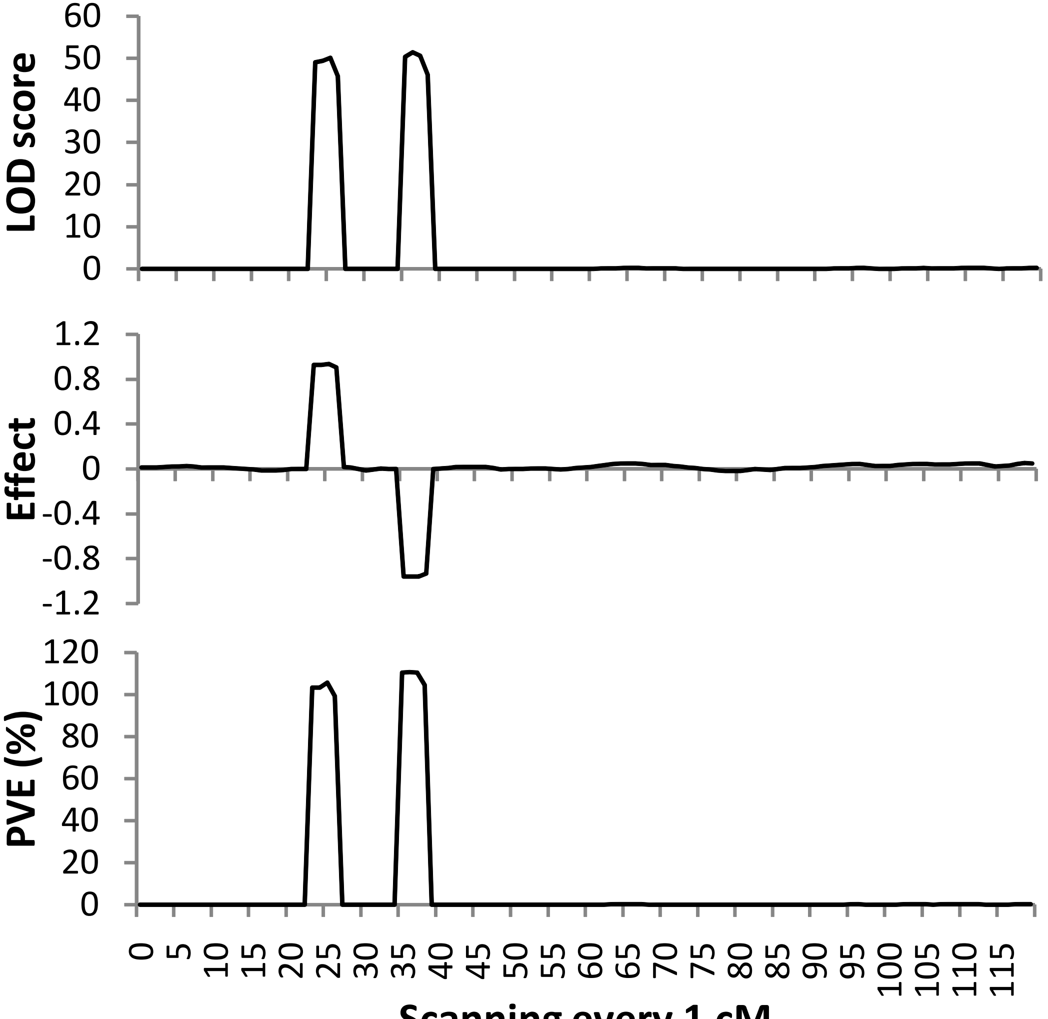
PVE can be more than 100%

Genotype	Frequency	Genotypic value
AABB	(1-r)/2	$m+a_1+a_2$
AAbb	r/2	$m+a_1-a_2$
aaBB	r/2	$m-a_1+a_2$
aabb	(1-r)/2	$m-a_1-a_2$

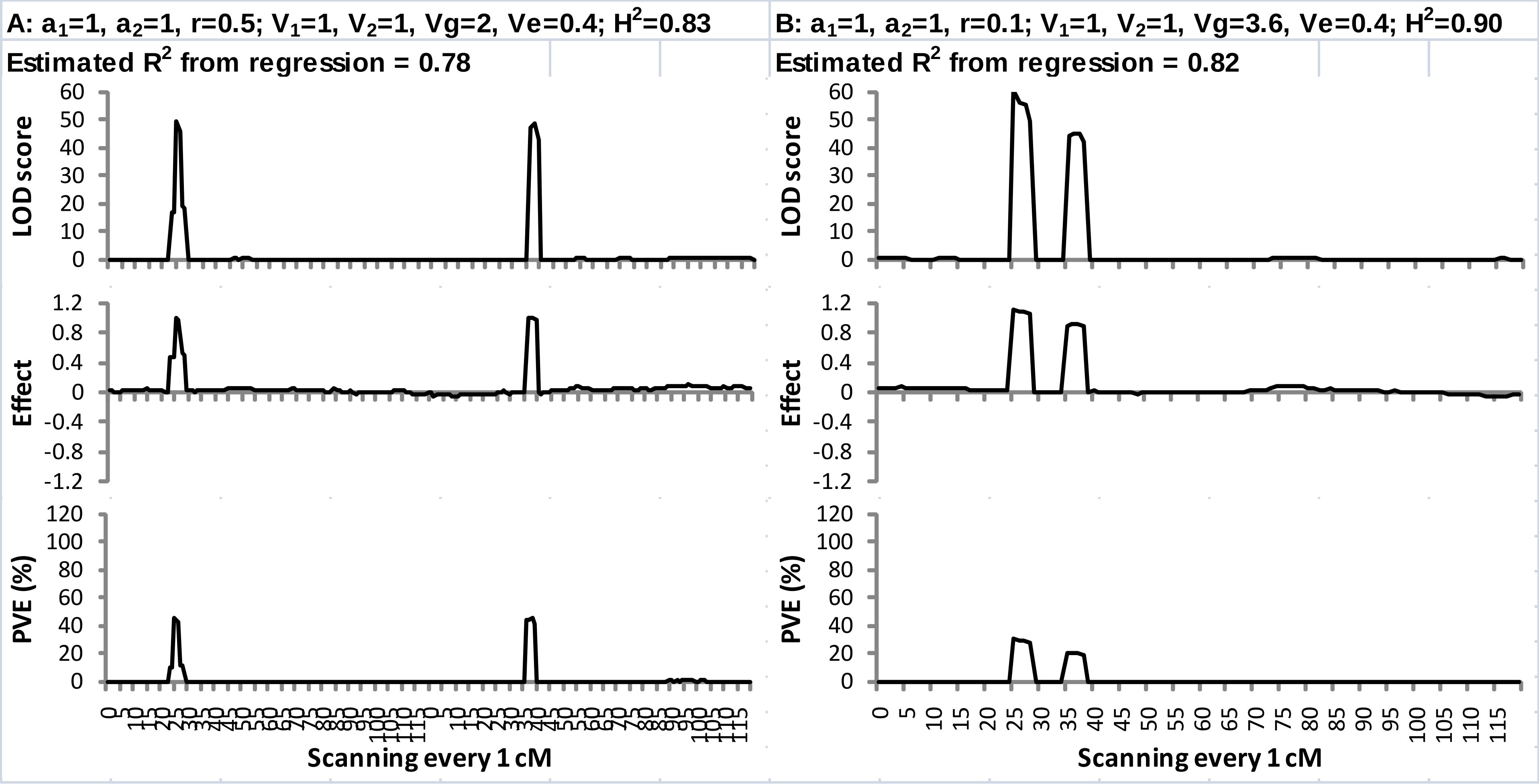
- Two loci A-a and B-b with a recombination frequency r
- In the DH population
  - Genetic variance of A-a: a_1^2
  - Genetic variance of B-b: a_2^2
  - Total genetic variance : $a_1^2 + a_2^2+2(1-2r)a_1a_2$

A simulated population of 200 DH lines

Two QTL are located at 25 cM and 36 cM on a chromosome of 120 cM. Their additive effects are 1.0 and –1.0. Random error variance is 0.4. Marker interval is 2 cM.



Linkage in coupling



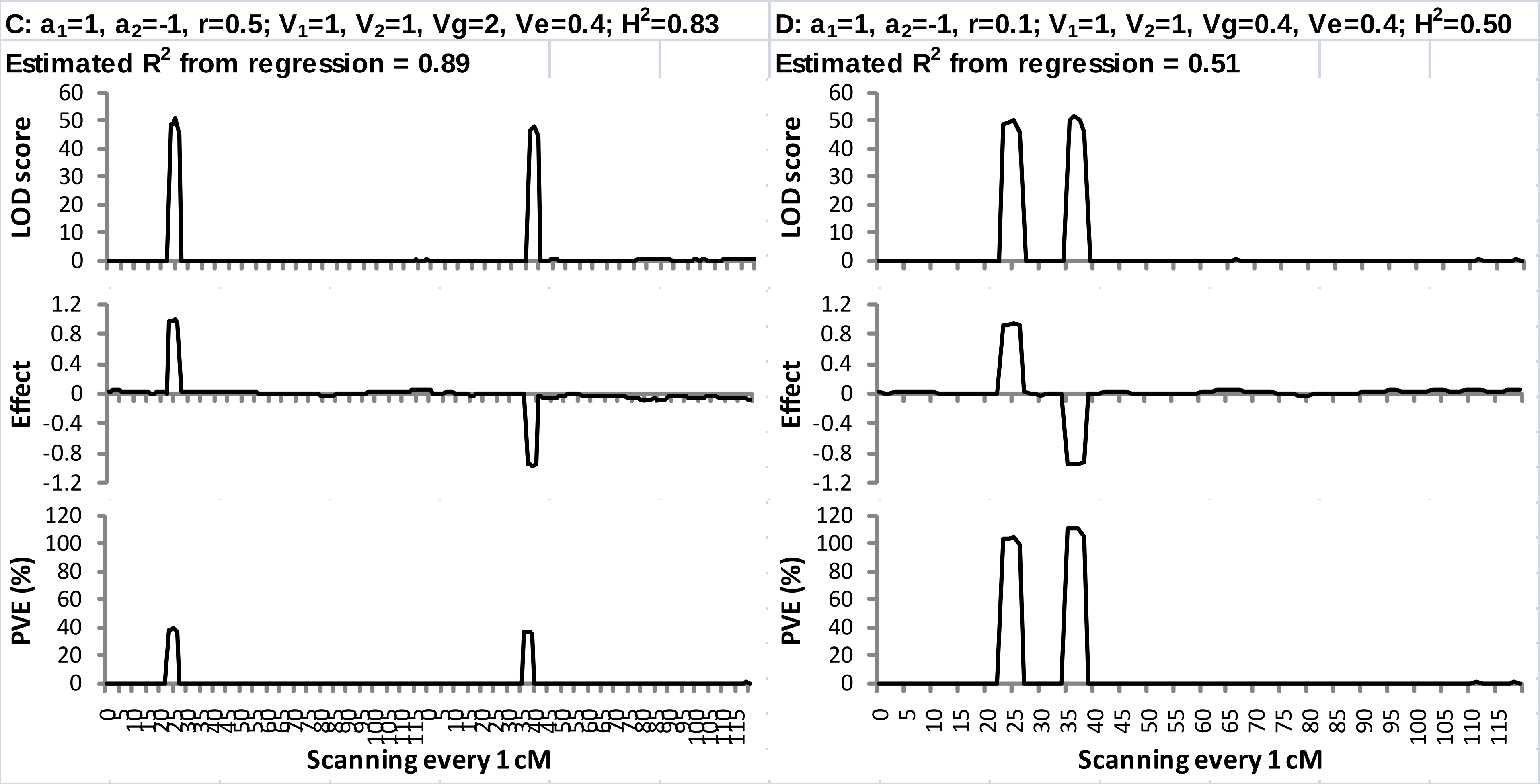
A: a_1=1, a_2=1, r=0.5; V_1=1, V_2=1, Vg=2, Ve=0.4; H^2=0.83

