The 9th Workshop on QTL Mapping and Breeding Simulation
The University of Sydney, Cobbitty NSW, 7-9 March 2012

Linkage Analysis and Linkage Map Construction

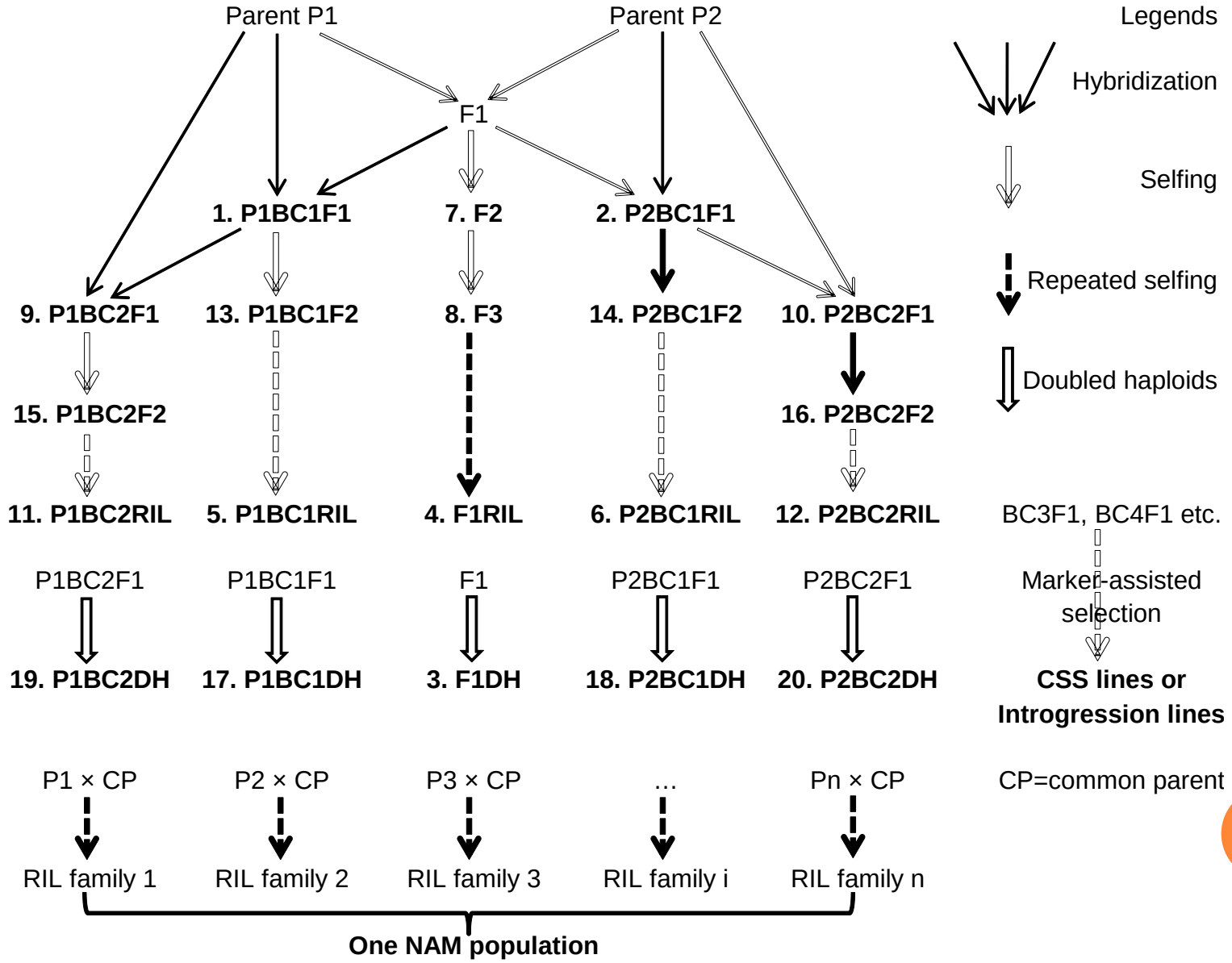
Jiankang Wang
CAAS and CIMMYT China
E-mail: wangjk@caas.net.cn; jkwang@cgiar.org

Outlines

- Genetic populations and pair-wise linkage analysis
- Three-point analysis and linkage map construction
- The MAP functionality in QTL IciMapping

