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

LOD Threshold and QTL Detection Power Simulation

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

Outlines

- Hypothesis testing and two types of associated error
- LOD threshold in QTL mapping
- QTL detection power simulation
- Avoid the over fitting problem in ICIM

Hypothesis testing and two types of associated error

Hypothesis testing

- A hypothesis is a statement that something is true.
- Null hypothesis: A hypothesis to be tested. We use the symbol H_0 to represent the null hypothesis
- Alternative hypothesis: A hypothesis to be considered as an alternative to the null hypothesis. We use the symbol H_a to represent the alternative hypothesis.
- The alternative hypothesis is the one believe to be true, or what you are trying to prove is true.

Two types of error

- We may make mistakes in the test.
- **Type I error:** reject the null hypothesis when it is true.
- Probability of type I error is denoted by α
- **Type II error:** accept the null hypothesis when it is wrong.
- Probability of type II error is denoted by β

Power of a statistical test

- *P*(reject the null hypothesis when it is false)=1-β
- (1-α) is the probability we accept the null when it was in fact true
- **(1-β) is the probability we reject when the null is in fact false - this is the power of the test.**
- The power changes depending on what the actual population parameter is.

Factors affecting power

- For example: H_0: $\mu = \mu_0$, H_a: $\mu > \mu_0$
- Test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

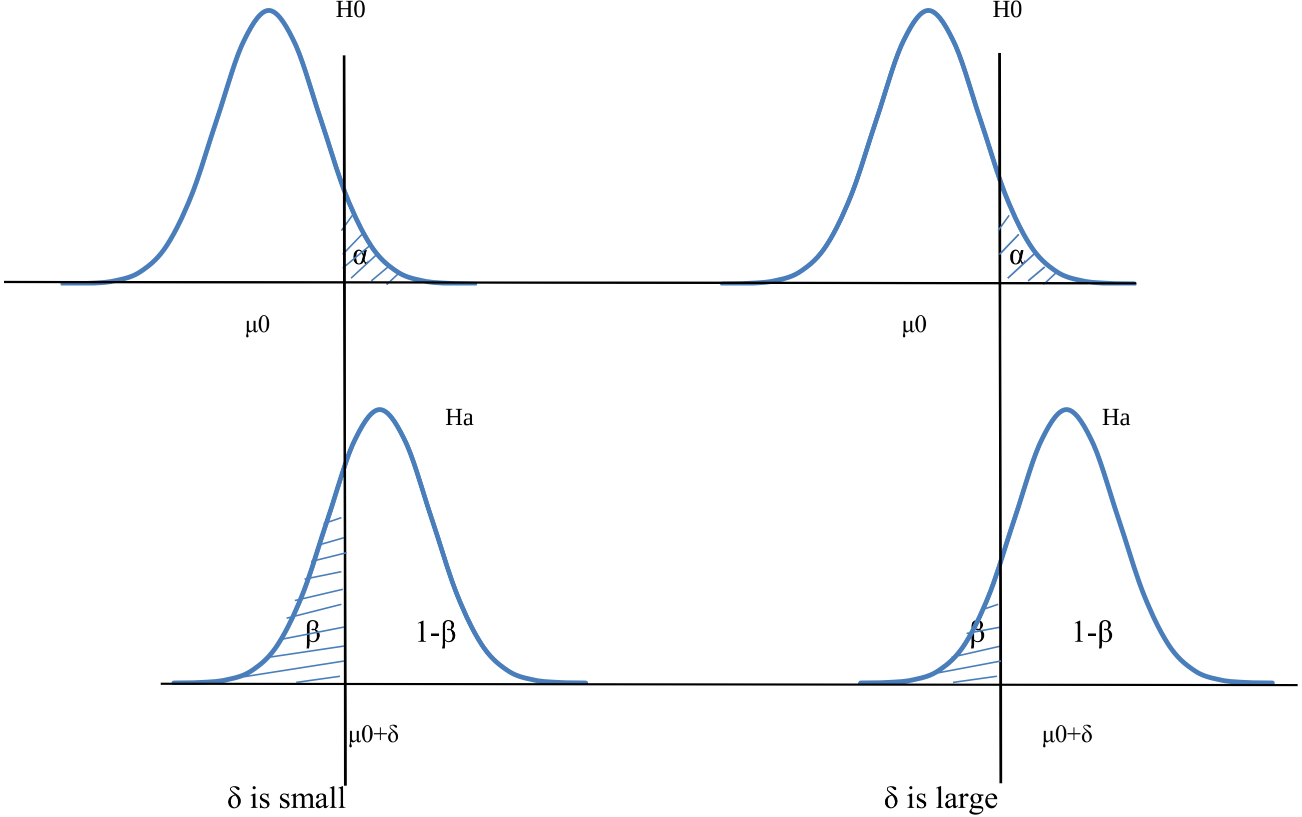
- If we want to reject H_0, we need

$$\frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \geq Z_\alpha$$

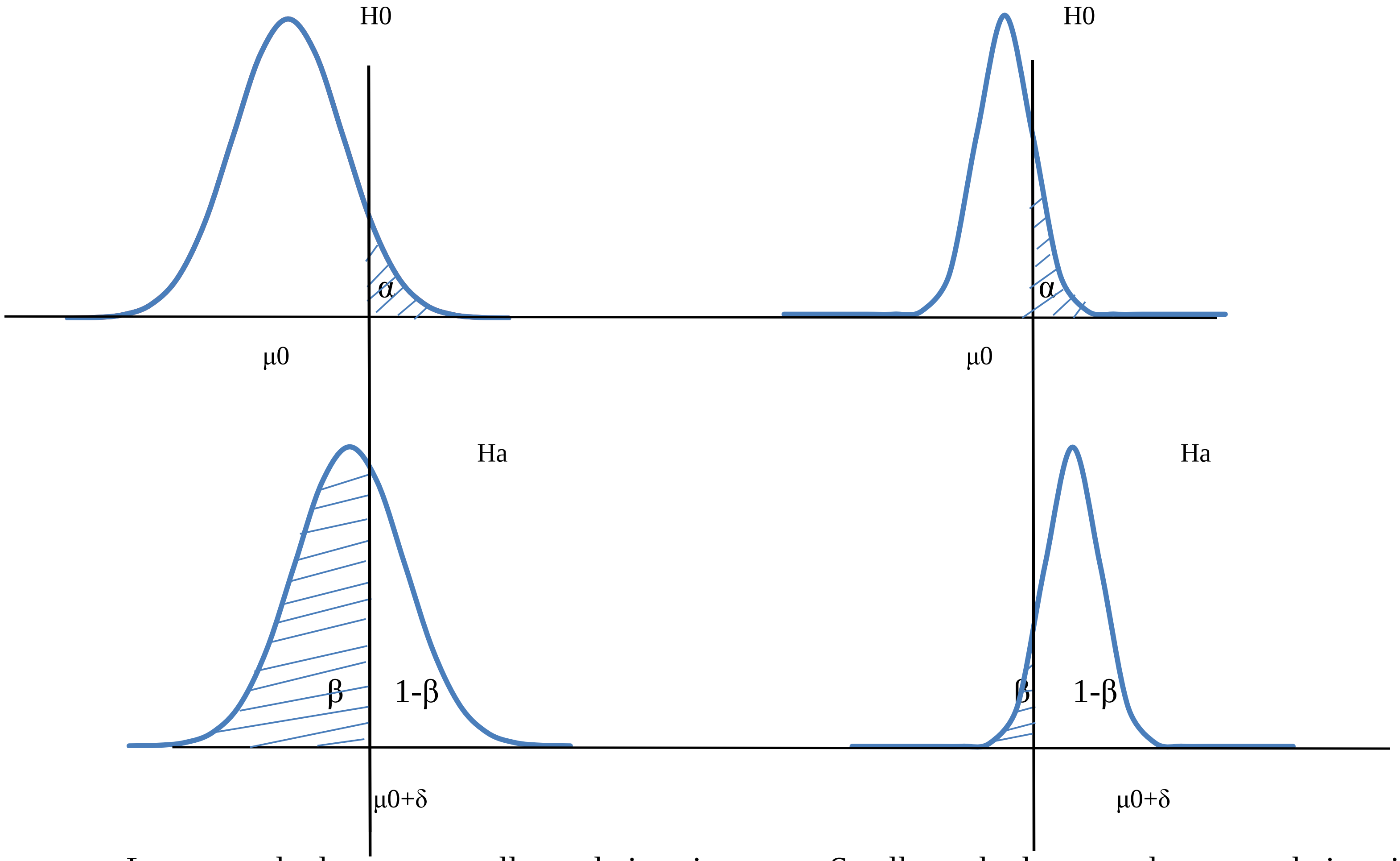
- So the power depends on $\delta = \overline{x} - \mu_0$, σ, n, and α

