Three-point Linkage Analysis and Map Construction

Jiankang Wang

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

Web: http://www.isbreeding.net

Outlines of the presentation

- Comparison of the Estimated Recombination Frequency in Bi-Parental Populations
- Linkage analysis of three markers and map construction
- Marker categories and coding criteria
- Removal of redundancy

Comparison of the Estimated Recombination Frequency in Bi-Parental Genetic Populations

Sun Z., H. Li*, L. Zhang, J. Wang. 2012. Estimation of recombination frequency in biparental genetic populations. **Genetics Research** 94: 163-177

The genetic analysis can be very complicated even with biparental populations!

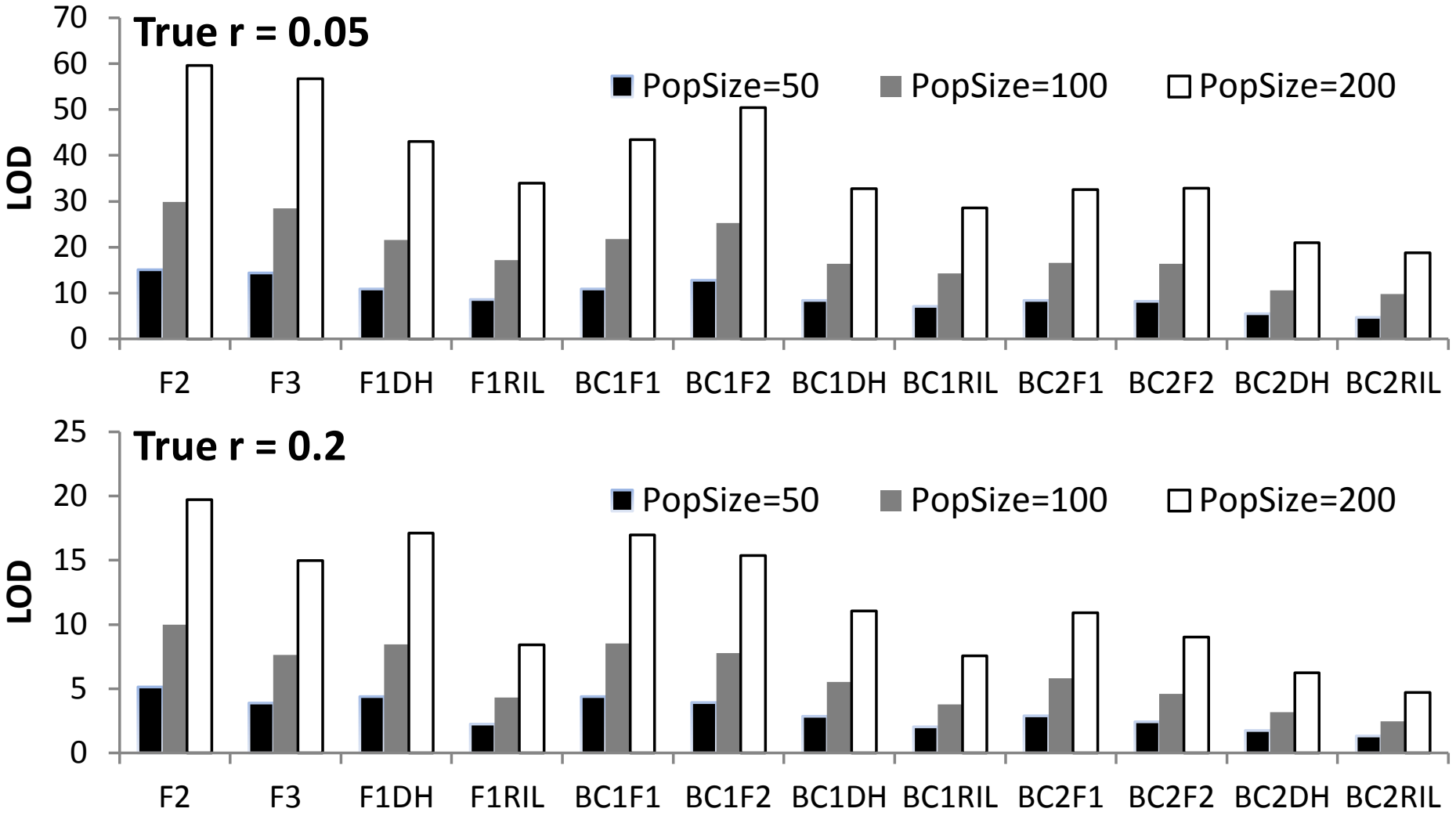
- F1-derived populations
  - F2, F3, DH, RIL: p=0.5, q=0.5 at each locus
- P1BC1-derived population
  - F2, F3, DH, RIL: p=0.75, q=0.25 at each locus
- P2BC1-derived population
  - F2, F3, DH, RIL: p=0.25, q=0.75 at each locus
- P1BC2-derived population
  - F2, F3, DH, RIL: p=0.875, q=0.125 at each locus
- P2BC2-derived population
  - F2, F3, DH, RIL: p=0.125, q=0.875 at each locus

Twenty biparental populations

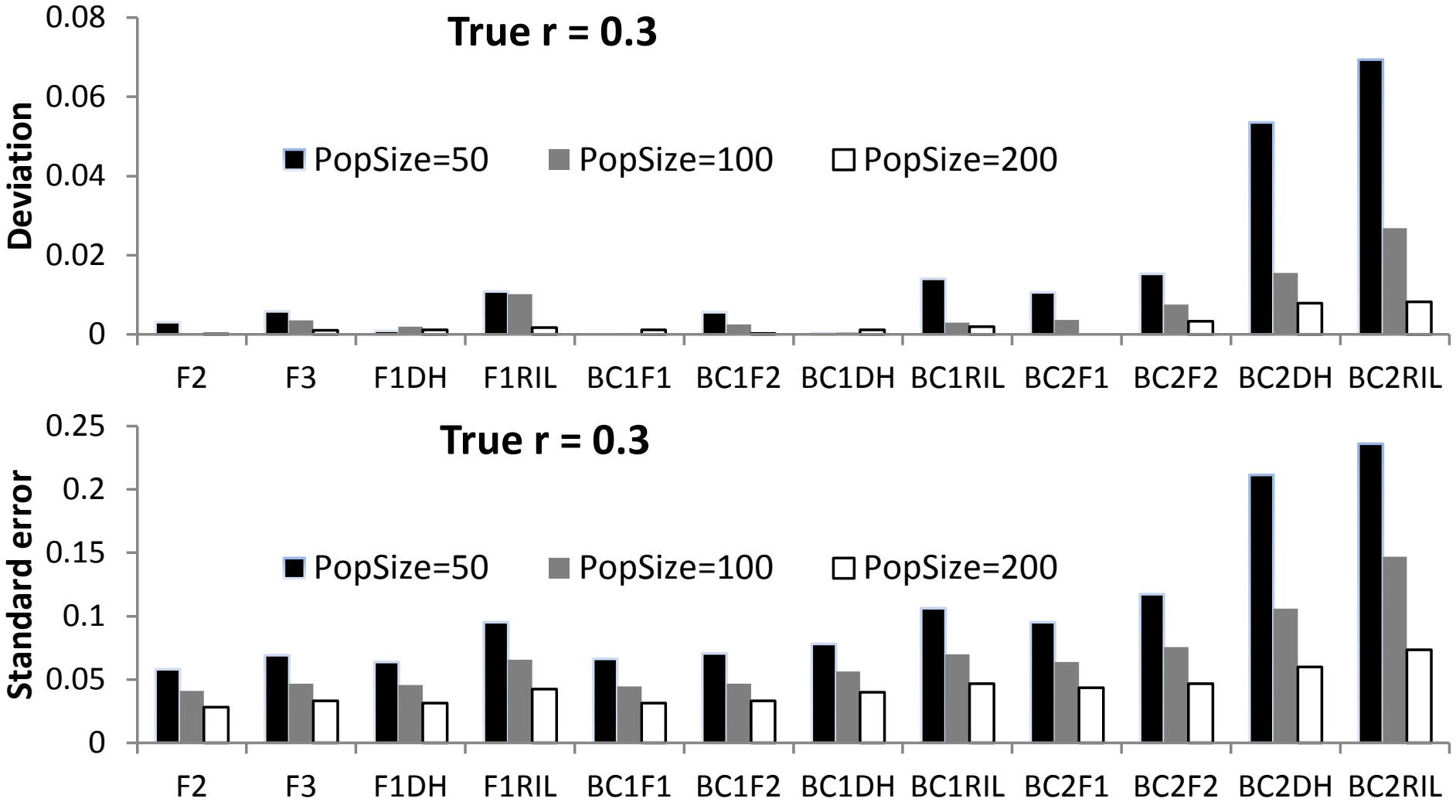
- Allele frequencies can be different
- Genotypes and their frequencies are much different
- Are they equal good in estimating the recombination frequency between two linked loci?

