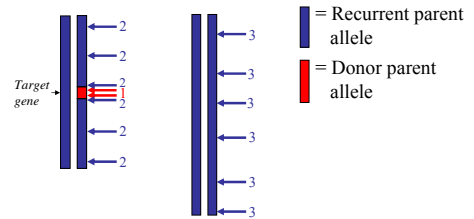


Marker-Assisted Selection

CS741
2009
Jim Holland

Marker-Assisted Backcrossing



1. Select donor allele at markers linked to target gene.
2. Select recurrent parent allele at other linked markers.
3. Select recurrent parent allele at unlinked markers throughout genome.

1. Select donor alleles at markers flanking target gene

- Flanking markers reduce loss of target allele to recombination.
- Useful if target allele phenotype is not easily observed:
 - recessive alleles
 - multiple resistance gene pyramids
 - environmentally-sensitive genes
 - expensive phenotypes (e.g., grain quality)

Probability of losing target allele in marker-assisted backcrossing

- Recombination between marker and target genes will result in selection of wrong allele.
- How often is this a problem?
- Model selection on one marker (M) linked to target gene (Q, may be QTL or major effect gene).
- Backcross heterozygous F_1 to parent 1:

$$\frac{M_1 \text{---} Q_1}{M_1 \text{---} r \text{---} Q_1} \times \frac{M_1 \text{---} Q_1}{M_2 \text{---} r \text{---} Q_2}$$

F_1 Gamete and BC_1F_1 Genotypic Frequencies

Gametes from F_1 :

Gamete	Frequency
$M_1 \text{---} Q_1$	$1/2(1-r)$
$M_1 \text{---} Q_2$	$1/2(r)$
$M_2 \text{---} Q_1$	$1/2(r)$
$M_2 \text{---} Q_2$	$1/2(1-r)$

Gametes from P1:

Gamete	Frequency
$M_1 \text{---} Q_1$	1

×

BC_1F_1 genotypes:

Genotype	Frequency
$M_1M_1Q_1Q_1$	$1/2(1-r)$
$M_1M_1Q_1Q_2$	$1/2(r)$
$M_1M_2Q_1Q_1$	$1/2(r)$
$M_1M_2Q_1Q_2$	$1/2(1-r)$

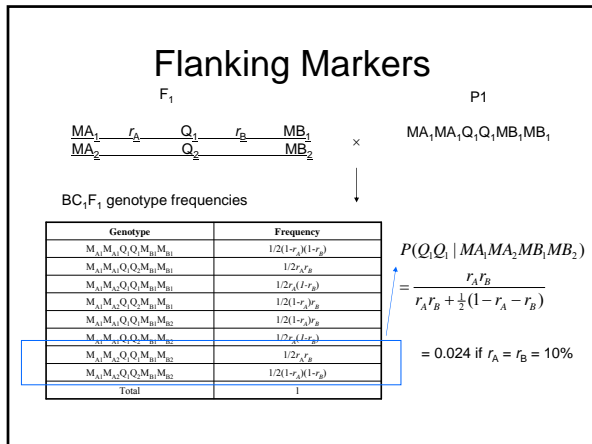
Select these.
Prob. of losing Q_2 allele =
Prob. of selecting Q_1Q_1 among M_1M_2 types:

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

What happens if M_2 allele is recessive to M_1 allele?