Genetic populations and pair-wise linkage analysis

Populations handled in QTL IciMapping



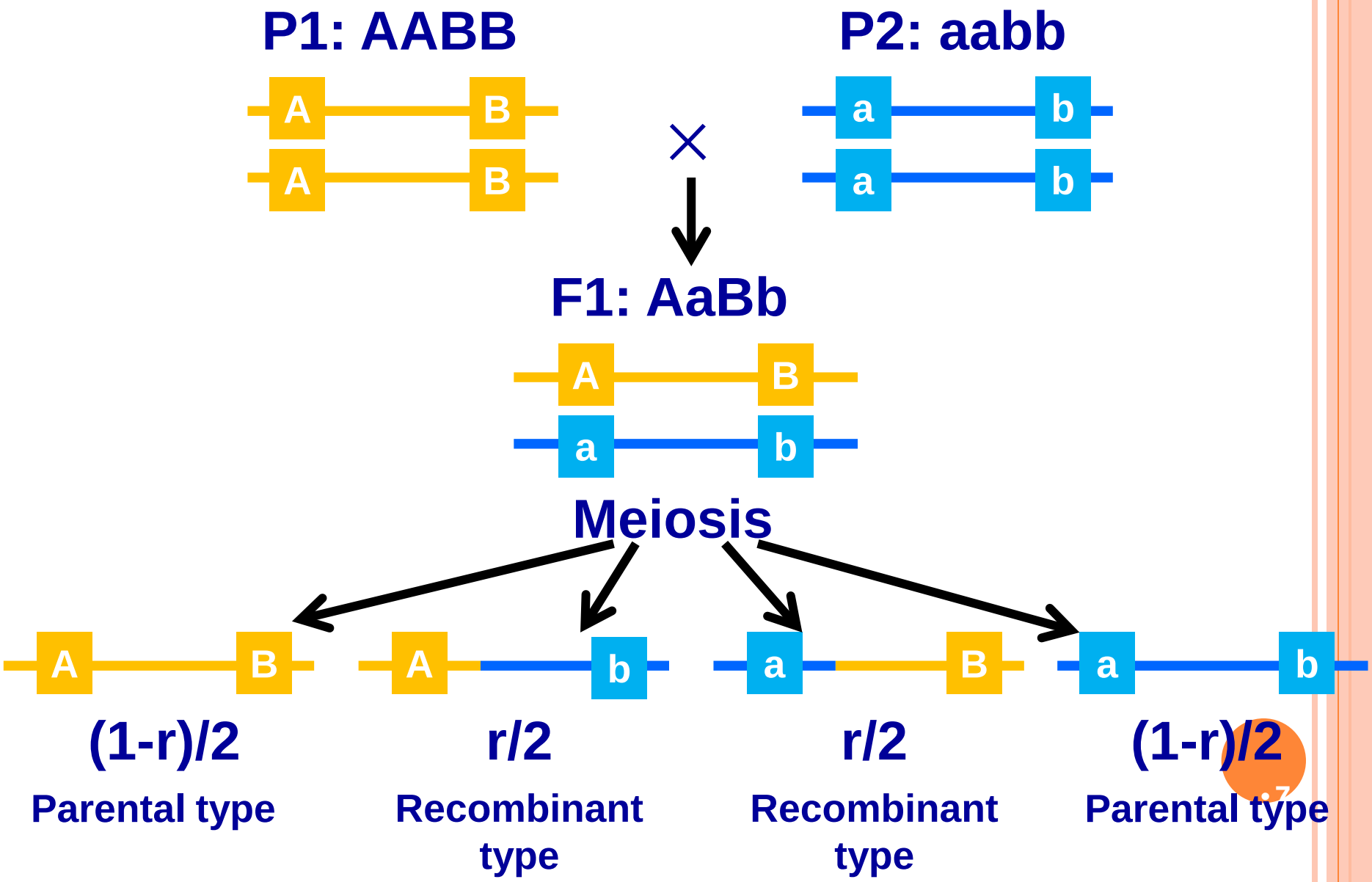
Example: 10 RILs in a rice population (Linkage map of Chr. 5)

Marker	C263	R830	R3166	XNpb387	R569	R1553	C128	C1402	XNpb81	C246	R2953	C1447	Grain width (mm)
Position (cM)	0.0	3.5	8.5	19.5	32.0	66.6	74.1	78.6	81.8	91.9	92.7	96.8	
RIL1	0	0	0	0	0	0	0	0	0	0	0	0	2.33
RIL2	2	2	2	2	2	0	0	0	0	2	2	2	1.99
RIL3	0	2	2	2	2	2	2	2	2	2	2	2	2.24
RIL4	0	0	0	0	0	0	2	2	2	2	2	2	1.94
RIL5	0	0	0	0	0	2	2	0	0	0	0	0	2.76
RIL6	0	0	0	2	2	2	2	2	2	2	2	2	2.32
RIL7	0	0	0	0	0	0	0	0	0	0	0	0	2.32
RIL8	2	2	0	2	2	0	0	0	0	2	2	2	2.08
RIL9	0	0	0	0	2	2	0	0	0	0	0	0	2.24
RIL10	0	0	0	0	2	2	0	0	0	0	0	0	2.45

Genetic markers in linkage analysis

- **Morphological traits**
  - **Qualitative traits used in Mendel's hybridization experiments**
- **Cytogenetic and bio-chemistry markers (e.g. isozyme)**
- **DNA molecular markers**
  - **RFLP, SSR, SNP etc.**

The four gametes (haplotypes) of an F1



Expected genotypic frequency in backcross and DH populations

P1: AABB; P2: aabb

P1BC1	P2BC1	DH	Samples	Theoretical frequency
AABB	AaBb	AABB	n_1	$f_1=(1-r)/2$
AABb	Aabb	AAbb	n_2	$f_2=r/2$
AaBB	aaBb	aaBB	n_3	$f_3=r/2$
AaBb	aabb	aabb	n_4	$f_4=(1-r)/2$

MLE of recombination frequency

- **Likelihood function**

$$L = \frac{n!}{n_1!n_2!n_3!n_4!}\left[\frac{1}{2}(1-r)\right]^{n_1}\left[\frac{1}{2}r\right]^{n_2}\left[\frac{1}{2}r\right]^{n_3}\left[\frac{1}{2}(1-r)\right]^{n_4} = C(1-r)^{n_1+n_4}(r)^{n_2+n_3}$$

- **Logarithm of likelihood**

$$\ln L = \ln C + (n_1 + n_4)\ln(1-r) + (n_2 + n_3)\ln r$$

- **MLE of r** $\hat{r} = \frac{n_2 + n_3}{n_1 + n_2 + n_3 + n_4} = \frac{n_2 + n_3}{n}$

- **Fisher information**

$$I = -E\left(\frac{d^2 \ln L}{d^2 r}\right) = -E\left[-\frac{n_1 + n_4}{(1-r)^2} - \frac{n_2 + n_3}{r^2}\right] = \frac{n}{r(1-r)}$$

- **Variance of estimated r** $V_{\hat{r}} = \frac{1}{I} = \frac{\hat{r}(1-\hat{r})}{n}$

Significance test of linkage

- Null hypothesis H_0: r = 0.5 (no genetic linkage, or locus A-a and B-b are independent)
- Alternative hypothesis H_A: r ≠ 0.5