The larger the difference δ is, the higher the power is

- $\overline{X} \sim N(\mu, \sigma^2/n)$
- If H_0 is true, $\overline{X} \sim N(\mu_0, \sigma^2/n)$
- If H_a is true, $\overline{X} \sim N(\mu_0+\delta, \sigma^2/n)$



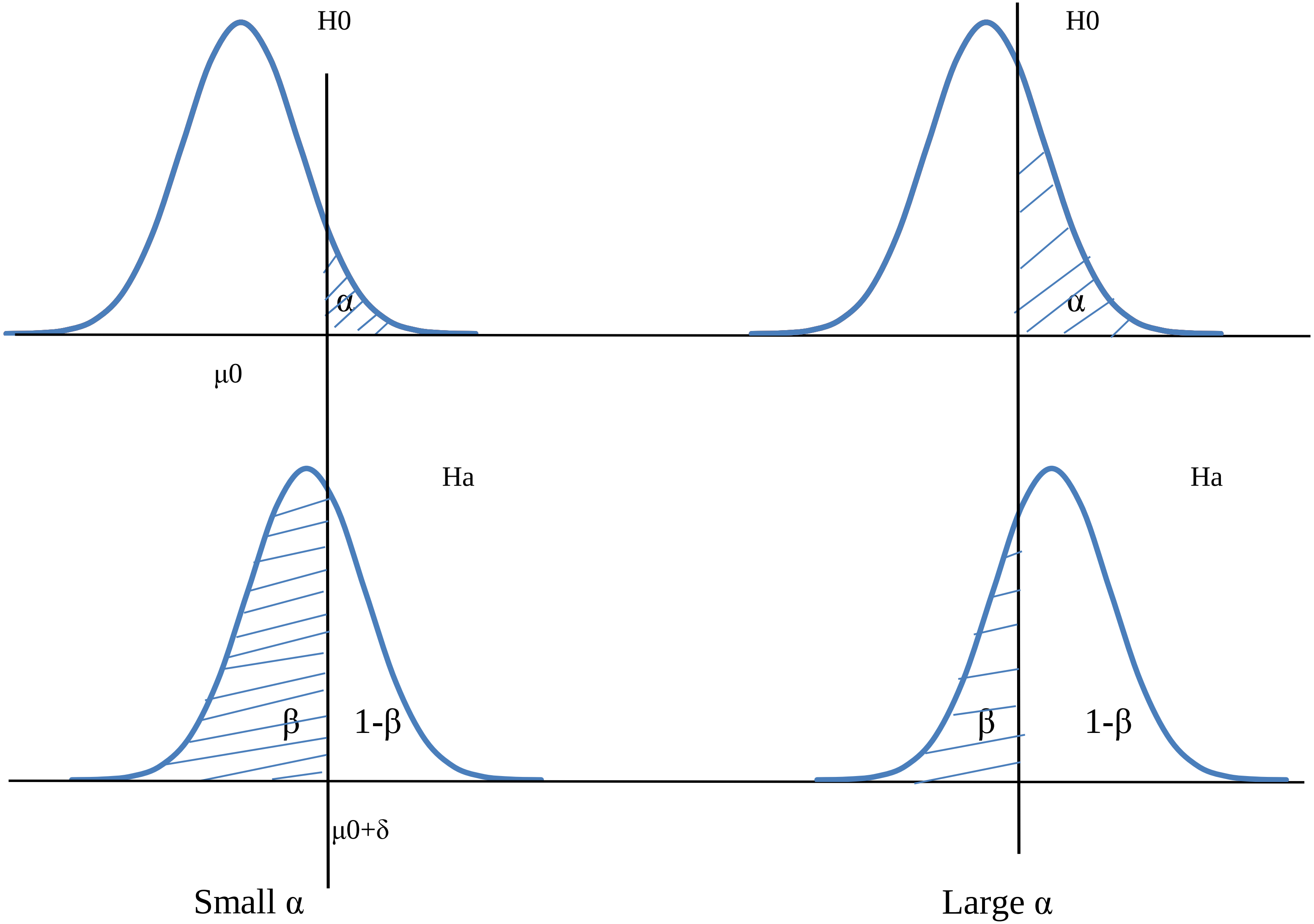
The smaller the standard error or the larger the population size is, the higher the power is



