Computational Methods in Genetics

Pr. Laurent Gentzbittel
Pr. Cécile Ben

Project Center for Agro Technologies
Skolkovo Institute of Science and Technology

MSc LS, 2021

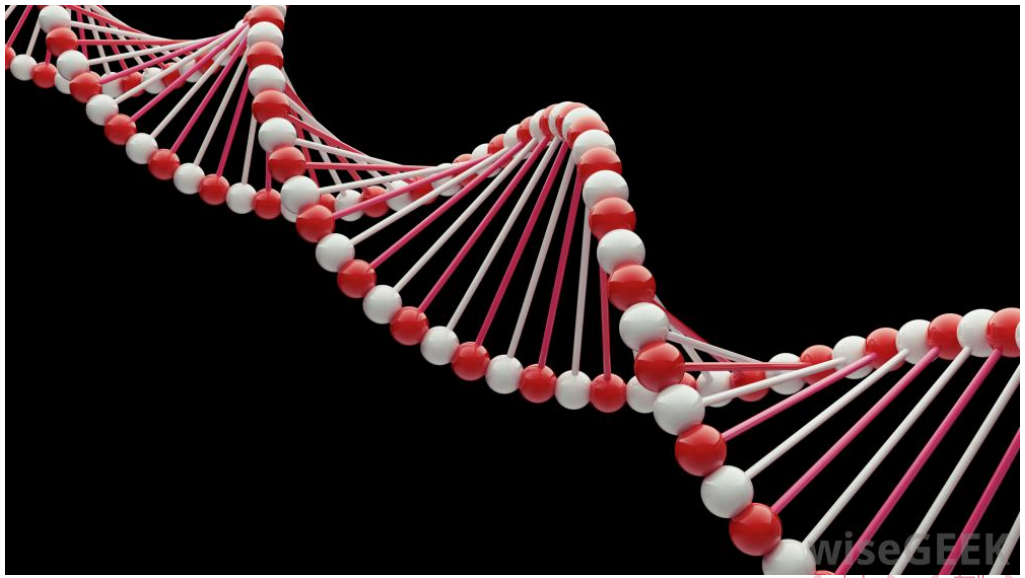
Skills

1. Understanding what is linkage and its use in plant and animal genetics
2. Estimates of recombination fraction :
   - Experimental segregating populations in plants
   - Computation methods: Maximum Likelihood, *Expectation-Maximization*
3. Building genetic maps
   - Practical using **R**

Genetics ? for this course, "Genetics" is:

Understanding some biological facts by using

- the linkage between one trait and another trait
- the linkage between one trait and a genetic polymorphism
- the linkage between one genetic polymorphism and another genetic polymorphism



How to conduct a genetical analysis

Three main steps :

1. Are the studied traits genetically controlled? What is their heritability?
2. Are the traits genetically linked ?
3. If we have information on the structure of the genome, can we locate the characters in relation to each other?

These analyzes are based on the concept of ***Linkage Disequilibrium – LD***

Outline :

Outline:

Some 'classical' data

Hundred fungal strains collected in a restricted geographical area were studied. The following results are obtained :

	B_1	B_2
A_1	50	10
A_2	20	20

1. locus A : Color mycelium : orange ou grey
2. locus B : Compatible (causes a disease) or uncompatible (does not cause a disease) with respect to a plant genotype

A basic analysis :

A significant association between mycelium morphology – pathogenicity ?

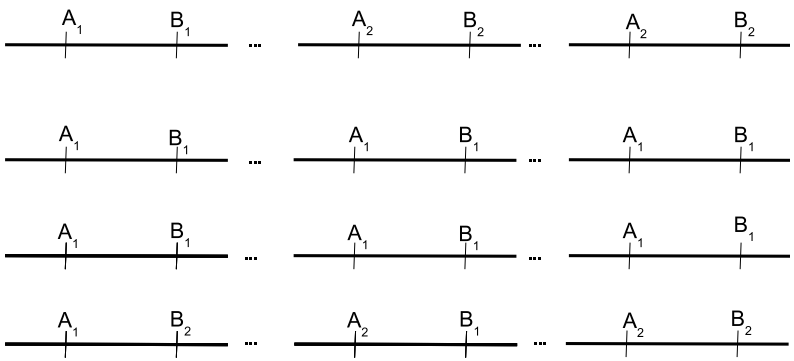
	B_1	B_2
A_1	50	10
A_2	20	20

```
Pearson's Chi-squared test
with Yates' continuity correction
data:  Comptages
X-squared = 11.1607,
df = 1,
p-value = 0.0008355
```

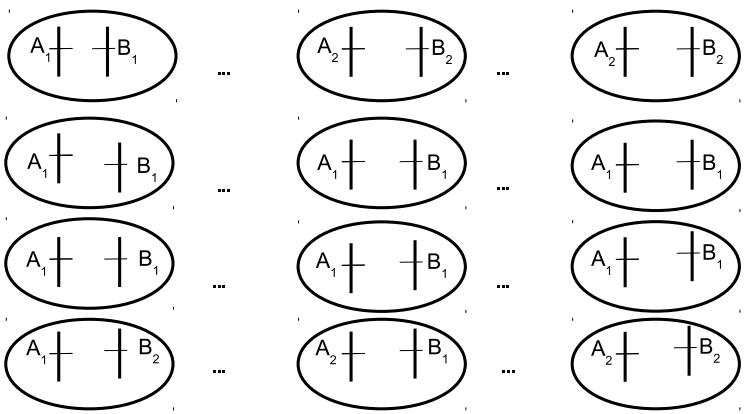
It exists a significant LD betwen the alleles at locus **A** and the alleles at locus **B**.

Biological Interpretation:

"Genetic linkage"

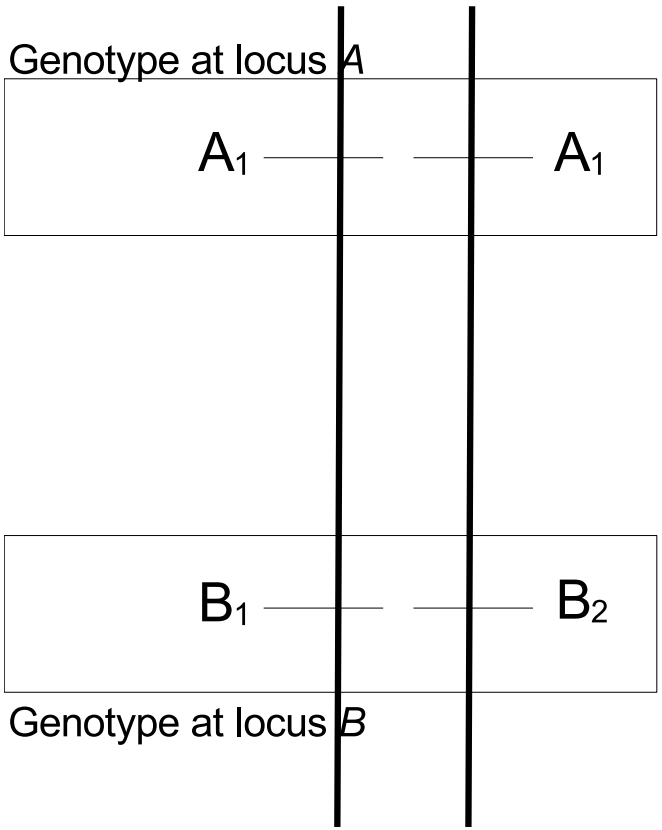


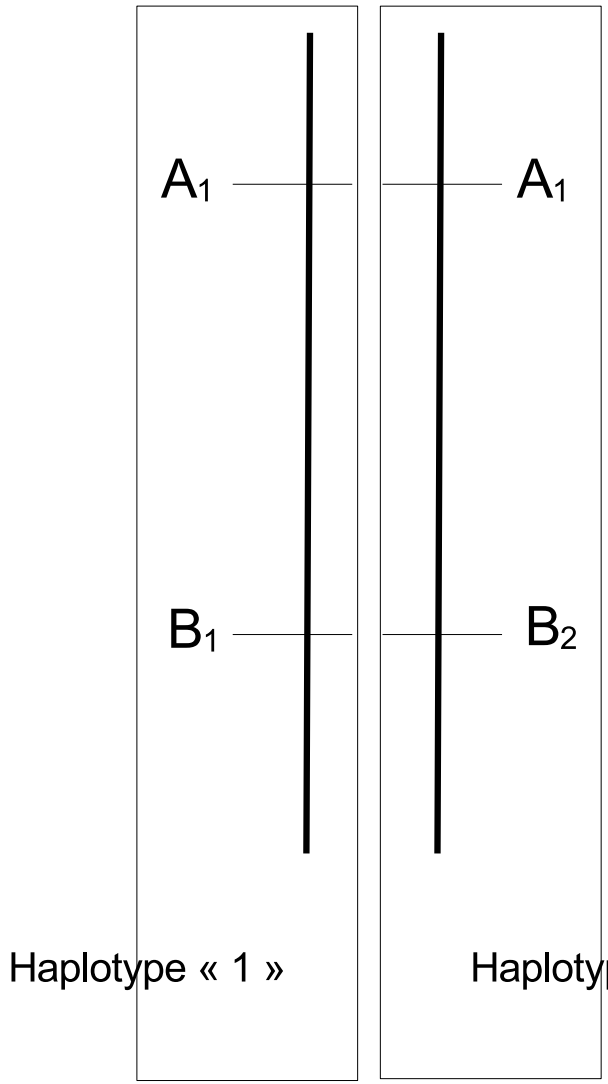
"Population structure" : within-species structure of biodiversity



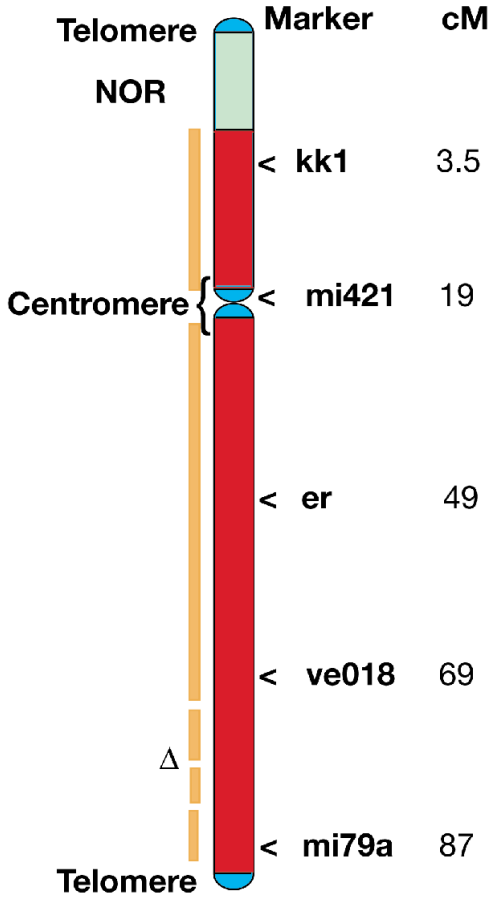
Both cases are valid, before any additional information.