Estimated R^2 from regression = 0.78

B: a_1=1, a_2=1, r=0.1; V_1=1, V_2=1, Vg=3.6, Ve=0.4; H^2=0.90

Estimated R^2 from regression = 0.82

Linkage in repulsion



C: a_1=1, a_2=-1, r=0.5; V_1=1, V_2=1, Vg=2, Ve=0.4; H^2=0.83

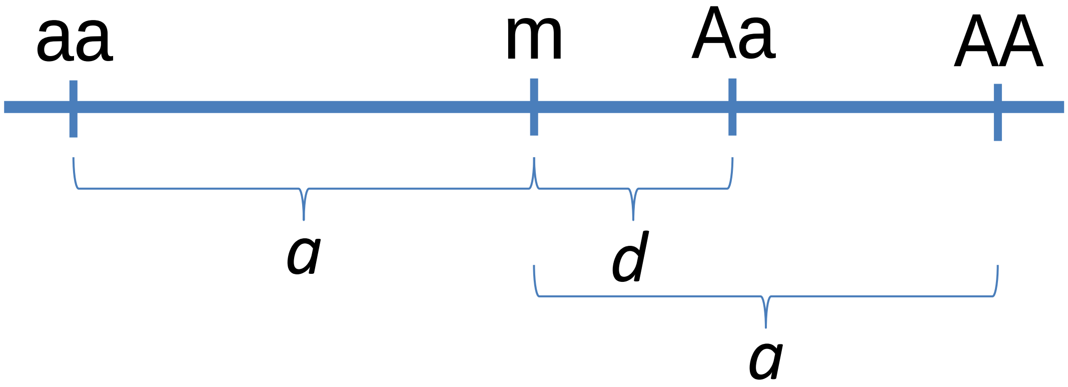
Estimated R^2 from regression = 0.89

D: a_1=1, a_2=-1, r=0.1; V_1=1, V_2=1, Vg=0.4, Ve=0.4; H^2=0.50

Estimated R^2 from regression = 0.51

Q6: How to determine the source of favorable alleles?

Source of favorable alleles

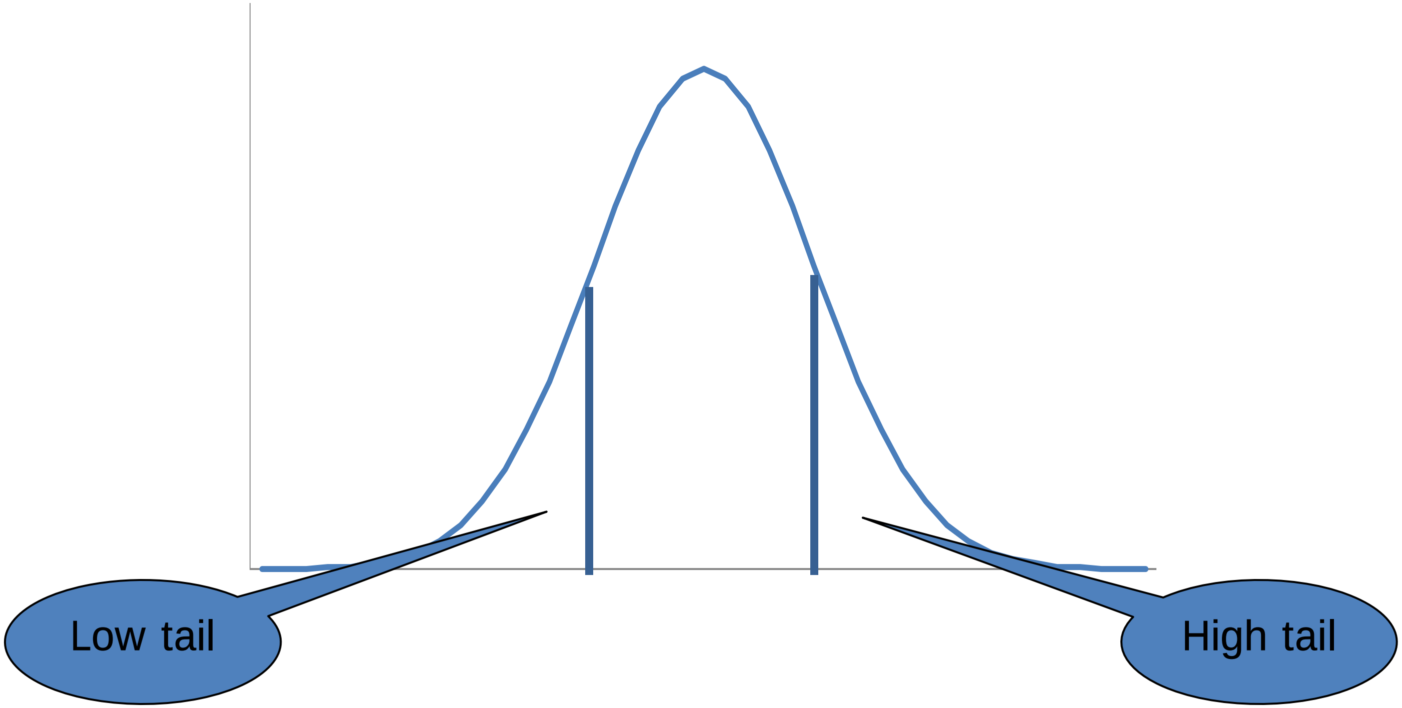


- Definition of additive and dominance genetic effects
  - Coding in QTL mapping: 2 (P1), 0 (P2), 1 (F1)
  - P1: m+a; F1: m+d; P2: m-a
  - When higher value is favored
    - If *a* is positive, the favorable allele is carried by P1
    - If *a* is negative, the favorable allele is carried by P2
  - When lower value is favored
    - If *a* is negative, the favorable allele is carried by P1
    - If *a* is positive, the favorable allele is carried by P2

Q7: Is selective genotyping still useful?

Sun, Y., **J. Wang**, J. H. Crouch, and Y. Xu. * 2010. Efficiency of selective genotyping for genetic analysis and crop improvement of complex traits. **Mol. Breed.** 26: 493-511.

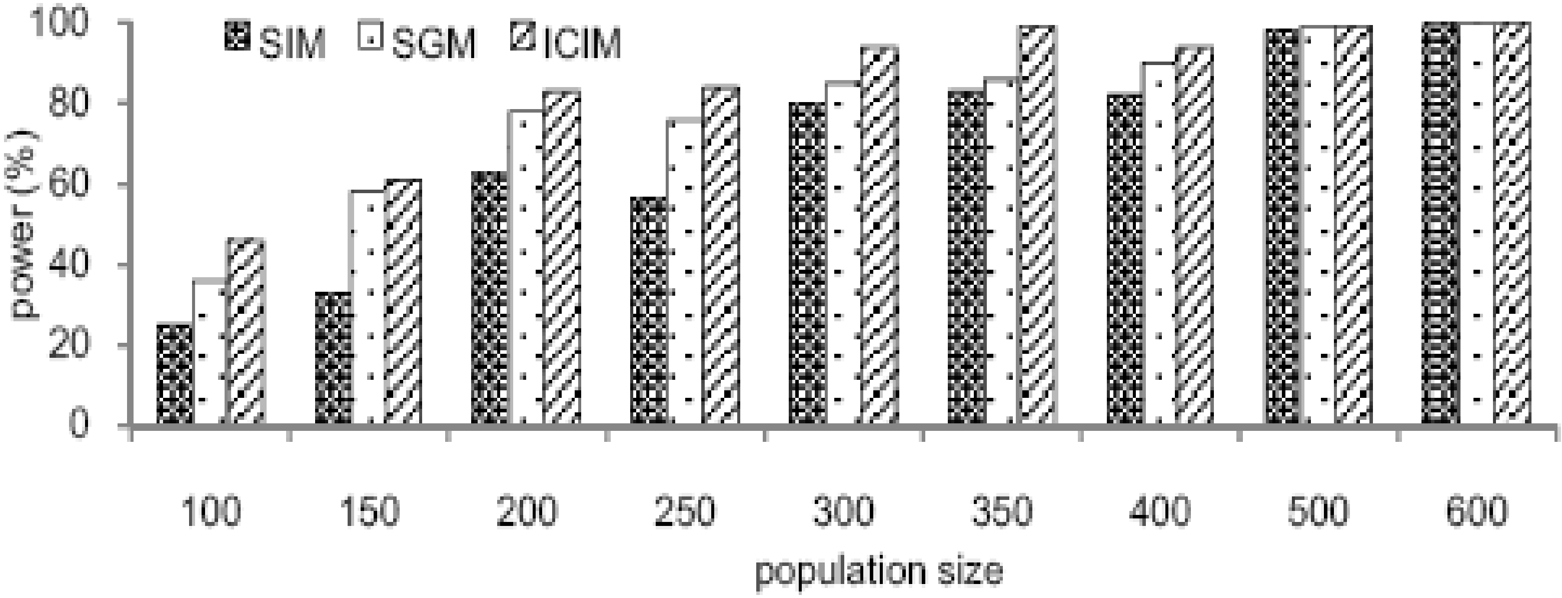
Selective genotyping



$$t = \frac{p_H - p_L}{\sqrt{\frac{p_H(1-p_H)}{2N_H} + \frac{p_L(1-p_L)}{2N_L}}}$$

Comparison of SGM with IM and ICIM

(PVE=5%, MD=5cM and both tails have the selected proportion of 10%)



SGM has higher detection power than the conventional IM but lower detection power than ICIM

SGM may still be useful!

