

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

QTL Mapping in CSS Lines and Other Mapping Methods

Jiankang Wang, CIMMYT China and CAAS

E-mail: jkwang@cgiar.org; wangjiankang@caas.cn

Web: <http://www.isbreeding.net>

Outlines

- **Chromosome segment substitution lines (CSSL)**
- **Selective genotyping and bulk segregant analysis (BSA)**
- **Association mapping**
- **The CSL functionality in QTL IciMapping**

Chromosome segment substitution lines (CSSL)

Wang, J., H. Li, X. Wan, W. Pfeiffer, J. Crouch, and J. Wan*. 2007. Application of identified QTL-marker associations in rice quality improvement through a design breeding approach.

Theor. Appl. Genet. 115: 87-100.

Wang, J., X. Wan, J. Crossa, J. Crouch, J. Weng, H. Zhai, and J. Wan*. 2006. QTL mapping of grain length in rice (*Oryza sativa* L.) using chromosome segment substitution lines. **Genetical**

Research 88: 93-104.

Ways to develop CSS lines

Standard of People's Republic of China [(NSPRC) 1999]. A chromosome segment substitution line (CSSL) population derived from cultivar Asominori (*japonica*)/IR24 (*indica*) backcrossed to Asominori and composed of 66 CSSLs was used for QTL identification. The CSSLs have several advantages over primary mapping populations such as F₂, F₃, recombinant inbred line (RIL), and double haploids in conducting QTL studies for complex traits. First, each CSSL carries a single or fewer donor segments in the near-isogenic background of a recurrent genotype. Interactions between donor alleles are limited to those between genes on homozygous substituted tracts, reducing the effects of interferences from genetic background (Howell et al. 1996; Yano 2001). Second, high-resolution mapping of putative QTLs as Mendelian factors and further map-based cloning will be feasible in many plants, using secondary F₂ population derived from a cross between a QTL-CSSL and the recurrent parent (Eshed and Zamir 1995; Frary et al. 2000; Takahashi et al. 2001; Yano et al. 2000). In addition, a secondary F₂ population between different target CSSLs can be used to precisely detect and confirm epistasis between QTLs (Lin et al. 2000; Yamamoto et al. 2000). Finally, the CSSLs can be used for simultaneous identification, mapping, and transfer of

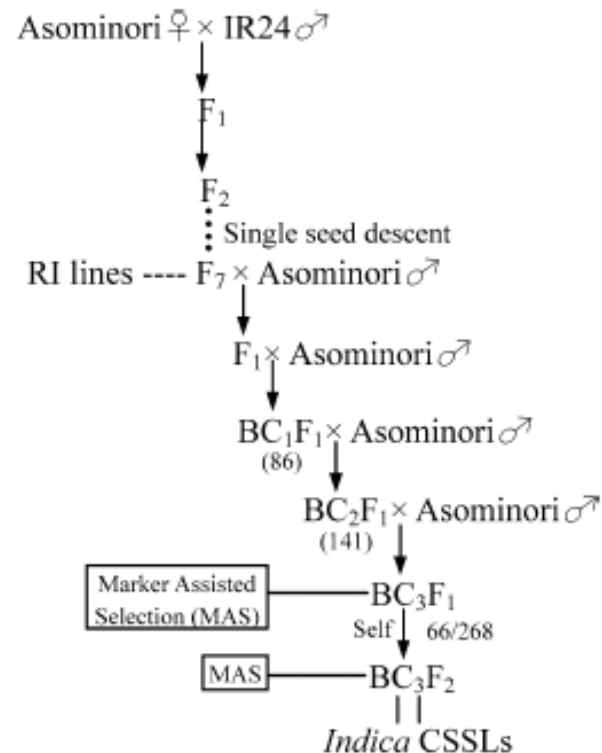
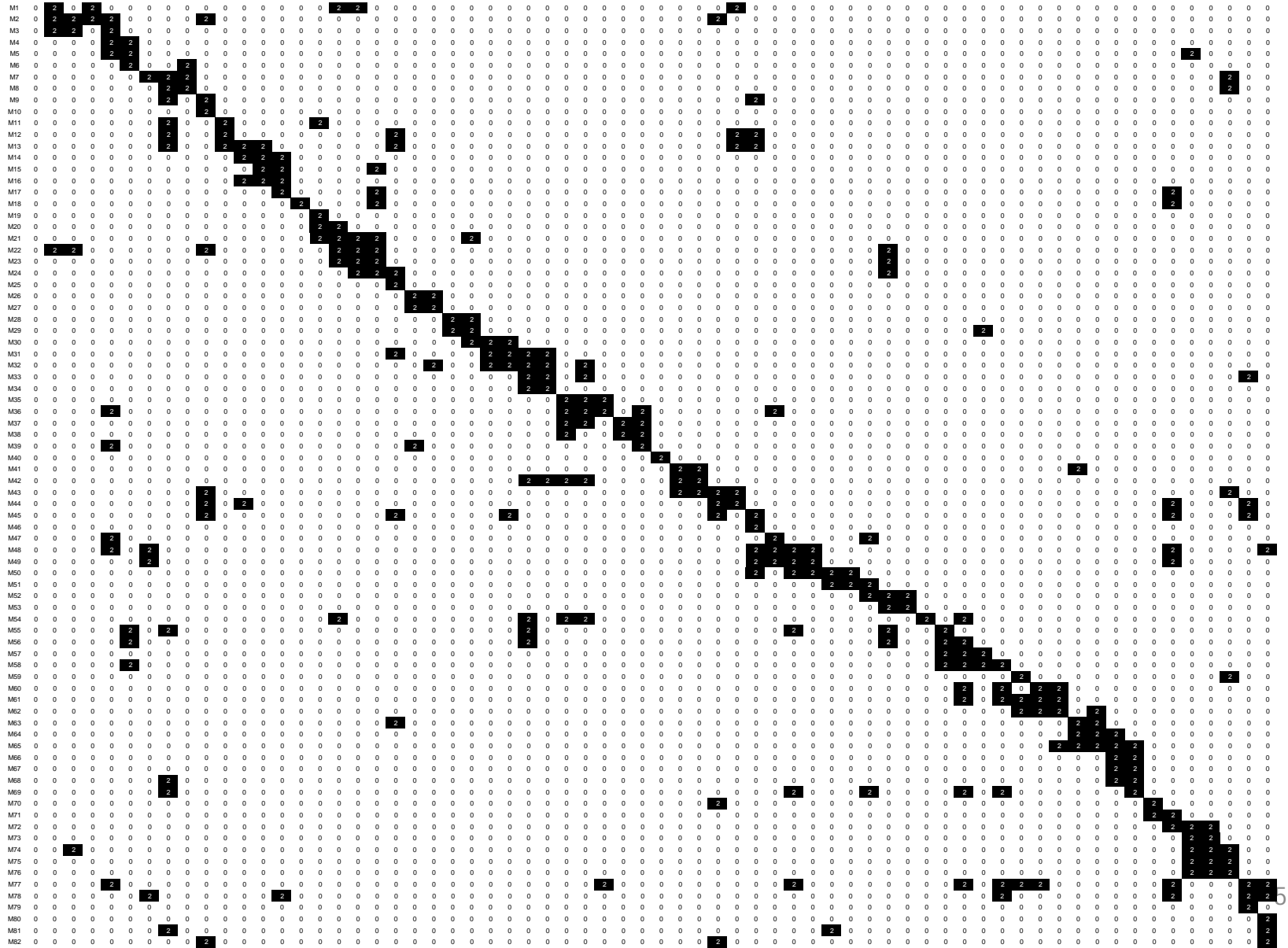


