

# Principles of marker-assisted selection

## I. Qualitative traits

H.G. Welz<sup>a</sup> and H.H. Geiger<sup>b</sup>

<sup>a</sup> *AgrEvo Seeds/Crop Improvement, D-65926 Frankfurt/M., Germany*

<sup>b</sup> *University of Hohenheim, Institute of Plant Breeding, Seed Science, and Population Genetics, D-70593 Stuttgart, Germany*

### Introduction

In this presentation, we distinguish between qualitative and quantitative traits because different breeding methods are used for their introgression or improvement. For the introgression of qualitative traits such as pathotype-specific disease resistances, which are typically controlled by single, dominant genes, backcross breeding has been used for a long time (Allard, 1960). It allows the transfer of one or few genes from a – mostly agronomically inferior – donor genotype into an elite recipient genotype, the recurrent parent (RP). Backcross (BC) breeding is not suited though for the improvement of quantitative traits (see Part II). We compare classical and marker-assisted BC schemes, distinguishing between marker-assisted foreground and background selection (**Slide 2**). In foreground selection, flanking markers around a target gene are used to guide selection whereas in background selection, markers dispersed throughout the genome are used to recover the RP genotype more efficiently than by phenotypic selection.

### Classical backcross breeding scheme

Classical BC breeding can be termed as phenotypic background selection (**Slide 3**). In each BC generation, carriers of the target gene would be directly identified by a phenotype-based assay and the portion of unwanted donor genes would be halved. For the transfer of a single dominant gene, six BC generations would normally be conducted to recover 99% of the RP genome. This procedure is too time-consuming for a modern maize hybrid breeding program where turnover times of new lines and hybrids are fast. In a BC<sub>1</sub> generation the proportion of the RP genome would be distributed normally around a mean of 75% (note that in later BC generations, the distribution would become increasingly skewed) but given a sufficient

---

In: B.I.G. Haussmann, H.H. Geiger, D.E. Hess, C.T. Hash, and P. Bramel-Cox (eds.). 2000. Application of molecular markers in plant breeding. Training manual for a seminar held at IITA, Ibadan, Nigeria, from 16-17 August 1999. International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502 324, Andhra Pradesh, India.

sample size, it would contain also plants with more than 85% RP genome (**Slide 4**). These plants can be identified with molecular markers to accelerate the breeding process (Tanksley et al., 1989). Without molecular markers flanking the target gene it is nearly impossible to remove the linkage drag coming as a “baggage” with the introgressed segment (**Slide 5**). This has been confirmed experimentally by Murray et al. (1988) who found, using DNA markers, a recovery of only 90% RP genome in two phenotypically selected BC<sub>10</sub>-equivalent conversions of the maize inbred line A632, equipped with resistance genes *Ht1* and *Rp1*, respectively. Thus the intention of marker-assisted backcross (MAB) breeding is to speed up the conversion, and to remove the linkage drag of the transferred gene.

### **Marker-assisted foreground selection**

Principally, one can conduct foreground and background selection with molecular markers (**Slide 6**). In the former case, marker alleles of the donor genotype, in the latter case, marker alleles of the recipient (RP) genotype would be employed. Marker-assisted foreground selection would be especially effective for the transfer of recessive genes since their classical transfer requires additional recurrent selfing generations, a procedure that is prohibitively slow for most commercial breeders. Melchinger (1990) developed equations for dimensioning a foreground selection program (**Slide 7**) but due to a lack of allele-specific markers practical applications in plant breeding are limited. In one MAB project of CIMMYT designed to transfer drought tolerance, three PCR primer pairs derived from RFLPs linked to major drought tolerance QTL were used to preselect BC<sub>2</sub>F<sub>1</sub> individuals, reducing the population size that would need to undergo subsequent phenotypic selection from 2300 to 300 (Ribaut et al., 1997). Final results have not been published though. In animal breeding, allele-specific markers are already being used on a large commercial scale to eliminate disease and stress-susceptibility genes from, e.g. pig, breeding populations (Plastow, 1999).