LOD scores from different populations

True r=0.05 (Upper), True r=0.2 (lower)



Deviations to the true value (upper) and standard errors (lower) of estimated recombination frequency



Observations

- When two alleles at each locus have equal frequency of 0.5, we had a better estimation.
- When a population has more genotypes, we had a better estimation.
- For F2 and F3 to be efficient, we need co-dominant markers.

Minimum population size to have at least one recombinant

Pop.	*r*=0.01	*r*=0.02	*r*=0.03	*r*=0.05	*r*=0.1	*r*=0.2	*r*=0.3
F_2 (C, C)	150	75	50	30	15	8	5
F_2 (C, D)	299	149	99	60	31	16	11
F_2 (C, R)	299	149	99	60	31	16	11
F_2 (D, D)	299	149	99	61	31	16	11
F_2 (D, R)	149786	29956	13616	4754	1197	299	132
F_2 (R, R)	299	149	99	61	31	16	11
DH	299	149	99	59	29	14	9
RIL	152	77	52	32	17	9	7

In the first column, C for co-dominant marker; D for dominant marker; R for recessive marker

Linkage analysis of three markers and map construction

Coefficient of interference

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

- **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**

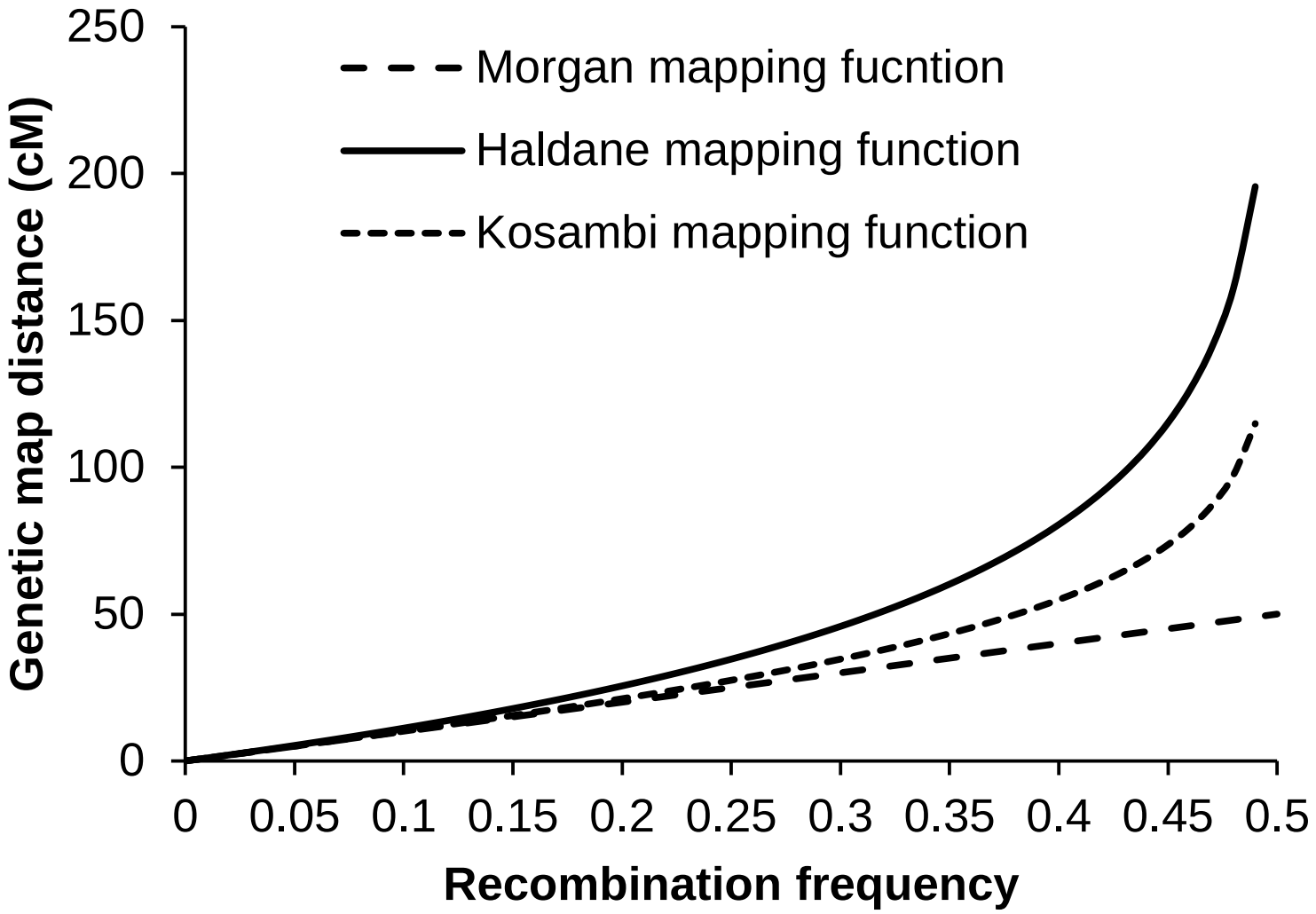
Mapping distance and recombination frequency

- **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 map construction

- **Step 1: Grouping.** Grouping can be based on
  - (i) a threshold of LOD score
  - (ii) a threshold of recombination frequency
  - (ii) a threshold of marker distance (cM)
  - (iv) anchor information
  - (v) a given number of group (to be added)