Large standard error or small population size

Small standard error or large population size

The larger α is, the higher the power is



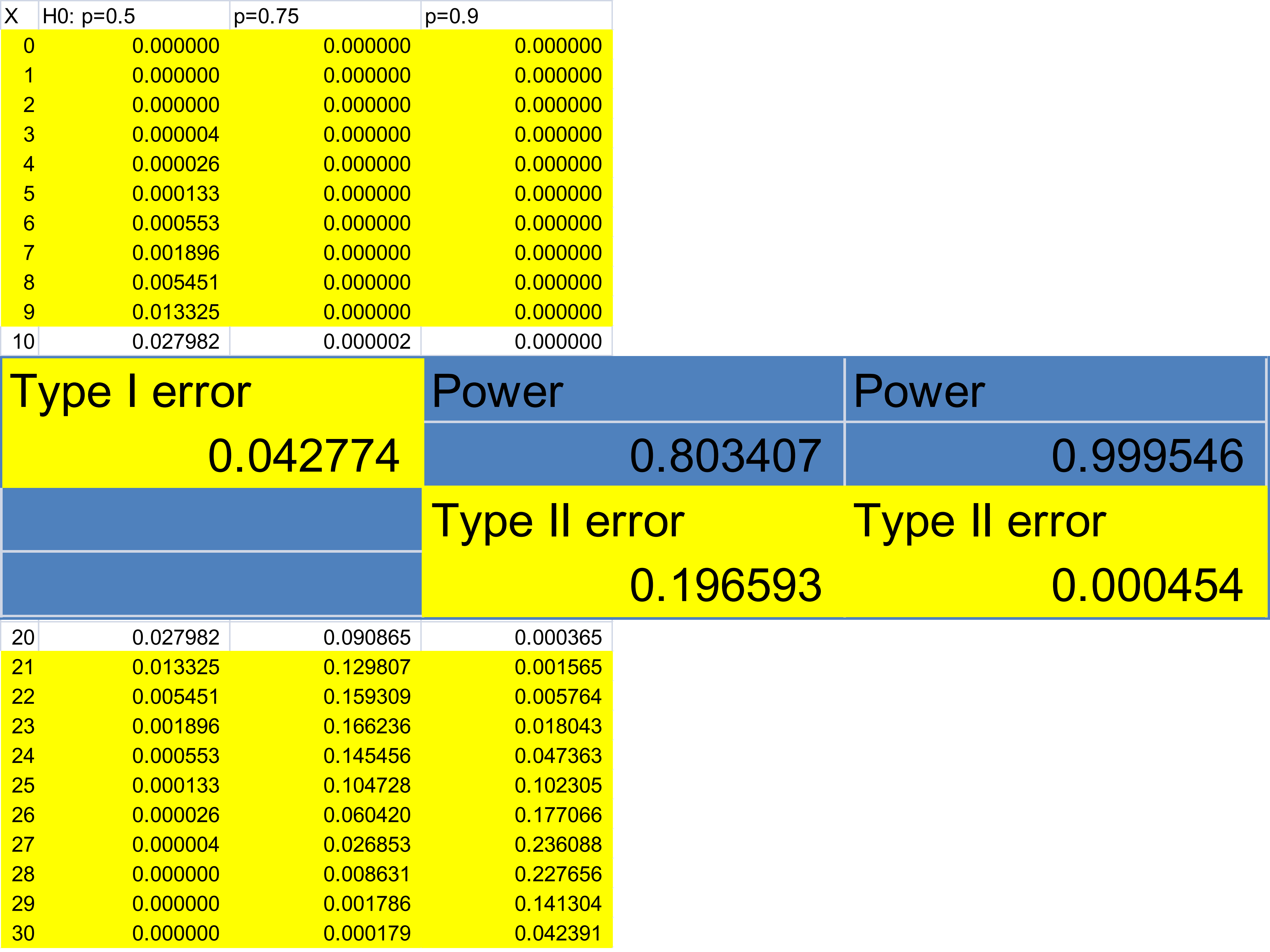
Let’s find some Type I and II errors

- Consider the Binomial distribution
- H0: $p=0.5$; Ha: $p\neq0.5$; $n=6$, $X\sim B(n=6, p)$
- Reject H0 when $X=0$, or 6
- Accept H0 when $X=1, \dots, 5$
- $\alpha = P(X=0|p=0.5) + P(X=6|p=0.5) = 0.0156+0.0156=0.0312$
- Given $p=0.75$, find the Type II error β, and the statistical power
  - $\beta= P(1<=X<=5|p=0.75)=0.8218$; Power=0.1782
- Given $p=0.90$, find the Type II error β, and the statistical power
  - $\beta= P(1<=X<=5|p=0.90)=0.4686$; Power=0.5314

	X~Binomial (n, p), n=6		
	{0, 5} is the rejection region, i.e. EstP=0.0 or 1.0		
	{1,2,3,4} is the acceptance region		
X	H0: p=0.5	p=0.75	p=0.9
0	0.015625	0.000244	0.000001
1	0.093750	0.004395	0.000054
2	0.234375	0.032959	0.001215
3	0.312500	0.131836	0.014580
4	0.234375	0.296631	0.098415
5	0.093750	0.355957	0.354294
6	0.015625	0.177979	0.531441
	Type I error	Power	Power
	0.031250	0.178223	0.531442
		Type II error	Type II error
		0.821777	0.468558

Let’s find some Type I and II errors

- Binomial distribution
- H0: $p=0.5$; Ha: $p\neq 0.5$; $n=30$, $X \sim B(30, p)$
- Reject H0 when $X<=9$, or $>=21$
- Accept H0 when $X=10, \ldots, 20$
- Find α
- Given $p=0.75$, find the Type II error β, and the statistical power
- Given $p=0.90$, find the Type II error β, and the statistical power



X	H0: p=0.5	p=0.75	p=0.9
0	0.000000	0.000000	0.000000
1	0.000000	0.000000	0.000000
2	0.000000	0.000000	0.000000
3	0.000004	0.000000	0.000000
4	0.000026	0.000000	0.000000
5	0.000133	0.000000	0.000000
6	0.000553	0.000000	0.000000
7	0.001896	0.000000	0.000000
8	0.005451	0.000000	0.000000
9	0.013325	0.000000	0.000000
10	0.027982	0.000002	0.000000

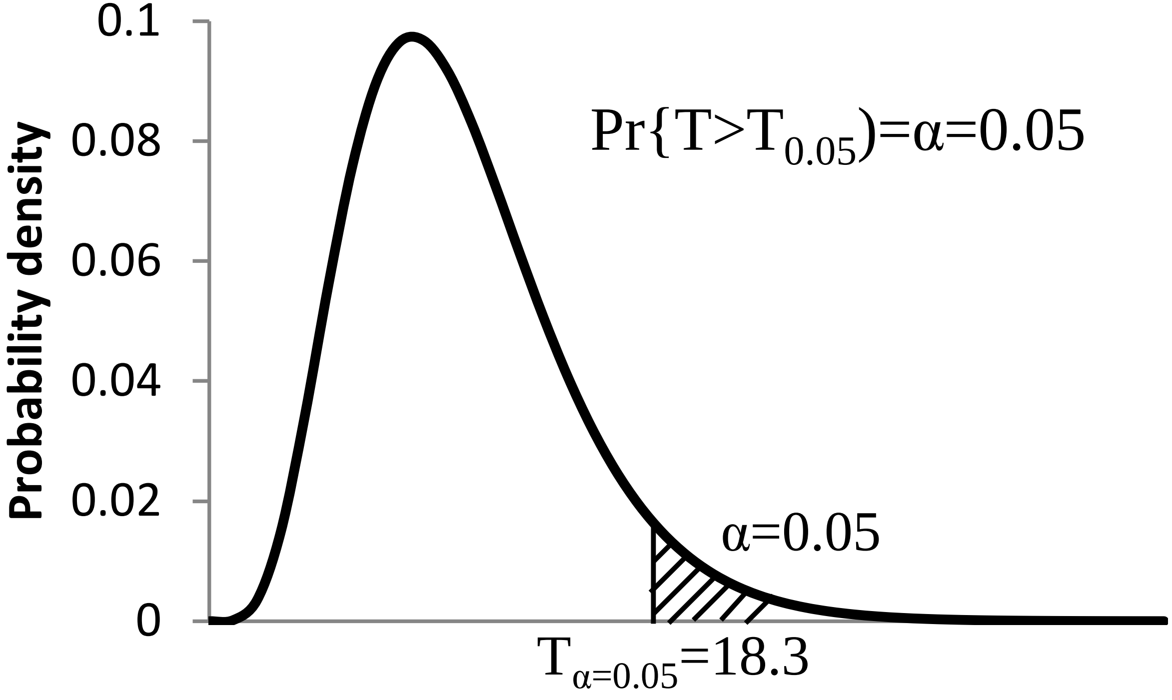
Type I error	Power	Power
0.042774	0.803407	0.999546
	Type II error	Type II error
	0.196593	0.000454

X	H0: p=0.5	p=0.75	p=0.9
20	0.027982	0.090865	0.000365
21	0.013325	0.129807	0.001565
22	0.005451	0.159309	0.005764
23	0.001896	0.166236	0.018043
24	0.000553	0.145456	0.047363
25	0.000133	0.104728	0.102305
26	0.000026	0.060420	0.177066
27	0.000004	0.026853	0.236088
28	0.000000	0.008631	0.227656
29	0.000000	0.001786	0.141304
30	0.000000	0.000179	0.042391

LOD threshold in QTL mapping

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)