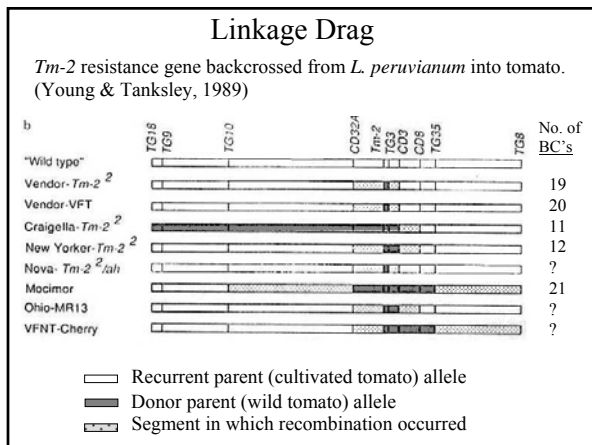
Losing the target allele

- Assume you maintain only one BC_1F_1 progeny per generation of backcrossing.
- Each generation prob. of loss = r .
- What is prob. after t generations of backcrossing?
- Each generation prob. of not loss = $1 - r$.
- So, after t generations, prob. of no loss = $(1 - r)^t$.
- So, after t generations, prob. of loss = $1 - (1 - r)^t$.
(= 41% if $r = 10\%$ and $t = 5$)



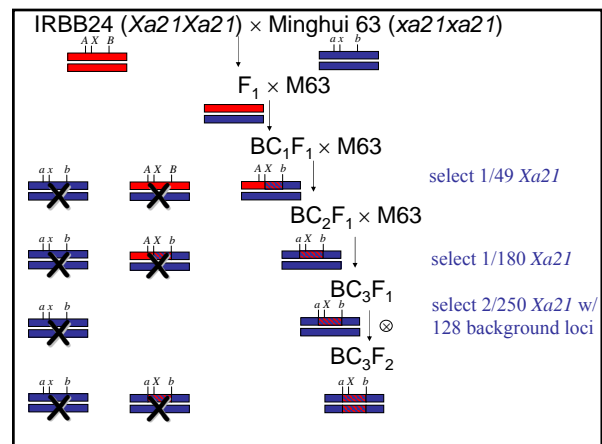
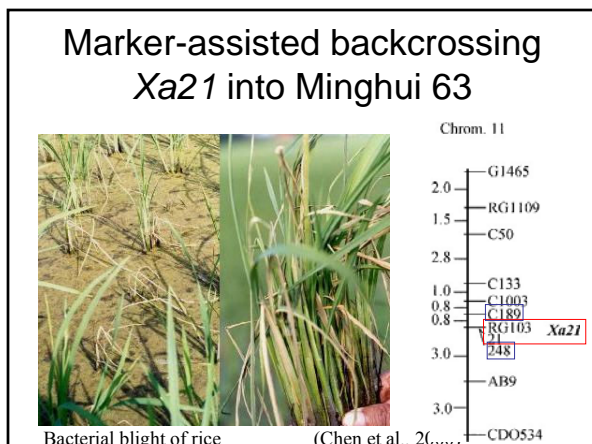
2. Select recurrent parent alleles at other linked markers

- This will reduce linkage drag around target gene (most important when introgressing wild or exotic germplasm).



3. Select for recurrent parent alleles in rest of genome.

- BC progeny vary by chance for amount of recurrent parent genome.
- "Background" markers can identify progeny most similar to recurrent parent.
- MAS requires 1 - 2 generations fewer to recover 97% of RPG (Frisch et al., 1998).
- Reasonable progeny sizes and numbers of markers suffice ($n=20, l=20$).

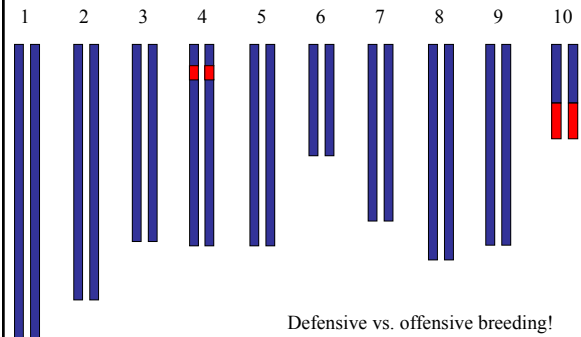


Effect of Xa21 Introgression

Minghui 63 × Zhenshan 97A = Shanyou 63 F₁ hybrid

Entry	Head date	Yield (no disease)	Yield (BB inoculated)
	d	g/plant	
Shanyou 63	100	13.6	9.5
Shanyou 63 Xa21	100	14.0	13.6

Backcrossing is fundamentally conservative!