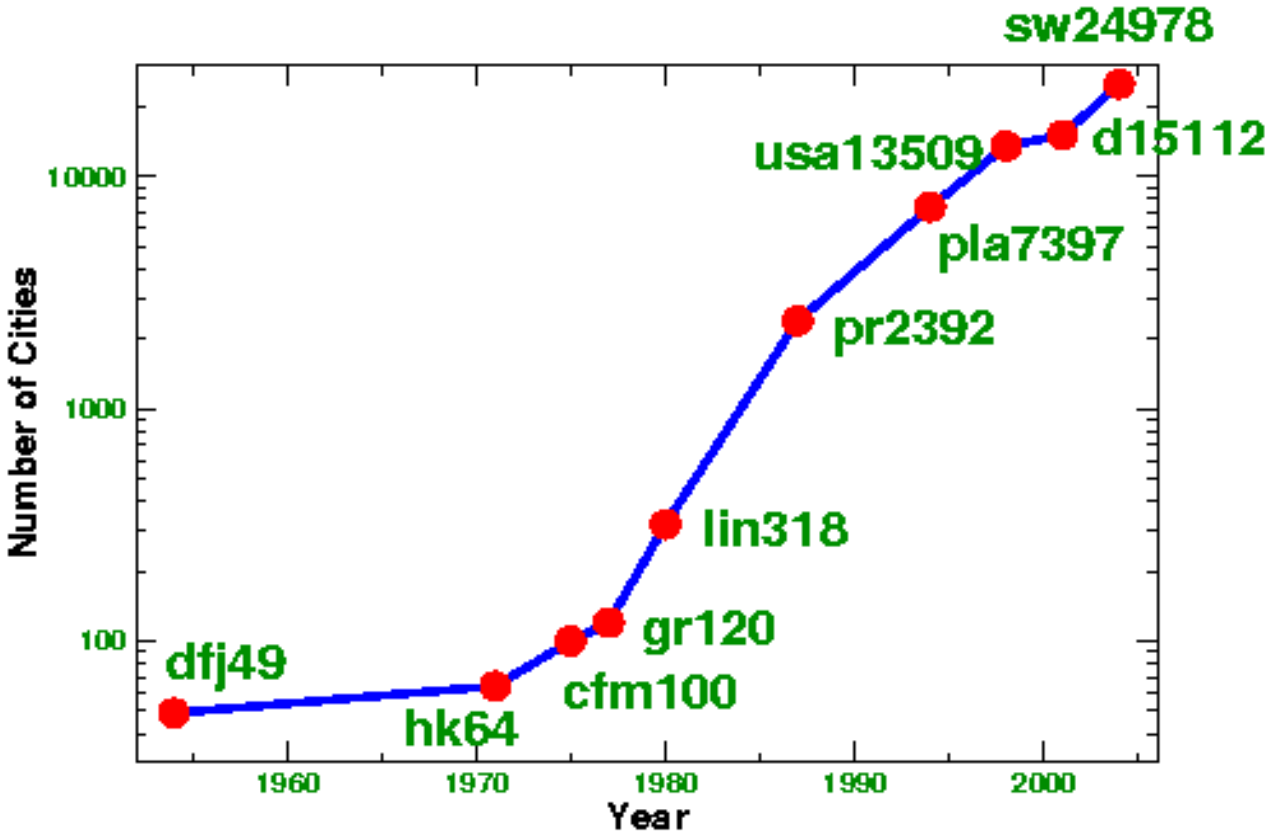
Three steps in map construction

- **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.
  - By Input, the order in input file will be used. Say we know the order from physical map for GBS markers
  - By Anchor Order: Only order those markers with no anchor information. No effect on the anchor marker order.

Travelling Salesman Problem (TSP)

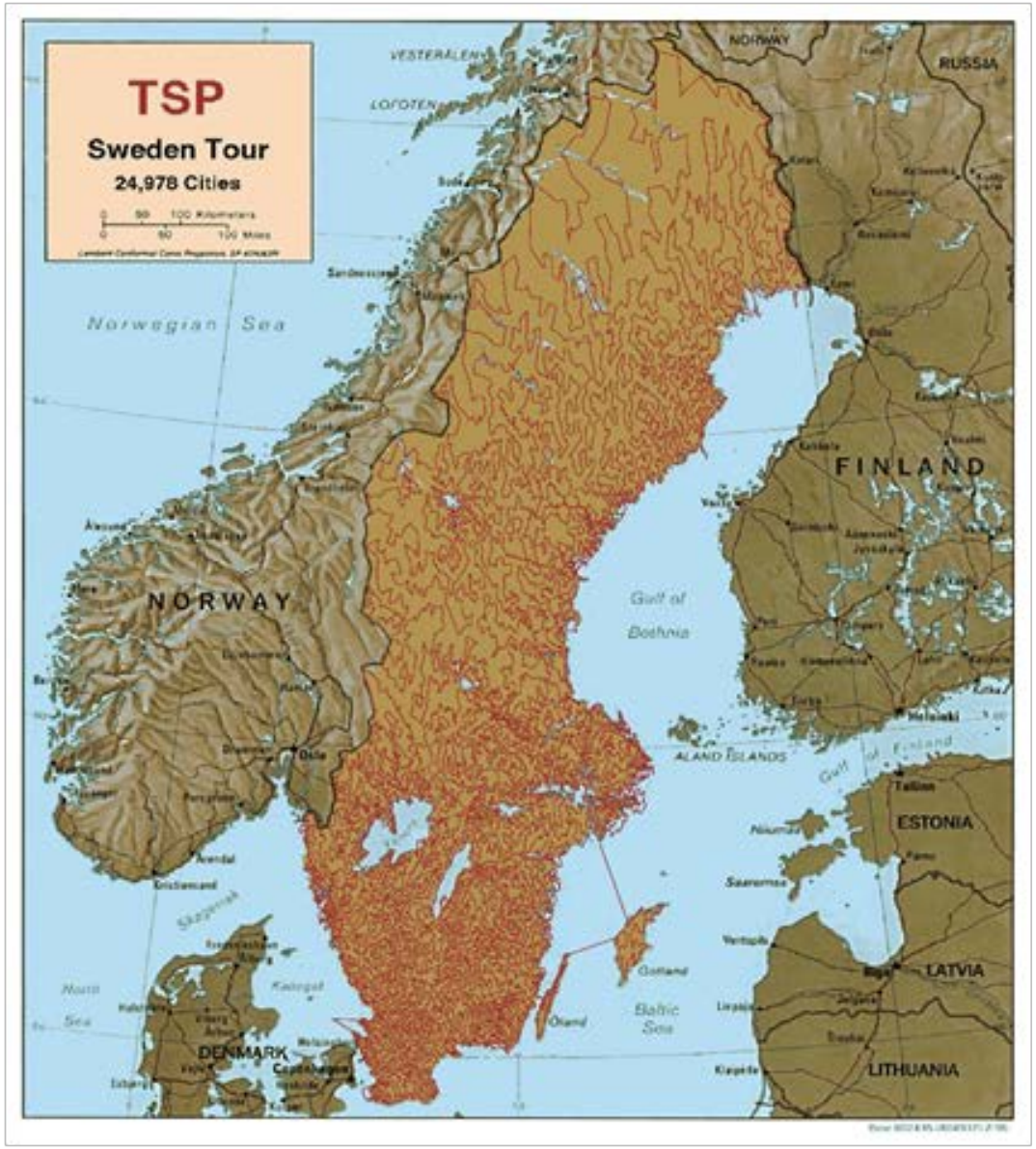
- A salesman is required to visit each of n given cities once and only once, starting from any city and returning to the original place of departure. What tour he choose in order to minimize his total travel distance?
- The distance between any pair of cities are assumed to be known by the salesman. Distance can be replaced by another notion, such as time or money.
- TSP is one of the most widely studied problems in combinatorial optimization. It is easy to state, but hard to solve! **TSP is an NP-hard problem, i.e. non-deterministic polynomial-time hard.**

Finding the solutions



- TSP is represented by some letters plus the number of cities. For example, there are 24978 Sweden cities in TSP “sw24978”.

The solution for TSP “sw24978”



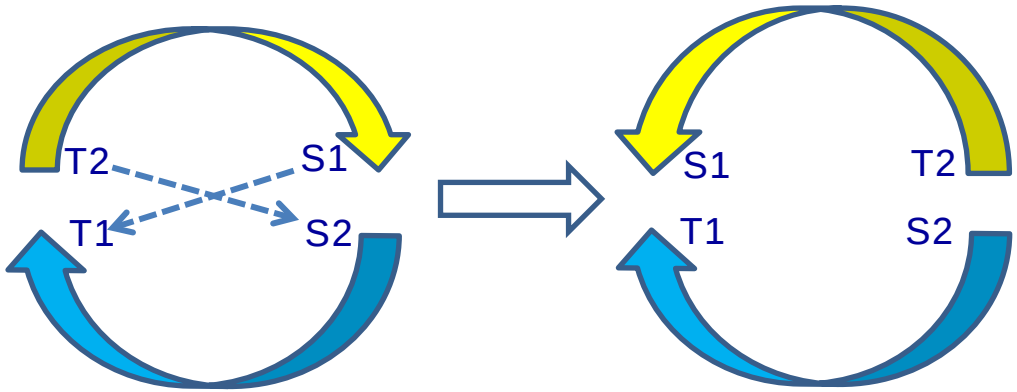
Approximate algorithm of TSP

- Tour construction algorithms
- Tour improvement algorithms
- Composite algorithms