### **Marker-assisted background selection**

Marker-assisted background selection (term coined by Hospital and Charcosset, 1997) is used extensively in commercial hybrid maize breeding programs in order to transfer herbicide tolerance or insect resistance genes. Ragot et al. (1995) described how the CryIA(b) Bt gene was transferred from a Lancaster to a Stiff Stalk maize inbred line. Since the Bt gene was fused to the *pat* gene, encoding glufosinate (herbicide) tolerance, foreground selection for Bt could be indirectly followed by spraying each BC generation with Basta™ (glufosinate). RFLP markers were successfully used for background selection saving two BC generations (**Slide 8**).

Several parameters can be optimized in a background selection program (**Slide 9**). Flanking markers for the target allele are needed to remove linkage drag. The optimal distance between target and flanking markers governs the selection intensity that can be exerted (**Slide 10**). While the model equation of Hospital et al. (1992) is based on the assumption of infinite population size, Frisch et al. (1999a) chose the method of probability analysis to arrive at practically more relevant conclusions. Their analysis was based on the principle that the probability of crossovers in a chromatid segment follows a Poisson distribution, under the assumption that (i) the average number of crossovers per chromatid is equal to its length in Morgans and (ii) that crossover events are uniformly and independently located on the chromatid [the same assumptions are valid for Haldane's frequently applied mapping function]. With the Poisson distribution function the probability of certain recombination events can be calculated (**Slide 11**). These values are used to calculate the probabilities of finding different genotypes (**Slide 12**) allowing to determine the best suited flanking molecular markers of a target gene for a given BC population size (**Slide 13**). By rewriting these equations, one can also determine how many BC plants must be generated and typed with a special set of flanking markers (**Slide 14**). Depending on the genotype of the selected individual in generation BC<sub>1</sub>, one can then choose the population size for BC<sub>2</sub>. One may also determine the population size in BC<sub>1</sub> such that the number of plants that must be genotyped in a two-generation BC program would be minimized (Frisch et al., 1999a).

Different selection strategies in a marker-assisted background selection program were further compared by computer simulation (Frisch et al., 1999b). Basically the authors considered four different selection schemes (**Slide 15**; note that a two-stage selection scheme is not necessarily equivalent with a two-generation selection scheme). From two-stage to four-stage selection, the genomic marker deployment would be extended in a stepwise manner (analyze target allele → flanking markers of target → all carrier chromosome markers → all genomic marker loci). They also varied population size (constant, increasing, decreasing) over the course of the BC program. Target parameters were the total number of marker data points (MDPs) required for each scheme and the 10%-percentile of the empirical distribution of the RP genome in 10,000 runs of the simulation. That percentile was used as an estimator of the proportion of RP genome recovered. The breeder would then know how to dimension his program to be confident that “with a probability of 90%, the RP proportion would be greater than 98% under the chosen selection scheme”. Employing, for example, a three-stage selection scheme with constant population size, the goal of a 10%-quantile of greater than 96.7% (Q<sub>10</sub> > 96.7, the same as expected after BC<sub>6</sub> in a classical BC scheme) was achieved with 80 individuals per generation (**Slide 16**). The number of MDPs required was 50% lower than in a two-stage selection scheme. Increasing the population size was advantageous. That is: focus on recovering recombinants around the target locus analyzing few individuals in early BC generations while ignoring the residual genome, and recover the residual RP

genome in later BC generations using larger population size. Four-stage selection was additionally saving resources: only about 10% of MDPs were required compared to two-stage selection (Frisch et al., 1999b). Computer software of the simulation program is available (Frisch et al., 1999c).

## **Outlook**

Marker-assisted background selection will become a standard breeding technique in as much as transgenic crop cultivars become more popular in the future. The marker-assisted transfer (e.g. pyramiding) of major disease resistance genes will remain a limited application as long as the costs for DNA marker analyses are high. Even with MAB, the conversion of elite inbred lines still takes about 2 years. This means the conversion breeder is always 2 years behind the “pipeline”, i.e. the best new lines emerging from the basic breeding program. In any case, the utility of MAB will therefore remain restricted to true major genes and exclude QTL with only small effects. The transfer of single, important QTL alleles from otherwise inferior genetic resources is a special case deserving more research. Marker-assisted foreground selection will certainly become more popular as allele-specific markers become available through genomics research programs. Biometrical tools are now available to optimize MAB breeding plans. These will help to predict costs and time lines of conversion programs more accurately.