Q8: Can mathematically derived traits be used in QTL mapping?

Wang, Y., H. Li, L. Zhang, W. Lu, **J. Wang***. 2012. On the use of mathematically-derived traits in QTL mapping. **Mol. Breed.** 29: 661–673

Genetic effects of composite traits

Effect	Trait I	Trait II	Addition	Subtraction	Multiplication	Division
Mean	25	20	45	5	500	1.2563
A_1	1	0	1	1	20	0.0503
A_2	1	0	1	1	20	0.0503
A_3	0	1	1	-1	25	-0.0631
A_4	0	1	1	-1	25	-0.0631
A_{12}	0	0	0	0	0	0
A_{13}	0	0	0	0	1	-0.0025
A_{14}	0	0	0	0	1	-0.0025
A_{23}	0	0	0	0	1	-0.0025
A_{24}	0	0	0	0	1	-0.0025
A_{34}	0	0	0	0	0	0.0063
A_{123}	0	0	0	0	0	0
A_{124}	0	0	0	0	0	0
A_{134}	0	0	0	0	0	0.0003
A_{234}	0	0	0	0	0	0.0003
A_{1234}	0	0	0	0	0	0

Composite traits reduced power and increased FDR

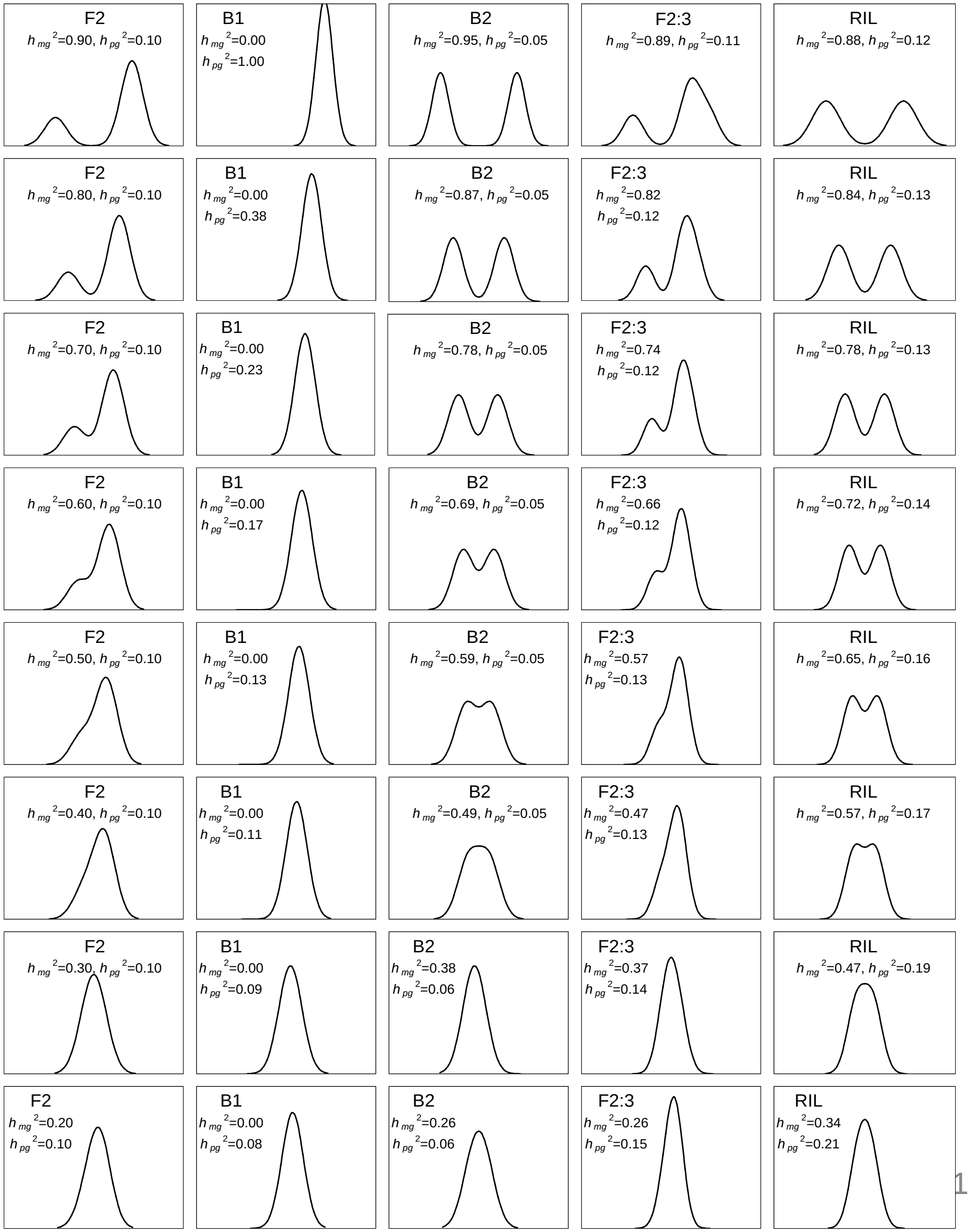
		QTL	Trait I	Trait II	Addition	Subtraction	Multiplication	Division
Model I	Power (%)	**Q1**	**95.10**		**69.60**	**69.30**	**55.20**	**50.50**
		Q2	**94.80**		**69.80**	**70.40**	**54.10**	**50.90**
		Q3		**92.50**	**67.20**	**65.30**	**76.90**	**75.20**
		Q4		**94.50**	**68.40**	**65.40**	**77.80**	**75.20**
	FDR (%)		**21.63**	**22.98**	**27.42**	**28.05**	**28.07**	**29.68**
Model II	Power (%)	**Q1**	**95.40**		**67.40**	**65.60**	**54.80**	**49.90**
		Q2	**92.90**		**62.40**	**66.00**	**50.00**	**49.90**
		Q3		**93.70**	**69.90**	**67.00**	**79.20**	**74.90**
		Q4		**91.90**	**62.40**	**64.90**	**73.50**	**72.90**
	FDR (%)		**21.35**	**22.18**	**28.76**	**28.59**	**28.07**	**28.89**
Model III	Power (%)	**Q1**	**95.20**		**66.60**	**52.40**	**53.60**	**37.70**
		Q2	**95.00**		**69.20**	**51.60**	**54.70**	**36.40**
		Q3		**92.90**	**63.40**	**47.80**	**69.70**	**56.20**
		Q4		**92.60**	**61.50**	**49.90**	**72.60**	**58.00**
	FDR (%)		**19.78**	**23.44**	**28.83**	**27.71**	**29.74**	**30.18**

Q9: Does the phenotype have to be normally distributed?

Quantitative traits are normally distributed under the polygene hypothesis

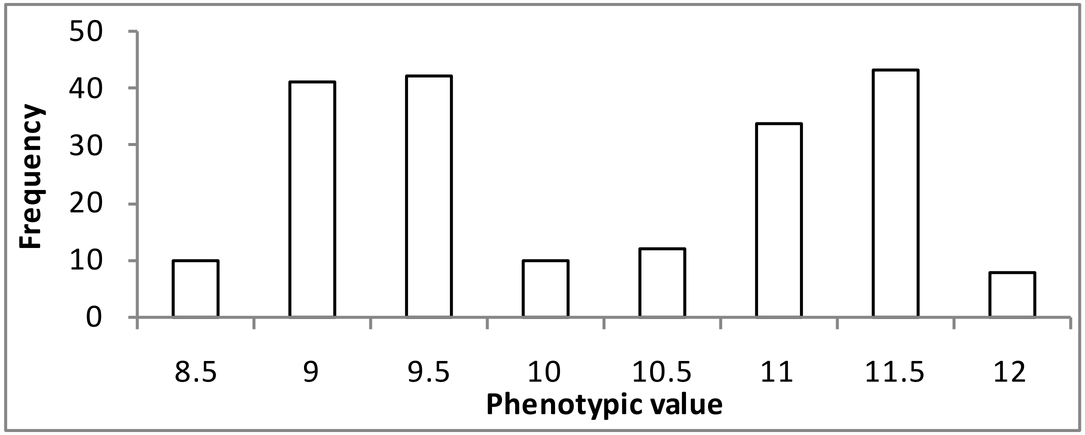
Phenotypic distribution under one major gene model

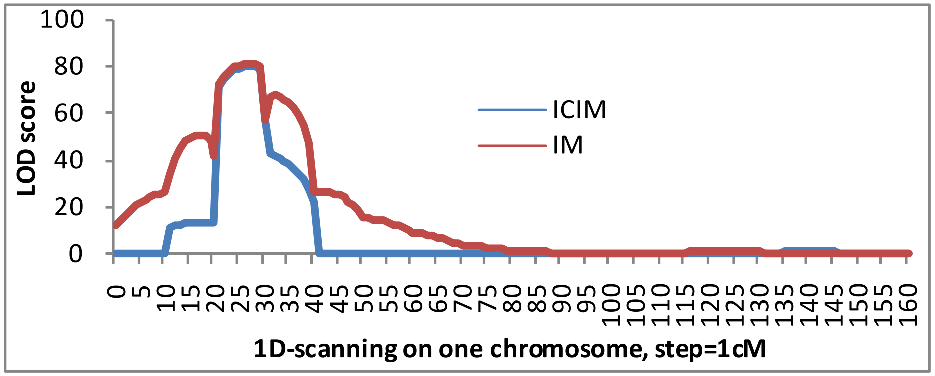
Random errors have to be normally distributed and independent!