Threshold is used to control Type I error, say no greater than 0.05

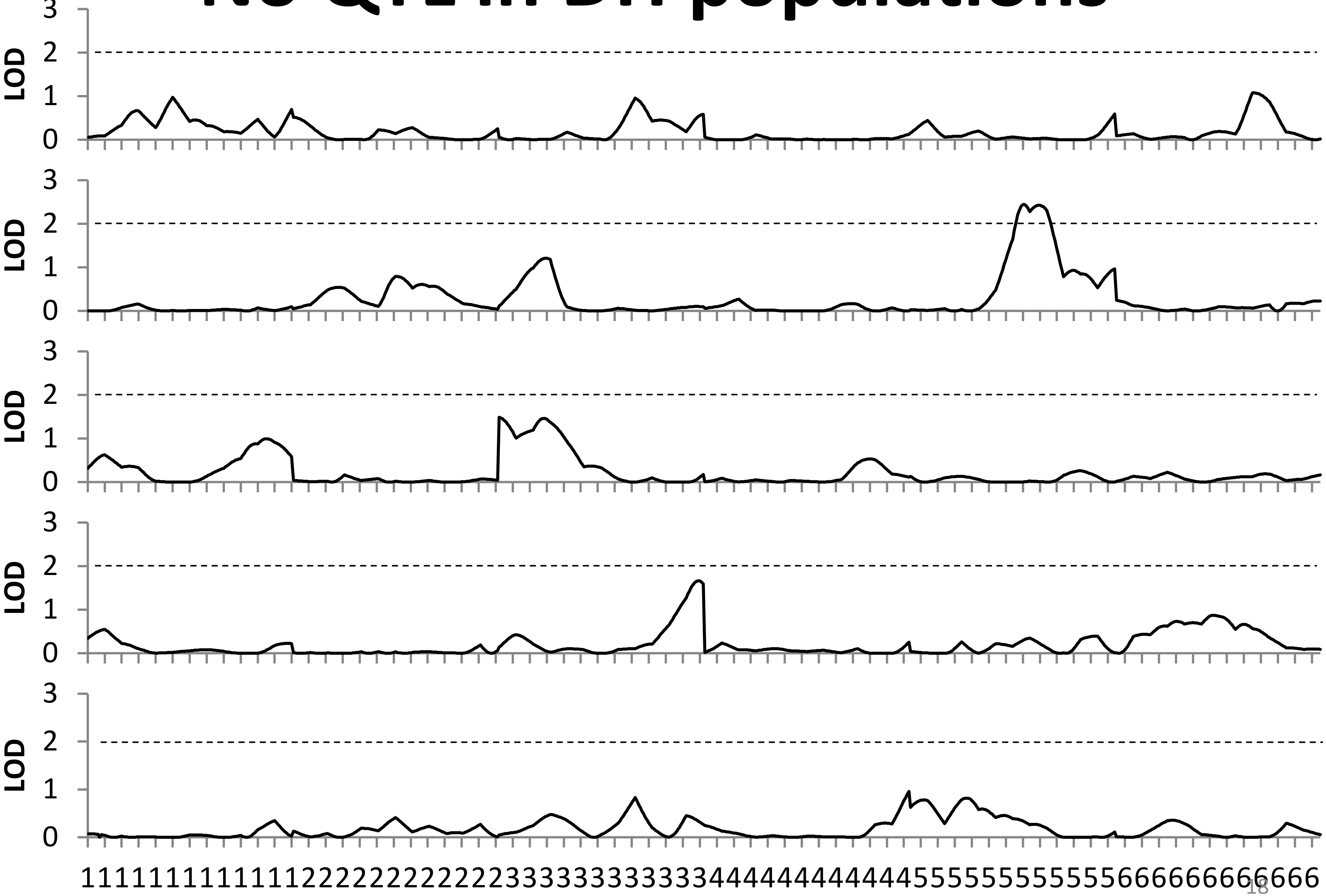


- Say we know a test under H_0 hypothesis has the χ^2(df=10) distribution, the use of threshold 18.3 can make sure the Type I error < 0.05

The reason to control Type I error

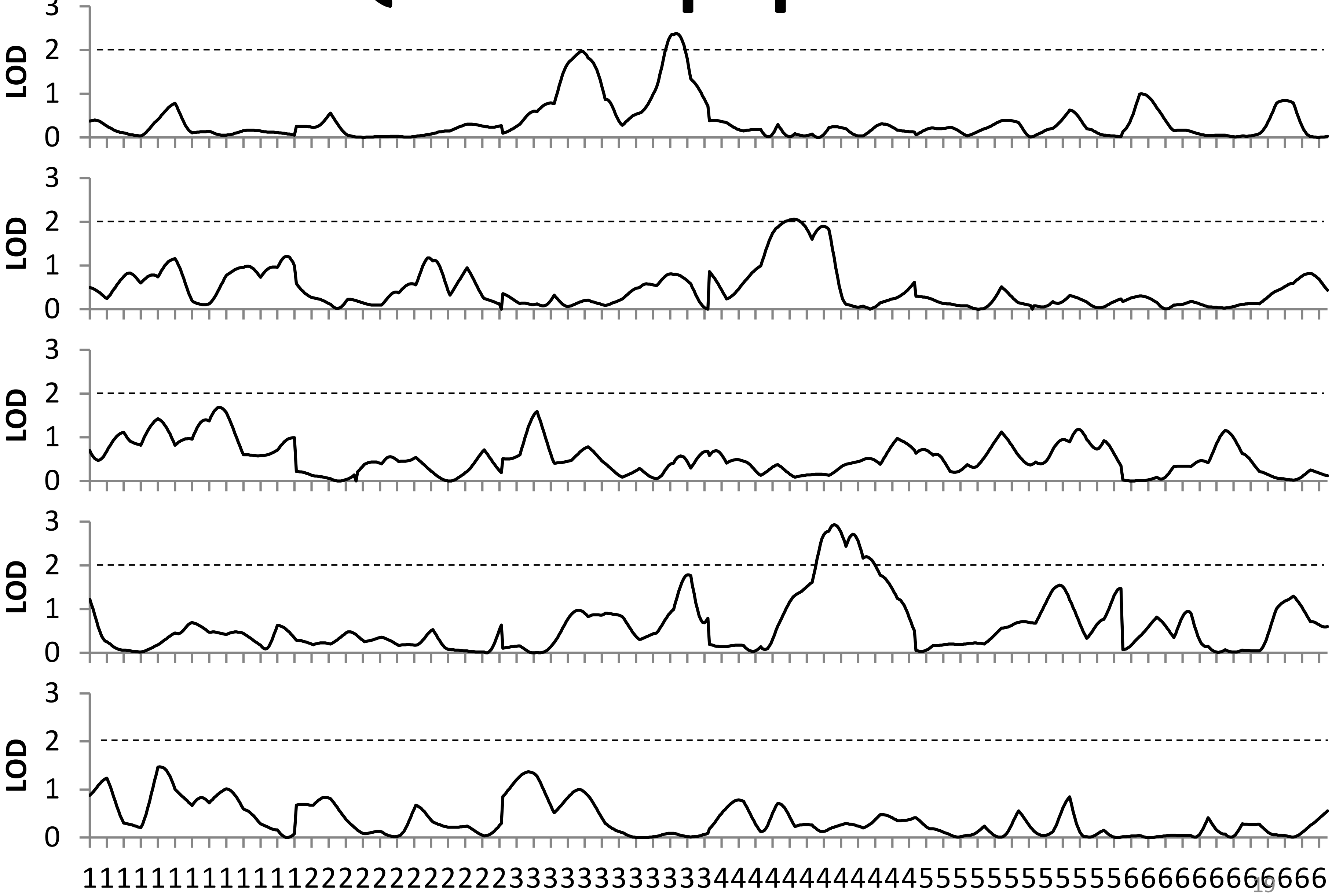
- High LOD score can be simply caused by chance!
- We simulated five DH populations and five F2 population. But we did not assume any QTL on the six chromosomes