Genotype, Haplotype



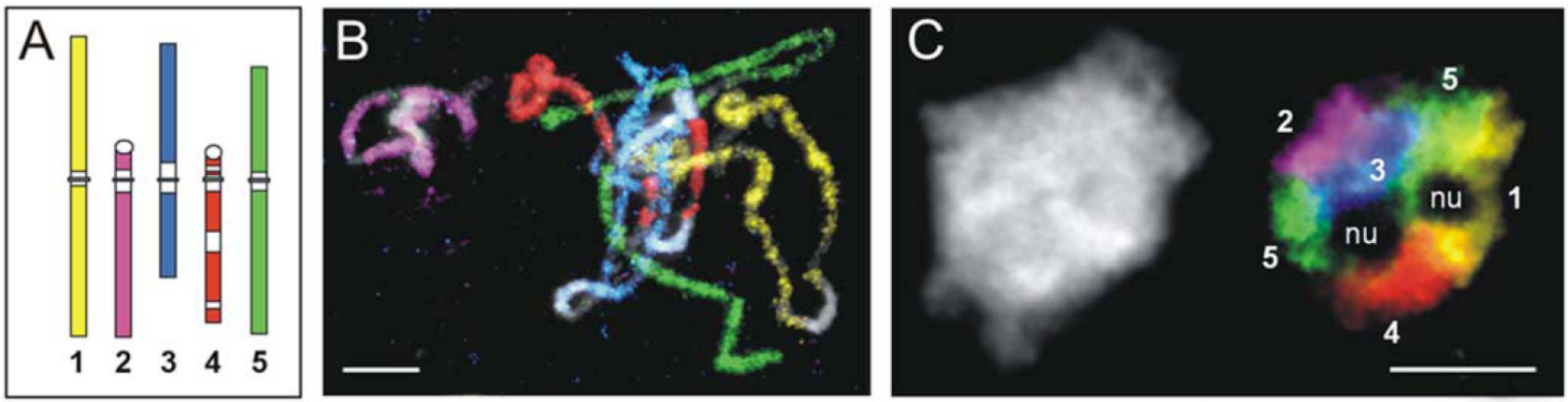


Constitutive elements of a genetic map



▷ Usually, a genetic map is colinear with a 'perfect' genome assembly.

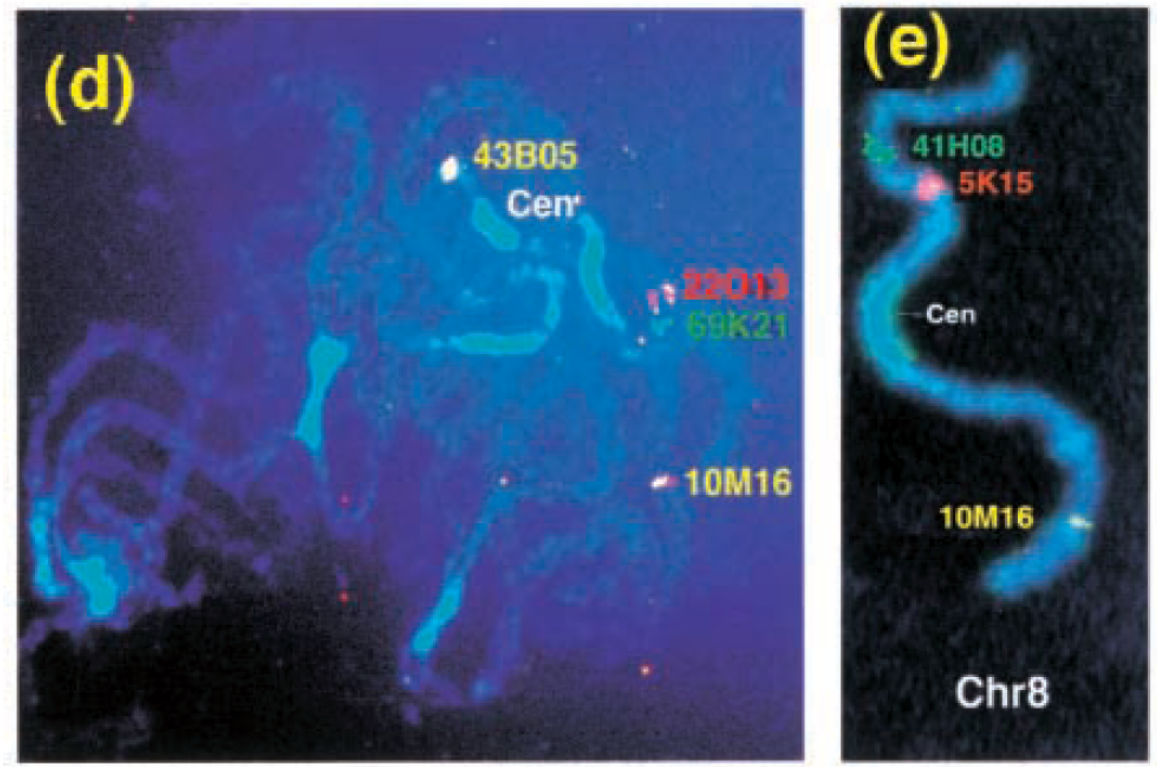
To see chromosomes 'in full colour'



Pecinka, A., Schubert, V., Meister, A., Kreth, G., Klatte, M., Lysak, M.A., Fuchs, J., Schubert, I., 2004. Chromosome territory arrangement and homologous pairing in nuclei of *Arabidopsis thaliana* are predominantly random except for NOR-bearing chromosomes. Chromosoma 113, 258269.

To see chromosomes 'in full colour'

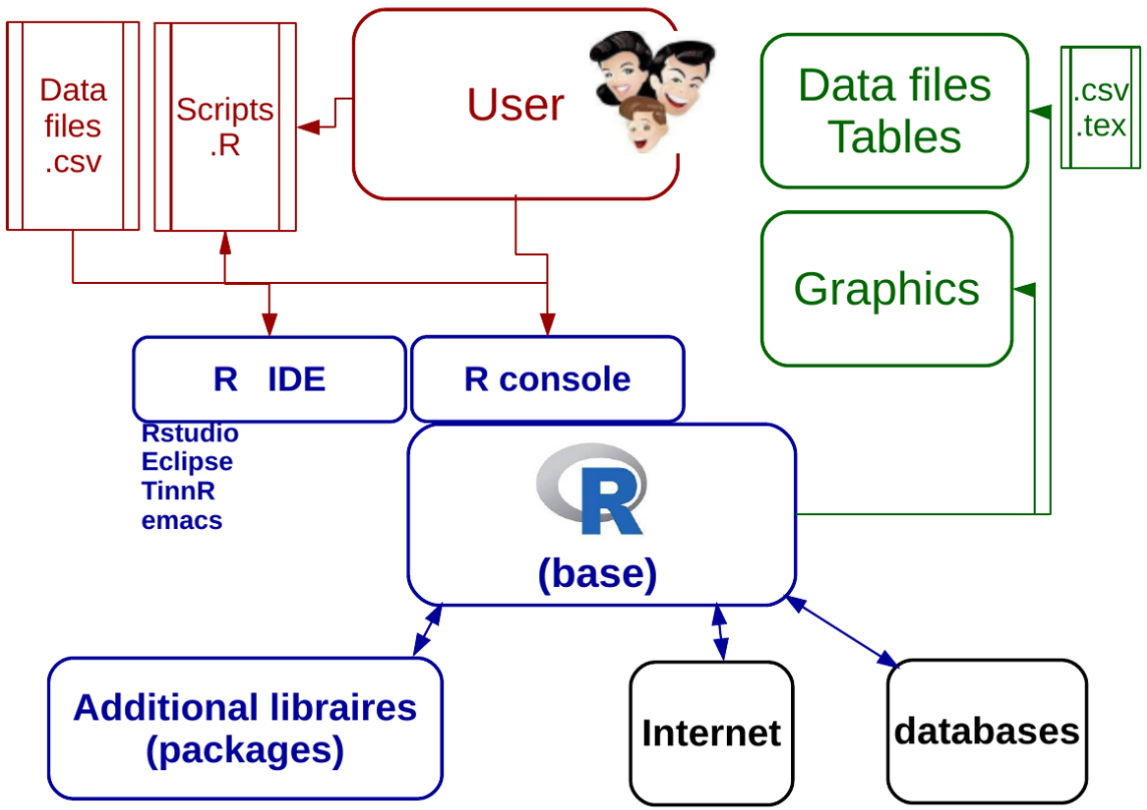
to visualize genetic recombinations



Choi, H.-K., Kim, D., Uhm, T., Limpens, E., Lim, H., Mun, J.-H., Kalo, P., Penmetsa, R.V., Seres, A., Kulikova, O., Roe, B.A., Bisseling, T., Kiss, G.B., Cook, D.R., 2004. A sequence-based genetic map of *Medicago truncatula* and comparison of marker colinearity with *M. sativa*. Genetics 166, 14631502.

Let's have a short overview of R

a Swiss-army knife tool for biologists



▷ `R-conference_Skoltech-2021.pdf`

Outline:

Barley, ear awns and resistance to rust

In a barley breeding program, the variety 'Thibault' with short awned ears, a high number of grains per ear but susceptible to the pathogenic fungus *Puccinia hordei* responsible for brown rust, was crossed with the mutant variety MU302, with long awned ears, low seed number per ear but resistant to the fungus,
to improve yield and pathogen resistance in this autogamous diploid species.

To evaluate the interest of the cross between these 2 varieties in a selection program, we need to compare the parental varieties and the F1 and F2 progenies with each other by statistical analysis.

Barley, ear awns and resistance to rust

A phenotypic analysis of the parental lines **and** the progeny of the crosses was carried out for 3 traits of interest under "controlled" environmental conditions and *identical for all* the individuals tested. We hypothesize that there is no genotype-environment interaction effect for the trait 'number of grains per ear'

Barley, ear awns and resistance to rust

The steps of the analysis using **R** :

1. Load data
2. Check data and describe data
3. Carry out analyses using different methods depending on questions

Questions:

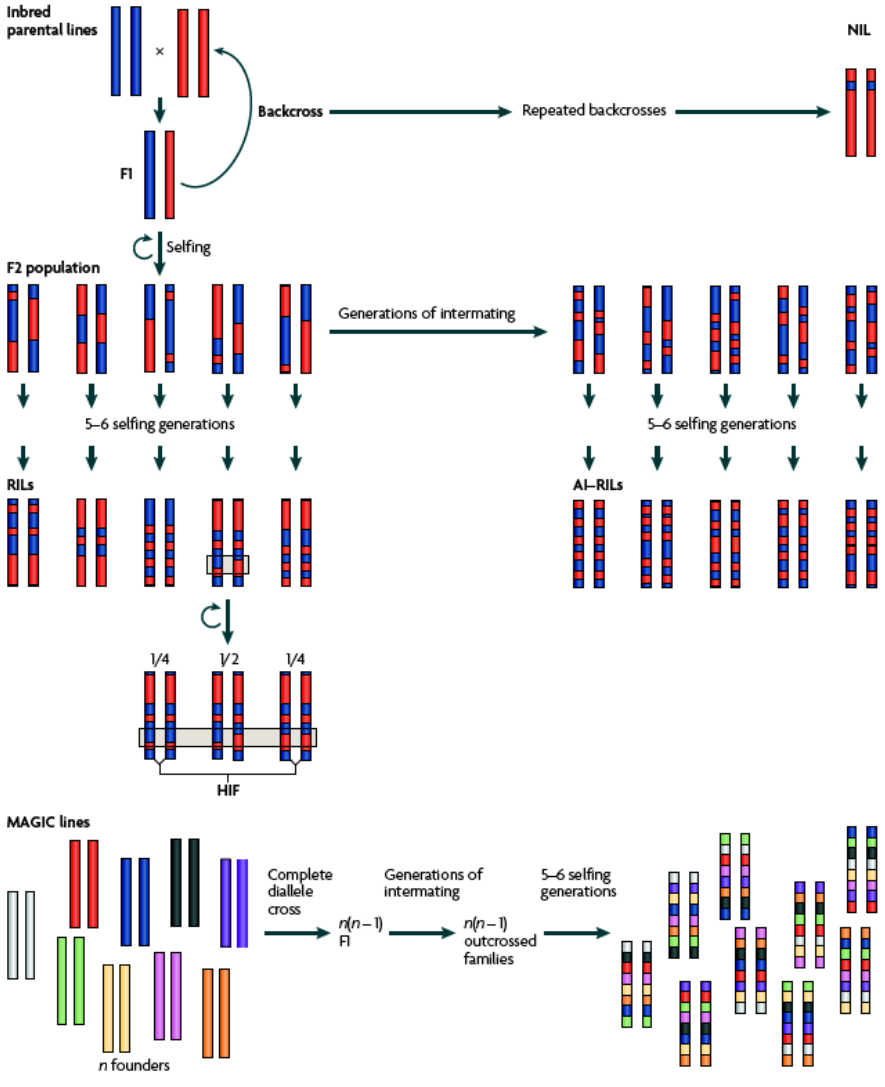
1. Is the possible difference observed in ‘Grain per Ear‘ between the parents, and between parents and F_1 or F_2, due to hereditary factors ? Why ‘Grain per Ear‘ is varying within each population ?
2. How much loci/genes are controling the morphologic traits ? Test for Mendelian inheritance. Are they linked ? Conclude
3. Is one of this morphologic traits linked to ‘Grain per Ear‘ ? Conclude.

Barley case study

- ▶ At your keyboard: A first simple example in Barley
- ▷ We will provide a PDF report of that analysis

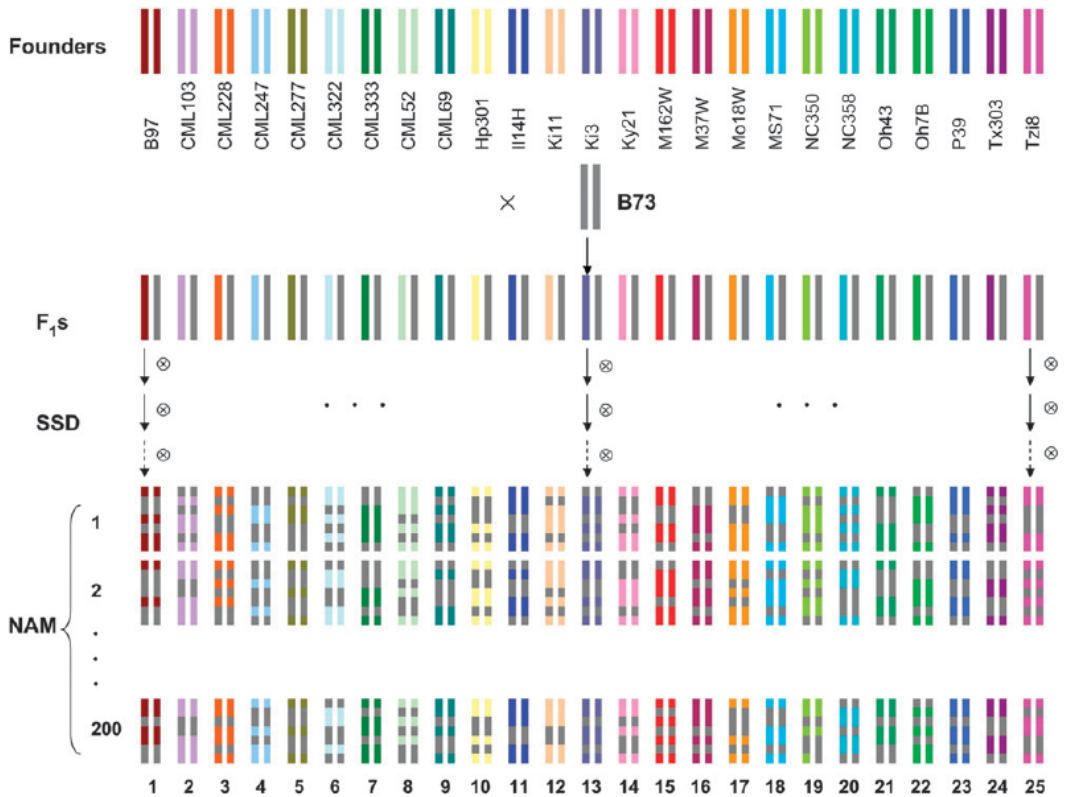
Outline:

Exp. pop. for linkage mapping (and breeding) – 1



Exp. pop. for linkage mapping (and breeding) – 2

Nested Association Mapping (NAM) populations



25 families for a total of 5.000 RILs

The basis of building a genetic map.

What is a genetic map :

- to order over each other numerous genetic markers
- using the genetic recombinations

BASIC hypothesis : two genes are located the closer to each other as the rate of genetic recombination between them is low.

Steps to construct a genetic map.

1. Compute recombination fraction between each pair of marker and decide if the two locus are linked or un-liked.
2. Marker are then grouped into linkage groups. They are then ordered ordered with respect to each other within each linkage group.
3. As a last step, recombination fractions are transformed into distances expressed in centiMorgans (cM) using the Kosambi or Haldane functions.
4. The numerical estimation of recombination fraction uses mainly the method of the *Maximum Likelihood*.

Steps to construct a genetic map.

1. Compute recombination fraction between each pair of marker and decide if the two locus are linked or un-liked.
2. Marker are then grouped into linkage groups. They are then ordered ordered with respect to each other within each linkage group.
3. As a last step, recombination fractions are transformed into distances expressed in centiMorgans (cM) using the Kosambi or Haldane functions.
4. The numerical estimation of recombination fraction uses mainly the method of the *Maximum Likelihood*.

Steps to construct a genetic map.

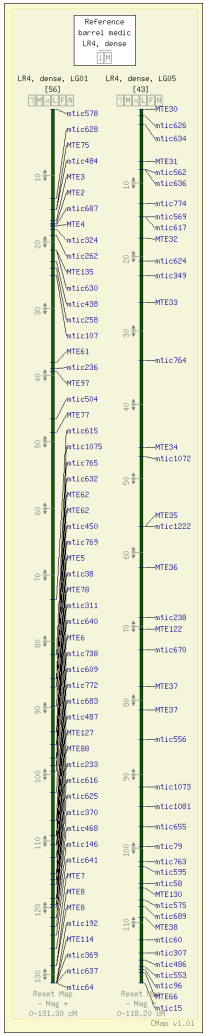
1. Compute recombination fraction between each pair of marker and decide if the two locus are linked or un-liked.
2. Marker are then grouped into linkage groups. They are then ordered ordered with respect to each other within each linkage group.
3. As a last step, recombination fractions are transformed into distances expressed in centiMorgans (cM) using the Kosambi or Haldane functions.
4. The numerical estimation of recombination fraction uses mainly the method of the *Maximum Likelihood*.

Steps to construct a genetic map.

1. Compute recombination fraction between each pair of marker and decide if the two locus are linked or un-liked.
2. Marker are then grouped into linkage groups. They are then ordered ordered with respect to each other within each linkage group.
3. As a last step, recombination fractions are transformed into distances expressed in centiMorgans (cM) using the Kosambi or Haldane functions.
4. The numerical estimation of recombination fraction uses mainly the method of the *Maximum Likelihood*.

The elements of a genetic map

example of two chromosomes of *Medicago truncatula*



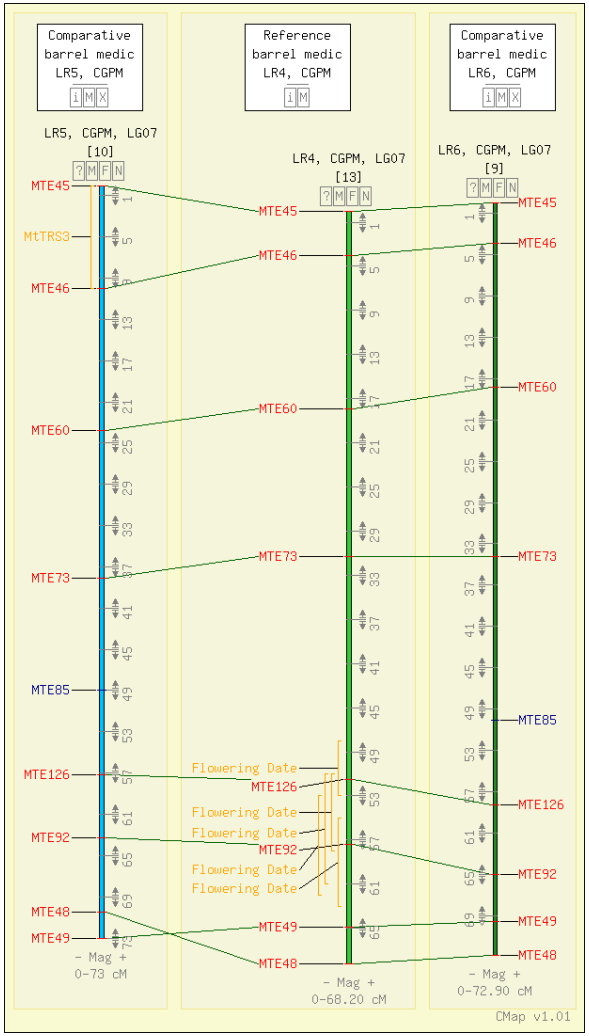
Exemple :

- Genetic map of the RILs population A17 x DZA315.16
- 370 SSRs loci, mapped onto 8 chromosomes
- It spans a total of 910 cM with an average distance between markers of 3 cM.