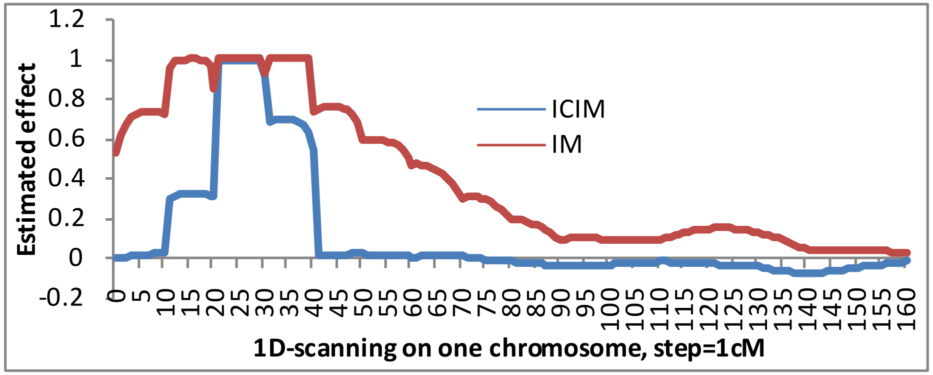


One QTL with PVE = 80%

Located at 25 cM and *a*=1.0







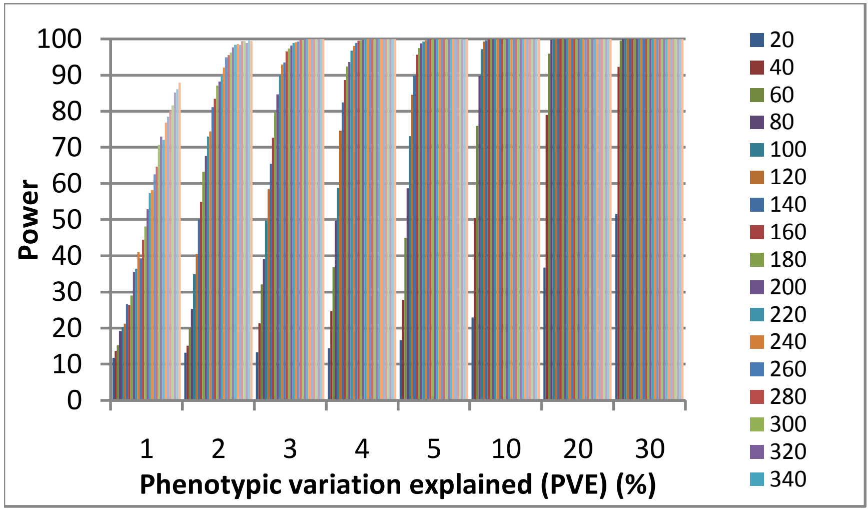
Q10: Can the precision be improved by adding more markers?

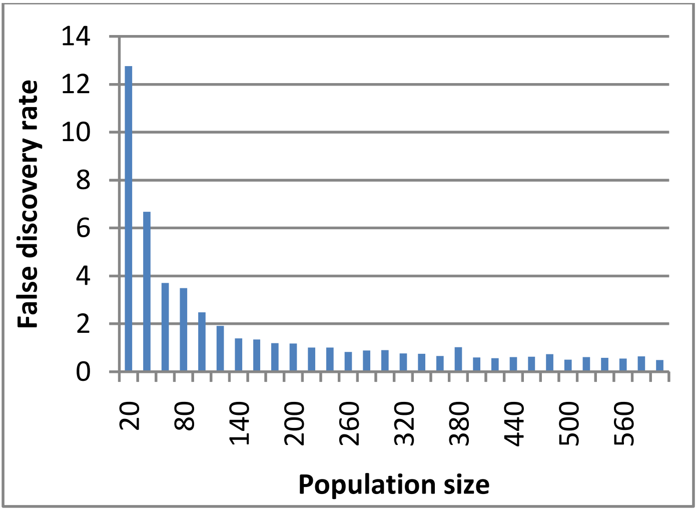
Empirical marker density

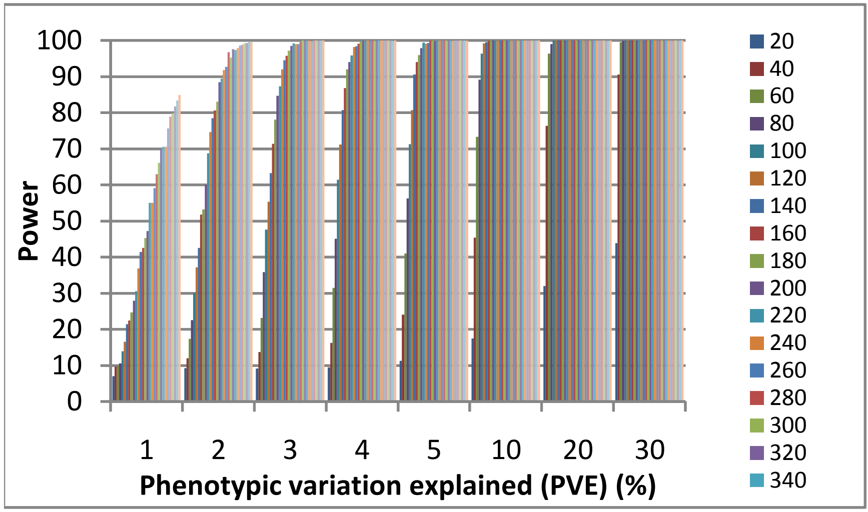
- **In linkage mapping**
  - – 10-20 cM, covering the whole genome
  - – Marker density + large population
- **In association mapping**
  - – The more, the better to exploit the remaining LD in the mapping population

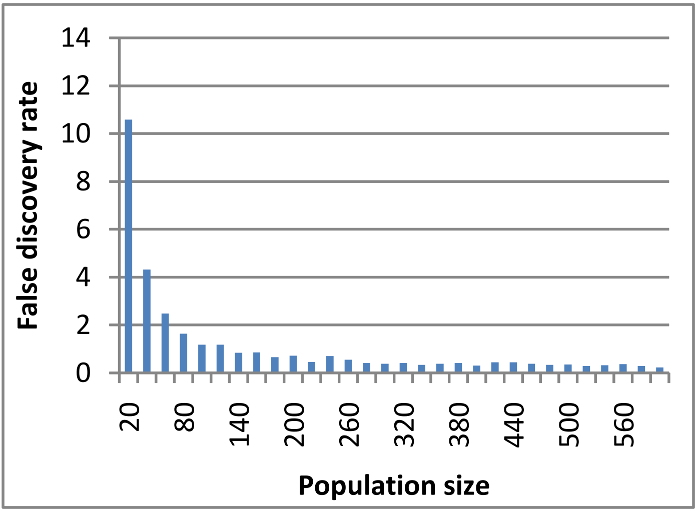
Power and FDR for two marker densities: 10 cM (up), and 20 cM (down)

(Confidence interval is the whole chromosome)



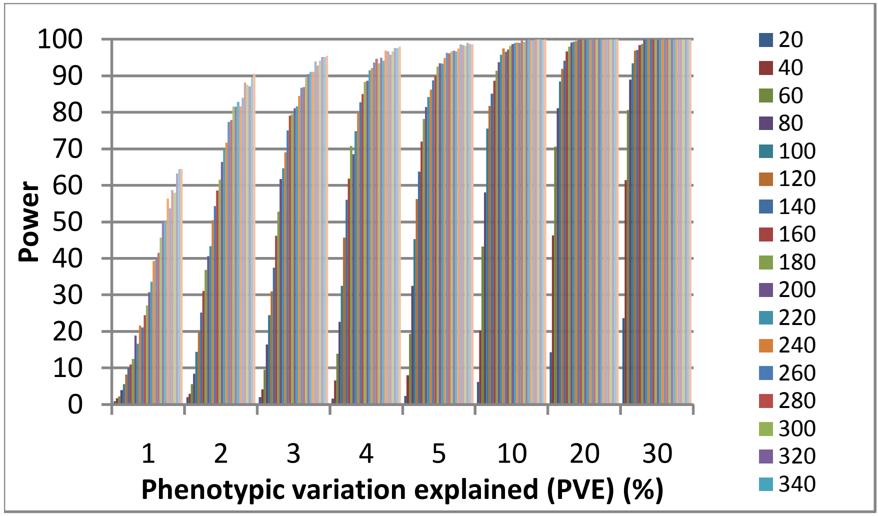


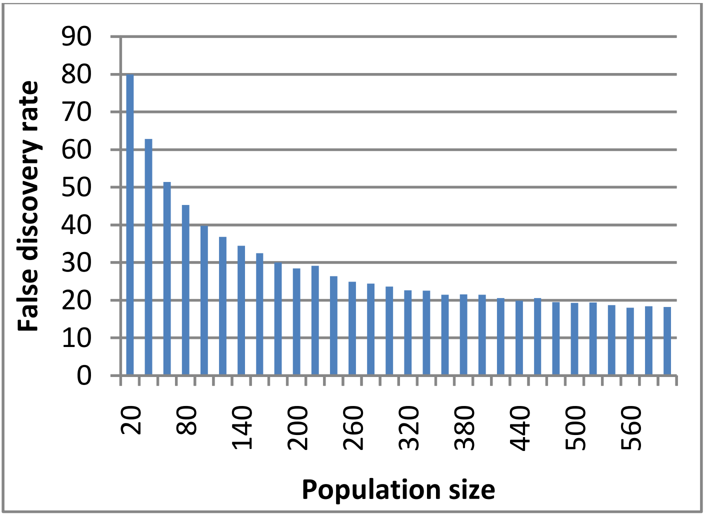


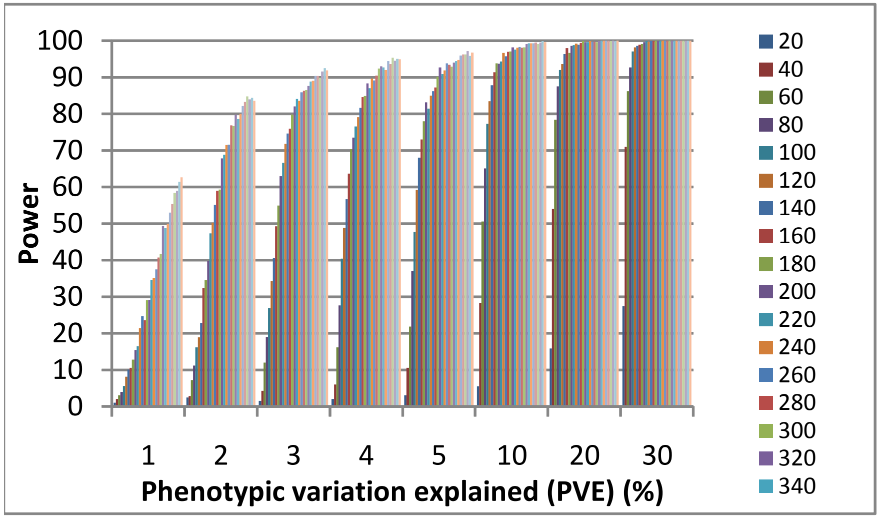


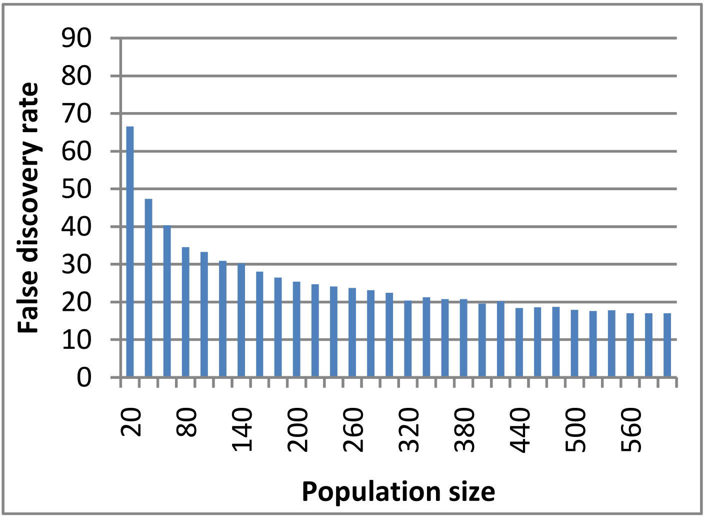
Power and FDR for two marker densities: 10 cM (up), and 20 cM (down)

(10 cM confidence interval, true QTL at the center of CI)









Q11: What is the effect of missing markers?