## Principles of Marker-Assisted Selection I. Qualitative Traits

H. Günter Welz<sup>1</sup> and Hartwig H. Geiger<sup>2</sup>

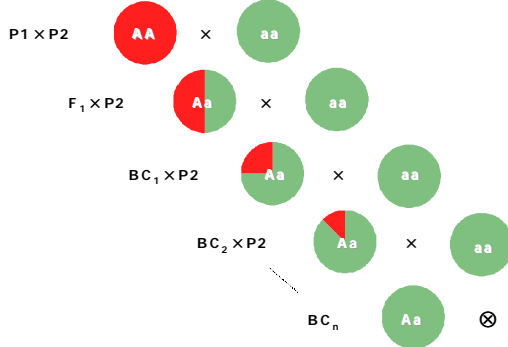
<sup>1</sup> AgrEvo Seeds/Crop Improvement, Frankfurt, Germany

<sup>2</sup> University of Hohenheim, Stuttgart, Germany

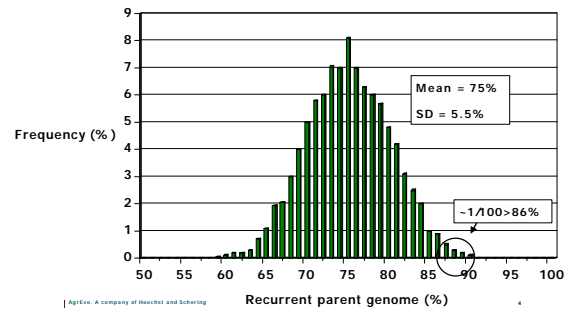
## Marker-assisted selection

- **Qualitative traits (H.G. Welz)**
  - Classical backcrossing scheme
  - Marker-assisted backcrossing (MAB)
    - Marker-assisted foreground selection
    - Marker-assisted background selection
- **Quantitative traits (H.H. Geiger)**
  - Reliability of QTL estimates
  - Theoretical studies on MAS efficiency
  - Empirical results
  - Integration of MAS into breeding programs

## Classical backcrossing scheme



## Distribution of genotypes in simulated BC<sub>1</sub> population: recovery of RP genome n = 10,000 plants; from Frisch (1999)



## Expected linkage drag around target locus in conventional BC programs

Generation	BC <sub>1</sub>	BC <sub>2</sub>	BC <sub>3</sub>	BC <sub>4</sub>	BC <sub>5</sub>	BC <sub>6</sub>	BC <sub>8</sub>	BC <sub>10</sub>	BC <sub>20</sub>
Length (cm)	78.7	63.2	51.8	43.2	36.7	31.7	24.5	19.9	10.0

Assuming target locus in the center of 100 cM chromosome.  
Calculations from Stam and Zeven (1981), modified after Frisch (1999).

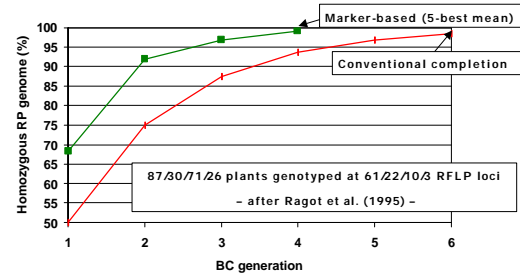
## Ways of using molecular markers in backcross breeding

- **Marker-assisted foreground selection**
  - Select for marker allele(s) of donor genotype
  - For identifying „difficult“ target alleles (recessive alleles, alleles expressed after flowering, alleles whose phenotypic identification is costly)
  - Close linkage between marker loci and target locus essential
- **Marker-assisted background selection**
  - Select for marker alleles of recurrent parent genotype
  - To select against donor genetic background
  - Loose or no linkage between marker loci and target locus, but good genome coverage desired
- **Fore- and background selection often conducted in combination**