No QTL in DH populations



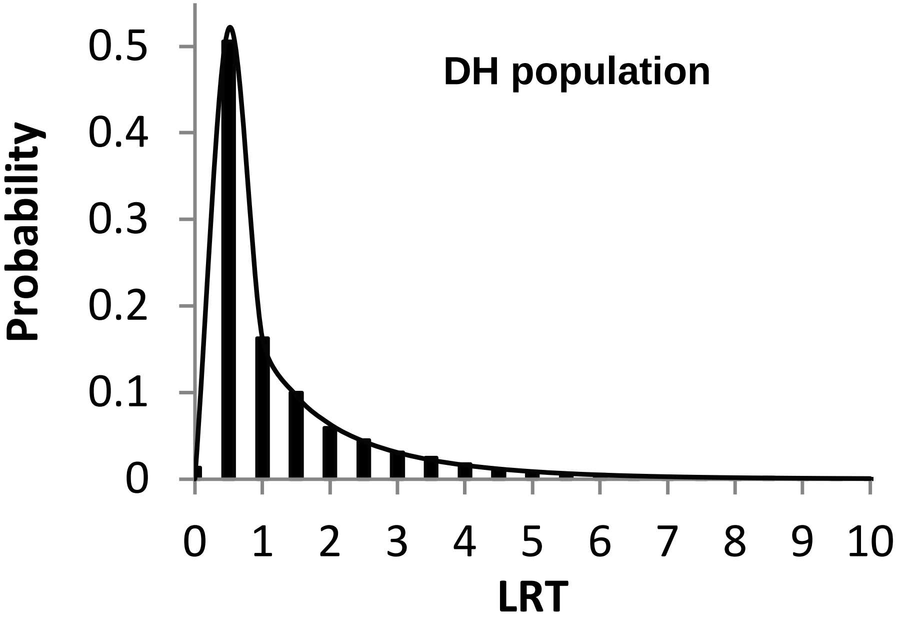
No QTL on the six chromosomes, DH population

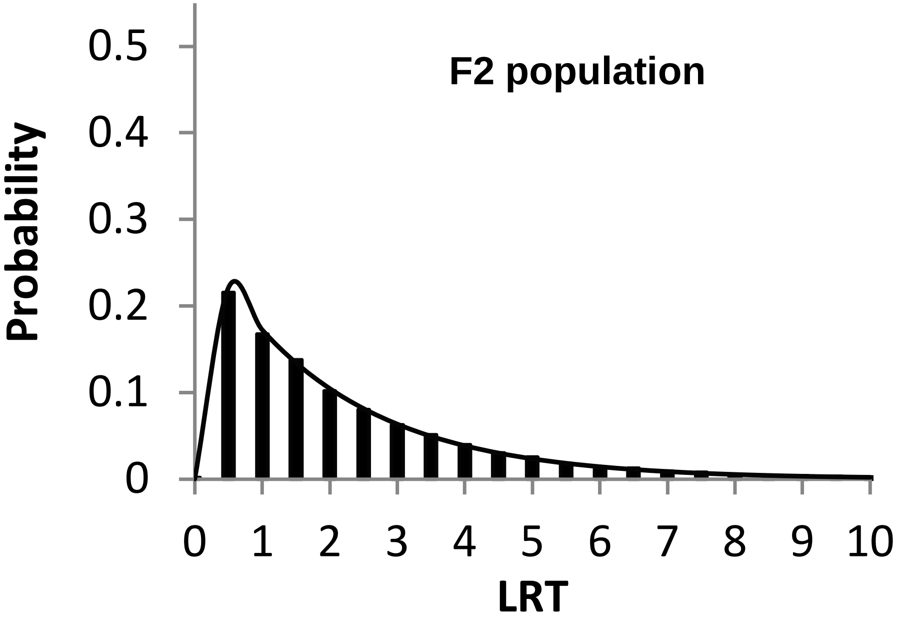
No QTL in F2 populations



No QTL on the six chromosomes, F2 population

Distribution of LRT under H_0 at each scanning position

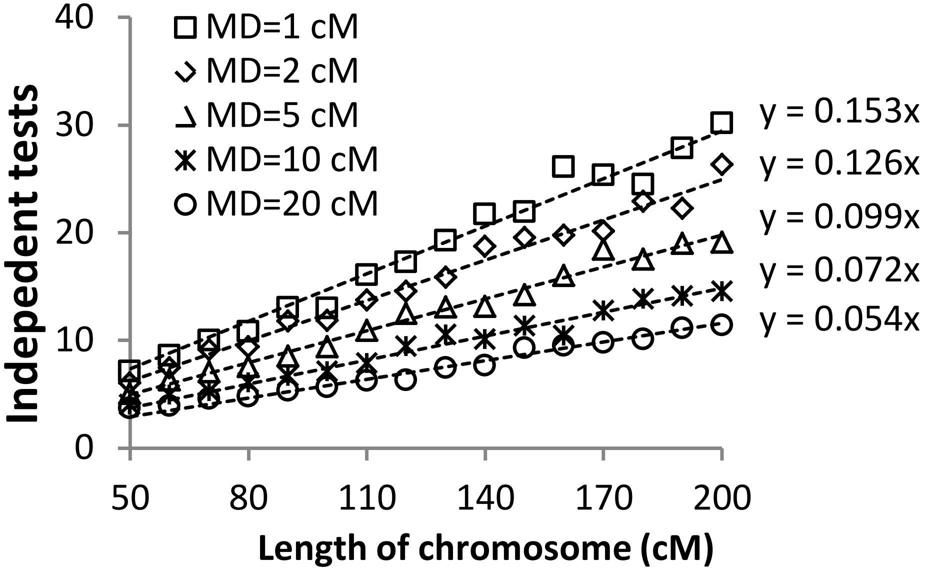




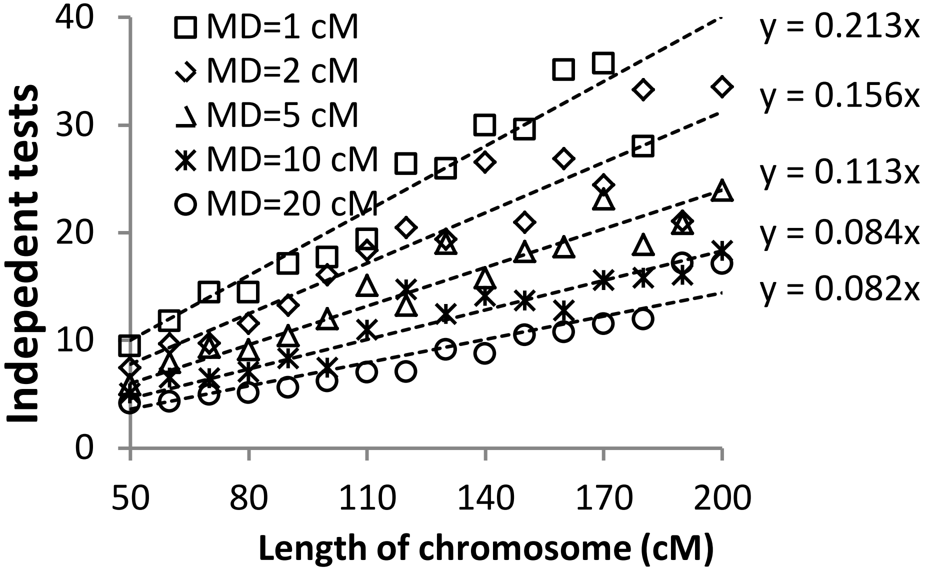
- In DH populations, LRT ~ χ^2(df=1)
- In F2 populations, LRT ~ χ^2(df=2)
- D.F. is equal to the number of independent genetic effects to be estimated