Fig. 1 The strategy for constructing the chromosome segment substitution lines population with the genetic background of a *japonica* variety, cultivar Asominori (quoted from Kubo et al. 1999)

QTL mapping with CSS lines

Background (-1): *japonica* Asominori; Donor (1): *indica* IR24



Idealized CSSL: SSSL (single segment substitution line)

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Background parent	0	0	0	0	0
CSSL1	2	0	0	0	0
CSSL2	0	2	0	0	0
CSSL3	0	0	2	0	0
CSSL4	0	0	0	2	0
CSSL5	0	0	0	0	2

21 Rice CSSL in chromosomes 1-3

M1	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0
M2	0	2	2	2	2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
M3	0	2	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M4	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M5	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M6	0	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
M7	0	0	0	0	0	0	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0
M8	0	0	0	0	0	0	0	2	2	2	0	0	0	0	0	0	0	0	0	0	0
M9	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0
M10	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
M11	0	0	0	0	0	0	0	2	0	0	2	0	0	0	0	2	0	0	0	0	0
M12	0	0	0	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	2	0
M13	0	0	0	0	0	0	0	2	0	0	2	2	2	2	0	0	0	0	0	2	0
M14	0	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0	0	0	0	0	0
M15	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	2	0	0
M16	0	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0	0	0	0	0	0
M17	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0
M18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2	0	0
M19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0
M20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0
M21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	2	0	0
M22	0	2	2	0	0	0	0	0	0	2	0	0	0	0	0	0	2	2	2	0	0
M23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0
M24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	0
M25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
M26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
M27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
T1E1R1	-3.702	-0.72	-0.002	-3.074	-2.069	-3.073	-4.309	-0.476	-0.25	-0.929	-0.647	-2.066	-3.5	-0.044	-4.033	-4.056	-6.404	-6.003	-3.32	-3.572	-4.949
T1E1R2	-3.805	-0.206	-0.943	-4.32	-2.504	-3.705	-2.963	-0.697	-0.989	-2.046	-0.747	-2.078	-3.249	-0.202	-3.934	-5.2	-6.706	-6.269	-3.626	-4.757	-5.755
T1E2R1	-2.785	-3.202	-0.806	-4.259	-0.698	-0.934	-3.357	-0.098	-0.805	-0.404	-2.509	-0.986	-3.362	-0.454	-3.388	-4.903	-6.2	-4.034	-4.433	-3.446	-6.702
T1E2R2	-2.988	-2.687	-0.637	-3.043	-2.279	-3.08	-4.4	-0.039	-2.02	-2.32	-0.978	-0.565	-2.66	-0.677	-3.886	-3.607	-4.302	-5.2	-3.066	-2.89	-5.476
T1E3R1	-2.6	-2.305	-0.044	-3.427	-3.009	-3.093	-2.453	-0.536	-2.7	-0.855	-0.855	-0.72	-3.304	-0.63	-2.752	-4.656	-5.795	-6.25	-3.257	-5.645	-4.85
T1E3R2	-4.205	-0.958	-2.054	-3.96	-3.404	-3.099	-4.558	0	-0.737	-0.99	-2.025	-2.023	-2.998	-2.098	-3.44	-4.734	-5.079	-5.522	-2.206	-3.729	-4.94
T1E4R1	-3.538	-0.936	-0.905	-0.995	-0.476	-2.939	-4.06	-0.623	-2.008	-2.074	-2.505	-2.349	-2.997	-0.075	-3.703	-4.9	-5.382	-6.202	-3.302	-4	-5.73
T1E4R2	-3.008	-0.76	-2.093	-2.823	-2.427	-0.044	-2.557	0.032	-2.034	-0.97	-2.043	-0.62	-3.492	-0.002	-3.08	-3.084	-5.327	-5.29	-4.005	-4.064	-6.084