## Description of a foreground selection breeding plan

- Cross donor X recurrent parent (RP). Backcross F<sub>1</sub> to RP, generating N BC<sub>1</sub> plants
- Genotype BC<sub>1</sub> plants at flanking markers of donor target allele
- Backcross all BC<sub>1</sub> plants with presumed target allele to RP
- Genotype K plants from each BC<sub>2</sub> family
- Per family, backcross one carrier of target allele to RP
- Repeat until generation BC<sub>i</sub> to obtain R carriers of target allele. Verify presence of the t. allele. In case of recessive gene self first.
- Determine N, K, the probability of success P, and M, the number of plants to be genotyped per BC generation with equations of Melchinger (1990). Example: A allele-specific marker (r=0), P=0.95, R=4, and K = 4 ⇒ N=14 in BC<sub>2</sub>

## Marker-assisted introgression of Bt gene into corn inbred line



### Parameters to be optimized in a marker-assisted background selection program

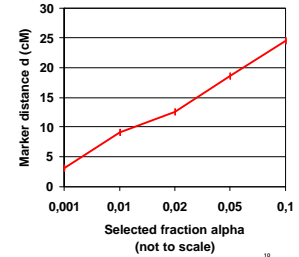
- Optimal distance between target locus and flanking markers for a given population size
- Minimal number of individuals for detecting recombinants in a given marker interval
- Minimal number of data points to achieve fast completion of BC program
- Allocation of marker analyses to different BC generations

### Optimal distance between target locus and flanking markers

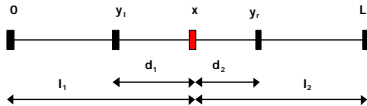
Model equation by Hospital et al. (1992)

- $\alpha$  = fraction of selected BC<sub>1</sub> plants
- $d_i$  = distance between flanking marker and target locus

$$d_1 = d_2 = \frac{1}{2} \ln(1 + 2\sqrt{\alpha})$$



### Probability of crossovers derived from Poisson distribution



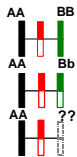
- A: No crossover in  $l_1 - d_1$        $p_A = e^{-(l_1 - d_1)}$   
 B: Odd no. of crossovers in  $d_1$        $p_B = \sinh(d_1)e^{-d_1}$   
 C: Odd no. of crossovers in  $d_2$        $p_C = \sinh(d_2)e^{-d_2}$   
 D: No crossover in  $l_2 - d_2$        $p_D = e^{-(l_2 - d_2)}$

### Genotypes occurring in BC generations

Plants would be heterozygous at target locus and otherwise be

- Type 1:** homozygous carrier of RP allele at both flanking markers  
**Type 2:** homozygous carrier of RP allele at one flanking marker, heterozygous at the other  
**Type 3:** homozygous carrier of RP allele at one flanking marker, homo- or heterozygous at the other

What is the minimum distance of flanking markers to find with probability  $q$  at least one plant of a certain type in a population of size  $n$ ?



Type 1:  $d_1 = -\ln(1 - 2\sqrt{2 - 2\sqrt{1 - q}}) / 2$

Type 2:  $d_1 = -\ln(4\sqrt{1 - q} - 3) / 4$

Type 3L/R:  $d_i = -\ln(4\sqrt{1 - q} - 3) / 2$

Assuming  $d_1 = d_2, l_1 = l_2$

After Frisch et al. (1999a)

### Choosing BC<sub>1</sub> population size based on distances of flanking markers from target gene

With a probability of 99% at least one individual should be generated that carries the target gene and at both flanking markers the RP allele (Type-1 case):  $q_1 = 0.99$

The distances between the target locus and the two flanking markers are 20 cM and 25 cM:  $d_1 = 0.2, d_2 = 0.25$

$$n = \ln(1 - q) / \ln[1 - \frac{1}{8}(1 - e^{-2d_1})(1 - e^{-2d_2})] = 282$$

At least 282 individuals must be generated and genotyped.