- Likelihood ratio test (LRT) or LOD score

$$LRT = -2\ln[\frac{L(r=0.5)}{L(\hat{r})}] \sim \chi^2(df=1)$$

$$LOD = \frac{L(\hat{r})}{L(r=0.5)}$$

An example P1BC1 population

- **Genotypes of two inbred parents P1 and P2 are AABB and aabb**
- **Observed samples of the four genotypes in P1BC1**
  - **AABB：162；AABb：40；AaBB：41；AaBb：158**

$$\hat{r} = \frac{40+41}{162+40+41+158} = \frac{81}{401} = 20.20\%$$

$$V_{\hat{r}} = \frac{\hat{r}(1-\hat{r})}{n} \approx 4.02\times10^{-4}$$

Test of linkage

- **Null hypothesis H_0: r = 0.5**
- **Alternative hypothesis H_A: r ≠ 0.5**

$$\frac{L(\hat{r})}{L(r=0.5)} = \frac{(1-r)^{n_1+n_4}\, r^{n_2+n_3}}{(\frac{1}{4})^{n_1+n_2+n_3+n_4}} = 6.3\times 10^{153}$$

- **Likelihood ratio test (LRT) (P<0.0001) and LOD score**

$$LRT = 2 * \ln[\frac{L(\hat{r})}{L(r=0.5)}] = 708.27$$

$$LOD = \log[\frac{L(\hat{r})}{L(r=0.5)}] = 153.80$$

Genotypic frequencies in RIL populations, compared with DH

DH population	Theoretical frequency	RIL population	Theoretical frequency
AABB	$f_1=(1-r)/2$	AABB	$f_1=(1-R)/2$
AAbb	$f_2=r/2$	AAbb	$f_2=R/2$
aaBB	$f_3=r/2$	aaBB	$f_3=R/2$
aabb	$f_4=(1-r)/2$	aabb	$f_4=(1-R)/2$

$$R = 2r/(1+2r)$$

10 RILs in a rice population

P1: 0 or A; P2: 2 or B; F1: 1 or H

RIL	Marker 1	Marker 2	Parent type or recombinant
	C263	XNpb387	
RIL1	0 or A	0 or A	P1 type
RIL2	2 or B	2 or B	P2 type
RIL3	0 or A	2 or B	Recombinant
RIL4	0 or A	0 or A	P1 type
RIL5	0 or A	0 or A	P1 type
RIL6	0 or A	2 or B	Recombinant
RIL7	0 or A	0 or A	P1 type
RIL8	2 or B	2 or B	P2 type
RIL9	0 or A	0 or A	P1 type
RIL10	0 or A	0 or A	P1 type

$n_1 = 6$

$n_2 = 2$

$n_3 = 0$

$n_4 = 2$

$R = 2/10 = 0.2$

$r = 0.125$

$LRT = 17.72,$
$(P = 2.56 \times 10^{-5})$

$LOD = 3.85$

Expected genotypic frequencies in F2 populations

Co-dominant markers		Dominant markers	
Marker type	Frequency	Marker type	Frequency
AABB	$(1-r)^2/4$	A_B_	$[2+(1-r)^2]/4$
AABb	$r(1-r)/2$		
AAbb	$r^2/4$	A_bb	$[1-(1-r)^2]/4$
AaBB	$r(1-r)/2$		
AaBb	$(1-2r+2r^2)/2$		
Aabb	$r(1-r)/2$		
aaBB	$r^2/4$	aaB_	$[1-(1-r)^2]/4$
aaBb	$r(1-r)/2$		
aabb	$(1-r)^2/4$	aabb	$(1-r)^2/4$

MLE of r in F2: dominant markers

- **Logarithm of the likelihood ratio** $k = (1-r)^2$

$$\ln L = C + n_1 \ln(3 - 2r + r^2) + (n_3 + n_7)\ln(2r - r^2) + n_9 \ln(1 - 2r + r^2)$$
$$= C + n_1 \ln(2 + k) + (n_3 + n_7)\ln(1 - k) + n_9 \ln k$$

- **MLE of r**

$$k = (1-r)^2 = \frac{-(2n - 3n_1 - n_9) \pm \sqrt{(2n - 3n_1 - n_9)^2 + n \times n_9}}{2n}$$

- **Variance of the estimated r**

$$V_{\hat{r}} = \frac{(1-k)(2-k)}{2n(1+2k)} = \frac{(2r - r^2)(3 - 2r + r^2)}{2n(3 - 4r + 2r^2)}$$

MLE of r in F2: co-dominant markers (Newton-Raphson algorithm)

- **Log-likelihood function**

$$\ln L = \ln C + (2n_1 + 2n_9 + n_2 + n_4 + n_6 + n_8)\ln(1-r)$$
$$+ (n_2 + n_4 + n_6 + n_8 + 2n_3 + 2n_7)\ln r + n_5 \ln(1 - 2r + 2r^2)$$

- **The first-order derivative of LogL**
- f'(r) $= \frac{d \ln L}{dr} = \frac{2n_1 + 2n_9 + n_2 + n_4 + n_6 + n_8}{r-1} + \frac{n_2 + n_4 + n_6 + n_8 + 2n_3 + 2n_7}{r} + \frac{n_5(4r-2)}{1-2r+2r^2}$
- **The second-order derivative of LogL**
- f''(r) $= \frac{d^2 \ln L}{d^2 r} = -\frac{2n_1 + 2n_9 + n_2 + n_4 + n_6 + n_8}{(r-1)^2} - \frac{n_2 + n_4 + n_6 + n_8 + 2n_3 + 2n_7}{r^2} + \frac{n_5(4r - 4r^2)}{(1-2r+2r^2)^2}$
- **The iteration algorithm:**

$$r_{i+1} = r_i - f'(r_i)/f''(r_i)$$

MLE of r in F2: co-dominant markers (EM algorithm)