see ▸ *M. truncatula* web server for genetic map

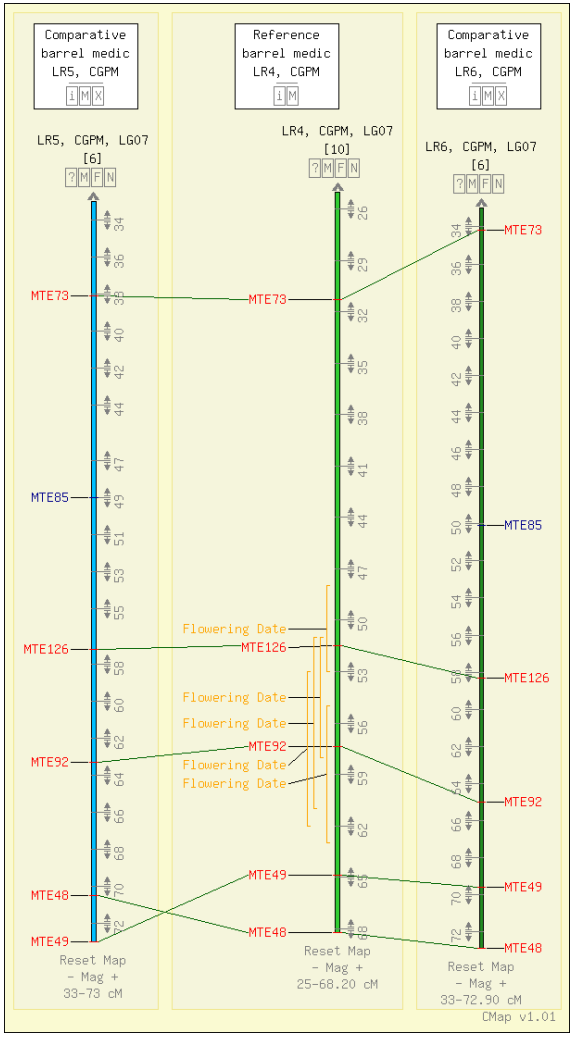
The elements of a genetic map

It is possible to compare the genetic maps of several crosses



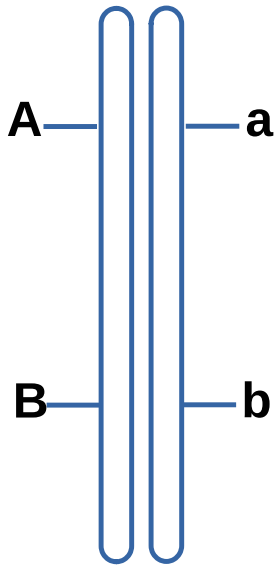
The elements of a genetic map

It is possible to postulate the likely position of traits in different crosses

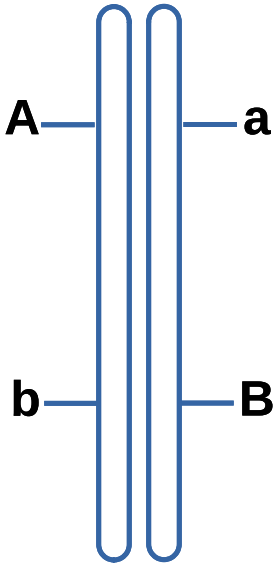


Outline:

Locus in Coupling - Repulsion



two locus in
coupling

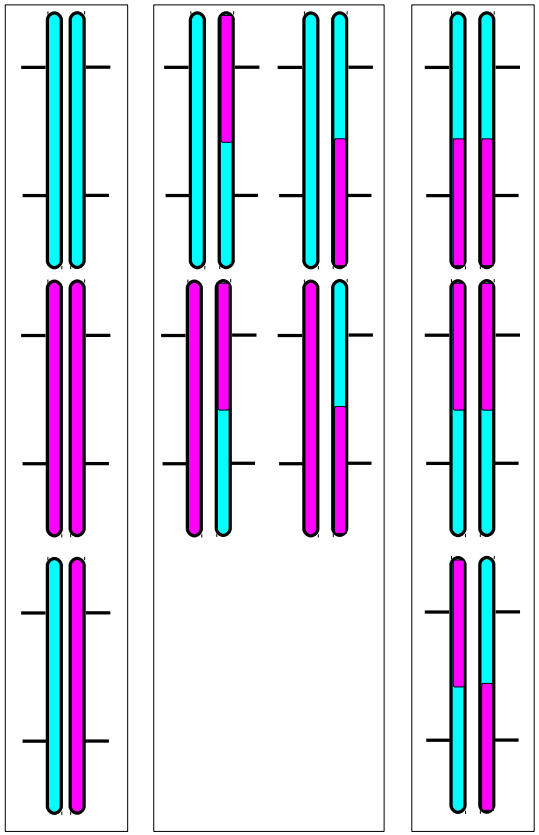


two locus in
repulsion

Outline:

What if we can't count all crossovers?

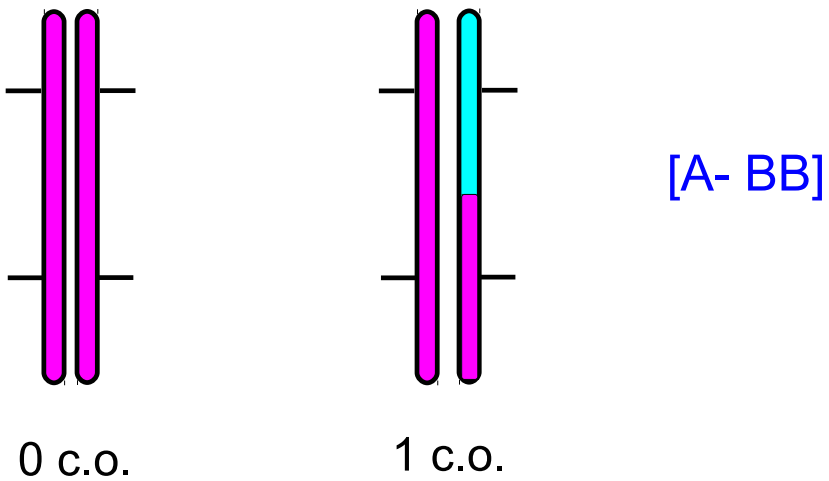
In an F2 design, possible progenies have 0, 1 or 2 cross overs



And what if some markers are dominant?

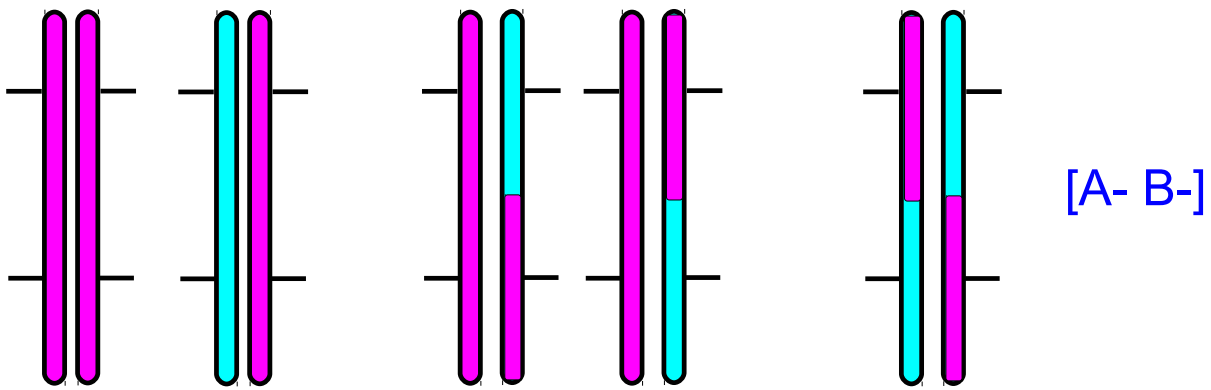
Dominance: one of the homozygotes has the same phenotype as the heterozygote

e.g. AA looks like Aa, so [A- BB] may hide 0 or 1 crossover



And what if some markers are dominant?

and it gets worse if two mar“kers are dominant...
[A- B-] may hide 0,..... 1,..... or 2 crossovers



Moral: life is tough.

We face the problem of more possible categories (**genotypes**) than can be distinguished (**phenotypes**)

- So, we often can't just count crossovers.
- We need a probability model that describes the observations in terms of θ, the recombination fraction
- We then need a numerical method that allows to estimate θ from the observations

1. the **Scoring Method** for maximum likelihood estimate of θ
2. the **Expectation - Maximization method** for maximum likelihood estimate of θ

Moral: life is tough.

We face the problem of more possible categories (**genotypes**) than can be distinguished (**phenotypes**)

- So, we often can't just count crossovers.
- We need a probability model that describes the observations in terms of θ, the recombination fraction
- We then need a numerical method that allows to estimate θ from the observations

1. the **Scoring Method** for maximum likelihood estimate of θ
2. the **Expectation - Maximization method** for maximum likelihood estimate of θ

Moral: life is tough.

We face the problem of more possible categories (**genotypes**) than can be distinguished (**phenotypes**)

- So, we often can't just count crossovers.
- We need a probability model that describes the observations in terms of θ, the recombination fraction
- We then need a numerical method that allows to estimate θ from the observations

1. the **Scoring Method** for maximum likelihood estimate of θ
2. the **Expectation - Maximization method** for maximum likelihood estimate of θ

Moral: life is tough.

We face the problem of more possible categories (**genotypes**) than can be distinguished (**phenotypes**)

- So, we often can't just count crossovers.
- We need a probability model that describes the observations in terms of θ, the recombination fraction
- We then need a numerical method that allows to estimate θ from the observations

1. the **Scoring Method** for maximum likelihood estimate of θ
2. the **Expectation - Maximization method** for maximum likelihood estimate of θ

Moral: life is tough.

We face the problem of more possible categories (**genotypes**) than can be distinguished (**phenotypes**)

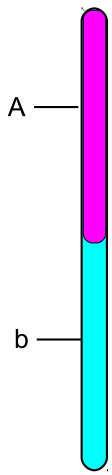
- So, we often can't just count crossovers.
- We need a probability model that describes the observations in terms of θ, the recombination fraction
- We then need a numerical method that allows to estimate θ from the observations

1. the **Scoring Method** for maximum likelihood estimate of θ
2. the **Expectation - Maximization method** for maximum likelihood estimate of θ

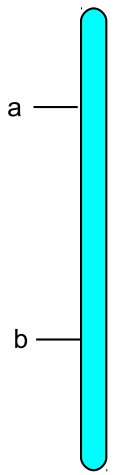
The probability model : What F2 segregation do we expect?

If the parent genotype is AB/ab, the gamete probabilities are

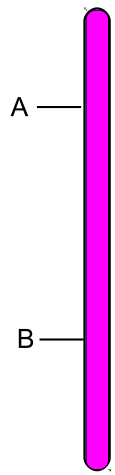
$\theta/2$



$\theta/2$



$(1-\theta)/2$



$(1-\theta)/2$

and the progeny genotypes have probabilities ...

Genotype probabilities from selfing an AB/ab genotype

		$(1-\theta)/2$ ab	$(1-\theta)/2$ AB	$\theta/2$ Ab	$\theta/2$ aB
$(1-\theta)/2$	ab	$(1-\theta)^2/4$	$(1-\theta)^2/4$	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$
$(1-\theta)/2$	AB	$(1-\theta)^2/4$	$(1-\theta)^2/4$	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$
$\theta/2$	Ab	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$	$\theta^2/4$	$\theta^2/4$
$\theta/2$	aB	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$	$\theta^2/4$	$\theta^2/4$

"Marker Phenotype" probabilities from selfing an AB/ab genotype

Combining symetrical genotypes

		$(1-\theta)/2$ ab	$(1-\theta)/2$ AB	$\theta/2$ Ab	$\theta/2$ aB
$(1-\theta)/2$	ab	$(1-\theta)^2/4$	$(1-\theta)^2/4$	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$
$(1-\theta)/2$	AB	$(1-\theta)^2/4$	$(1-\theta)^2/4$	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$
$\theta/2$	Ab	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$	$\theta^2/4$	$\theta^2/4$
$\theta/2$	aB	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$	$\theta^2/4$	$\theta^2/4$

$$\Pr(\mathrm{Aabb}) = \frac{\theta(1-\theta)}{4} + \frac{\theta(1-\theta)}{4} = \frac{2\theta(1-\theta)}{4} = \theta(1-\theta)/2$$

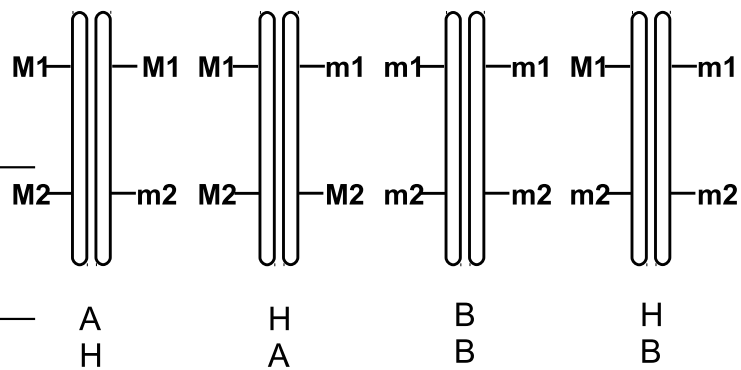
Outline:

Genotype's shortcut rules

the three-letters code for genotypes

- Female parent : A code, (big letters)
- Male parent : B code , (small letters)

code :	A	H	B
M1	M1M1	M1m1	m1m1
M2	M2M2	M2m2	m2m2



Bases of the EM algorithm

Statistical strategy for estimating θ :
Consider a set of starting parameters

1. Use these to estimate the complete (unobserved, ambiguous) data, keeping unambigous data fixed :
   What segregation would we expect for non-observed genotypes if θ were equal to our guess?
2. Use estimates of complete data to update parameters : Adjust θ
3. Repeat until our expectation matches the segregation!