After Frisch (1999)

### Different selection strategies in MAB

	Stage 1	Stage 2	Stage 3	Stage 4
<b>Two-stage selection</b>	Select carriers of target allele	Select for RP genome at all loci	—	—
<b>Three-stage selection</b>	Select carriers of target allele	Select for recombinants at flanking markers	Select for RP genome at all loci	—
<b>Four-stage selection</b>	Select carriers of target allele	Select for recombinants at flanking markers	Select for RP genome on carrier chromosome	Select for RP genome at all loci

### Three-stage selection scheme with constant population size

	Number of individuals per BC generation							
	20	40	60	80	100	125	150	200
----- Q10 of RP genome (%) -----								
BC <sub>1</sub>	71.2	72.7	73.4	73.6	73.3	73.2	72.8	72.2
BC <sub>2</sub>	86.1	87.2	88.5	89.3	90.2	90.7	91.3	91.8
BC <sub>3</sub>	94.4	95.7	96.5	<b>96.9</b>	97.2	97.3	97.5	97.6
----- Total no. of MDPs required -----								
BC <sub>1</sub>	250	320	420	510	590	690	750	840
BC <sub>2</sub>	440	610	830	1100	1390	1780	2210	3110
BC <sub>3</sub>	550	820	1130	<b>1470</b>	1810	2260	2740	3740

## References

- Allard, R.W. 1960. Principles of plant breeding. Wiley, New York.
- Frisch, M. 1999. Marker-assisted selection in recurrent backcrossing for transfer of a target gene. Dissertation thesis, University of Hohenheim, Stuttgart, Germany.
- Frisch, M., M. Bohn, and A.E. Melchinger. 1999a. Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. *Crop Science*: in press.
- Frisch, M., M. Bohn, and A.E. Melchinger. 1999b. Comparison of selection strategies for marker-assisted backcrossing of a gene. *Crop Science*: in press.
- Frisch, M., M. Bohn, and A.E. Melchinger. 1999c. PLABSIM: software for simulation of marker-assisted backcrossing. *Journal of Heredity*: in press.
- Hospital, F. and A. Charcosset. 1997. Marker-assisted introgression of quantitative trait loci. *Genetics* 147:1469-1485.
- Hospital, F., C. Chevalet, and P. Mulsant. 1992. Using markers in gene introgression breeding programs. *Genetics* 132: 1119-1120
- Melchinger, A.E. 1990. Use of molecular markers in breeding for oligogenic disease resistance. *Plant Breeding* 104:1-19.
- Murray, M.G., Y. Ma, J. Romero-Severson, D.P. West, and J.H. Cramer. 1988. Restriction fragment length polymorphisms: what are they and how can breeders use them? Pages 72-87 in: Proc. 43<sup>rd</sup> Annual Corn and Sorghum research Conference, American Seed Trade Association, eds. D. Wilkinson et al., Washington, DC.
- Plastow, G.S. 1999. First applications of molecular tools in pig breeding. Book of Abstracts, 7<sup>th</sup> Agrogene Seminar, 25-26 Feb 1999, Paris, France.
- Ragot, M., M. Biasioli, M.F. Delbut, ..., and G. Gay. 1995. Marker-assisted backcrossing: a practical example. Pages 45-56 in: Techniques et utilisations de marqueurs moléculaires. Les Colloques No.72, INRA, Paris, France.
- Ribaut, J.M., X. Hu, D. Hoisington, and D. González-de-Léon. 1997. Use of STSs and SSRs as rapid and reliable preselection tools in a marker-assisted selection-backcross scheme. *Plant Molecular Biology Reporter* 15:154-162.
- Stam, P. and A.C. Zeven. 1981. The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. *Euphytica* 30:227-238.
- Tanksley, S.D., N.D. Yound, A.H. Paterson, and M.W. Bonierbale. 1989. RFLP mapping in plant breeding: new tools for an old science. *Bio/Technology* 7: 257-263.