The linear model

$$y_j = b_0 + \sum_{i=1}^m b_i x_{ji} + e_j$$

- $i=1, \dots, m$ for marker or segment
- $j=1, \dots, n$ for each CSSL

The equivalence between the traditional t -test and the regression model for idealized CSS lines

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Background parent	0	0	0	0	0
CSSL1	2	0	0	0	0
CSSL2	0	2	0	0	0
CSSL3	0	0	2	0	0
CSSL4	0	0	0	2	0
CSSL5	0	0	0	0	2

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 & 2 \end{bmatrix}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 6 & -1 & -1 & -1 & -1 & -1 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} y_{P1} \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}$$

Multicollinearity

- Caused by the correlation between variables
- Results in unreliable estimation of variable effects
- Measured by
 - Variable inflation factor: $VIF=1/(1-R^2)$
 - Condition number: $k = \lambda_{\max} / \lambda_{\min}$
 - λ_{\max} : maximum eigenvalue of the correlation matrix between markers
 - λ_{\min} : minimum eigenvalue

To include the donor can increase multicollinearity

- Background parent + n SSSL, correlation between two variables is $r = -\frac{1}{n}$

- Background parent + donor + n SSSL, correlation between two variables is $r = \frac{n-3}{2n-2}$

- No need to include the donor parent in QTL mapping

The sequential process for decreasing the level of multicollinearity among markers

Step	Condition	N	FirstMarker	SampleSize	SecondMarker	SampleSize	Coefficient Deleted
1	Infinity		M14	3	M16	3	1.0000 M16
2	Infinity		M26	2	M27	2	1.0000 M27
3	Infinity		M66	2	M67	2	1.0000 M67
4	Infinity		M75	3	M76	3	1.0000 M76
5	Infinity		M60	4	M61	5	0.8872 M61
20	Infinity		M23	4	M24	4	0.7339 M24
21	6020.		M31	5	M32	6	0.7062 M32
22	1819.		M55	6	M56	5	0.7062 M55
23	1766.		M19	1	M20	2	0.7016 M20
24	1725.		M33	4	M34	2	0.6960 M33
25	1393.		M2	6	M3	3	0.6901 M2
26	1340.		M35	3	M36	6	0.6901 M36
27	1293.		M14	3	M15	3	0.6508 M15
27	758.						

A likelihood ratio test combined with stepwise regression (RSTEP-LRT)