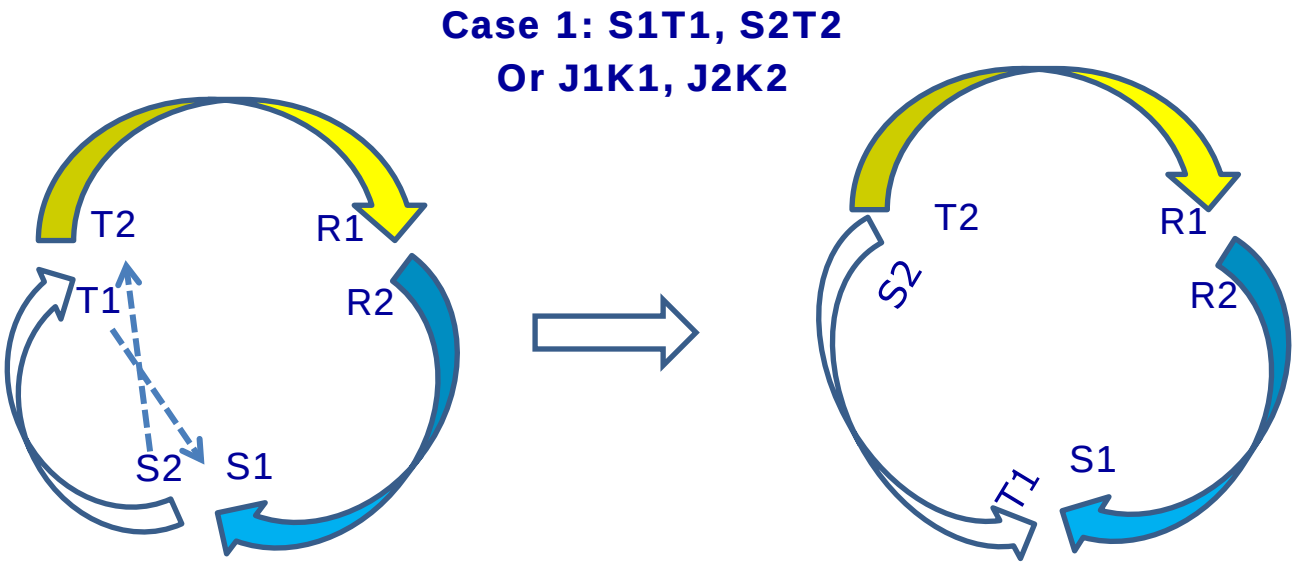
Nearest-neighbor (nn) algorithm for tour construction

- A simple algorithm for tour construction
- Start in an arbitrary city. As long as there are cities, that have not yet been visited, visit the nearest city that still not appeared in the tour. Finally, return to the first city.
- This approach is simple, but often too greedy!
- The first distances in the construction process are reasonable short, whereas the distance at the end of the process usually will be rather long.

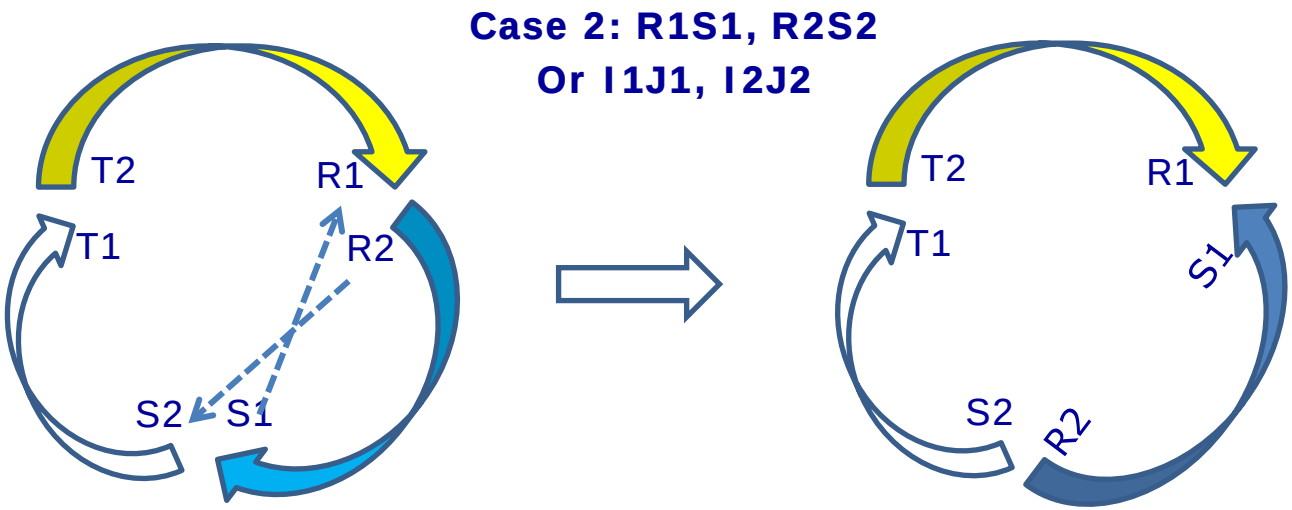
Two-Opt algorithm for tour improvement (Lin and Kernighan, 1973)



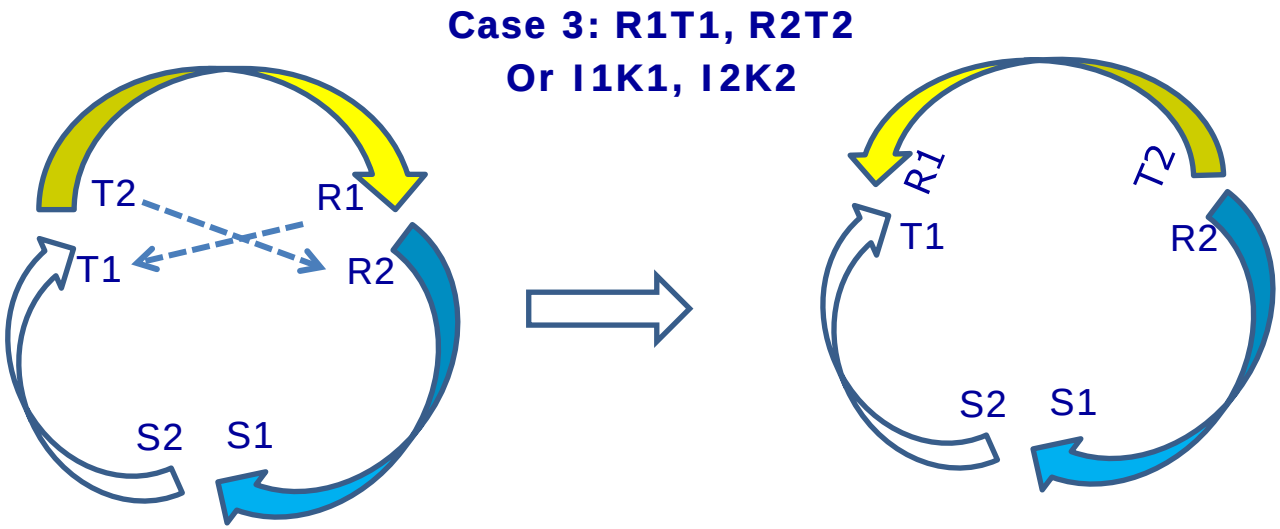
Three-Opt algorithm for tour improvement: Case 1



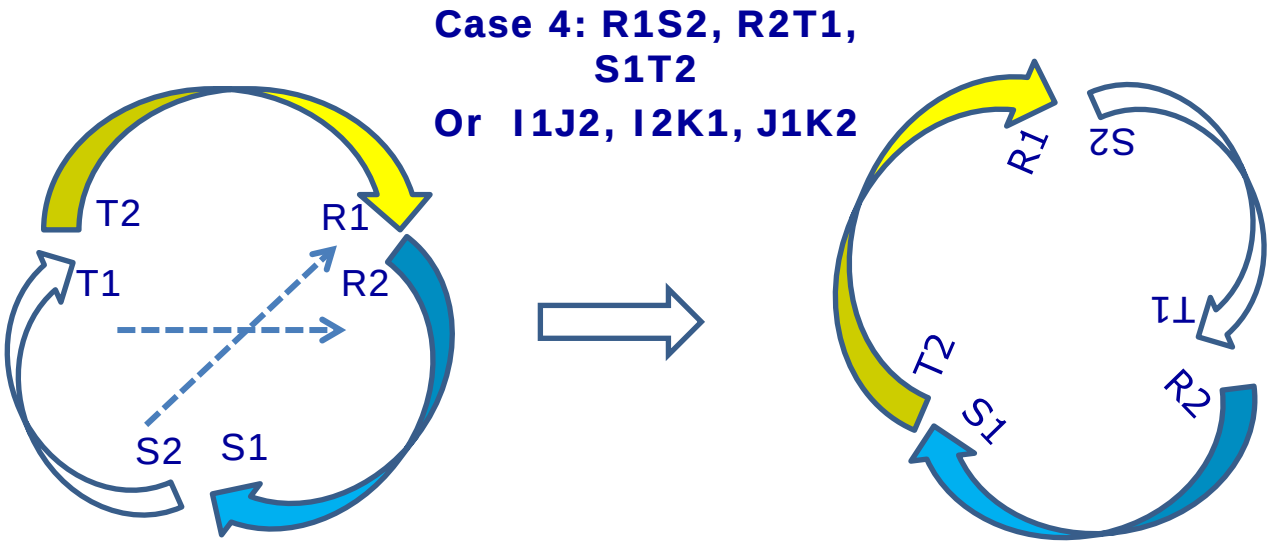
Three-Opt algorithm for tour improvement: Case 2