F2 genotype probabilities. An example

Let's compute the genotype frequencies for $\theta = 0.28$

Genotype	Phenotype	Cross. overs	Probability	at $\theta = 0.28$
aabb	BB	0	$(1-\theta)^2/4$	0.1296
Aabb	HB	1	$\theta(1-\theta)/2$	0.1008
AAbb	AB	2	$\theta)^2/4$	0.0196
aaBb	BH	1	$\theta(1-\theta)/2$	0.1008
AaBb	HH	0	$(1-\theta)^2/2$	0.2592
aABb	HH	2	$\theta^2/2$	0.0392
AABb	AH	1	$(1-\theta)/2$	0.1008
aaBB	BA	2	$\theta^2/4$	0.0196
AaBB	HA	1	$\theta(1-\theta)/2$	0.1008
AABB	AA	0	$(1-\theta)^2/4$	0.1296

F2 genotype probabilities. An example

Let's compute the genotype frequencies for $\theta = 0.28$

Genotype	Phenotype	Cross. overs	Probability	at $\theta = 0.28$
aabb	BB	0	$(1-\theta)^2/4$	0.1296
Aabb	HB	1	$\theta(1-\theta)/2$	0.1008
AAbb	AB	2	$\theta)^2/4$	0.0196
aaBb	BH	1	$\theta(1-\theta)/2$	0.1008
AaBb	HH	0	$(1-\theta)^2/2$	0.2592
aABb	HH	2	$\theta^2/2$	0.0392
AABb	AH	1	$(1-\theta)/2$	0.1008
aaBB	BA	2	$\theta^2/4$	0.0196
AaBB	HA	1	$\theta(1-\theta)/2$	0.1008
AABB	AA	0	$(1-\theta)^2/4$	0.1296

F2 genotype probabilities. An example

Let's compute the genotype frequencies for $\theta = 0.28$

Genotype	Phenotype	Cross. overs	Probability	at $\theta = 0.28$
aabb	BB	0	$(1-\theta)^2/4$	0.1296
Aabb	HB	1	$\theta(1-\theta)/2$	0.1008
AAbb	AB	2	$\theta)^2/4$	0.0196
aaBb	BH	1	$\theta(1-\theta)/2$	0.1008
AaBb	HH	0	$(1-\theta)^2/2$	0.2592
aABb	HH	2	$\theta^2/2$	0.0392
AABb	AH	1	$(1-\theta)/2$	0.1008
aaBB	BA	2	$\theta^2/4$	0.0196
AaBB	HA	1	$\theta(1-\theta)/2$	0.1008
AABB	AA	0	$(1-\theta)^2/4$	0.1296

F2 genotype probabilities. An example

Let's compute the genotype frequencies for $\theta = 0.28$

Genotype	Phenotype	Cross. overs	Probability	at $\theta = 0.28$
aabb	BB	0	$(1-\theta)^2/4$	0.1296
Aabb	HB	1	$\theta(1-\theta)/2$	0.1008
AAbb	AB	2	$\theta)^2/4$	0.0196
aaBb	BH	1	$\theta(1-\theta)/2$	0.1008
AaBb	HH	0	$(1-\theta)^2/2$	0.2592
aABb	HH	2	$\theta^2/2$	0.0392
AABb	AH	1	$(1-\theta)/2$	0.1008
aaBB	BA	2	$\theta^2/4$	0.0196
AaBB	HA	1	$\theta(1-\theta)/2$	0.1008
AABB	AA	0	$(1-\theta)^2/4$	0.1296

F2 genotype probabilities. An example

Let's compute the genotype frequencies for $\theta = 0.28$

Genotype	Phenotype	Cross. overs	Probability	at $\theta = 0.28$
aabb	BB	0	$(1-\theta)^2/4$	0.1296
Aabb	HB	1	$\theta(1-\theta)/2$	0.1008
AAbb	AB	2	$\theta)^2/4$	0.0196
aaBb	BH	1	$\theta(1-\theta)/2$	0.1008
AaBb	HH	0	$(1-\theta)^2/2$	0.2592
aABb	HH	2	$\theta^2/2$	0.0392
AABb	AH	1	$(1-\theta)/2$	0.1008
aaBB	BA	2	$\theta^2/4$	0.0196
AaBB	HA	1	$\theta(1-\theta)/2$	0.1008
AABB	AA	0	$(1-\theta)^2/4$	0.1296

F2 genotype probabilities

Let's compute the genotype counts for 100 plants, with $\theta = 0.28$

Genotype	Phenotype	Cross. overs	Prob.	Expec. counts
aabb	BB	0	0.1296	13
Aabb	HB	1	0.1008	10
AAbb	AB	2	0.0196	2
aaBb	BH	1	0.1008	10
AaBb	HH	0	0.2592	26
aABb	HH	2	0.0392	4
AABb	AH	1	0.1008	10
aaBB	BA	2	0.0196	2
AaBB	HA	1	0.1008	10
AABB	AA	0	0.1296	13

Segregation data we'd see if all F2 genotypes were known

Now, imagine that we did'nt know θ, and were able to observe all data

Genotype	Phenotype	Obs. counts	X.overs	# of X.overs
aabb	BB	13	0	0
Aabb	HB	10	1	10
AAbb	AB	2	2	4
aaBb	BH	10	1	10
AaBb	HH	26	0	0
aABb	HH	4	2	8
AABb	AH	10	1	10
aaBB	BA	2	2	4
AaBB	HA	10	1	10
AABB	AA	13	0	0
Sum		100		56

# of meisoses : 2×100

$\hat{\theta} = 56/200 = 0.28$ RIGHT!

Segregation data we'd see if all F2 genotypes were known

Now, imagine that we did'nt know θ, and were able to observe all data

Genotype	Phenotype	Obs. counts	X.overs	# of X.overs
aabb	BB	13	0	0
Aabb	HB	10	1	10
AAbb	AB	2	2	4
aaBb	BH	10	1	10
AaBb	HH	26	0	0
aABb	HH	4	2	8
AABb	AH	10	1	10
aaBB	BA	2	2	4
AaBB	HA	10	1	10
AABB	AA	13	0	0
Sum		100		56

# of meisoses : 2×100

$\hat{\theta} = 56/200 = 0.28$ RIGHT!

Segregation data we'd see if all F2 genotypes were known

Now, imagine that we did'nt know θ, and were able to observe all data

Genotype	Phenotype	Obs. counts	X.overs	# of X.overs
aabb	BB	13	0	0
Aabb	HB	10	1	10
AAbb	AB	2	2	4
aaBb	BH	10	1	10
AaBb	HH	26	0	0
aABb	HH	4	2	8
AABb	AH	10	1	10
aaBB	BA	2	2	4
AaBB	HA	10	1	10
AABB	AA	13	0	0
Sum		100		56

# of meisoses : 2×100

$\hat{\theta} = 56/200 = 0.28$ **RIGHT!**

Real segregation data : all F2 genotypes are NOT known

Starting estimate $\theta = 0.5$?

Unfortunately in real life, we see only marker phenotype

Phenotype	Obs. counts	X.overs	Prob	Cond. counts	Expec. X.overs
BB	13	0	0.0625	13	0
HB	10	1	0.125	10	10
AB	2	2	0.0625	2	4
BH	10	1	0.125	10	10
HH	30	0	0.125	15	0
HH	-	2	0.125	15	30
AH	10	1	0.125	10	10
BA	2	2	0.0625	2	4
HA	10	1	0.125	10	10
AA	13	0	0.0625	13	0
Sum	100				78

Real segregation data : all F2 genotypes are NOT known

Starting estimate $\theta = 0.5$?

Unfortunately in real life, we see only marker phenotype

Phenotype	Obs. counts	X.overs	Prob	Cond. counts	Expec. X.overs
BB	13	0	0.0625	13	0
HB	10	1	0.125	10	10
AB	2	2	0.0625	2	4
BH	10	1	0.125	10	10
HH	30	0	0.125	15	0
HH	-	2	0.125	15	30
AH	10	1	0.125	10	10
BA	2	2	0.0625	2	4
HA	10	1	0.125	10	10
AA	13	0	0.0625	13	0
Sum	100				78

First estimate of θ

- With a starting estimate of $\theta = 0.5$,
  - we estimated 78 crossovers in 200 chances
  - for an updated estimate of $\theta = 0.39$
- Let's plug in the new θ...

Second estimate of θ

starting estimate $\theta = 0.39$

Phenotype	Obs. counts	X.overs	Prob	Cond. counts	Expec. X.overs
BB	13	0	0.093	13	0
HB	10	1	0.119	10	10
AB	2	2	0.038	2	4
BH	10	1	0.119	10	10
HH	30	0	0.186	21.3	0
HH	-	2	0.076	8.7	17.4
AH	10	1	0.119	10	10
BA	2	2	0.038	2	4
HA	10	1	0.119	10	10
AA	13	0	0.093	13	0
Sum	100				65.4

Second estimate of θ

starting estimate $\theta = 0.39$

Phenotype	Obs. counts	X.overs	Prob	Cond. counts	Expec. X.overs
BB	13	0	0.093	13	0
HB	10	1	0.119	10	10
AB	2	2	0.038	2	4
BH	10	1	0.119	10	10
HH	30	0	0.186	21.3	0
HH	-	2	0.076	8.7	17.4
AH	10	1	0.119	10	10
BA	2	2	0.038	2	4
HA	10	1	0.119	10	10
AA	13	0	0.093	13	0
Sum	100				65.4

Second estimate of θ

- With a starting estimate of $\theta = 0.39$,
  - we estimated 65.4 crossovers in 200 chances
  - for an updated estimate of $\theta = 0.327$
- Let's plug in the new θ...