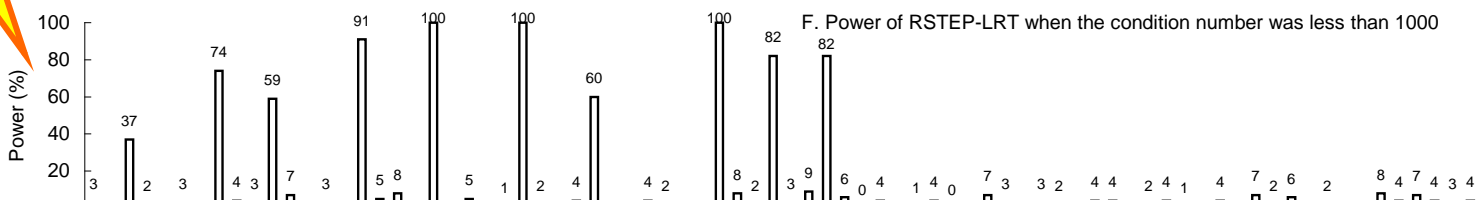
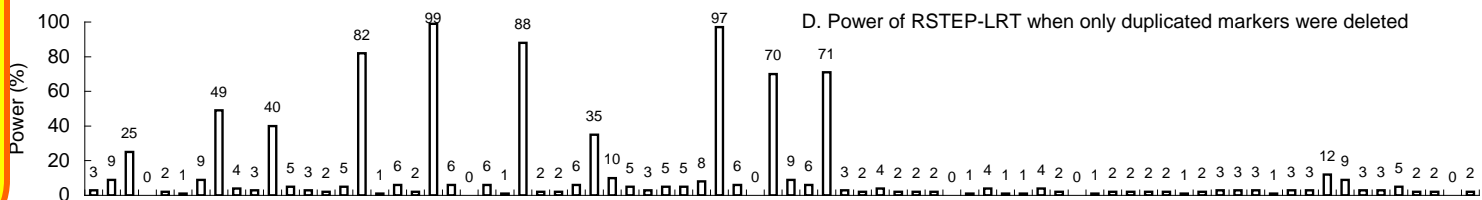
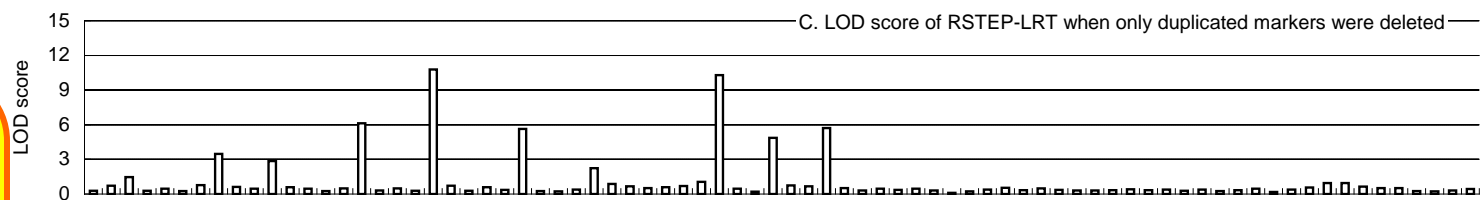
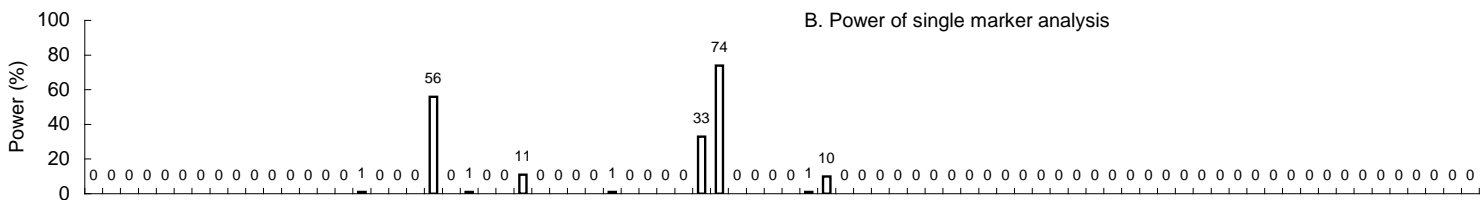
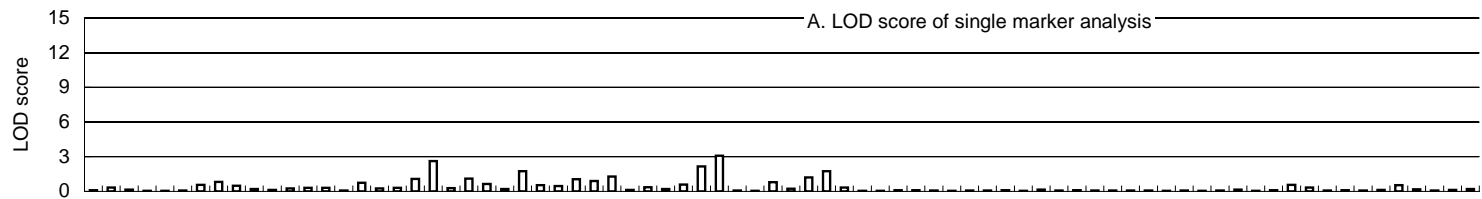
$$\Delta y_i = y_i - \sum_{k \neq j} b_k x_{ik}$$

$$H_A : \mu_1 \neq \mu_2 \quad L_A = \sum_{i=0}^{n_1} \ln f(\Delta y_i; \mu_1, \sigma_A^2) + \sum_{i=n_1+1}^n \ln f(\Delta y_i; \mu_2, \sigma_A^2)$$

$$H_0 : \mu_1 = \mu_2 \quad L_0 = \sum_{i=0}^n \ln f(\Delta y_i; \mu_0, \sigma_0^2)$$

RSTEP-LRT for QTL mapping

- Chromosome segment substitution (CSS) lines have great potential for use in QTL fine mapping and map-based cloning
- The standard t -test used in the idealized case that each CSS line has a single segment from the donor parent is not suitable for non-idealized CSS lines carrying several substituted segments
- RSTEP-LRT: a likelihood ratio test based on stepwise regression for QTL mapping in a population consisting of non-idealized CSS lines
 - Stepwise regression was used to select the most important segments for the trait of interest
 - Likelihood ratio test was used to calculate the LOD score of each chromosome segment
 - To further improve the power of QTL mapping, we have also used a method to decrease the effects of multicollinearity among chromosome segments.