Implementing MAS in forward crossing

- Forward crossing – each generation, intermate many combinations of complementary elite lines.
- Identify transgressive segregants for multiple traits.
- Result: different combinations of target and marker alleles segregating in different populations.
- So, markers must be **DIAGNOSTIC** across numerous crosses (populations).

Implementing MAS in Australian wheat breeding

- 19 markers regularly used:
- disease resistance,
 - abiotic stress resistance,
 - grain quality
- (Eagles et al., 2001)



A tale of two wheat traits

Cereal cyst nematode resistance

- Difficult phenotype.
- Markers for 2 genes.
- Absolute linkage.
- Sources of resistance – a wild relative and a landrace.
- Markers diagnostic across populations.

Soil boron toxicity tolerance

- Difficult phenotype.
- Markers for 2 genes.
- Tight linkage.
- Source of tolerance – Australian cultivar Federation.
- Markers not always useful across populations

Linkage disequilibrium (LD) between target and marker genes sufficient to implement MAS.

Linked markers can be diagnostic: MAS for SCN resistance in soy



- > Phenotype hard to score reliably.
- > Oligogenetic.
- > Linked markers identified for most important genes.
- > Markers diagnostic across most crosses.

Why are markers linked to SCN resistance diagnostic?

SCN resistance genes and their linked marker alleles are distinct from N. American gene pool

N. America = narrow germplasm pool, lacking SCN resistance

Asia = center of diversity, source of SCN resistance

All SCN resistance genes are derived from outside elite N. American germplasm pool. Novel marker alleles linked to resistance genes have strong LD to resistance genes. → In contrast, LD between wheat boron tolerance genes and their markers has broken down because they have coexisted in common gene pool long enough.

Utility of markers across populations increased by:

- Tight linkage to target gene (ensured by allele-specific markers created from causal gene sequence) – cre genes in Australian wheat.
- Introduction of novel target gene alleles from distinct gene pool – SCN resistance in USA soybean, FHB resistance in USA wheat.

Evolution of MAS:

- Large investment up front in genotyping and phenotyping to mark genes.
- Payoff later if markers are easier to assay than phenotypes and are diagnostic in many populations.
- Start by marker-assisted backcrossing novel alleles from exotic germplasm.
- Once target alleles are introgressed into adapted lines, forward crossing can commence.
- MAS will be self-reinforcing – initial marker selection followed by phenotypic evaluation will maintain marker-target gene LD.
- Industrial scale-up possible.

MAS for polygenic traits

- MAS for yield → mixed results
- Impediments include:
 - Accurate estimation of location and effects of underlying Quantitative Trait Loci is hard.
 - Different QTLs important in different populations.
 - Phenotypic selection is already efficient for moderate to high heritability traits – diminishing returns for MAS.
 - QTL mapping methods not yet integrated into efficient breeding procedures.

Mapping population size limits accuracy of QTL mapping

The Beavis effect!

# QTL	h^2	n	Power	Bias (σ_G^2)
10	.30	100	9%	+559%
10	.95	100	39%	+197%
10	.30	500	57%	+144%
40	.95	500	46%	+165%

(Beavis, 1998)

(more markers do not solve this problem!)

Reports of QTL with large yield effects from small populations are unreliable:

QTL mapping with 990 maize lines in 19 environments: 28 QTL, each with < 0.2 T/ha effect (Openshaw & Frascaroli, 1997).

Empirical Evaluations of MAS

- Early generation selection for hybrid maize yield compared with and without markers.
- Not much advantage to using markers... why?
- Experiment 1: G × E interactions impeded BOTH phenotypic and marker-assisted selection → QTLs mapped in environments not representative of target production environments (Stromberg et al., 1994).
- Experiment 2: line mean heritability for grain yield was high ($h^2 = 0.77$) → phenotypic selection hard to beat (Eathington et al., 1997).