Number of independent tests

DH population, genome-wide Type I error = 0.05



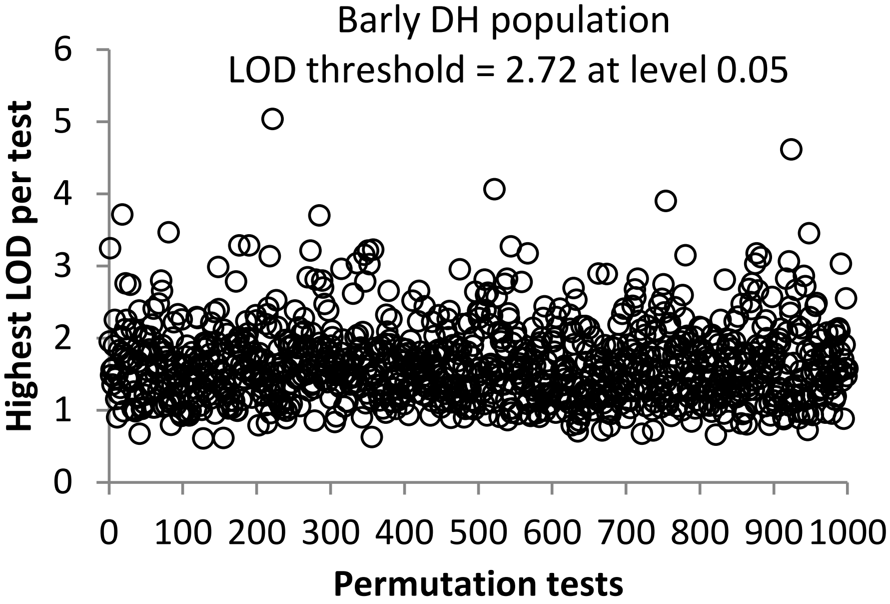
DH population, genome-wide Type I error = 0.01

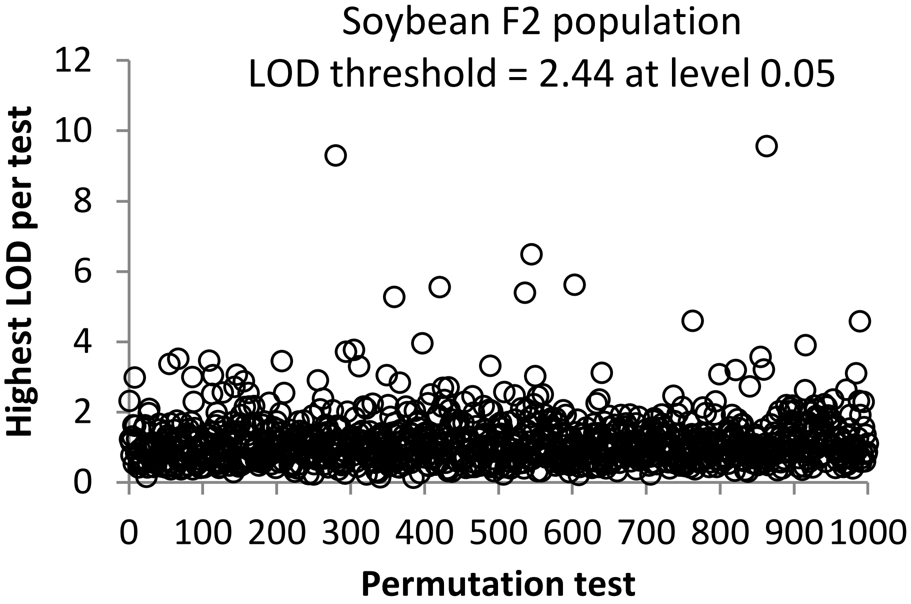


LOD threshold, assuming marker density is 1 cM

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


Barly DH population
LOD threshold = 2.72 at level 0.05


Soybean F2 population
LOD threshold = 2.44 at level 0.05

QTL detection power simulation

Zhang, L., H. Li, **J. Wang***. 2012. Statistical power of inclusive composite interval mapping in detecting digenic epistasis showing common F2 segregation ratios. **Journal of Integrative Plant Biology** 54: 270-279

Li, H., S. Hearne, M. Bänziger, Z. Li, and **J. Wang*.** 2010. Statistical properties of QTL linkage mapping in biparental genetic populations. **Heredity** 105: 257-267.

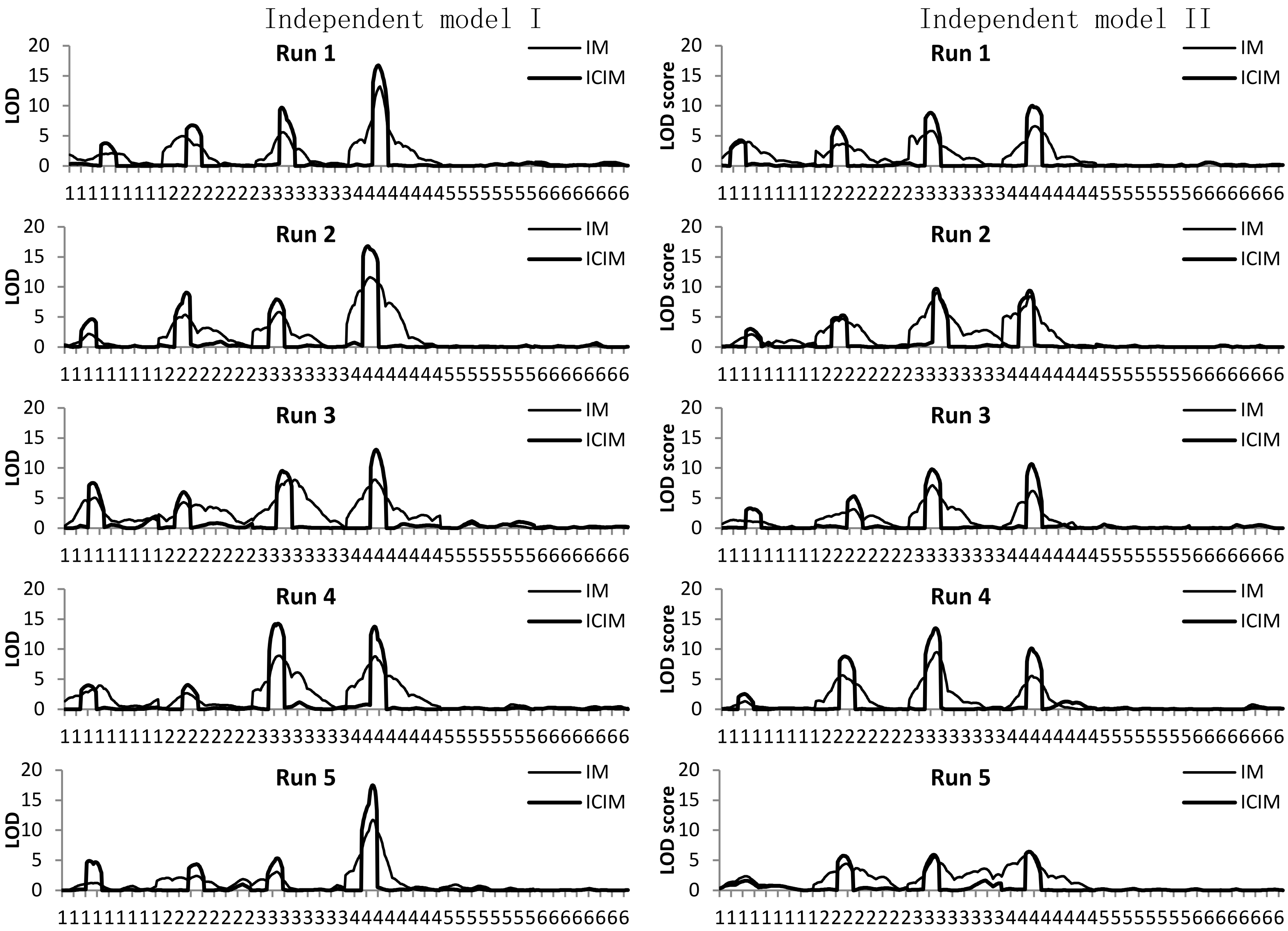
Two independent QTL models

	Chromosome	Position (cM)	Additive	PVE (%)
Independent model I				
Q1	1	35	0.316	5.0
Q2	2	35	**0.447**	10.0
Q3	3	35	0.548	15.0
Q4	4	35	**0.633**	20.0
Genetic variance	1.000			
Error variance	1.000	Heritability	0.500	
Independent model II				
Q1	1	35	0.316	5.0
Q2	2	35	**-0.447**	10.0
Q3	3	35	0.548	15.0
Q4	4	35	**-0.633**	20.0
Genetic variance	1.000			
Error variance	1.000	Heritability	0.500	