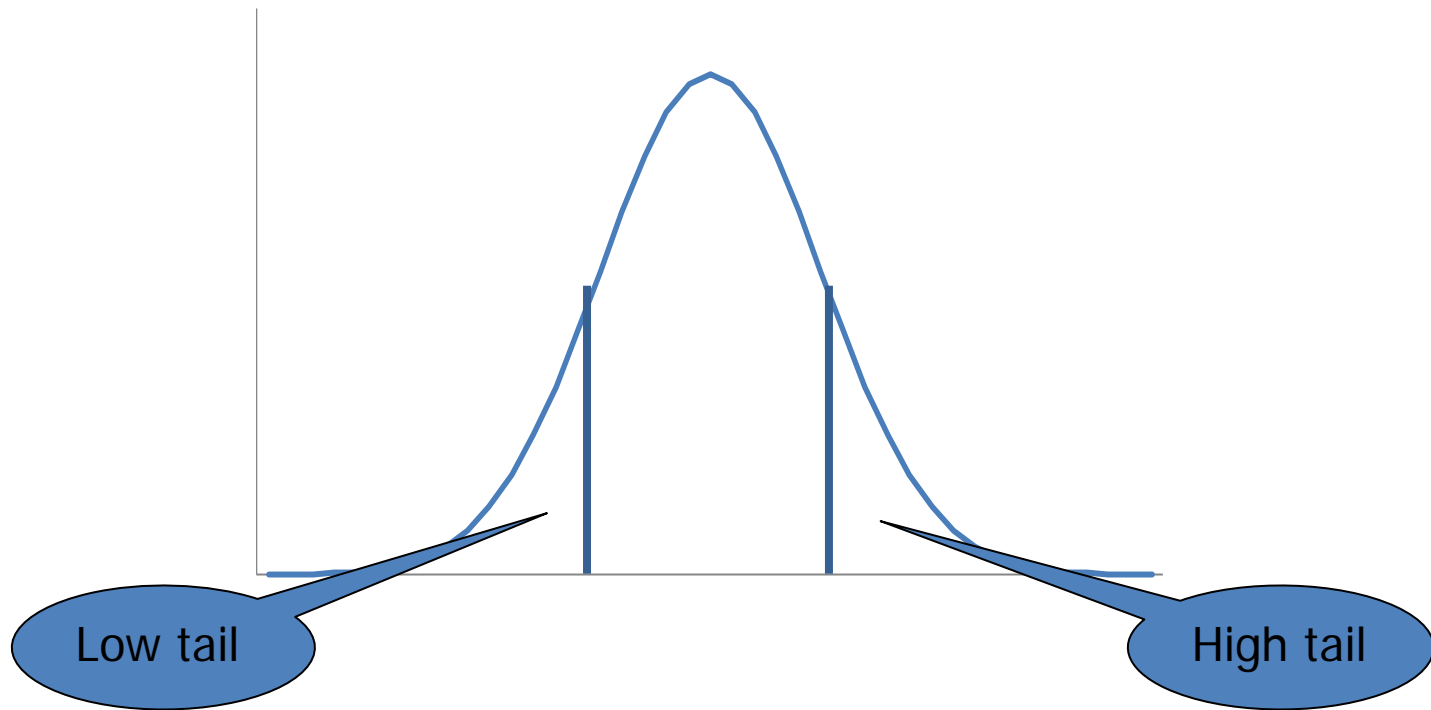


The recommended method for QTL mapping with non-idealized CSS lines

M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17 M18 M19 M20 M21 M22 M23 M24 M25 M26 M27 M28 M29 M30 M31 M32 M33 M34 M35 M36 M37 M38 M39 M40 M41 M42 M43 M44 M45 M46 M47 M48 M49 M50 M51 M52 M53 M54 M55 M56 M57 M58 M59 M60 M61 M62 M63 M64 M65 M66 M67 M68 M69 M70 M71 M72 M73 M74 M75 M76 M77 M78 M79 M80 M81 M82

Selective genotyping and bulk segregation analysis (BSA)