Zhang, L., S. Wang, H. Li, Q. Deng, A. Zheng, S. Li, P. Li, Z. Li, **J. Wang***. 2010. Effects of missing marker and segregation distortion on QTL mapping in F_2 populations. **Theor. Appl. Genet.** 121:1071-1082.

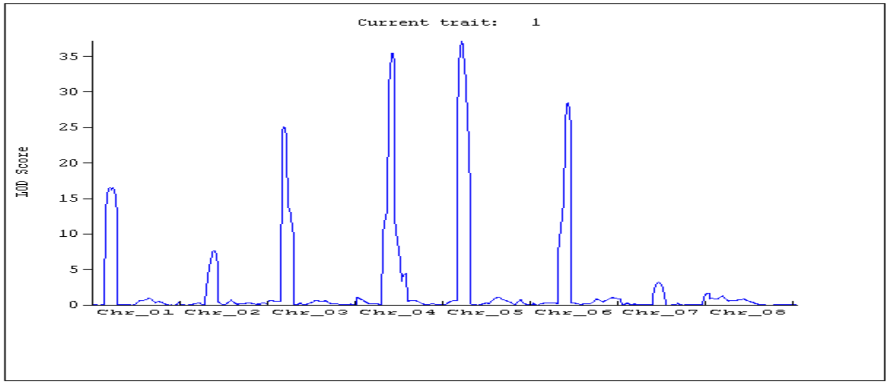
Missing data in QTL mapping

- Missing markers
  - Imputation using the linkage map
- Missing phenotype
  - Mean replacement
  - Deletion

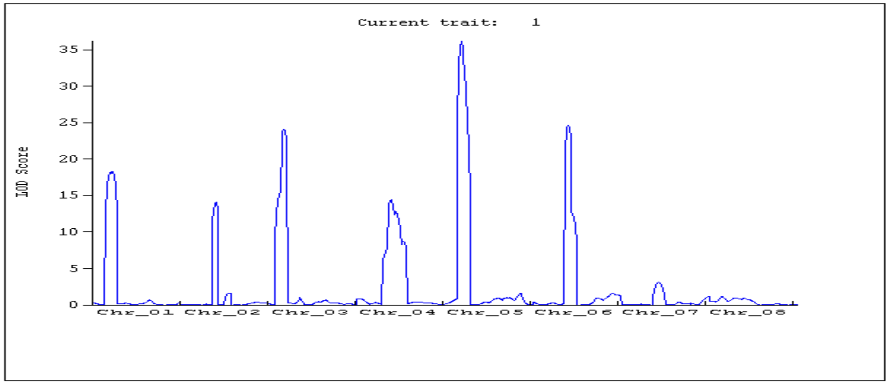
Effect of missing markers

(First simulated F_2 population from QTL distribution model I and population size 500)

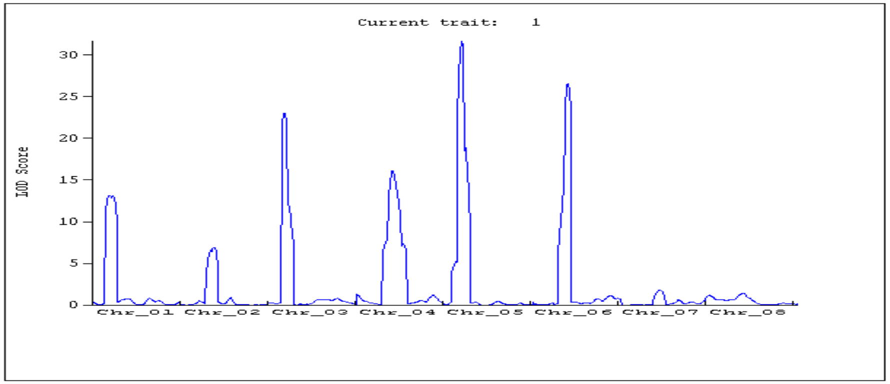
No missing markers



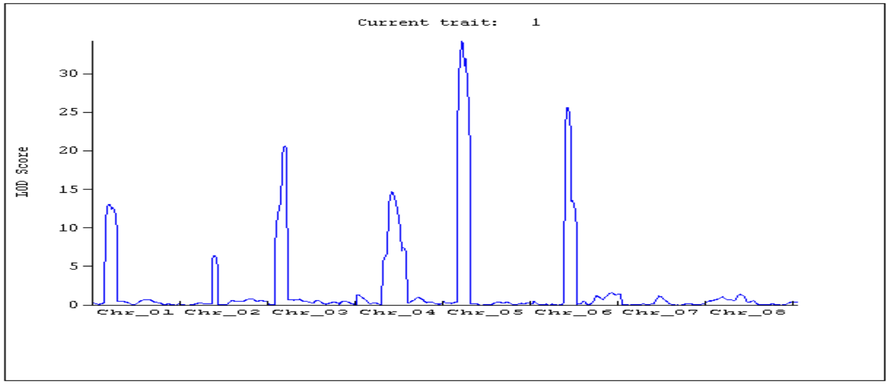
5% of missing



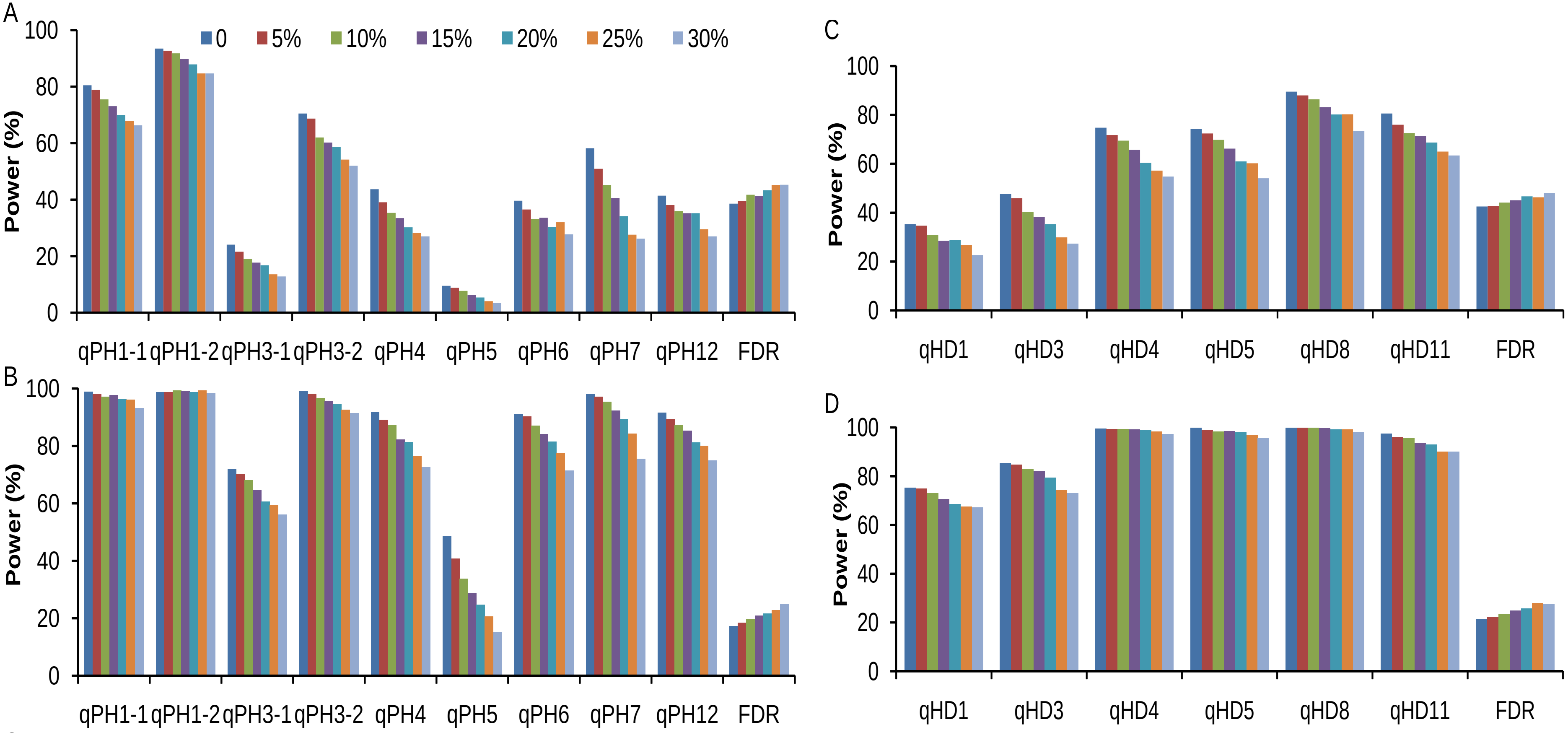
10% of missing



15% of missing



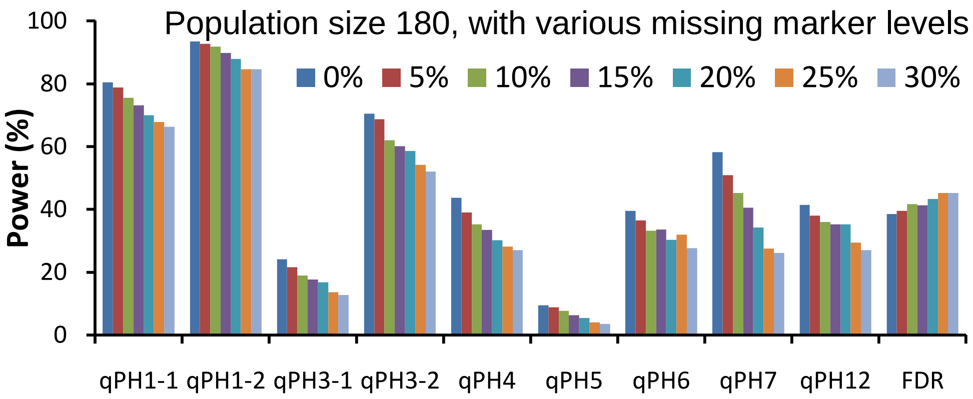
Power analysis of various levels of missing markers