- EM for expectation and maximization

- E-step: for an initial r_0, calculate the probability of crossover in each marker type
- M-step: Update r, and repeat from the E-step

$$r' = \frac{1}{n} \sum_k n_k P_k(R \mid G)$$

Expected probability of crossover

Marker type	Frequency	Expected sample size	P(R|G)
AABB	$f_1=(1-r)^2/4$	$n_1 = nf_1$	0
AABb	$f_2=r(1-r)/2$	$n_2 = nf_2$	0.5
AAbb	$f_3=r^2/4$	$n_3 = nf_3$	1
AaBB	$f_4=r(1-r)/2$	$n_4 = nf_4$	0.5
AaBb	$f_5=(1-2r+2r^2)/2$	$n_5 = nf_5$	$r^2/(1-2r+2r^2)$
Aabb	$f_6=r(1-r)/2$	$n_6 = nf_6$	0.5
aaBB	$f_7=r^2/4$	$n_7 = nf_7$	1
aaBb	$f_8=r(1-r)/2$	$n_8 = nf_8$	0.5
aabb	$f_9=(1-r)^2/4$	$n_9 = nf_9$	0

$$r = [n_1 \times 0 + n_2 \times 0.5 + n_3 \times 1 + \ldots + n_8 \times 0.5 + n_9 \times 0]/n$$

Estimated r after 3 EM iterations (r_0=0.5)

Geno.	Size	r0	Exp. Freq.	P(R | G)	r1	Exp. Freq.	P(R | G)	r2	Exp. Freq.	P(R | G)	r3
AABB	**30**	**0.5**	**0.063**	**0**	**0.313**	**0.118**	**0**	**0.198**	**0.161**	**0**	**0.159**
AABb	**7**	**0.5**	**0.125**	**0.5**	**0.313**	**0.107**	**0.5**	**0.198**	**0.080**	**0.5**	**0.159**
AAbb	**1**	**0.5**	**0.063**	**1**	**0.313**	**0.024**	**1**	**0.198**	**0.010**	**1**	**0.159**
AaBB	**9**	**0.5**	**0.125**	**0.5**	**0.313**	**0.107**	**0.5**	**0.198**	**0.080**	**0.5**	**0.159**
AaBb	**50**	**0.5**	**0.250**	**0.5**	**0.313**	**0.285**	**0.1712**	**0.198**	**0.341**	**0.0577**	**0.159**
Aabb	**12**	**0.5**	**0.125**	**0.5**	**0.313**	**0.107**	**0.5**	**0.198**	**0.080**	**0.5**	**0.159**
aaBB	**0**	**0.5**	**0.063**	**1**	**0.313**	**0.024**	**1**	**0.198**	**0.010**	**1**	**0.159**
aaBb	**10**	**0.5**	**0.125**	**0.5**	**0.313**	**0.107**	**0.5**	**0.198**	**0.080**	**0.5**	**0.159**
aabb	**25**	**0.5**	**0.063**	**0**	**0.313**	**0.118**	**0**	**0.198**	**0.161**	**0**	**0.159**
	144		**1**			**1**			**1**		

Estimated r after 3 EM iterations (r_0=0.25)

Geno.	Size	r0	Exp. Freq.	P(R | G)	r1	Exp. Freq.	P(R | G)	r2	Exp. Freq.	P(R | G)	r3
AABB	**30**	**0.25**	**0.141**	**0**	**0.174**	**0.171**	**0**	**0.154**	**0.179**	**0**	**0.150**
AABb	**7**	**0.25**	**0.094**	**0.5**	**0.174**	**0.072**	**0.5**	**0.154**	**0.065**	**0.5**	**0.150**
AAbb	**1**	**0.25**	**0.016**	**1**	**0.174**	**0.008**	**1**	**0.154**	**0.006**	**1**	**0.150**
AaBB	**9**	**0.25**	**0.094**	**0.5**	**0.174**	**0.072**	**0.5**	**0.154**	**0.065**	**0.5**	**0.150**
AaBb	**50**	**0.25**	**0.313**	**0.1**	**0.174**	**0.357**	**0.0423**	**0.154**	**0.370**	**0.0319**	**0.150**
Aabb	**12**	**0.25**	**0.094**	**0.5**	**0.174**	**0.072**	**0.5**	**0.154**	**0.065**	**0.5**	**0.150**
aaBB	**0**	**0.25**	**0.016**	**1**	**0.174**	**0.008**	**1**	**0.154**	**0.006**	**1**	**0.150**
aaBb	**10**	**0.25**	**0.094**	**0.5**	**0.174**	**0.072**	**0.5**	**0.154**	**0.065**	**0.5**	**0.150**
aabb	**25**	**0.25**	**0.141**	**0**	**0.174**	**0.171**	**0**	**0.154**	**0.179**	**0**	**0.150**
	144		**1**			**1**			**1**		

Estimated r after 3 EM iterations (r_0=0.0)

Geno.	Size	r0	Exp. Freq.	P(R | G)	r1	Exp. Freq.	P(R | G)	r2	Exp. Freq.	P(R | G)	r3
AABB	**30**	**0**	**0.250**	**0**	**0.139**	**0.185**	**0**	**0.148**	**0.182**	**0**	**0.149**
AABb	**7**	**0**	**0.000**	**0.5**	**0.139**	**0.060**	**0.5**	**0.148**	**0.063**	**0.5**	**0.149**
AAbb	**1**	**0**	**0.000**	**1**	**0.139**	**0.005**	**1**	**0.148**	**0.005**	**1**	**0.149**
AaBB	**9**	**0**	**0.000**	**0.5**	**0.139**	**0.060**	**0.5**	**0.148**	**0.063**	**0.5**	**0.149**
AaBb	**50**	**0**	**0.500**	**0**	**0.139**	**0.380**	**0.0253**	**0.148**	**0.374**	**0.0292**	**0.149**
Aabb	**12**	**0**	**0.000**	**0.5**	**0.139**	**0.060**	**0.5**	**0.148**	**0.063**	**0.5**	**0.149**
aaBB	**0**	**0**	**0.000**	**1**	**0.139**	**0.005**	**1**	**0.148**	**0.005**	**1**	**0.149**
aaBb	**10**	**0**	**0.000**	**0.5**	**0.139**	**0.060**	**0.5**	**0.148**	**0.063**	**0.5**	**0.149**
aabb	**25**	**0**	**0.250**	**0**	**0.139**	**0.185**	**0**	**0.148**	**0.182**	**0**	**0.149**
	144		**1**			**1**			**1**		