Third estimate of θ

starting estimate $\theta = 0.327$

Phenotype	Obs. counts	X.overs	Prob	Cond. counts	Expec. X.overs
BB	13	0	0.113	13	0
HB	10	1	0.110	10	10
AB	2	2	0.027	2	4
BH	10	1	0.110	10	10
HH	30	0	0.226	24.3	0
HH	-	2	0.053	5.7	11.5
AH	10	1	0.110	10	10
BA	2	2	0.027	2	4
HA	10	1	0.110	10	10
AA	13	0	0.113	13	0
Sum	100				59.5

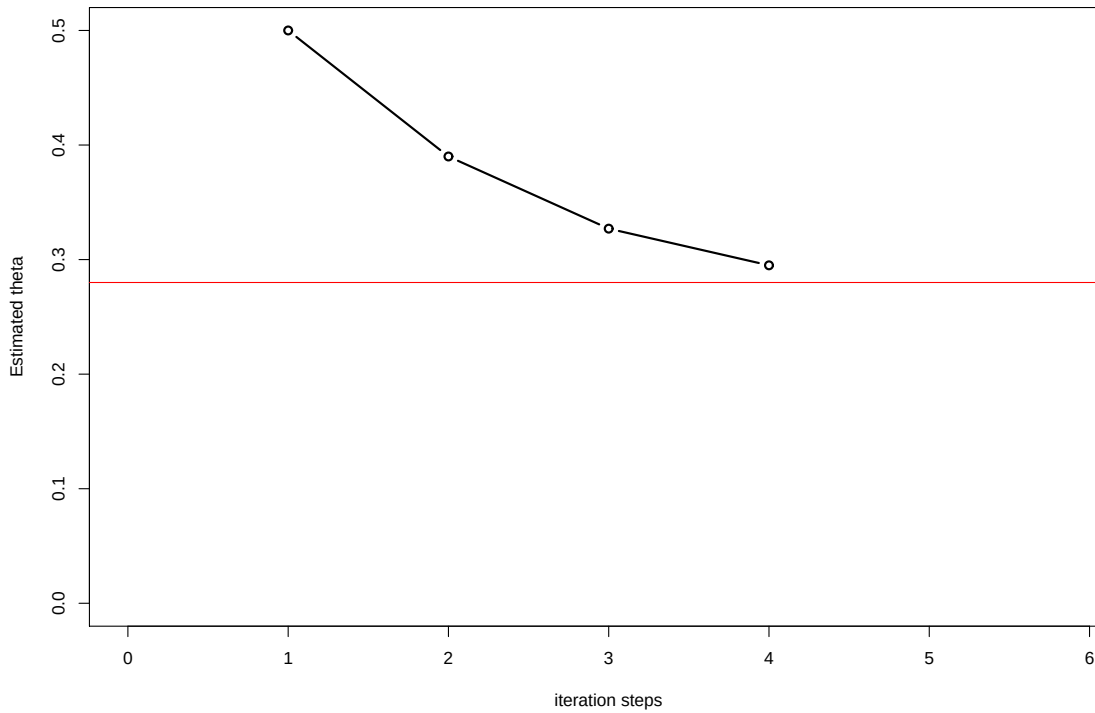
Third estimate of θ

starting estimate $\theta = 0.327$

Phenotype	Obs. counts	X.overs	Prob	Cond. counts	Expec. X.overs
BB	13	0	0.113	13	0
HB	10	1	0.110	10	10
AB	2	2	0.027	2	4
BH	10	1	0.110	10	10
HH	30	0	0.226	24.3	0
HH	-	2	0.053	5.7	11.5
AH	10	1	0.110	10	10
BA	2	2	0.027	2	4
HA	10	1	0.110	10	10
AA	13	0	0.113	13	0
Sum	100				59.5

Third estimate of θ

- We've adjusted our θ estimate from
  - 0.5 to
  - 0.39 to
  - 0.327 to
  - $59.5/200 = 0.295$
- and it will eventually get to 0.279



This is the EM algorithm in action

Outline:

How to order markers ?

Historically, the first formal linkage analysis involving more than two loci was given by Fisher (1922).
He showed how to combine data from a number of two-point analyses in order to obtain efficient estimates of a set of recombination fractions

The TravelSales Man problem

To find a tour that passes in all cities only one time and that has the minimum total distance.

The TSP problem uses a distance matrix

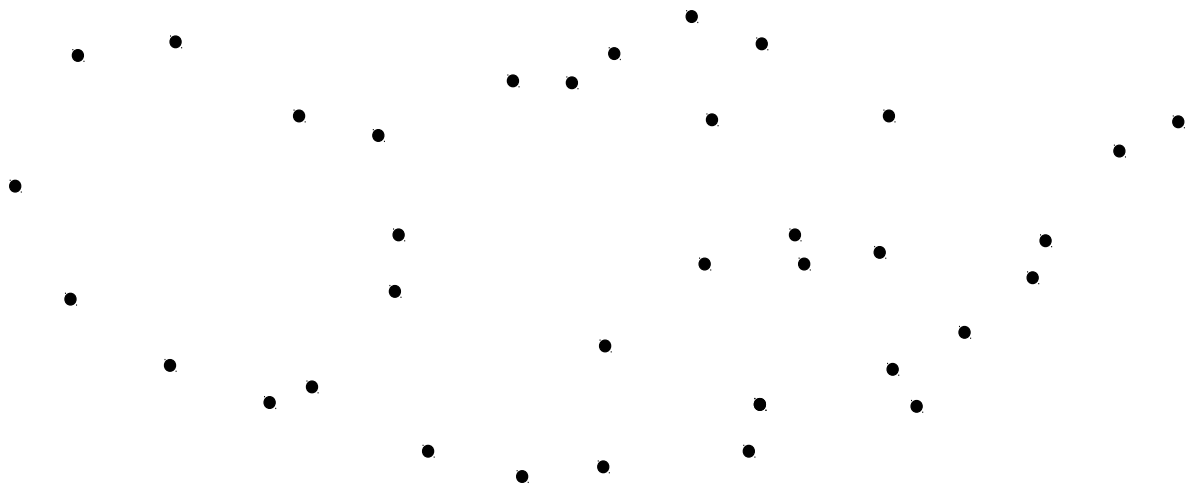
Distances	**Camara**	**Caniço**	**Funchal**	**...**
Camara	0	15	7	...
Caniço	15	0	8	...
Funchal	7	8	0	...
...	...	...	...	...

We already HAVE a distance matrix

Distances	**M1**	**M2**	**M3**	**...**
M1	0	14	7	...
M2	14	0	8	...
M3	7	8	0	...
...	...	...	...	...

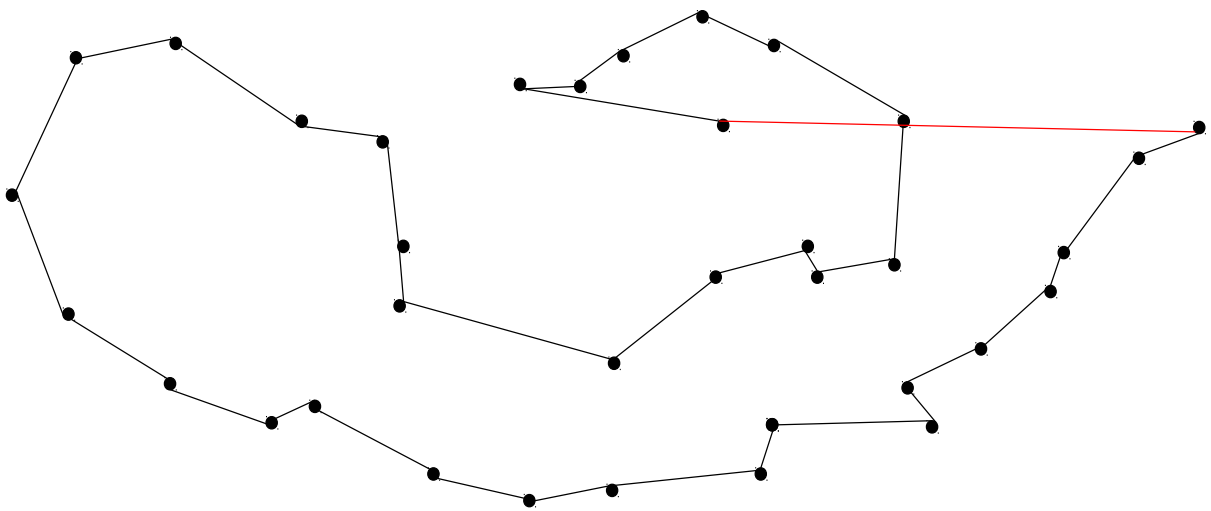
How to build a tour

the map



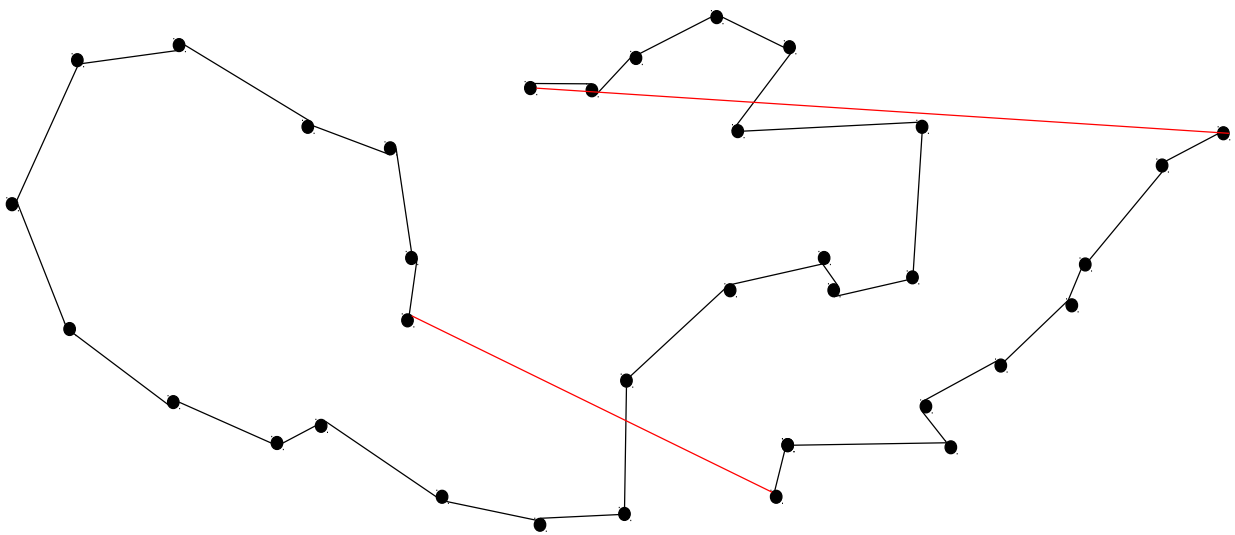
How to build a tour

several algorithms exist – example : Nearest Neighbor



How to build a tour

several algorithms exist – example : smaller 2 points distance, 'Greedy'



Outline:

Building a linkage map

1. experimentally obtain the genotypes at several locus in plants of a segregating population
2. compute two-point recombination fraction for each pair of locus
3. define linkage groups : by transitivity, group locus that are pairwise linked
4. for each linkage group, find the optimal order of locus
5. recompute distances on the basis of order and draw the map
6. Proudly present the results to your colleagues

Outline:

Linkage mapping with “qtl” – first dataset

- At your keyboard : linkage mapping using codominant markers !!