Catch-22 of MAS

If phenotypic data are poor indicators of genotypes, you cannot adequately map QTLs to implement MAS.



If phenotypic data are good, you do not need MAS.

The Catch-22 can be avoided if:

- A small number of QTLs explain most of the genetic variation, in which case:
- High heritability in the QTL mapping phase is optimal to identify QTL markers, then:
- Markers can be implemented more easily/cheaply than phenotyping in future selection cycles.
- BUT – yield variation is not likely to be explained by a few QTLs! And underlying QTLs will vary across populations.

Limits to the Effectiveness of MAS

- QTLs for polygenic traits vary across populations. Mapping populations may not be representative of many breeding populations.
- Breeders typically make 10s - 100s of new crosses (breeding populations) each year.
- QTL mapping in each population is not feasible for public programs... but for private industry?

Applying markers in industrial-scale programs

- Scale of programs is increasing:

Scope of Monsanto's global maize program

Breeding populations:	5,000
Segregating lines:	500,000
Finished lines:	30,000
Test crosses:	500,000
Nursery rows:	2,000,000
Yield trial plots:	4,000,000

Crosbie et al., 2005

Applying markers in industrial-scale programs

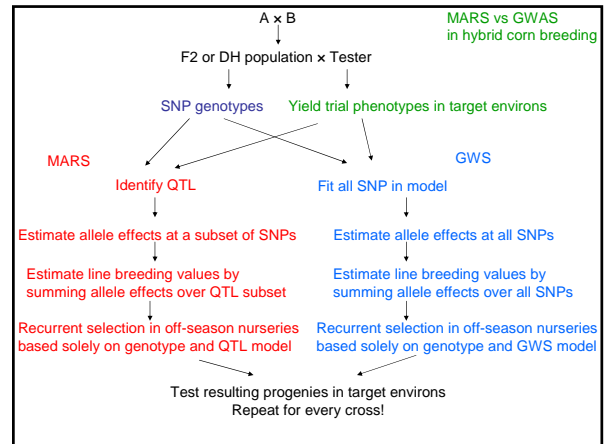
- Costs of genotyping are dropping rapidly:
- ~1,000 SNPs per line/progeny can be collected on a routine basis NOW
- 60k SNP chips cost ~\$150/sample in public sector.
- Next-gen sequencing costs will soon make nearly whole-genome sequence available routinely.
- Genotyping now cheaper than phenotyping.

MARS

- Marker-assisted recurrent selection
- Map QTL based on F2 topcrosses
- Assign each line a molecular-based breeding value = sum of QTL allelic effects for each line
- Recovery of a single line with all favorable alleles not possible in 1 generation.
- Recurrent selection based on molecular breeding values for 2 or more off-season nursery generations to increase favorable allele frequencies.
- Evaluate lines from later cycle of MARS for hybrid development.
- Re-do this separately for each breeding cross!

GWS

- Genome-wide selection
- We know that QTL effect estimates are poor unless sample sizes are huge, this hinders MARS.
- Do not perform QTL mapping. Instead, fit all markers simultaneously in model and estimate breeding value for each marker allele. Then sum over all markers to obtain line breeding value.
- Again, conduct recurrent selection in off-season nurseries for 2 or more generations and evaluate resulting lines.
- We learn nothing about QTL, but this provides better estimates of line breeding values. Black box approach.
- Re-do this separately for each breeding cross!



And finally...

- Genome-wide association studies (GWAS)
- Using 60k SNP information across very large panels of diverse breeding lines, conduct association analysis across whole genome.
- Model population structure to avoid false-positives.
- Identify QTN (quantitative trait nucleotides) with effects on complex traits, then fit these effect estimates in breeding values across families.