Co-dominant markers in other populations

Marker type	Population						
	F2	P1B1F1	P2B1F1	F1DH	P1BC1DH	P2BC1DH	F1-RIL
AABB	$(1-r)^2/4$	$(1-r)/2$		$(1-r)/2$	$\frac{1}{2}+(1-r)^2/4$	$(1-r)^2/4$	$(1-R)/2$
AABb	$r(1-r)/2$	$r/2$					
AAbb	$r^2/4$			$r/2$	$r/2-r^2/4$	$r/2-r^2/4$	$R/2$
AaBB	$r(1-r)/2$	$r/2$					
AaBb	$(1-2r+2r^2)/2$	$(1-r)/2$	$(1-r)/2$				
Aabb	$r(1-r)/2$		$r/2$				
aaBB	$r^2/4$			$r/2$	$r/2-r^2/4$	$r/2-r^2/4$	$R/2$
aaBb	$r(1-r)/2$		$r/2$				
aabb	$(1-r)^2/4$		$(1-r)/2$	$(1-r)/2$	$(1-r)^2/4$	$\frac{1}{2}+(1-r)^2/4$	$(1-R)/2$

$$R = 2r/(1+2r)$$

More populations (e.g. BC1F2, F3 etc): Generation transition matrix of

Parent	Genotype and frequency in self-pollinated progeny									
	AABB	AABb	AAbb	AaBB	AB/ab	Ab/aB	Aabb	aaBB	aaBb	aabb
AABB	1									
AABb	0.25	0.5	0.25							
AAbb			1							
AaBB	0.25			0.5				0.25		
AB/ab	$(1-r)^2/4$	$r(1-r)/2$	$r^2/4$	$r(1-r)/2$	$(1-r)^2/2$	$r^2/2$	$r(1-r)/2$	$r^2/4$	$r(1-r)/2$	$(1-r)^2/4$
Ab/aB	$r^2/4$	$r(1-r)/2$	$(1-r)^2/4$	$r(1-r)/2$	$r^2/2$	$(1-r)^2/2$	$r(1-r)/2$	$(1-r)^2/4$	$r(1-r)/2$	$r^2/4$
Aabb			0.25				0.5			0.25
aaBB								1		
aaBb								0.25	0.5	0.25
aabb										1

Distortion has little effect on linkage analysis!

DH pop	Theo. Freq.	Distortion	Freq. in distortion
AABB	$f_1=(1-r)/2$	$(1-r)/2$	$(1-r)/(1+s)$
AAbb	$f_2=r/2$	$r/2$	$r/(1+s)$
aaBB	$f_3=r/2$	$s\times r/2$	$r\times s/(1+s)$
aabb	$f_4=(1-r)/2$	$s\times(1-r)/2$	$(1-r)\times s/(1+s)$
Sum	1	$(1+s)/2$	1

$$\hat{r} = r/(1+s) + r\times s/(1+s) = r(1+s)/(1+s) = r$$

Three-point analysis and linkage map construction

Linkage analysis of three markers

$$r_{13} = r_{12} + r_{23} - 2(1-\delta)\ \ r_{12}r_{23}$$

- **When $\delta = 0$ (no interference),**

$$(1-r_{13}) = (1-r_{12})(1-r_{23}) + r_{12}r_{23}$$

$$r_{13} = r_{12}(1-r_{23}) + (1-r_{12})r_{23} = r_{12} + r_{23} - 2r_{12}r_{23}$$

- **When $\delta = 1$ (complete interference),**

$$r_{13} = r_{12} + r_{23}$$

- **The order of the three loci can be determined after linkage analysis (3!/2=3 potential orders)**
  - **1—2—3, or 1—3—2, or 2—1—3**

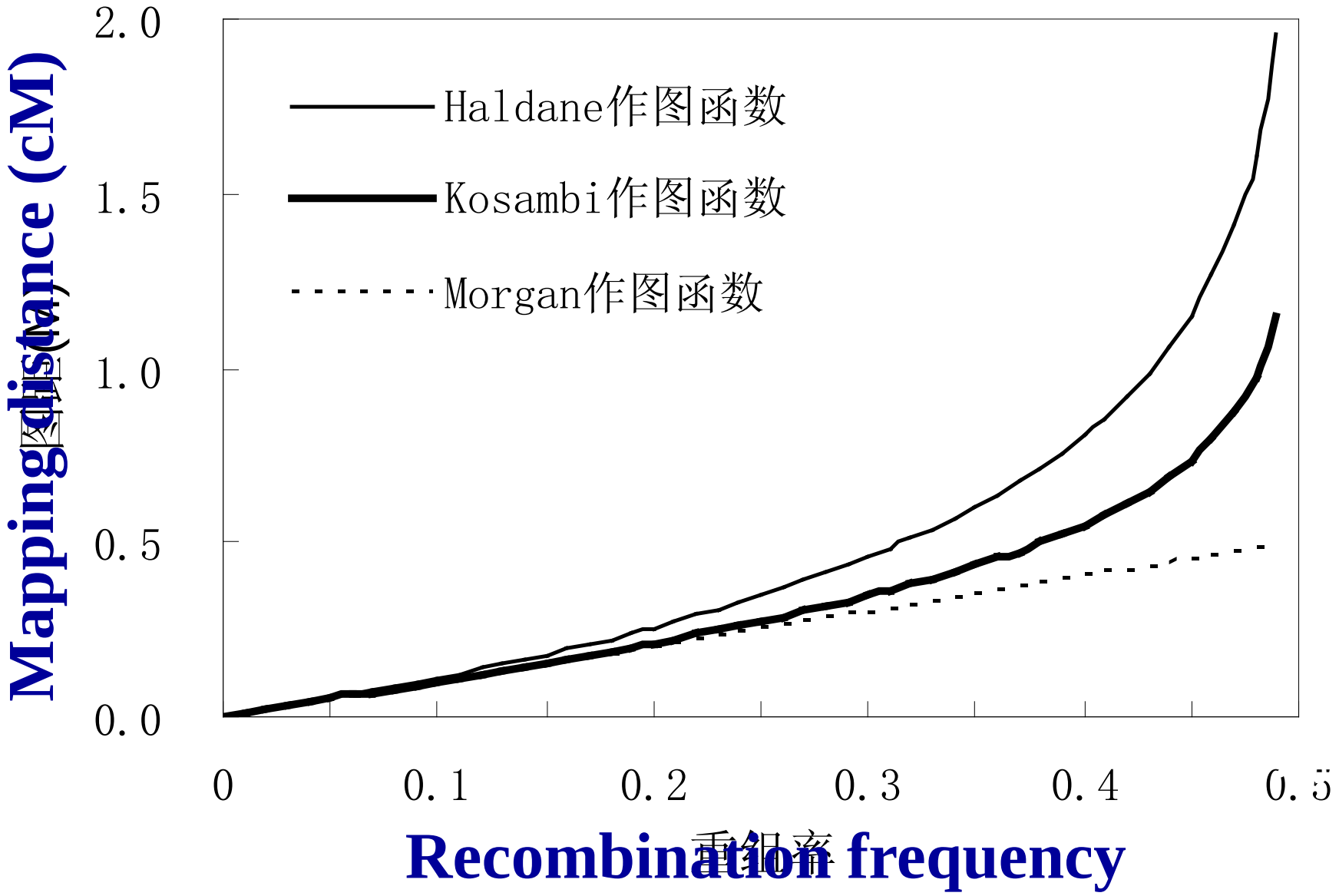
Mapping distance and recombination frequency