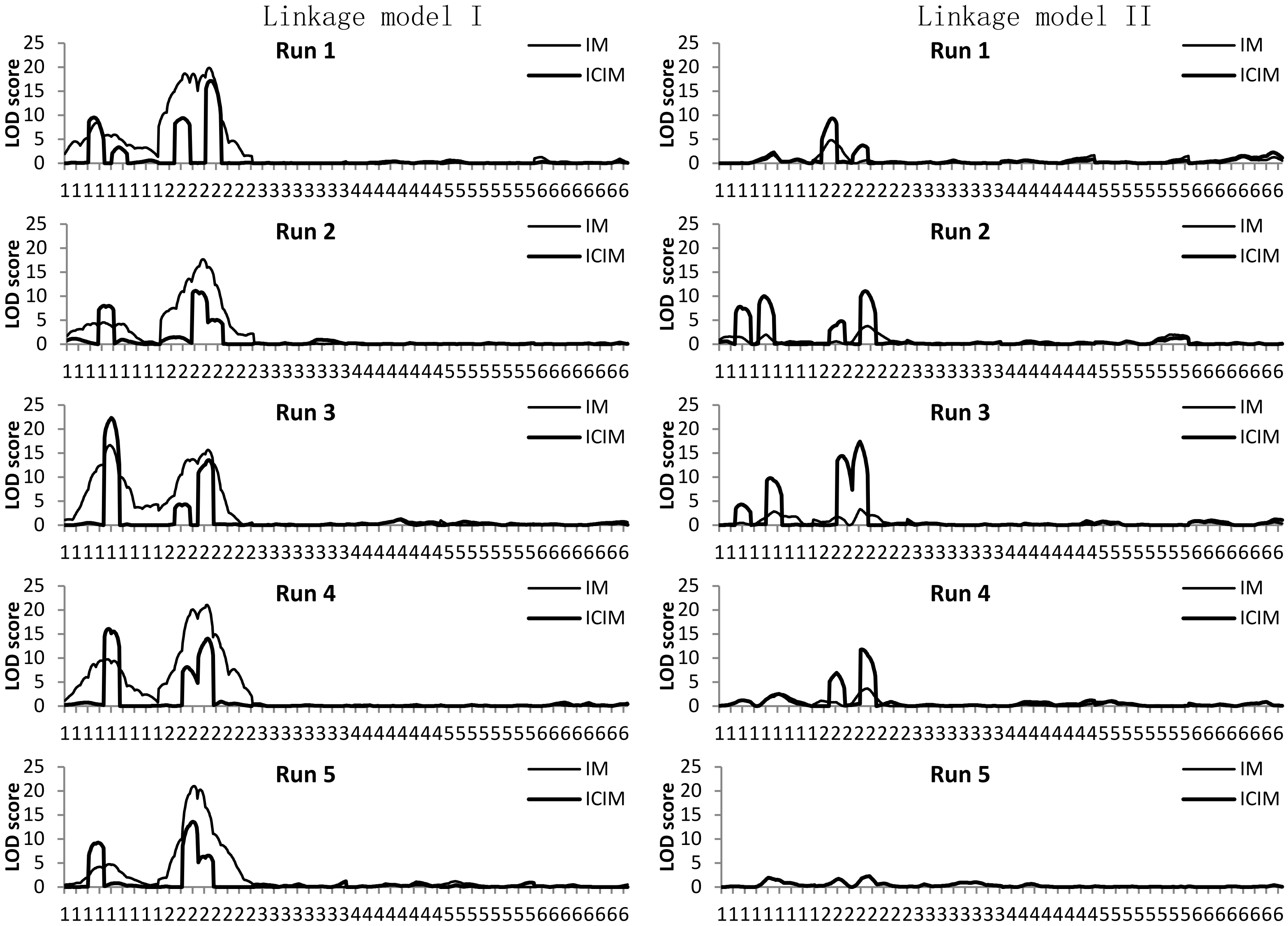
Two linked QTL models

	Chromosome	Position (cM)	Additive	PVE (%)
Linkage model I				
Q1	1	35	0.316	3.9
Q2	1	65	**0.447**	7.9
Q3	2	35	0.548	11.8
Q4	2	65	**0.633**	15.8
Genetic variance	1.535			
Error variance	1.000	Heritability	0.606	
Linkage model II				
Q1	1	35	0.316	6.8
Q2	1	65	**-0.447**	13.7
Q3	2	35	0.548	20.5
Q4	2	65	**-0.633**	27.3
Genetic variance	0.465			
Error variance	1.000	Heritability	0.317	

QTL mapping in 5 simulation runs of the two independent models



QTL mapping in 5 simulation runs of the two linkage models



Count of power and false QTL for IM

Run	QTL identified					Support interval	
	Chrom.	Position	LOD	PVE (%)	Additive	10 cM	10 cM
1	2	25	4.97	11.44	0.503	False	Q2
	3	35	5.61	13.35	0.541	Q3	Q3
	4	40	13.21	26.22	0.761	Q4	Q4
2	2	34	5.36	13.01	0.509	Q2	Q2
	3	34	5.82	13.72	0.521	Q3	Q3
	4	30	11.59	23.43	0.682	Q4	Q4
3	1	39	5.05	11.22	0.508	Q1	Q1
	2	32	4.30	10.09	0.482	Q2	Q2
	3	54	8.03	18.42	0.651	False	False
	4	36	8.06	18.55	0.653	Q4	Q4
4	1	45	3.97	10.21	0.420	False	Q1
	2	36	2.69	6.81	0.343	Q2	Q2
	3	34	8.92	19.66	0.583	Q3	Q3
	4	36	8.79	20.15	0.591	Q4	Q4
5	3	33	3.08	8.16	0.389	Q3	Q3
	4	35	11.71	26.65	0.701	Q4	Q4

Count of power and false QTL for ICIM

Run	QTL identified					Support interval	
	Chrom.	Position	LOD	PVE (%)	Additive	10 cM	10 cM
1	1	47	3.80	5.06	0.335	False	False
	2	38	6.79	9.11	0.448	Q2	Q2
	3	33	9.70	13.81	0.551	Q3	Q3
	4	38	16.72	25.50	0.753	Q4	Q4
2	1	35	4.65	6.26	0.352	Q1	Q1
	2	36	9.07	12.56	0.500	Q2	Q2
	3	31	7.93	10.41	0.454	Q3	Q3
	4	27	16.77	24.93	0.703	False	Q4
3	1	36	7.52	10.23	0.486	Q1	Q1
	2	32	6.00	8.10	0.432	Q2	Q2
	3	38	9.52	13.63	0.560	Q3	Q3
	4	38	13.05	19.18	0.664	Q4	Q4
4	1	30	3.99	5.13	0.298	Q1	Q1
	2	37	4.04	5.89	0.319	Q2	Q2
	3	33	14.21	21.68	0.613	Q3	Q3
	4	36	13.73	21.23	0.607	Q4	Q4
5	1	35	4.91	8.04	0.384	Q1	Q1
	2	51	4.35	6.87	0.356	False	False
	3	34	5.35	9.45	0.419	Q3	Q3
	4	35	17.46	31.65	0.764	Q4	Q4

Power and false QTL from the 5 runs

Method	QTL	Times to be detected		Detection power (%)	
		10cM	20cM	10cM	20cM
IM	Q1	1	2	20	40
	Q2	3	4	60	80
	Q3	4	4	80	80
	Q4	5	5	100	100
	False QTL	3	1	19	6
ICIM	Q1	4	4	80	80
	Q2	4	4	80	80
	Q3	5	5	100	100
	Q4	4	5	80	100
	False QTL	3	2	15	10