Selective genotyping



Genotyping selected individuals in both tails

QTL mapping from the difference on marker frequency

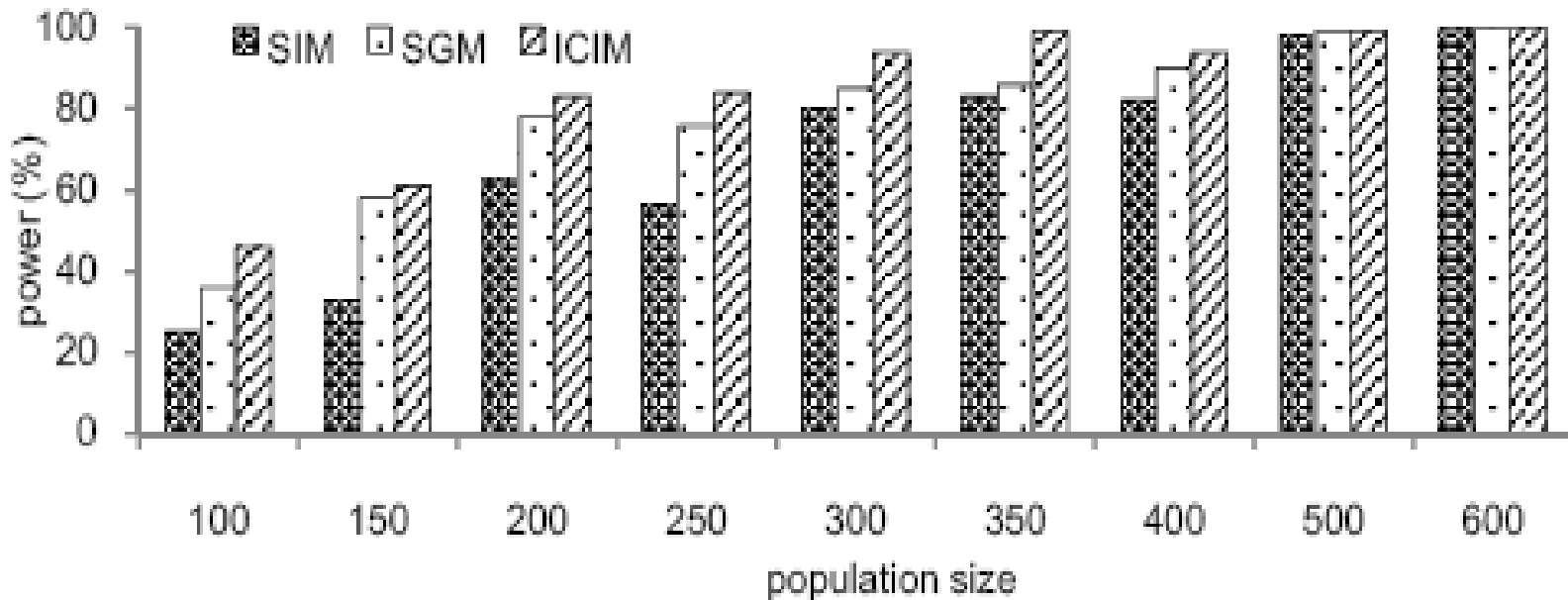
- Significance test of allele frequencies in both tails

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

- Useful when phenotyping is cheaper/easier than genotyping
- Disadvantages
 - Cannot be used for other traits
 - Difficulty to estimate QTL effects

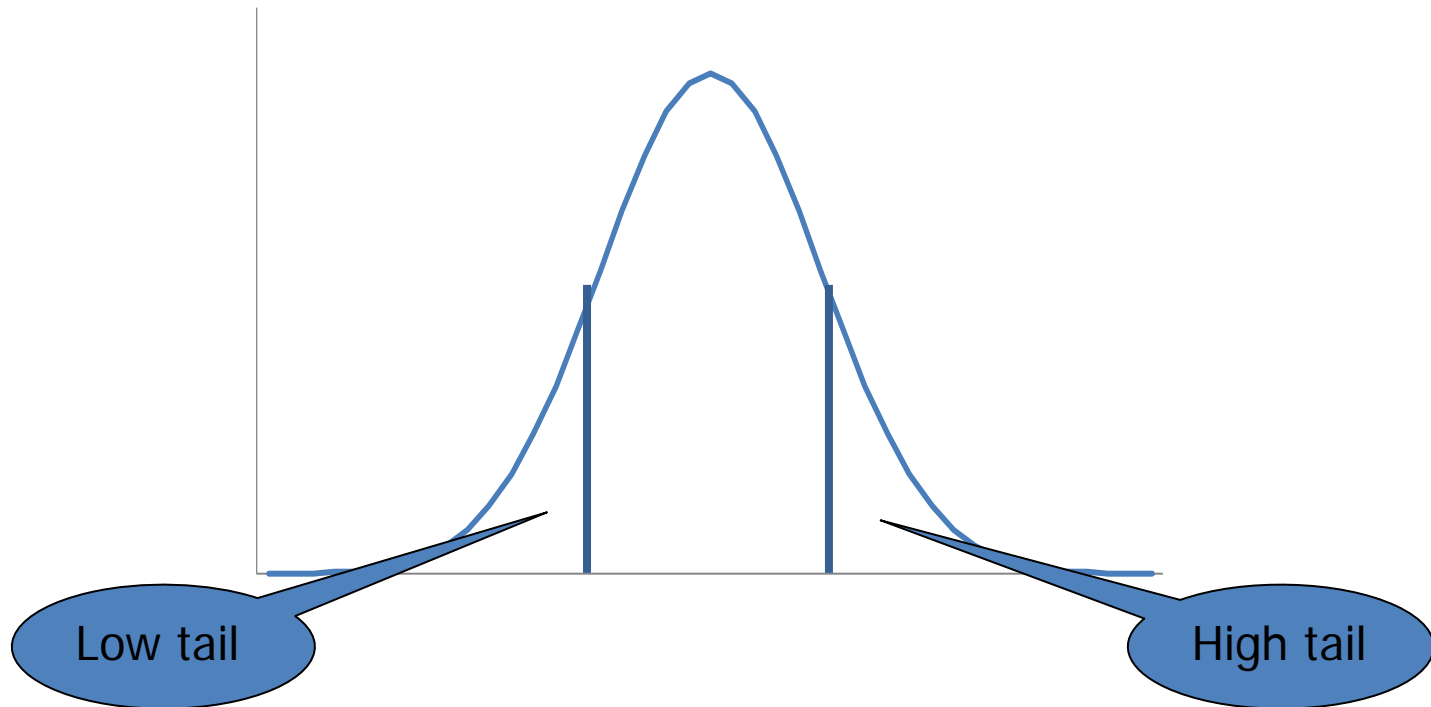
Comparison of SGM with IM and ICIM

(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

Bulked segregant analysis (BSA)



Genotyping two DNA pools of the two tails.
For polymorphism markers in two pools, conduct marker screening in original population.