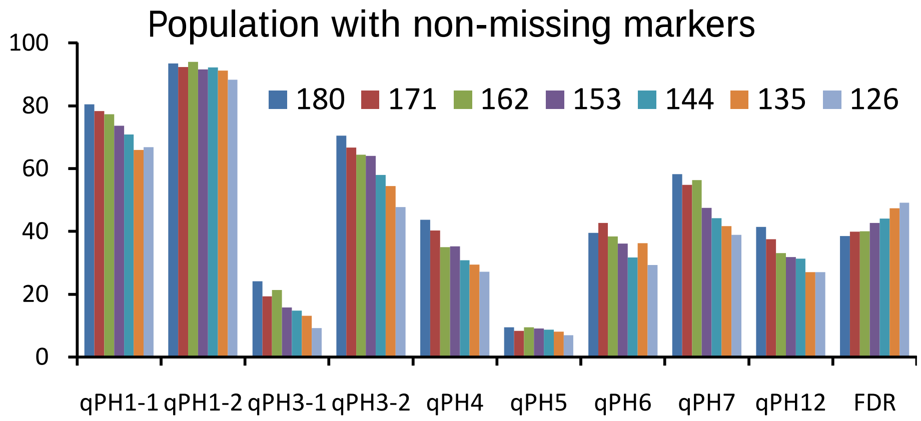


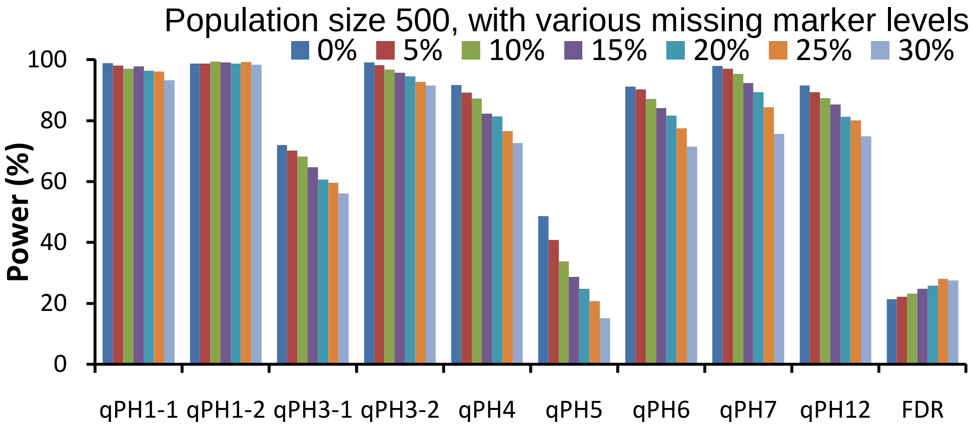
A,QTL for plant height, population size=180
B,QTL for plant height, population size=500

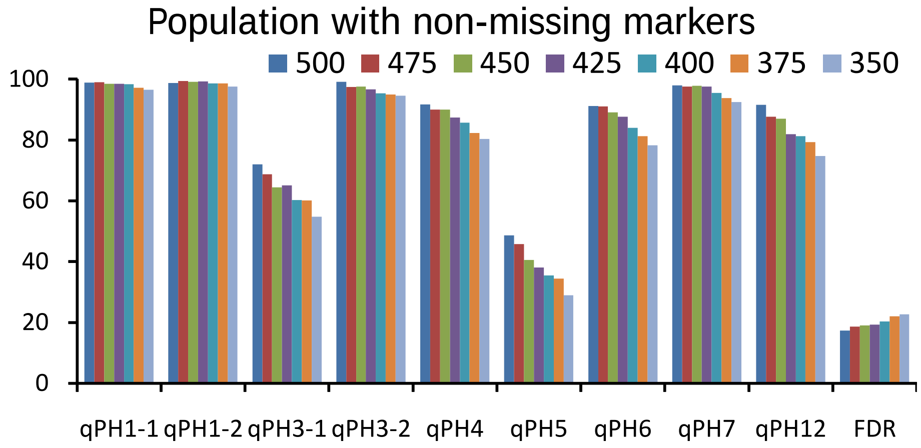
C, QTL for heading days, population size=180
D, QTL for heading days, population size=500

Effect of missing markers is similar to the reduction in population size


Population size 180, with various missing marker levels


Population with non-missing markers


Population size 500, with various missing marker levels


Population with non-missing markers

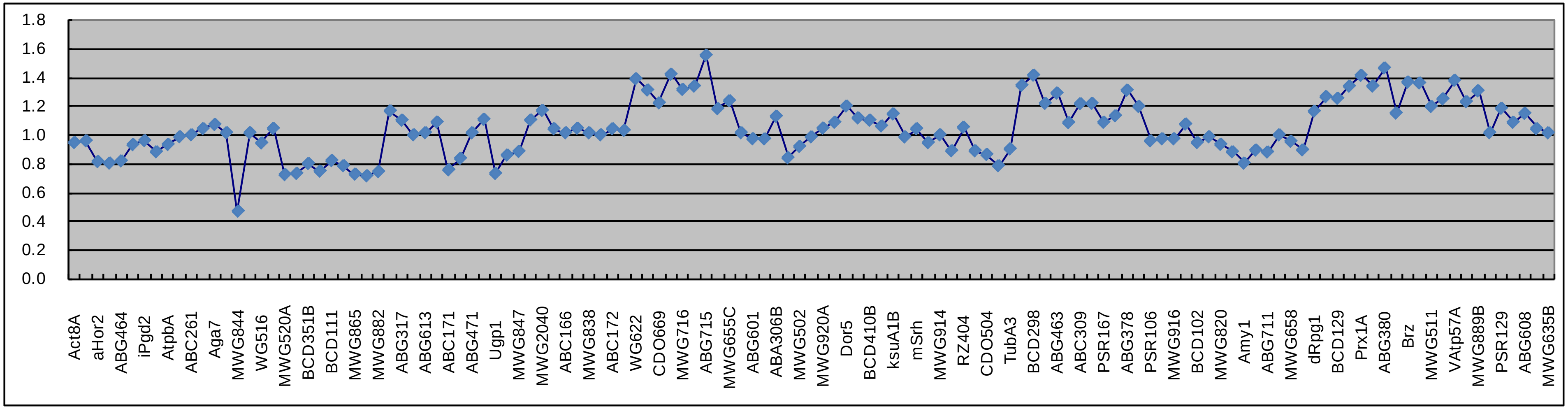
Q12: What is the effect of segregation distortion?

Zhang, L., S. Wang, H. Li, Q. Deng, A. Zheng, S. Li, P. Li, Z. Li, **J. Wang***. 2010. Effects of missing marker and segregation distortion on QTL mapping in F_2 populations. **Theor. Appl. Genet.** 121:1071-1082.

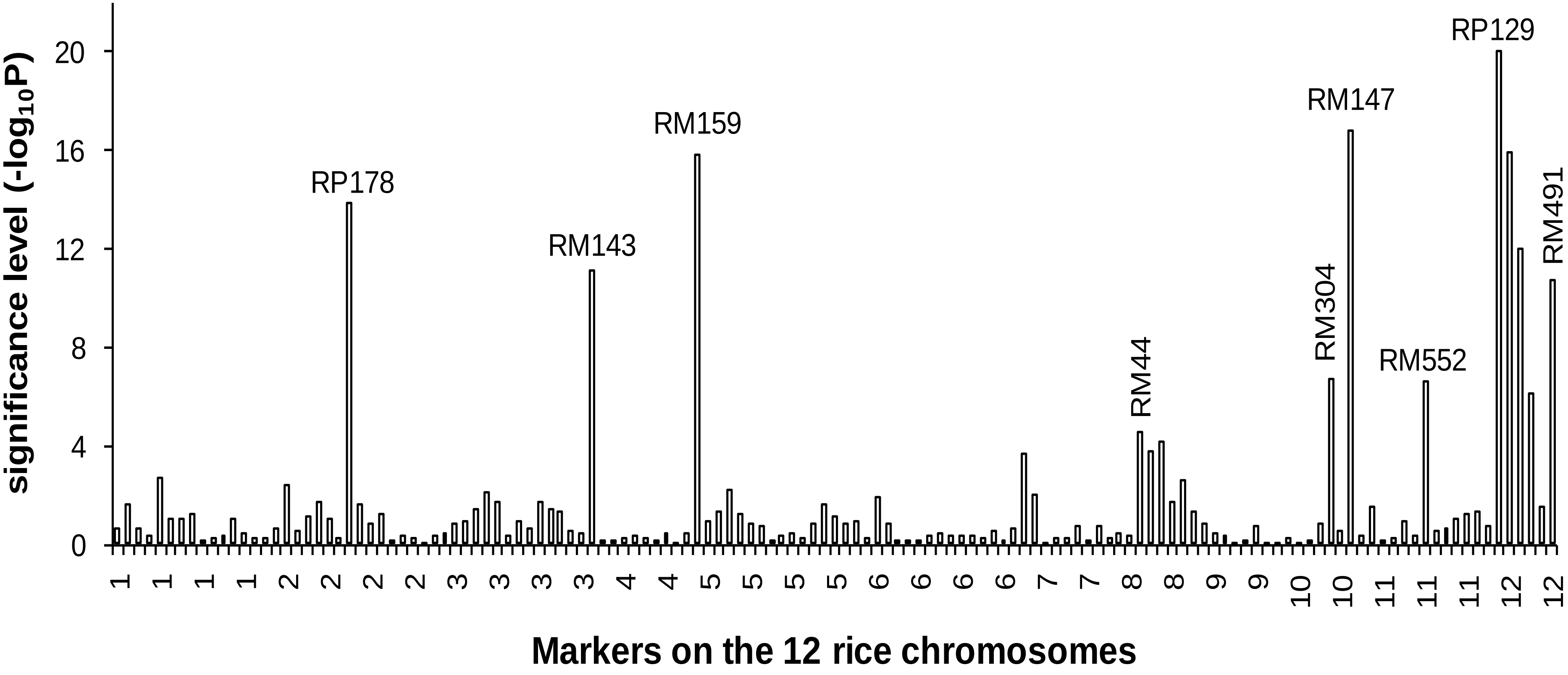
Segregation distortion

- P1 (AA) X P2 (aa), no distortion
  - P1BC1: AA:Aa=1:1
  - P2BC1: Aa:aa=1:1
  - F2: AA:Aa:aa=1:2:1
  - DH, RIL: AA:aa=1:1
- Reasons for distortion
  - Random drift
  - Selection in gametes and zygotes