- **Mapping distance** $m_{13} = m_{12} + m_{23}$
- **Unit of mapping distance**
  - **M (Morgan) or cM (centi-Morgan), 1M=100cM**
- **The function of mapping distance on recombination frequency (Mapping function):** $m = f(r)$

Common mapping functions

- **Morgan function (complete interference)**
  - ✓**In M: *m* =*r* (M)**
  - ✓**In cM: *m* =*r* ×100 (cM)**
- **Haldane function (no interference)**
  - ✓**In M:** $m = f(r) = -\frac{1}{2}\ln(1-2r) \quad r = \frac{1}{2}(1-e^{-2m})$
  - ✓**In cM:** $m = f(r) = -50\ln(1-2r) \quad r = \frac{1}{2}(1-e^{-m/50})$
- **Kosambi function (interference depends on length of interval)**
  - ✓**In M:** $m = \frac{1}{4}\ln\frac{1+2r}{1-2r} \quad r = \frac{1}{2}\frac{e^{4m}-1}{e^{4m}+1}$
  - ✓**In cM:** $m = 25\ln\frac{1+2r}{1-2r} \quad r = \frac{1}{2}\frac{e^{m/25}-1}{e^{m/25}+1}$

Comparison of the three functions



Three steps in linkage map construction

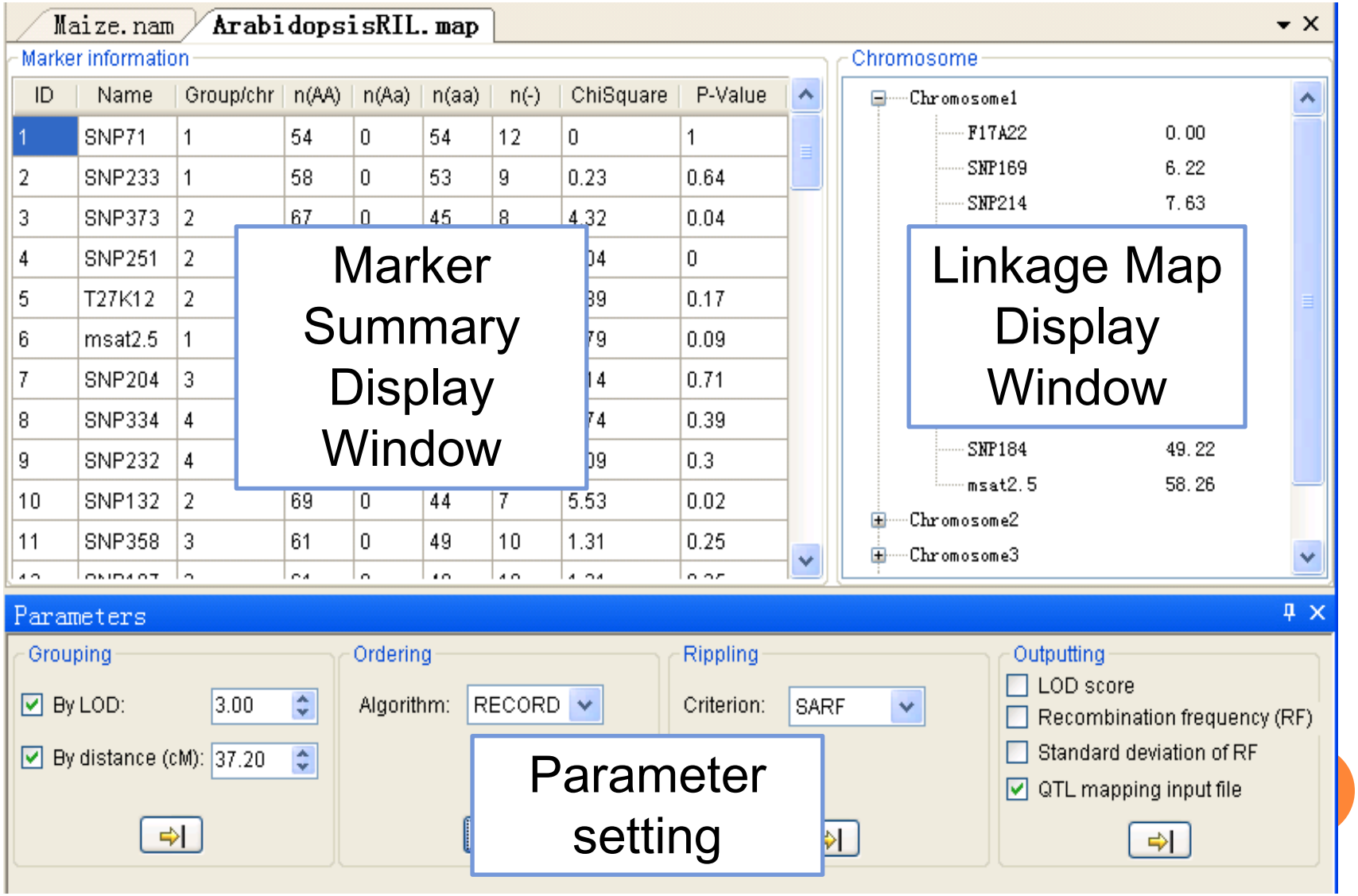
- **Step 1: Grouping.** Grouping can be based on
  - (i) a threshold of LOD score
  - (ii) a threshold of marker distance (cM)
  - (iii) anchor information
- **Step 2: Ordering.** Three ordering algorithms are
  - (i) SER: SERiation (Buetow and Chakravarti, 1987. Am J Hum Genet 41:180–188)
  - (ii) RECORD: REcombination Counting and ORDering (Van Os et al., 2005. Theor Appl Genet 112: 30–40)
  - (iii) nnTwoOpt: nearest neighbor was used for tour construction, and two-opt was used for tour improvement, similar to Travelling Salesman Problem (TSP) (Lin and Kernighan, 1973. Oper. Res. 21: 498–516.