Steps of BSA

- To form two DNA pools
- To screen for polymorphism in the two pools
- For polymorphism markers, screen all individuals in original population
- To use the standard QTL linkage mapping
- Disadvantages: Two DNA pools cannot be used for other traits

Association mapping

Linkage disequilibrium (LD)

- Linkage
 - A (A1, A2) — B (B1, B2) with recom. freq. r
 - Linkage in coupling
 - A1B1, A2B2: $(1-r^2)/2$
 - Linkage in repulsion
 - A1B2, A2B1, $r^2/2$
- Linkage disequilibrium
 - In F1, $P(A1)=P(A2)=P(B1)=P(B2)=0.5$
 - But $P(A1B1) \neq P(A1) \times P(B1)$, unless $r=0.5$

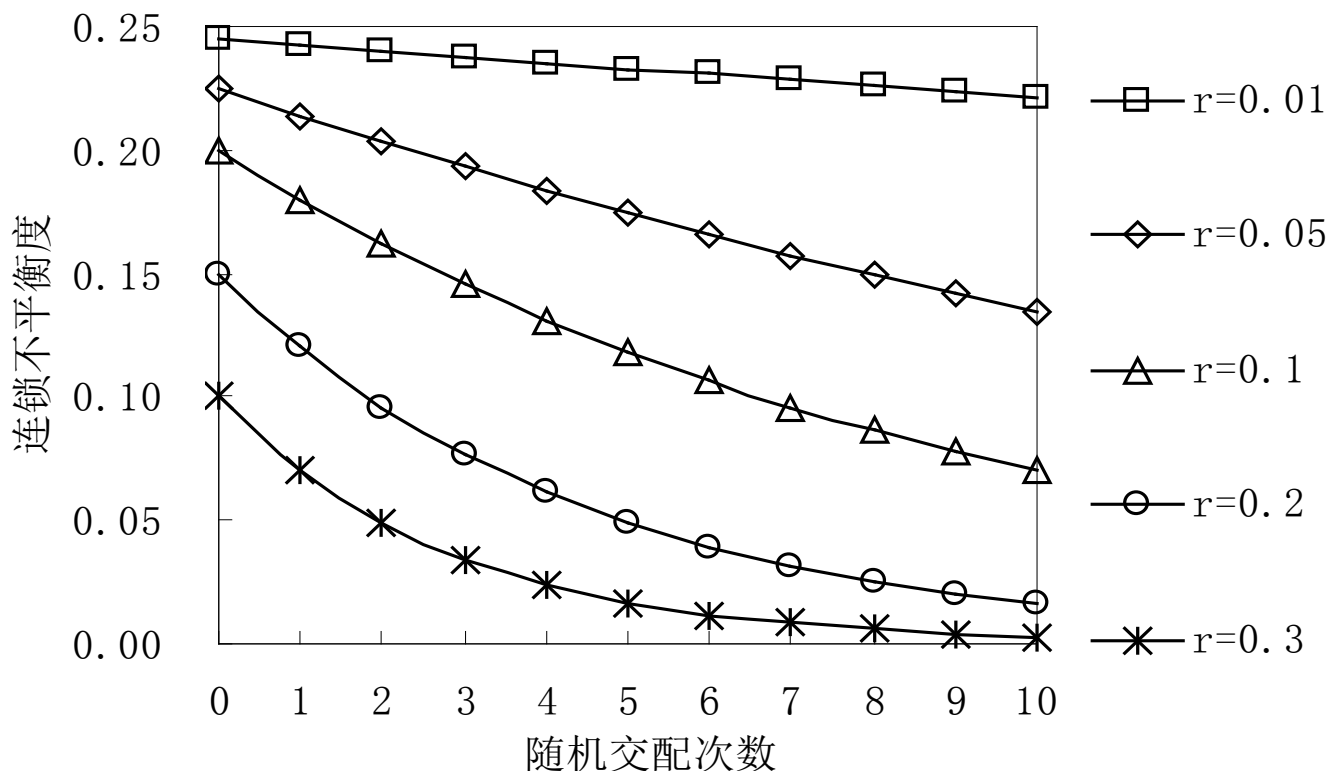
General definition of LD

$$D_{ij} = p(A_i B_j) - p(A_i) p(B_j)$$

Attention 1: Linkage does not really see high LD

- Random mating breaks LD

$$D_t = D_0(1-r)^t$$



Locus A and B are apart by 17.8 kb in fruit fly

Haplotype		OBS.	Frequency	Expected frequency	EXP.
A	B				
+	+	4	0.085	0.06	2.8
+	-	4	0.085	0.11	5.2
-	+	13	0.277	0.30	14.1
-	-	26	0.533	0.53	24.9

- LD=0.023
- Chi-square=0.93, df=1, $P > 0.5$

Attention 2: High LD value does not really mean linkage

- Population structure can result in high value of LD

Population	Allele frequency				Genotype frequency
	A1	A2	B1	B2	A1A1B1B1
Pop 1	0.7	0.3	0.7	0.3	0.2401
Pop 2	0.3	0.7	0.3	0.7	0.0081
Admixture	0.5	0.5	0.5	0.5	OBS: 0.0625 EXP: 0.1241

Linkage mapping or association mapping in plants?