Ratio of AA:aa in a barley DH population

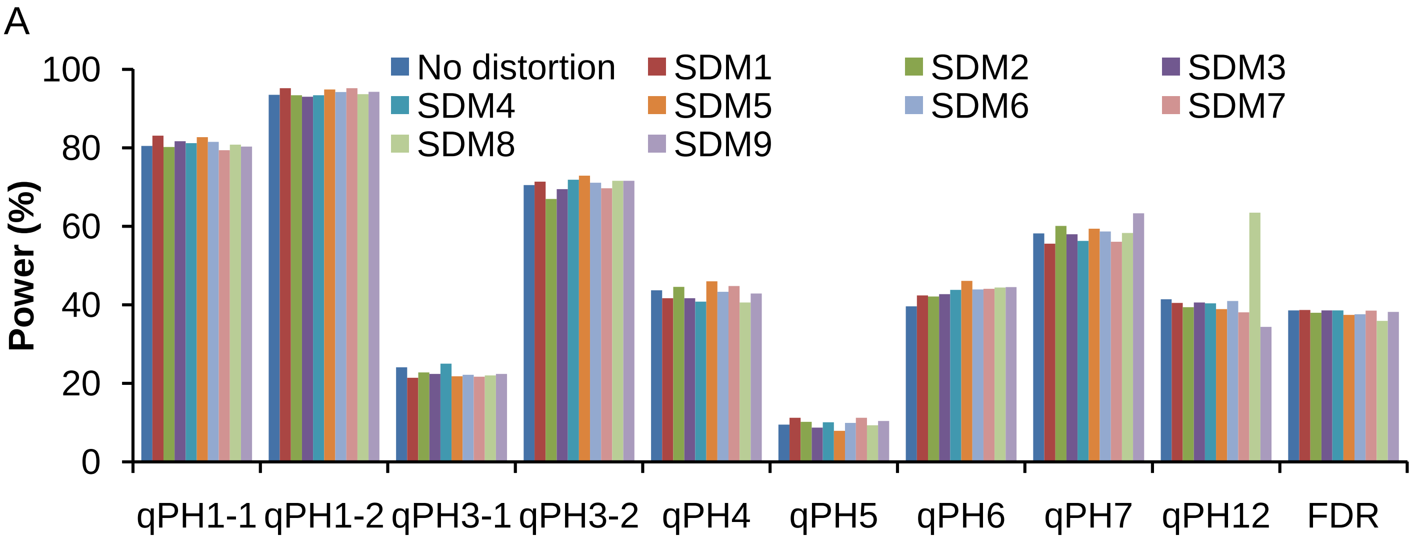


Segregation distortion in an actual rice F2 population

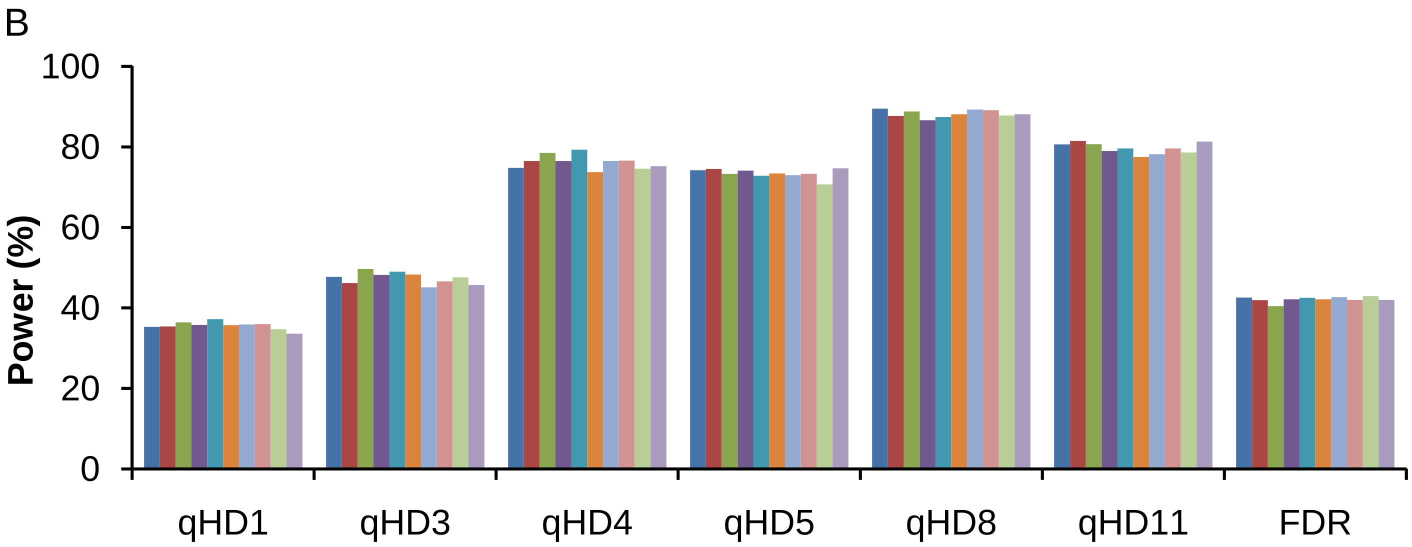


When segregation distortion markers are not linked with QTL

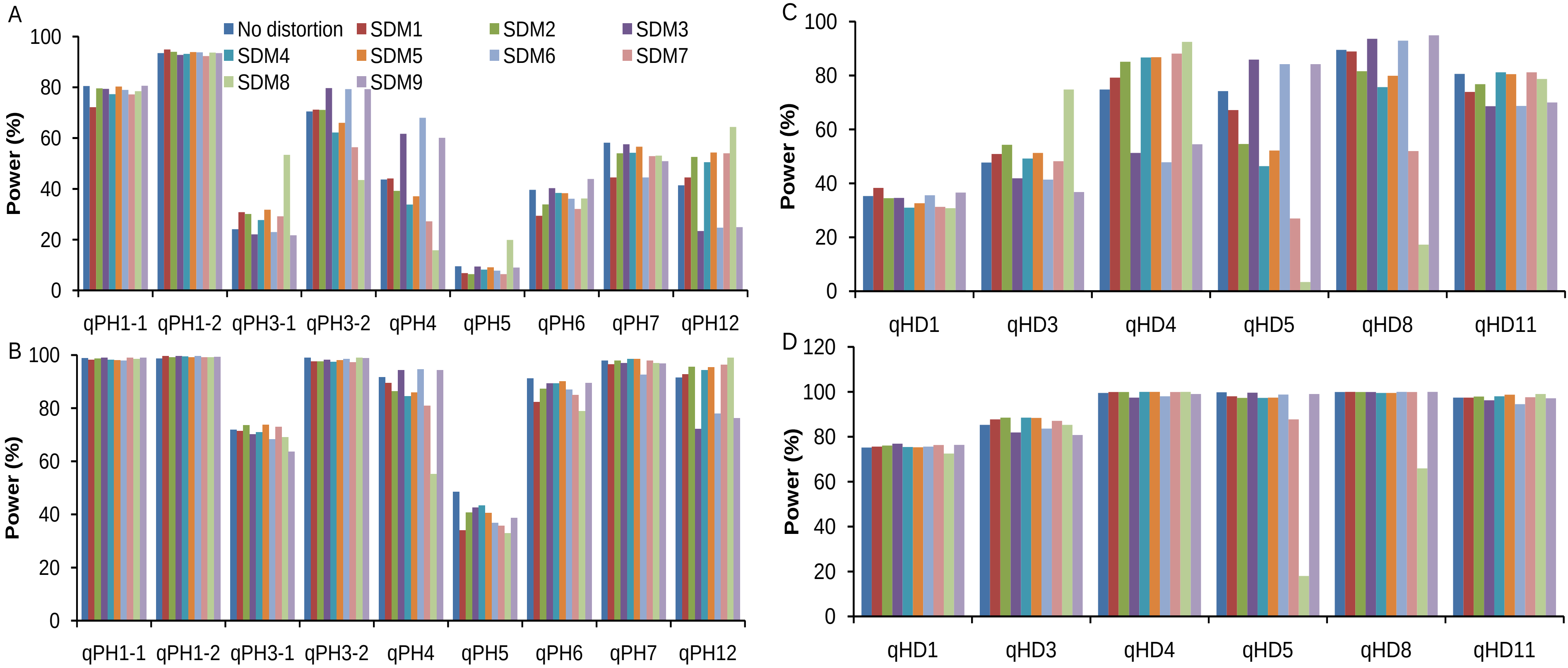
A,QTL for plant height, population size=180



B,QTL for heading days, population size=180



When segregation distortion markers are linked with QTL



A,QTL for plant height, population size=180
B,QTL for plant height, population size=500