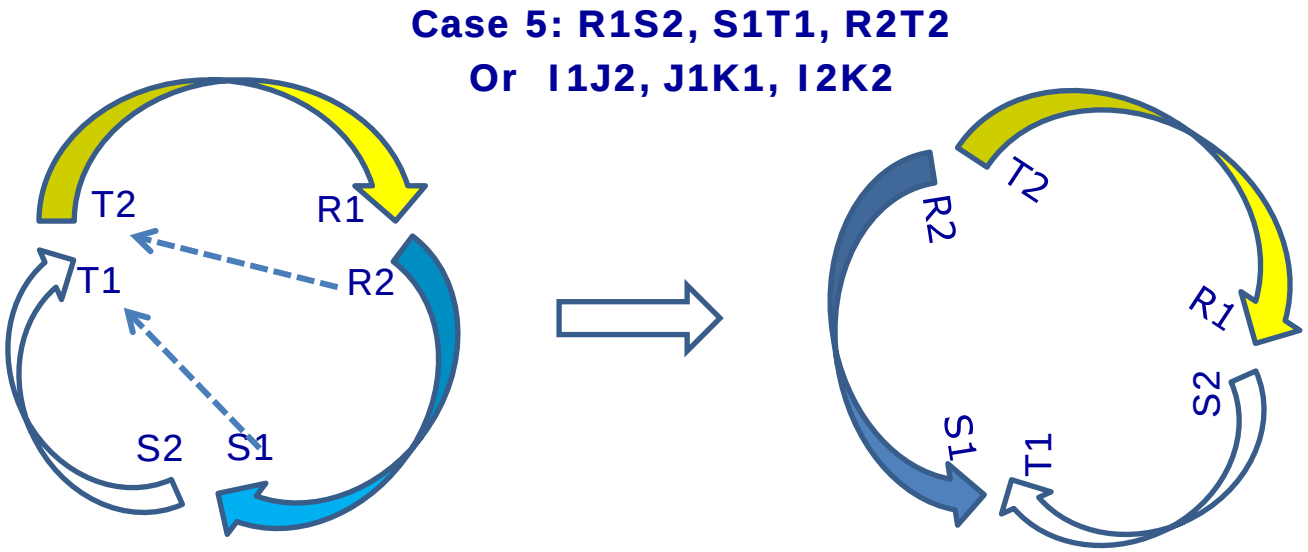
Three-Opt algorithm for tour improvement: Case 3



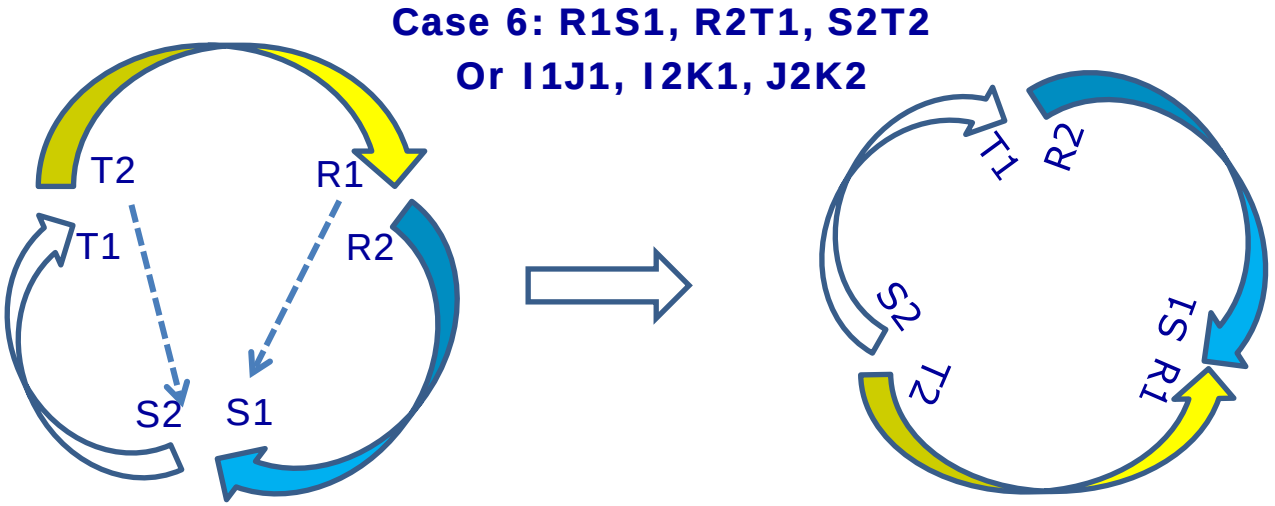
Three-Opt algorithm for tour improvement: Case 4



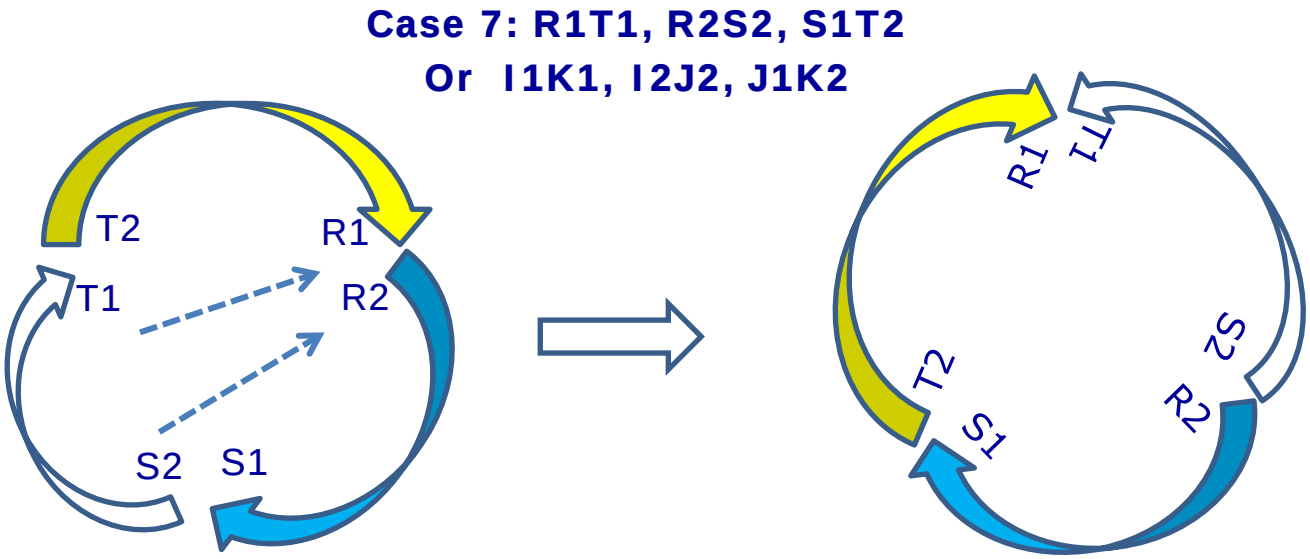
Three-Opt algorithm for tour improvement: Case 5