Three steps in linkage map construction

- Due to the large number of markers (n), it is impossible to compare all possible orders (say n=50, possible orders are $n!/2=1.52x10^{64}$). Orders from the above algorithms are regional optimizations.
- **Step 3: Rippling.** Five rippling criteria are
  - (i) SARF (Sum of Adjacent Recombination Frequencies)
  - (ii) SAD (Sum of Adjacent Distances)
  - (iii) SALOD (Sum of Adjacent LOD scores)
  - (iv) COUNT (number of recombination events)

The MAP functionality in QTL IciMapping

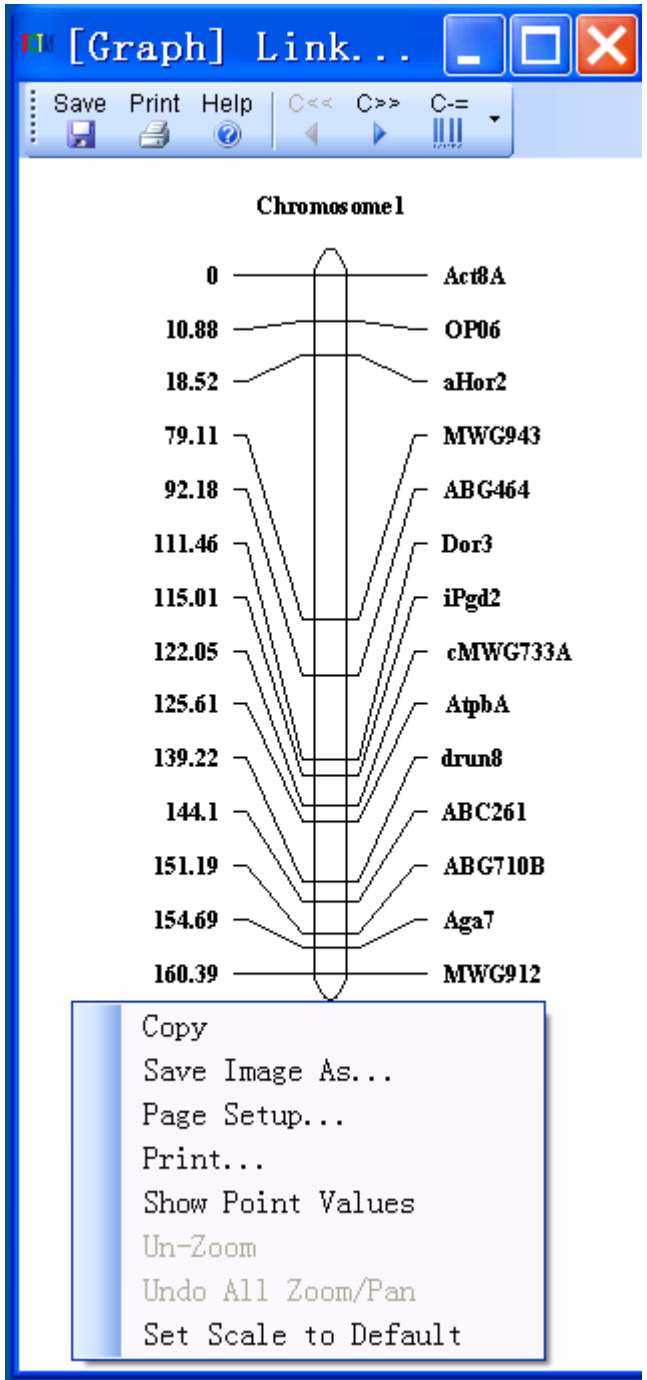
Interface of the MAP functionality