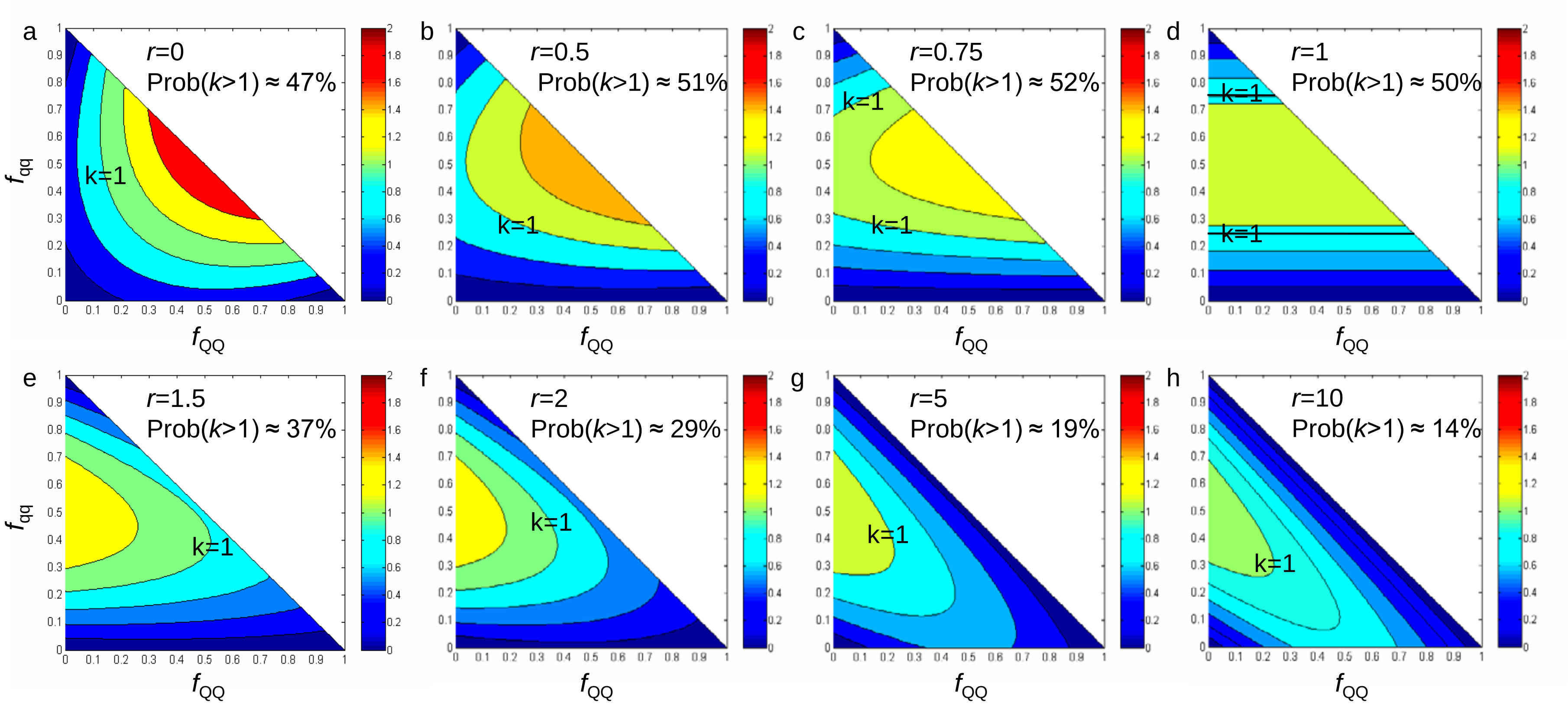
C, QTL for heading days, population size=180
D, QTL for heading days, population size=500

Effect of segregation distortion markers (SDM) on QTL mapping

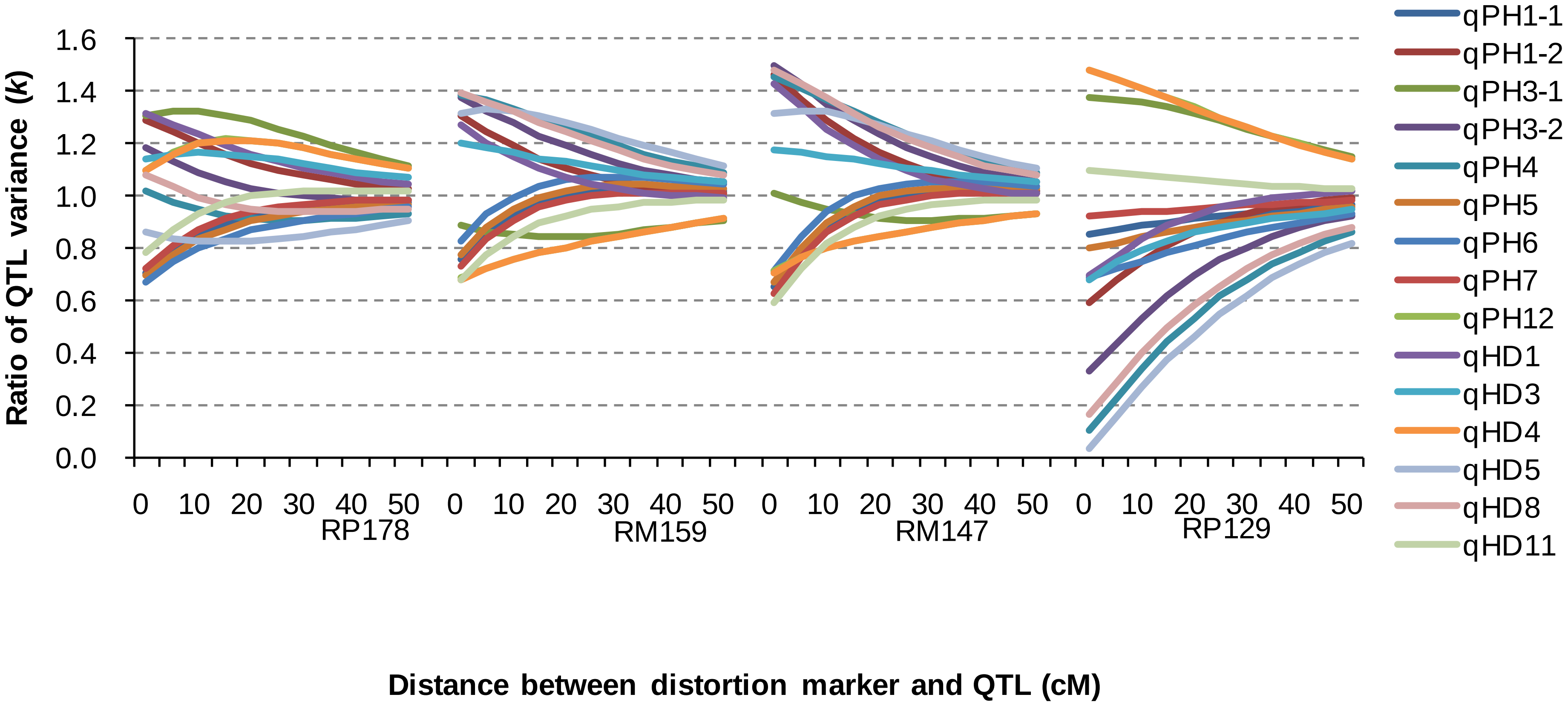
- If the SDM is not closely linked with any plant height or heading date QTL, no significant effects were observed on the detection power.
- Otherwise, SDM may increase or decrease the QTL detection power.
- In large-size populations, say size of 500, the effect of SDM was minor even the SDM was closely linked with QTL.

Genetic variance determines the effect of segregation distortion!

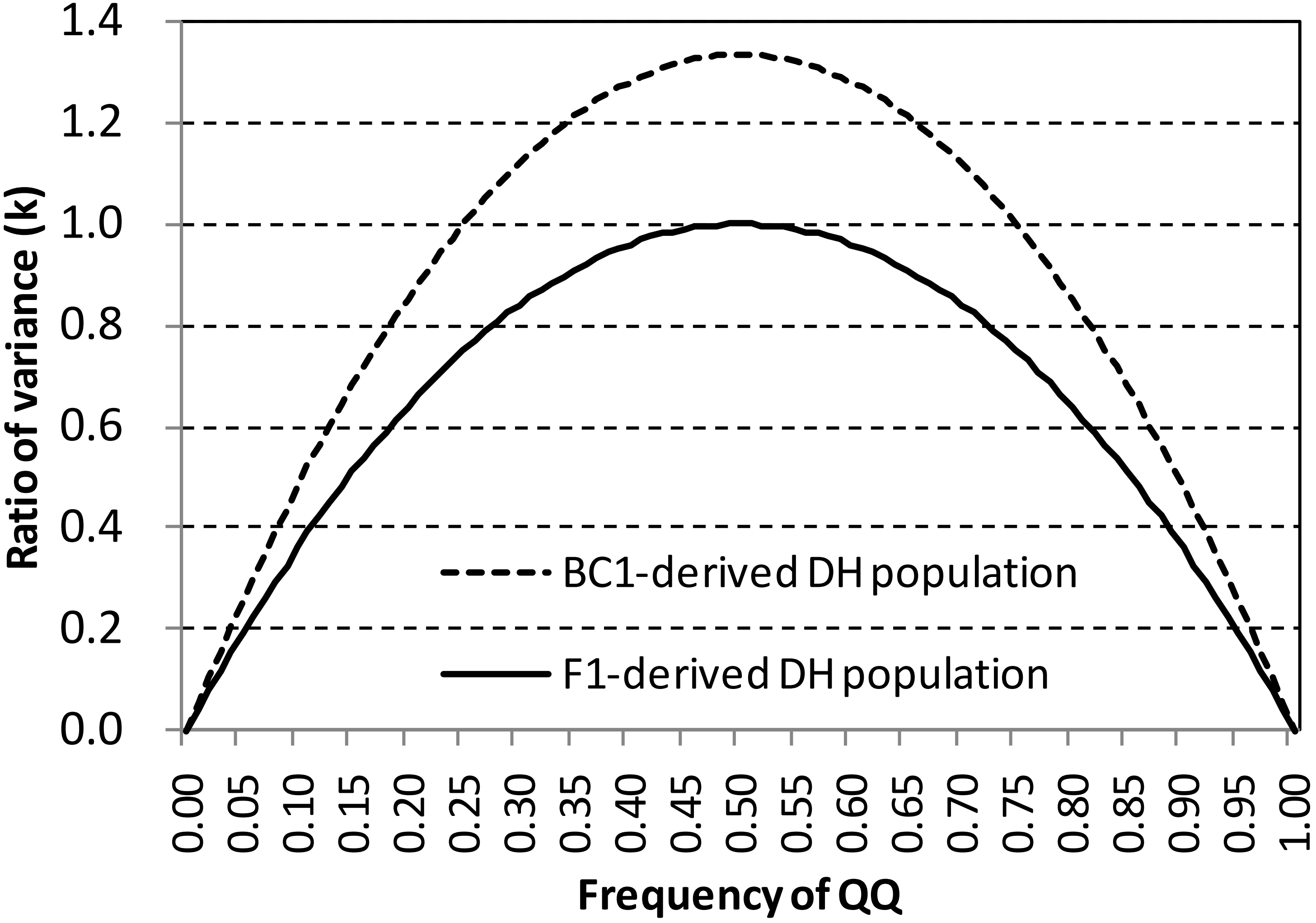
$$V_G = [f_2 + f_0 - (f_2 - f_0)^2]a^2 - 2f_1(f_2 - f_0)ad + (f_1 - f_1^2)d^2$$



How far can one SDL affect?



In F1 and BC derived DH populations



Do you have more question?

Please add.