Three-Opt algorithm for tour improvement: Case6



Three-Opt algorithm for tour improvement: Case 7



Difference between TSP route and genetic linkage map

- TSP is a closed route. There is no starting and ending points.
- Linkage map is an open route. It has a starting point and an ending point.
- Criteria for the shortest map
  - Convert a TSP route to a map by breaking the TSP route from the longest interval
  - Map length = TSP route length – the longest interval
- The above criteria are used to determine whether a shorter map is found

Three steps in 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 only criteria: shortest map!**

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

Marker/Gene Categories and Coding Criteria

Coding Criteria of Co-dominant Markers

(Two parental bands are both present in F1)

	Two homozygous parental lines A and B and their F1 hybrid			**Populations with heterozygousity, i.e., 1, 2, 7, 8, 9, 10, 13, 14, 15, 16**			**Populations with no heterozygousity, i.e., 3, 4, 5, 6, 11, 12, 17, 18, 19, 20**	
	Parent A	Parent B	F1	Type A	Type B	Type H	Type A	Type B
Coding by numbers	2	0	1	2	0	1	2	0
Coding by letters	A AA	B BB	H AB	A AA	B BB	H AB BA	A AA	B BB

Coding by numbers: The two parental bands are coded as 2 and 0. Both are present in F1, which is coded as 1. The coding number can be viewed as the number of parent A allele. When heterozygote is present in a population, all three numbers 2, 1, and 0 could be present. When heterozygote is absent, only numbers 2, and 0 are present.

Coding by letters: Parent A is coded as A or AA; Parent B is coded as B or BB; Their F1 hybrid is coded as H, AB or BA.

Missing values: Missing values of marker type are coded as -1, X, XX, *, or **.

Mixed coding: We recommend either numbers or letters (not both) be used in coding. But, mixed coding is acceptable in QTL IciMapping software. Taking F2 population as an example , some individuals could be coded as 2, some coded as A, some coded as BA, and some coded as AA etc.

When letters are used in coding, only the capital case is acceptable.

Coding Criteria of Dominant Markers

(F1 shows the same band as Parent A)

	Two homozygous parental lines A and B and their F1 hybrid			**Populations with heterozygousity, i.e., 1, 2, 7, 8, 9, 10, 13, 14, 15, 16**			**Populations with no heterozygousity, i.e., 3, 4, 5, 6, 11, 12, 17, 18, 19, 20**	
	Parent A	Parent B	F1	Type A	Type B	Type H	Type A	Type B
Coding by numbers	2	0	1	12	0	12	2	0
Coding by letters	A AA	B BB	H AB	AH AX A* A_	B BB	HA XA *A _A	A AA	B BB

Coding by numbers: F1 shows the same band as Parent A. When heterozygote is present, Type A and Type H cannot be separated, and are coded as 12. When heterozygote is absent, Type A can be clearly identified and coded as 2. Similar to co-dominant markers, Type B is coded as 0.

Coding by letters: When Type A and Type H cannot be separated, they are coded as AH, HA, AX, XA, A*, *A, A_ or _A. Similar to co-dominant markers, Type B is coded as B or BB.

Missing values: Missing values of marker type are coded as -1, X, XX, *, or **. Please be noted that missing values may not be separated from Type B when markers are dominant .

Mixed coding: Taking F2 population as an example, some individuals could be coded as 12, some coded as AX, some coded as BB, some coded as A*, and some coded as _A, etc.

When letters are used in coding, only the capital case is acceptable.

Coding Criteria of Recessive Markers

(F1 shows the same band as Parent B)

	Two homozygous parental lines A and B and their F1 hybrid			Populations with heterozygousity, i.e., 1, 2, 7, 8, 9, 10, 13, 14, 15, 16			Populations with no heterozygousity, i.e., 3, 4, 5, 6, 11, 12, 17, 18, 19, 20	
	Parent A	Parent B	F1	Type A	Type B	Type H	Type A	Type B
Coding by numbers	2	0	1	2	10	10	2	0
Coding by letters	A AA	B BB	H AB	A AA	BH BX B* B_	HB XB *B _B	A AA	B BB

Coding by numbers: F1 shows the same band as Parent B. When heterozygote is present, Type B and Type H cannot be separated, and are coded as 10. When heterozygote is absent, Type B can be clearly identified and coded as 0. Similar to co-dominant markers, Type A is coded as 2.

Coding by letters: When Type B and Type H cannot be separated, they are coded as BH, HB, BX, XB, B*, *B, B_ or _B. Similar to co-dominant markers, Type A is coded as A or AA.

Missing values: Missing values of marker type are coded as -1, X, XX, *, or **. Please be noted that missing may not be separated from Type A when markers are recessive.

Mixed coding: Taking F2 population as an example, some individuals could be coded as 10, some coded as BX, some coded as AA, some coded as B*, and some coded as _B, etc.

When letters are used in coding, only the capital case is acceptable.

Summary of Marker Coding Criteria

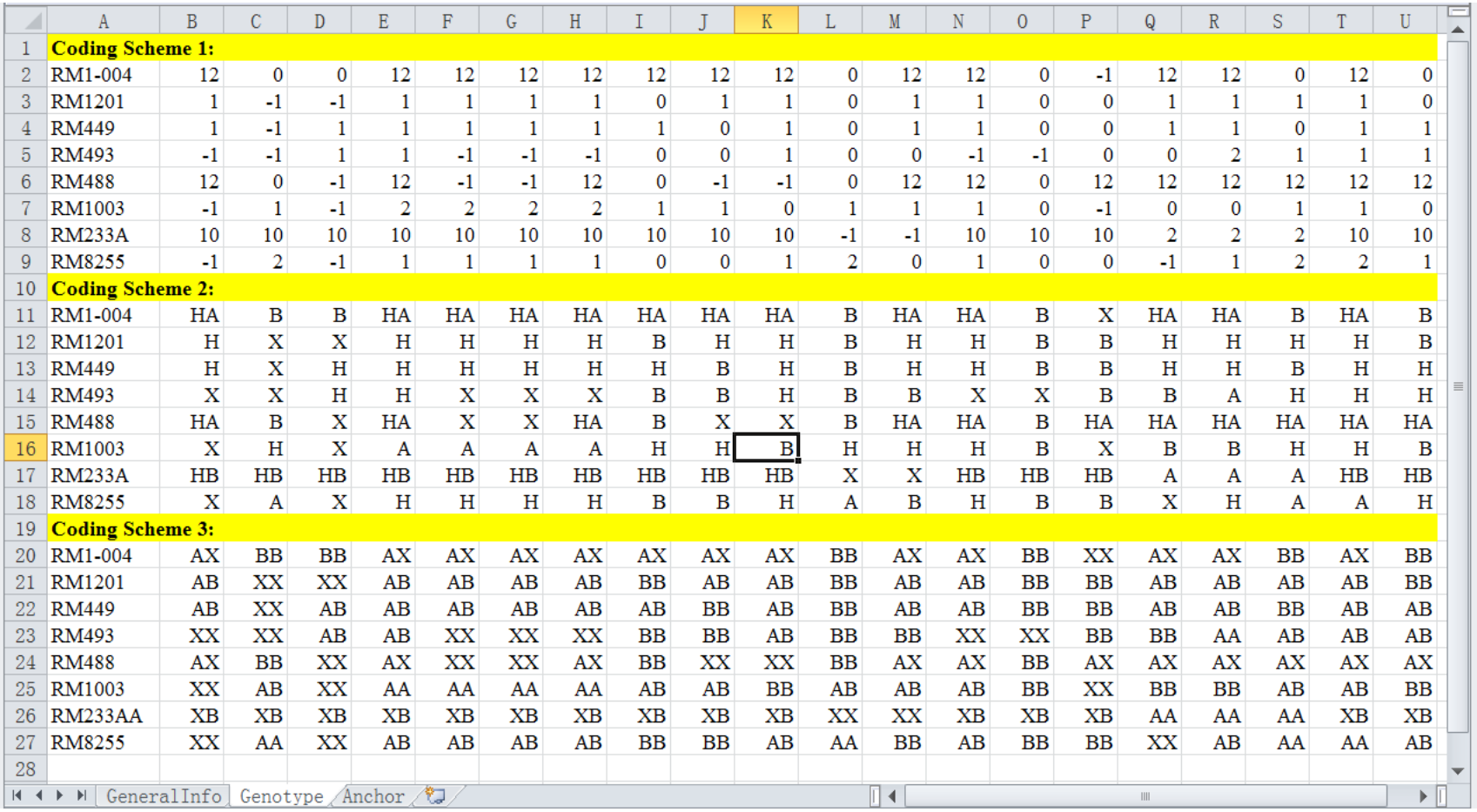
Coding summary: Assuming AA is the genotype of Parent A, BB is the genotype of Parent B, and AB is the genotype of their F1 hybrid. When one marker is dominant, AB+AA represents the two non-separated genotypes AB and AA. When one marker is recessive, AB+BB represents the two non-separated genotypes AB and BB. In total, we may have six possible types in bi-parental populations. Accepted coding options for the six types are summarized in the following table.