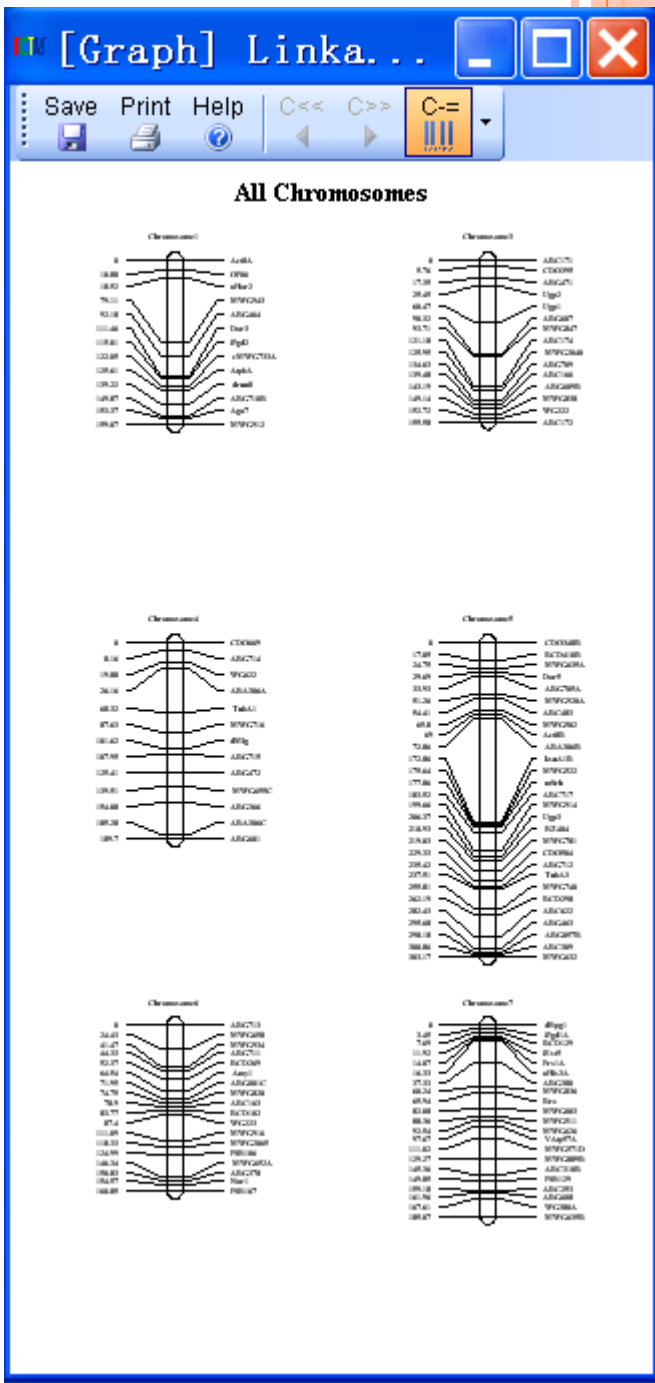
Map outputs:

Linkage map for each chromosome (A) or all chromosomes (B)

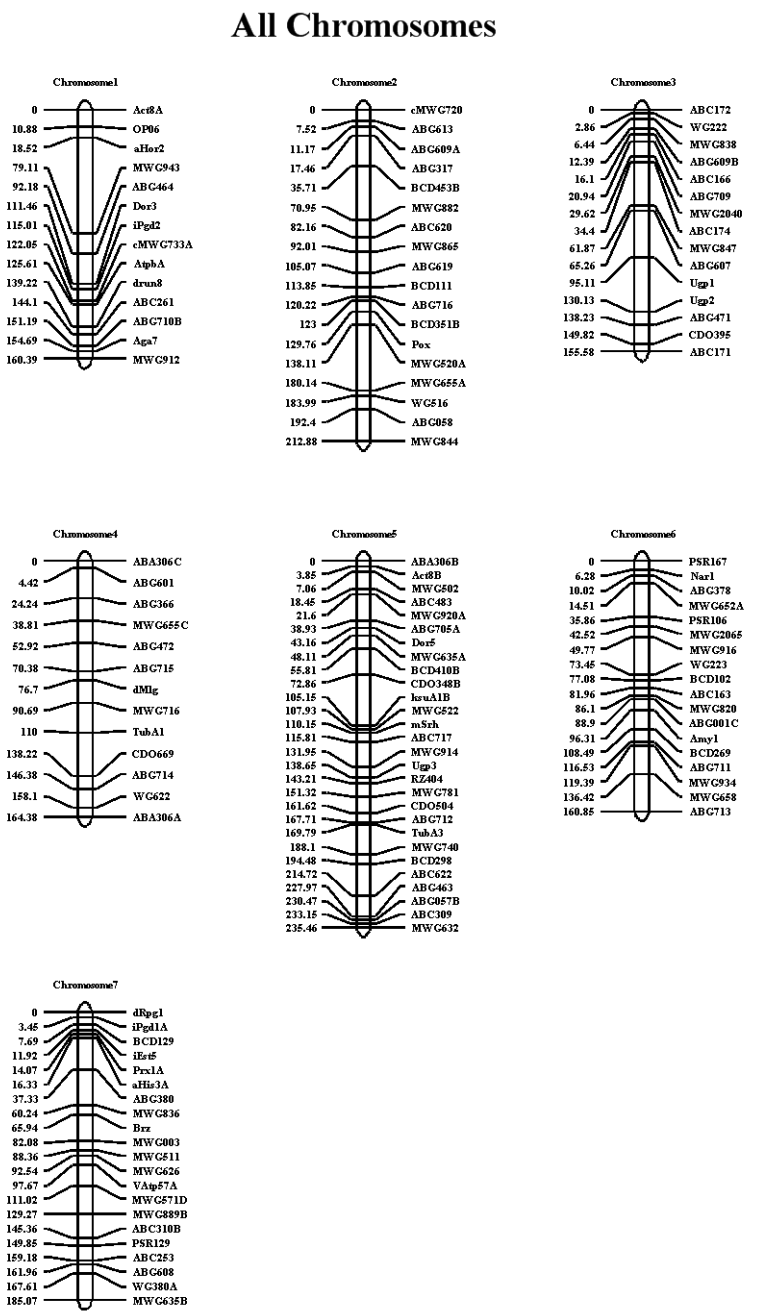
A. Map of one chromosome



B. Map of all chromosomes



An example map of seven chromosomes or groups



Linkage map and physical map

Species	Size of haploid genome (kb)	Size of linkage map (cM)	kb/cM
Yeast	2.2×10^4	3700	6
Neurospora	4.2×10^4	500	80
Arabidopsis	7.0×10^4	500	140
Drosophila	2.0×10^5	290	700
Tomato	7.2×10^5	1400	510
Human	3.0×10^6	2710	1110
Wheat	1.6×10^7	2575	6214
Rice	4.4×10^5	1575	279
Corn	3.0×10^6	1400	2140