The best method

- Has the highest power
- Has the lowest false discovery rate

Independent model I

Method	QTL	Power (%)	Pos. (cM)	SE	LOD	SE	Additive	SE
IM	Q1	25.8	35.182	3.461	3.849	1.003	0.421	0.056
	Q2	62.6	34.861	3.014	5.084	1.692	0.480	0.082
	Q3	77.7	35.006	2.669	7.013	2.178	0.560	0.092
	Q4	85.4	35.067	2.464	9.205	2.584	0.635	0.095
	FDR (%)	32.4						
ICIM	Q1	49.5	34.867	3.184	4.667	1.656	0.354	0.062
	Q2	73.9	34.874	2.769	7.156	2.295	0.450	0.077
	Q3	82.8	34.958	2.521	10.161	2.710	0.548	0.078
	Q4	89.0	35.160	2.278	13.087	3.229	0.632	0.083
	FDR (%)	22.6						

Independent model II

Method	QTL	Power (%)	Pos. (cM)	SE	LOD	SE	Additive	SE
IM	Q1	27.3	34.835	3.414	3.865	1.151	0.424	0.062
	Q2	64.2	35.062	3.035	5.006	1.607	-0.478	0.078
	Q3	78.6	34.956	2.706	6.963	2.143	0.558	0.090
	Q4	84.6	34.865	2.481	9.374	2.441	-0.640	0.089
	FDR (%)	31.1						
ICIM	Q1	49.2	34.831	3.204	4.589	1.640	0.352	0.063
	Q2	76.1	35.030	2.861	7.142	2.328	-0.448	0.076
	Q3	85.6	35.051	2.484	10.193	2.755	0.548	0.081
	Q4	90.0	34.939	2.325	13.203	3.221	-0.634	0.082
	FDR (%)	21.3						

Linkage model I

Method	QTL	Power (%)	Pos. (cM)	SE	LOD	SE	Additive	SE
IM	Q1	24.1	35.448	2.757	6.944	2.092	0.625	0.099
	Q2	49.0	64.790	2.549	7.859	2.278	0.665	0.104
	Q3	40.2	35.759	1.648	17.017	3.105	0.937	0.090
	Q4	56.5	64.165	1.624	18.682	3.359	0.971	0.093
	FDR (%)	53.1						
ICIM	Q1	26.9	35.353	3.051	7.335	3.466	0.449	0.118
	Q2	55.5	64.872	2.701	10.519	4.184	0.558	0.133
	Q3	77.0	34.952	2.618	10.560	3.890	0.559	0.113
	Q4	84.2	64.828	2.533	13.668	4.761	0.649	0.130
	FDR (%)	26.4						

Linkage model II

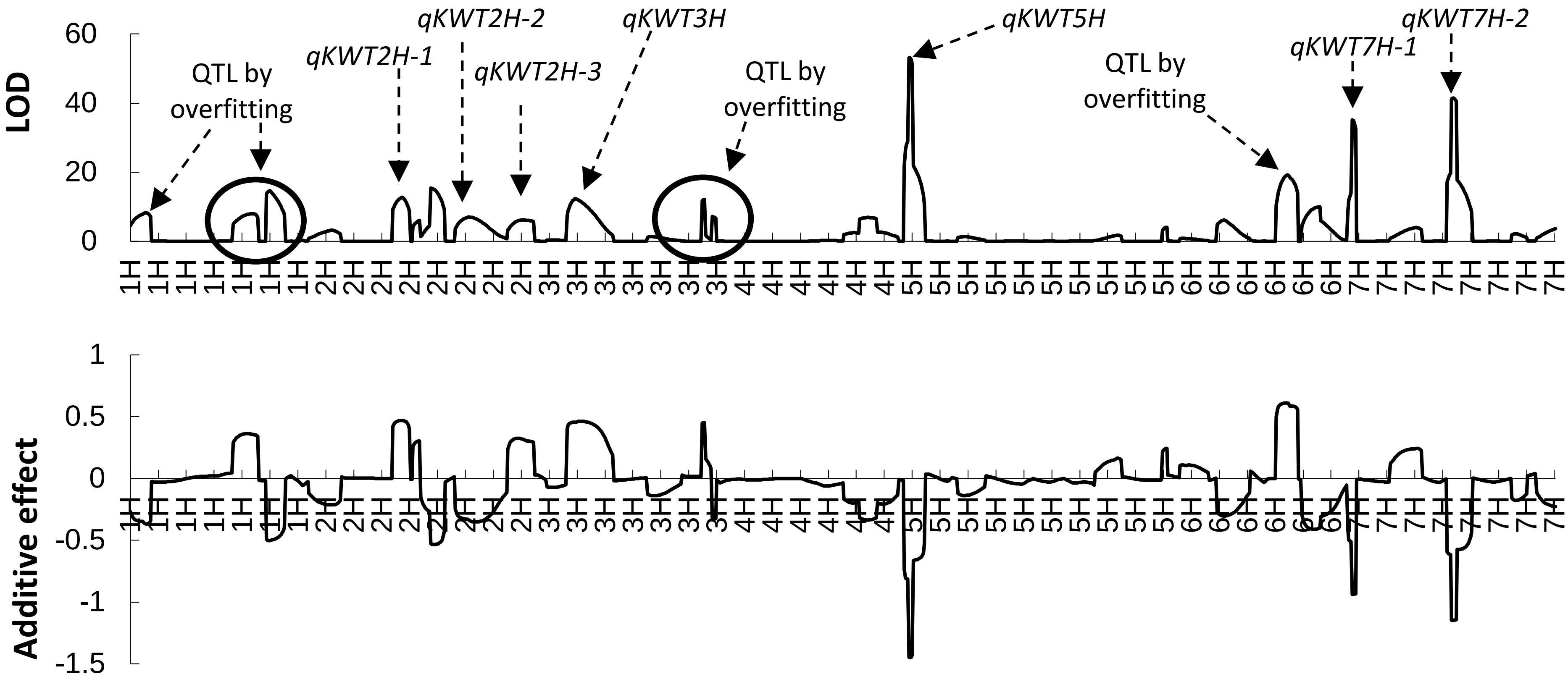
Method	QTL	Power (%)	Pos. (cM)	SE	LOD	SE	Additive	SE
IM	Q1	0.3	33.333	4.714	2.965	0.431	0.313	0.013
	Q2	25.3	66.534	3.220	3.667	1.029	-0.354	0.050
	Q3	6.8	31.691	2.608	3.176	0.553	0.334	0.031
	Q4	40.6	67.746	2.680	4.169	1.302	-0.373	0.061
	FDR (%)	38.9						
ICIM	Q1	11.6	34.216	3.615	5.100	1.915	0.370	0.066
	Q2	33.0	66.179	3.053	5.872	3.188	-0.402	0.108
	Q3	56.2	34.383	2.894	8.332	3.381	0.492	0.104
	Q4	60.9	65.984	2.429	11.413	4.131	-0.591	0.114
	FDR (%)	23.8						

Size of the mapping population

PVE (%)	Marker density 5 cM		Marker density 10 cM	
	Power 0.8	Power 0.9	Power 0.8	Power 0.9
1	300	560	540	>600
2	160	300	280	320
3	110	200	180	200
4	100	160	140	180
5	80	140	120	140
10	50	80	70	80
20	40	60	50	60
30	40	40	40	40

Avoid the over fitting problem in ICIM

Over-fitting can cause fake QTL



One-dimensional scaning on the barley genome, step = 1 cM

How can I know there is an over-fitting problem?

- R^2 in step-wise regression exceeds the broad-sense heritability
- There are closely linked QTL identified, especially the QTL are linked in repulsion

PIN in step-wise regression	**0.001**	**0.01**	**0.05**	**0.1**
R^2	0.7289	0.7963	0.8131	0.8886

- Use smaller PIN to avoid over-fitting problem