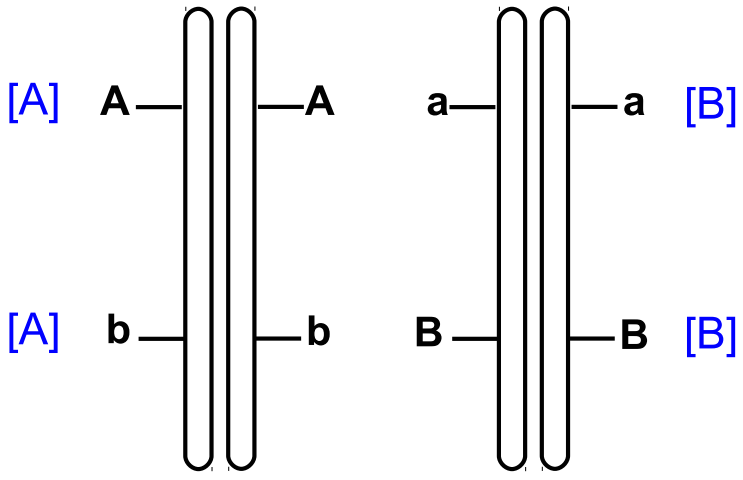
Outline:

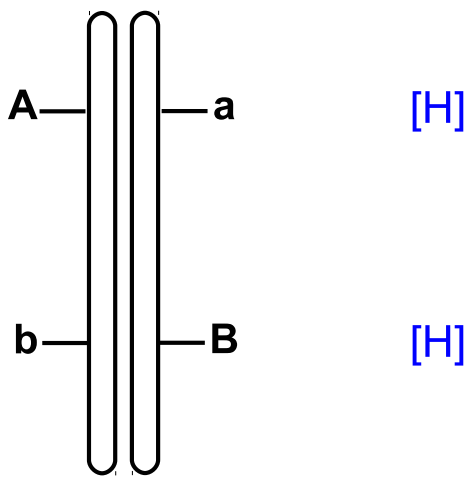
Dominance for markers : same idea of filling in partial information using EM

but watch out!

- Filling in unknown data always risks in filling in the wrong data!
- Estimating linkage between dominant markers is always less precise...

Dominant markers : Genotypes and marker Phenotypes





The worst situation : linking repulsion-phase markers.

- Let one marker show maternal dominance
  We observe only phenotypes A- and aa
  (with one-letter codes D (notB) and B)
- and the other marker paternal dominance
  with only phenotypes bb and B-
  (with one-letter codes A and C (notA)
- We now have only four two-marker phenotypes
  - BA (includes only BA) (how many crossovers?)
  - BC (includes BH, BB)
  - DA (includes AA, HA)
  - DC (includes AH, AB, HH, HB)

Nine genotypes, four marker phenotypes ....

Starting estimate $\theta^{(0)} = 0.5$?

Phenotype	Un-obs gen.	Phen.	Counts	X.overs	Prob	cond. counts	Exp. X.overs
BB	10			0	0.0625	6	0
HB	12	BC	18	1	0.125	11.8	11.8
AB	4	DC	53	2	0.0625	5.9	11.8
BH	8	BA	0	1	0.125	12	12
HH	26	DA	29	0	0.125	11.8	0
HH	4			2	0.125	11.8	23.6
AH	7			1	0.125	11.8	11.8
BA	0			2	0.0625	0	0
HA	13			1	0.125	19.3	19.3
AA	16			0	0.0625	9.7	0
Sum	100						90.3

$$\theta^{(1)} = \frac{90.3}{200} = 0.4515$$

Nine genotypes, four marker phenotypes ....

Starting estimate $\theta^{(0)} = 0.5$?

Phenotype	Un-obs gen.	Phen.	Counts	X.overs	Prob	cond. counts	Exp. X.overs
BB	10			0	0.0625	6	0
HB	12	BC	18	1	0.125	11.8	11.8
AB	4	DC	53	2	0.0625	5.9	11.8
BH	8	BA	0	1	0.125	12	12
HH	26	DA	29	0	0.125	11.8	0
HH	4			2	0.125	11.8	23.6
AH	7			1	0.125	11.8	11.8
BA	0			2	0.0625	0	0
HA	13			1	0.125	19.3	19.3
AA	16			0	0.0625	9.7	0
Sum	100						90.3

$$\theta^{(1)} = \frac{90.3}{200} = 0.4515$$

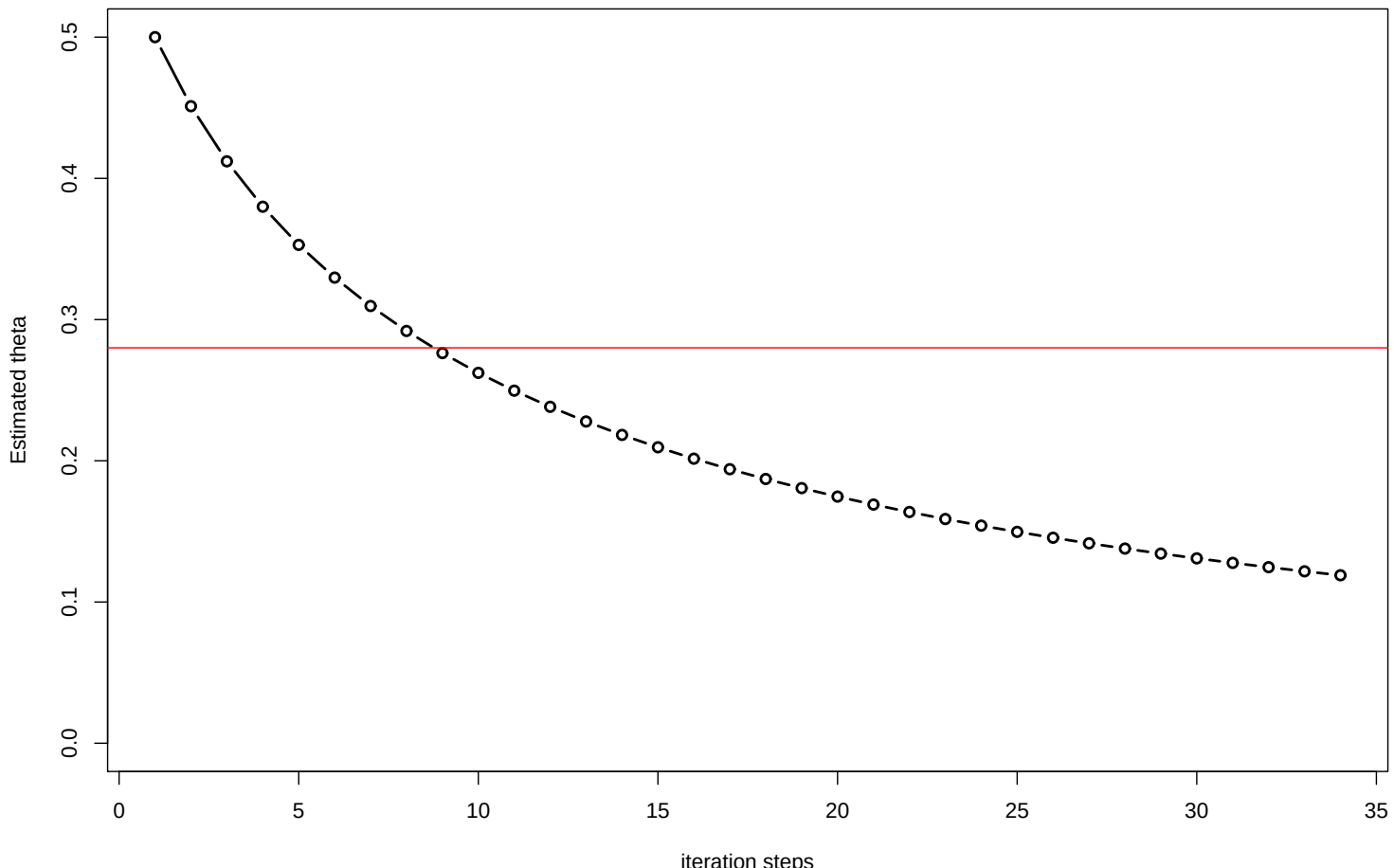
Nine genotypes, four marker phenotypes ....

Starting estimate $\theta^{(0)} = 0.5$?

Phenotype	Un-obs gen.	Phen.	Counts	X.overs	Prob	cond. counts	Exp. X.overs
BB	10			0	0.0625	6	0
HB	12	BC	18	1	0.125	11.8	11.8
AB	4	DC	53	2	0.0625	5.9	11.8
BH	8	BA	0	1	0.125	12	12
HH	26	DA	29	0	0.125	11.8	0
HH	4			2	0.125	11.8	23.6
AH	7			1	0.125	11.8	11.8
BA	0			2	0.0625	0	0
HA	13			1	0.125	19.3	19.3
AA	16			0	0.0625	9.7	0
Sum	100						90.3

$$\theta^{(1)} = \frac{90.3}{200} = 0.4515$$

the EM algorithm does'nt converge for this example !



To sum up 2-point linkage analysis

- We count crossovers, where we can : BC, DH
- Where we can't, we use the EM algorithm or Scoring method to estimate θ
- Uncertainty in genotype information always reduces precision *sometimes to the point of uselessness*
- Don't trust linkage estimates between dominant markers ; and *if they're in repulsion phase, divide the dataset*!

Outline:

Genotyping by Whole Genome Re-Sequencing

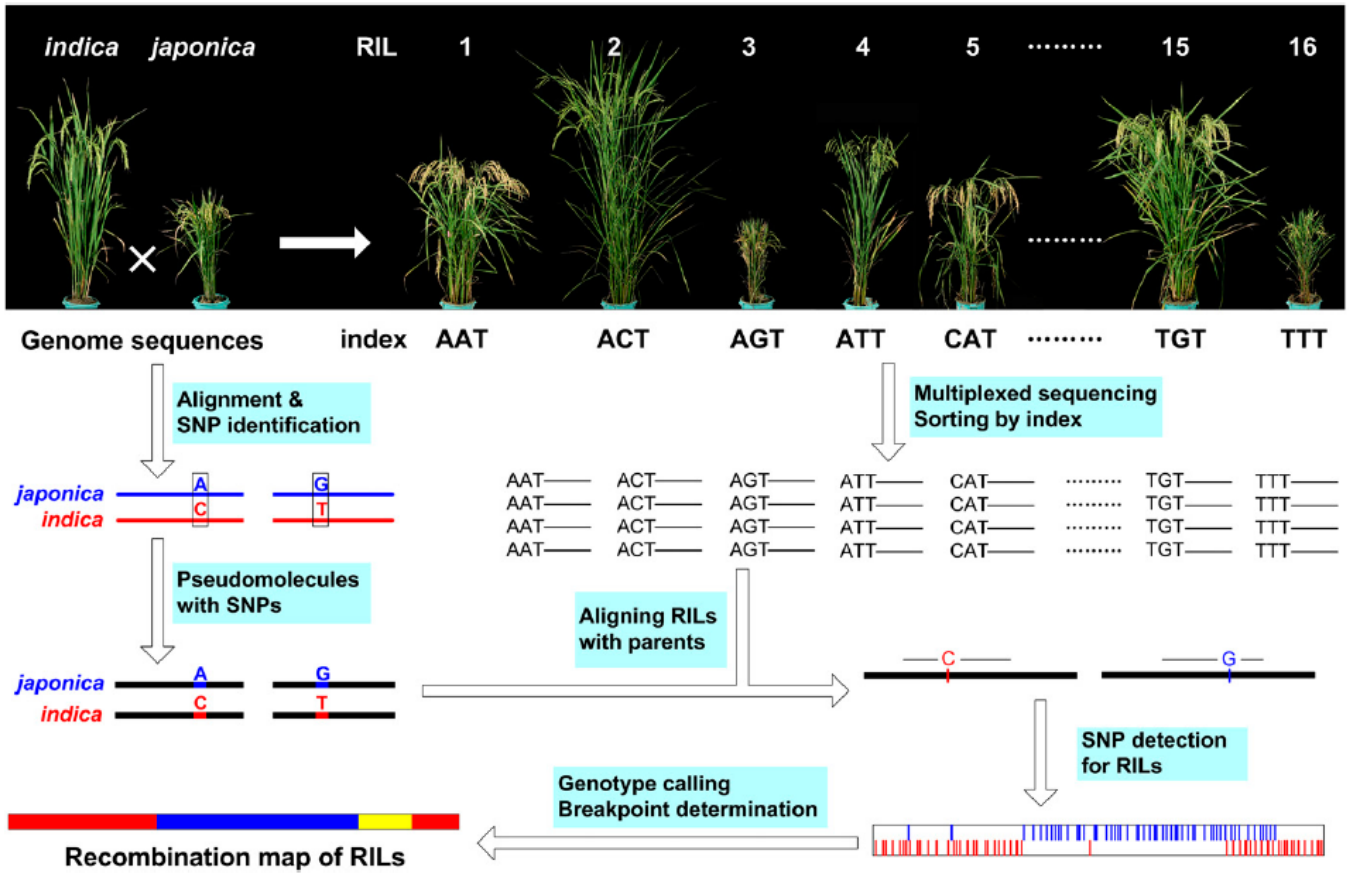
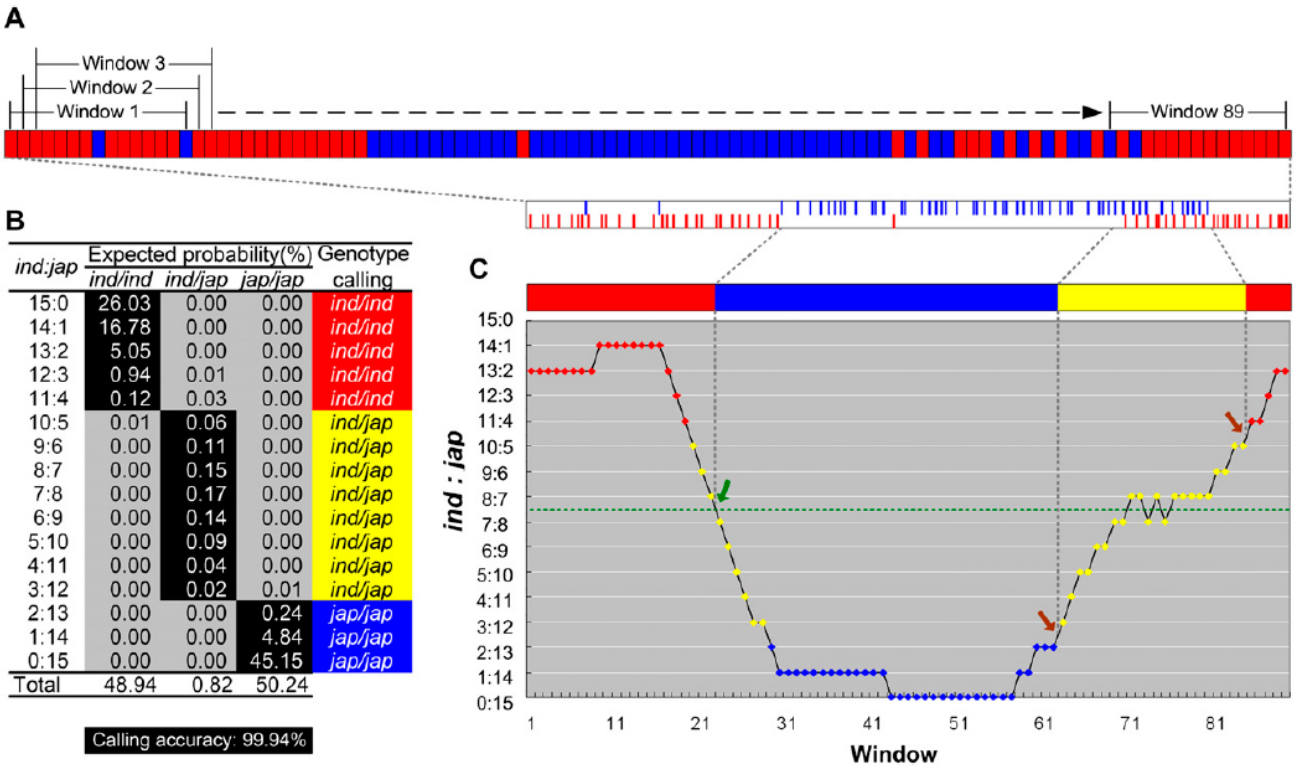


Figure 1. Sequence-based high-throughput genotyping. Rice RILs were developed from a cross between *indica* and *japonica* cultivars. Genome sequences of the parents were aligned and SNPs were identified. Genomes of the RILs were resequenced on the Illumina Genome Analyzer using the multiplexed sequencing strategy. Three-base indexed DNAs of 16 RILs were combined and sequenced in one lane. Sequences were sorted and aligned with the pseudomolecules of parental genome sequences for SNP detection. Detected SNPs were arranged along chromosomes according to their physical locations with genotypes indicated. A sliding window approach was used for genotype calling, recombination breakpoint determination, and map construction.

Genotyping by Whole Genome Re-Sequencing



ind:jap	Expected probability(%) *ind/ind*	*ind/jap*	*jap/jap*	Genotype calling
15:0	26.03	0.00	0.00	*ind/ind*
14:1	16.78	0.00	0.00	*ind/ind*
13:2	5.05	0.00	0.00	*ind/ind*
12:3	0.94	0.01	0.00	*ind/ind*
11:4	0.12	0.03	0.00	*ind/ind*
10:5	0.01	0.06	0.00	*ind/jap*
9:6	0.00	0.11	0.00	*ind/jap*
8:7	0.00	0.15	0.00	*ind/jap*
7:8	0.00	0.17	0.00	*ind/jap*
6:9	0.00	0.14	0.00	*ind/jap*
5:10	0.00	0.09	0.00	*ind/jap*
4:11	0.00	0.04	0.00	*ind/jap*
3:12	0.00	0.02	0.01	*ind/jap*
2:13	0.00	0.00	0.24	*jap/jap*
1:14	0.00	0.00	4.84	*jap/jap*
0:15	0.00	0.00	45.15	*jap/jap*
Total	48.94	0.82	50.24	

Calling accuracy: 99.94%

Figure 2. Sliding window approach for genotype calling and recombination breakpoint determination. (*A*) The *top* stripe of blocks represents SNPs along the hypothetical chromosomal region. This was redrawn from the two stripes of short vertical lines *below* illustrating SNPs detected by aligning 33-mers with the parental genome sequences. (Red) *Indica* genotype; (blue) *japonica* genotype. A sliding window covering 15 SNPs moves from *left* to *right* one base at a time. For each window, the ratio of the number of *indica* to *japonica* SNPs (*ind:jap*) is calculated. (*B*) Genotype calling based on the highest expected probabilities: Call homozygous *indica* genotype (*ind/ind*) when *ind:jap* ≥ 11:4; call heterozygous genotype (*ind/jap*) when 10:5 ≥ *ind:jap* ≥ 3:12; call homozygous *japonica* genotype (*jap/jap*) when *ind:jap* ≤ 2:13. Adding together the probabilities of these callings (shaded in black) gives the calling accuracy of 99.94%. (*C*) As the window slides, genotypes are called and recombination breakpoints are determined. Green and brown arrows point to breakpoints *between* two homozygous genotypes and *between* the heterozygous and homozygous genotypes, respectively. The resulting recombination map for this chromosomal region is illustrated in a solid bar, in which red, blue, and yellow represent genotypes *ind/ind*, *jap/jap*, and *ind/jap*, respectively. Identified breakpoints are indicated *between* SNPs.

Genotyping by Whole Genome Re-Sequencing

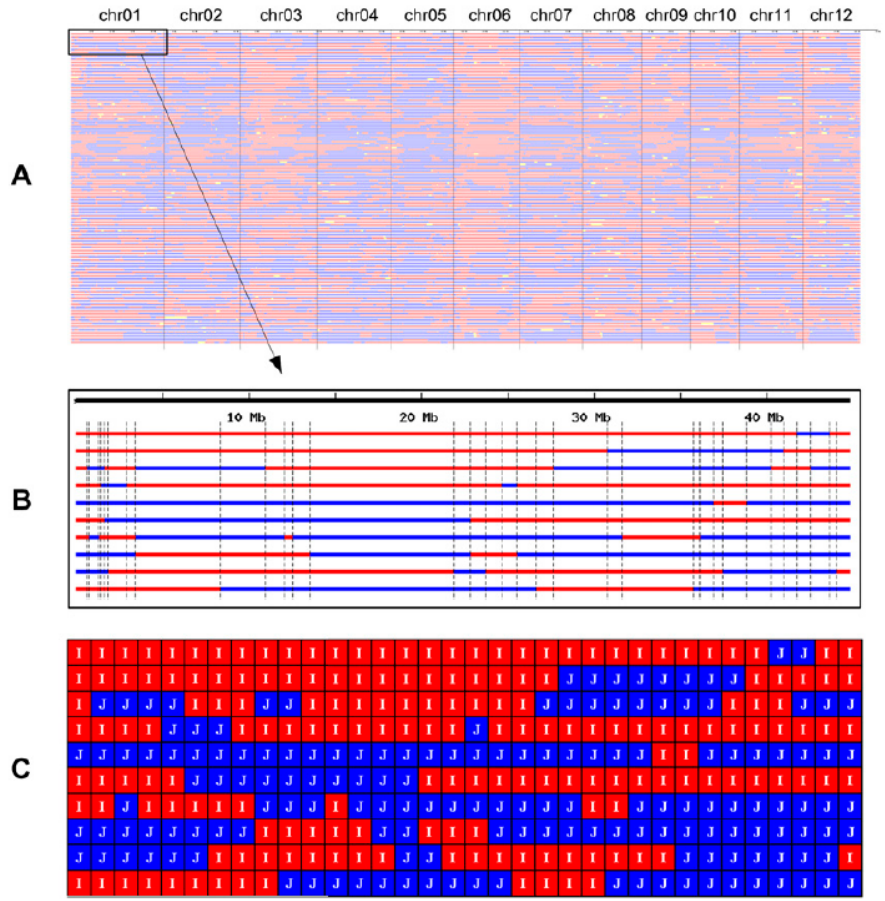


Figure 4. Recombination and bin maps. (*A*) Aligned recombination maps of 150 rice RILs. Red, *ind/ind*; blue, *jap/jap*; yellow, *ind/jap*. (*B*) Aligned chromosome 1 of the first ten RILs. Scale indicates physical distance. A *vertical* line labels a recombination breakpoint. A region *between* two *vertical* lines across all RILs is recognized as a recombination bin. (*C*) Bin map of the 10 RILs.

Genotyping by Whole Genome Re-Sequencing

Application to Mtr.: 1778 markers on chrom. 1 !