- Both methods utilize LD in QTL mapping
- For association mapping based on natural populations
 - Population structure causes false LD
 - Random mating during evolution reduce LD
- For linkage mapping based on biparental populations
 - Maximum LD can be used
 - Population structure is clear

Contingency table test

- When each observation in our sample is a bivariate discrete random vector (a pair of discrete random variables), then there is a simple way to test the hypothesis that the two random variables are independent. The test is another form of χ^2 test.

Two-way contingency table

- A table in which each observation is classified in two or more ways is called a contingency table.
- For example, a two-way contingency table.

	Candidate preferred			
Curriculum	A	B	Undecided	Totals
Engineering and science	24	23	12	59
Humanities and social sciences	24	14	10	48
Fine arts	17	8	13	38
Industrial and public administration	27	19	9	55
Totals	92	64	44	200

The χ^2 test of independence

- Let \hat{E}_{ij} denote the MLE of the expected number of observations that will be classified in the i th row and the j th column of the table when H_0 is true.

$$Q = \sum_{i=1}^R \sum_{j=1}^C \frac{(N_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

- Q has the property that if H_0 is true and the sample size $n \rightarrow \infty$, then Q converges in the distribution to the χ^2 distribution with $RC-1-s=(R-1)(C-1)$ degrees of freedom.

Simpson's Paradox

- When tabulating discrete data, we need to be careful about aggregating groups.
- Suppose that a survey has two questions. If we construct a single table of responses to the two questions that includes both men and women, we might get a very different picture than if we construct separate tables for the responses of men and women.

An example of the paradox

- Disaggregated by sex

Men only	Improved	Not improved	Percent improved
New treatment	12	18	40
Standard treatment	3	7	30

Women only	Improved	Not improved	Percent improved
New treatment	8	2	80
Standard treatment	21	9	70

- Aggregated by sex

All patients	Improved	Not improved	Percent improved
New treatment	20	20	50
Standard treatment	24	16	60

An example of the paradox

- According to the first table, the new treatment is superior to the standard treatment both for men and for women,
- According to the second and third tables, the new treatment is inferior to the standard treatment when all the subjects are aggregated.
- This type of result is known as Simpson's paradox.

The paradox explained

- In the example, women have a higher rate of improvement from the disease than men have, regardless of which treatment they receive.
- Furthermore, most of the women in the sample receive the standard treatment while most of the men received the new treatment.

A make-up example not to see LD

- Assume locus A-a is linked with locus B-b, with a genetic distance 1 cM
- We have the four genotypes AABB, AAbb, aaBB, and aabb in our hand.
- If we have a 1:1:1:1 mixture population of the 4 genotypes, we won't be able to see any LD between locus A-a and locus B-b
 - $p_A = p_a = p_B = p_b = 0.5$;
 - $P(AB) = 0.25$; $p_A * p_B = 0.25$; $P(AB) - p_A * p_B = 0$, so is true for Ab, aB, and ab

A make-up example to see fake LD

- Assume locus A-a is unlinked with locus B-b, say located on two chromosomes
- We have the four genotypes AABB, AAbb, aaBB, and aabb in our hand.
- If we have a 1:1 mixture population of genotypes AABB, and aabb, we are able to see LD=0 between locus A-a and locus B-b
 - $p_A=p_a=p_B=p_b=0.5$;
 - $P(AB)=0.5$; $p_A * p_B=0.25$; $P(AB)-p_A * p_B=0.25$
 - $P(Ab)=0$; $p_A * p_B=0.25$; $P(AB)-p_A * p_B=-0.25$

Two major problems with association mapping

- **Problem A: Random mating can break down true LD**
- **Problem B: Population structure can cause fake LD**
- The principle behind association mapping is simple: similar to single marker analysis. The more difficult work is to handle the two problems.
- To solve problem A, we need highly-dense markers. To solve problem B, we need to identify the structure of the mapping population.

The CSL functionality in QTL IciMapping

Three methods available in CSL

- SMA: single marker analysis (Soller et al., 1976. Theor. Appl. Genet. 47: 35-39)
- RSTEP-LRT-ADD: stepwise regression based likelihood ratio tests of additive QTL (Wang et al., 2006. Gen. Res. 88: 93-104)
- RSTEP-LRT-EPI: stepwise regression based likelihood ratio tests of digenic epistasis QTL (Wang et al., 2007. Theor. Appl. Genet. 115: 87-100)

LOD histogram of RSTEP-LRT-ADD