Type	AA	AB	BB	Missing	AB+AA	AB+BB
Coding options	2 A AA	1 H AB BA	0 B BB	-1 X XX * **	12 AH, HA AX, XA A*, *A A_, _A	10 BH, HB BX, XB B*, *B B_, _B
Scheme 1	2	1	0	-1	12	10
Scheme 2	A	H	B	X	HA	HB
Scheme 3	AA	AB	BB	XX	AX	XB

Three commonly-adopted coding schemes are given in the above table. Please be noted that not all the six types could be present in one population. One marker can only be either co-dominant, or dominant, or recessive. Assuming Scheme 2 is applied. For a co-dominant marker in an F2 population, A, H, B, and X are the possible values. When HA is present in this marker locus, the software will report an error, and ask the user to correct the error.

Separators: In EXCEL, each value takes one cell. Each marker is arranged in one row, first the marker name, followed by marker types of all lines in the population. In TEXT, values can be separated by Space, Comma, or Tab. An illustrated example can be found in the next slide.

Marker Types of the Three Coding Schemes in EXCEL Using 8 Marker Loci and 20 Individuals as an example



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	**Coding Scheme 1:**																				
2	RM1-004	12	0	0	12	12	12	12	12	12	12	0	12	12	0	-1	12	12	0	12	0
3	RM1201	1	-1	-1	1	1	1	1	0	1	1	0	1	1	0	0	1	1	1	1	0
4	RM449	1	-1	1	1	1	1	1	1	0	1	0	1	1	0	0	1	1	0	1	1
5	RM493	-1	-1	1	1	-1	-1	-1	0	0	1	0	0	-1	-1	0	0	2	1	1	1
6	RM488	12	0	-1	12	-1	-1	12	0	-1	-1	0	12	12	0	12	12	12	12	12	12
7	RM1003	-1	1	-1	2	2	2	2	1	1	0	1	1	1	0	-1	0	0	1	1	0
8	RM233A	10	10	10	10	10	10	10	10	10	10	-1	-1	10	10	10	2	2	2	10	10
9	RM8255	-1	2	-1	1	1	1	1	0	0	1	2	0	1	0	0	-1	1	2	2	1
10	**Coding Scheme 2:**																				
11	RM1-004	HA	B	B	HA	HA	HA	HA	HA	HA	HA	B	HA	HA	B	X	HA	HA	B	HA	B
12	RM1201	H	X	X	H	H	H	H	B	H	H	B	H	H	B	B	H	H	H	H	B
13	RM449	H	X	H	H	H	H	H	H	B	H	B	H	H	B	B	H	H	B	H	H
14	RM493	X	X	H	H	X	X	X	B	B	H	B	B	X	X	B	B	A	H	H	H
15	RM488	HA	B	X	HA	X	X	HA	B	X	X	B	HA	HA	B	HA	HA	HA	HA	HA	HA
16	RM1003	X	H	X	A	A	A	A	H	H	B	H	H	H	B	X	B	B	H	H	B
17	RM233A	HB	HB	HB	HB	HB	HB	HB	HB	HB	HB	X	X	HB	HB	HB	A	A	A	HB	HB
18	RM8255	X	A	X	H	H	H	H	B	B	H	A	B	H	B	B	X	H	A	A	H
19	**Coding Scheme 3:**																				
20	RM1-004	AX	BB	BB	AX	AX	AX	AX	AX	AX	AX	BB	AX	AX	BB	XX	AX	AX	BB	AX	BB
21	RM1201	AB	XX	XX	AB	AB	AB	AB	BB	AB	AB	BB	AB	AB	BB	BB	AB	AB	AB	AB	BB
22	RM449	AB	XX	AB	AB	AB	AB	AB	AB	BB	AB	BB	AB	AB	BB	BB	AB	AB	BB	AB	AB
23	RM493	XX	XX	AB	AB	XX	XX	XX	BB	BB	AB	BB	BB	XX	XX	BB	BB	AA	AB	AB	AB
24	RM488	AX	BB	XX	AX	XX	XX	AX	BB	XX	XX	BB	AX	AX	BB	AX	AX	AX	AX	AX	AX
25	RM1003	XX	AB	XX	AA	AA	AA	AA	AB	AB	BB	AB	AB	AB	BB	XX	BB	BB	AB	AB	BB
26	RM233AA	XB	XB	XB	XB	XB	XB	XB	XB	XB	XB	XX	XX	XB	XB	XB	AA	AA	AA	XB	XB
27	RM8255	XX	AA	XX	AB	AB	AB	AB	BB	BB	AB	AA	BB	AB	BB	BB	XX	AB	AA	AA	AB
28																					

GeneralInfo | Genotype | Anchor

Marker Types of the Three Coding Schemes in TEXT Using 8 Marker Loci and 20 Individuals as an example (Spaces are used to separate the values)

Coding Scheme 1:

RM1-004	12	0	0	12	12	12	12	12	12	12	0	12	12	0	-1	12	12	0	12	0
RM1201	1	-1	-1	1	1	1	1	0	1	1	0	1	1	0	0	1	1	1	1	0
RM449	1	-1	1	1	1	1	1	1	0	1	0	1	1	0	0	1	1	0	1	1
RM493	-1	-1	1	1	-1	-1	-1	0	0	1	0	0	-1	-1	0	0	2	1	1	1
RM488	12	0	-1	12	-1	-1	12	0	-1	-1	0	12	12	0	12	12	12	12	12	12
RM1003	-1	1	-1	2	2	2	2	1	1	0	1	1	1	0	-1	0	0	1	1	0
RM233A	10	10	10	10	10	10	10	10	10	10	-1	-1	10	10	10	2	2	2	10	10
RM8255	-1	2	-1	1	1	1	1	0	0	1	2	0	1	0	0	-1	1	2	2	1

Coding Scheme 2:

RM1-004	HA	B	B	HA	HA	HA	HA	HA	HA	HA	B	HA	HA	B	X	HA	HA	B	HA	B
RM1201	H	X	X	H	H	H	H	B	H	H	B	H	H	B	B	H	H	H	H	B
RM449	H	X	H	H	H	H	H	H	B	H	B	H	H	B	B	H	H	B	H	H
RM493	X	X	H	H	X	X	X	B	B	H	B	B	X	X	B	B	A	H	H	H
RM488	HA	B	X	HA	X	X	HA	B	X	X	B	HA	HA	B	HA	HA	HA	HA	HA	HA
RM1003	X	H	X	A	A	A	A	H	H	B	H	H	H	B	X	B	B	H	H	B
RM233A	HB	HB	HB	HB	HB	HB	HB	HB	HB	HB	X	X	HB	HB	HB	A	A	A	HB	HB
RM8255	X	A	X	H	H	H	H	B	B	H	A	B	H	B	B	X	H	A	A	H

Coding Scheme 3:

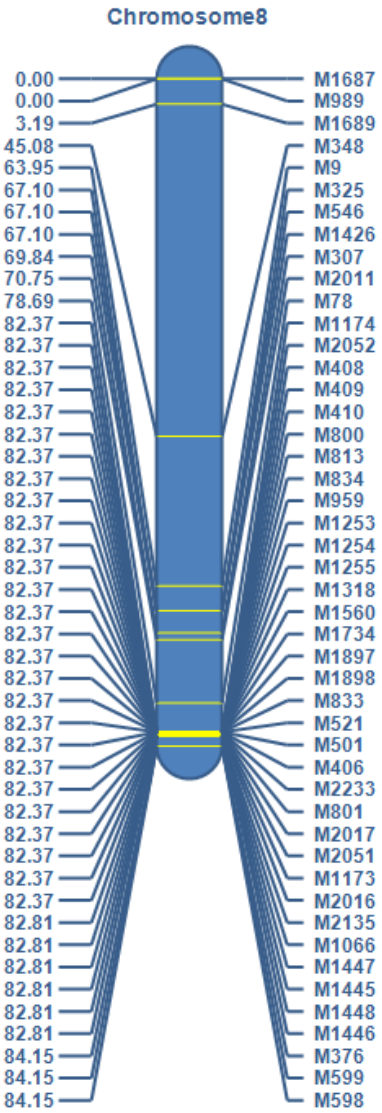
RM1-004	AX	BB	BB	AX	AX	AX	AX	AX	AX	AX	BB	AX	AX	BB	XX	AX	AX	BB	AX	BB
RM1201	AB	XX	XX	AB	AB	AB	AB	BB	AB	AB	BB	AB	AB	BB	BB	AB	AB	AB	AB	BB
RM449	AB	XX	AB	AB	AB	AB	AB	AB	BB	AB	BB	AB	AB	BB	BB	AB	AB	BB	AB	AB
RM493	XX	XX	AB	AB	XX	XX	XX	BB	BB	AB	BB	BB	XX	XX	BB	BB	AA	AB	AB	AB
RM488	AX	BB	XX	AX	XX	XX	AX	BB	XX	XX	BB	AX	AX	BB	AX	AX	AX	AX	AX	AX
RM1003	XX	AB	XX	AA	AA	AA	AA	AB	AB	BB	AB	AB	AB	BB	XX	BB	BB	AB	AB	BB
RM233AA	XB	XB	XB	XB	XB	XB	XB	XB	XB	XB	XX	XX	XB	XB	XB	AA	AA	AA	XB	XB
RM8255	XX	AA	XX	AB	AB	AB	AB	BB	BB	AB	AA	BB	AB	BB	BB	XX	AB	AA	AA	AB

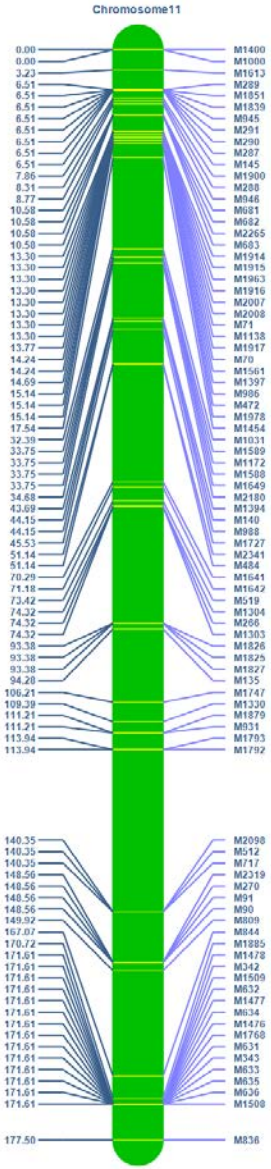
Removal of redundancy

Redundant markers

- Have a correlation of 1.0
- Have a recombination frequency of 0.0
- Locate at the same chromosomal position
- Cannot provide additional information in genetic analysis
- Only one is need in a bin of redundant markers

A map with redundant markers

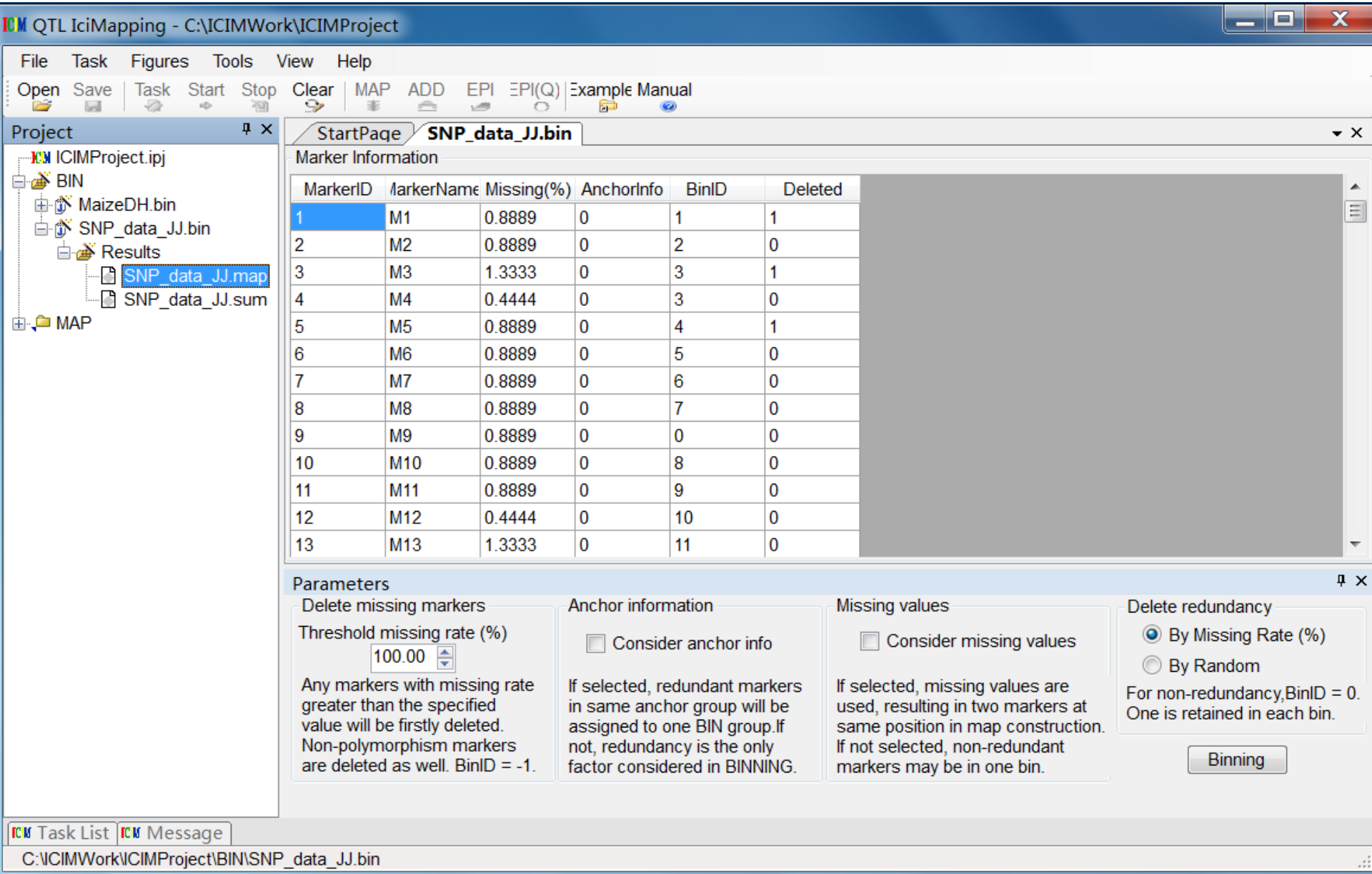




Define redundant markers

- No missing data (DH population, size =10)
  - Marker 1: A A A B B B A B B B
  - Marker 2: A A A B B B A B B B
  - Correlation between the two markers is 1.0
- Have missing data (DH population, size =10)
  - Marker 1: X X A B B B A X B B
  - Marker 2: A X A B B B A B B X
  - Correlation is not 1.0 if missing values considered
  - Correlation is 1.0 if missing values not considered
  - Recom. Freq. is 0 in linkage analysis
  - Marker 1 should be deleted due to its higher missing rate

BIN functionality by default setting



Now let's go for map construction!