Rapid and accurate identification of *in vivo*-induced haploid seeds based on oil content in maize

Albrecht E. Melchinger[1], Wolfgang Schipprack[1], Tobias Würschum[2], Shaojiang Chen[3], Frank Technow[1]

[1] Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70599 Stuttgart, Germany

[2] State Plant Breeding Institute, University of Hohenheim, 70599 Stuttgart, Germany

[3] National Maize Improvement Center of China, China Agricultural University, 100193 Beijing, China

Corresponding author: Albrecht E. Melchinger, e-mail: melchinger@uni-hohenheim.de,

SUPPLEMENTARY INFORMATION

Supplementary Table S1 | Numerical examples

	UH600			UH601
	PDH3 x PDH8	F103 x F087	P204 x P211	S072 x P213
Means[a] (OC in %)				
$\bar{x}_{♀}$	3.76a	4.15a	3.44a	3.35a
$\bar{x}_{♂}$	9.91d	9.91d	9.91d	11.63c
$\bar{x}_H$	4.16b	4.56b	3.89b	3.58a
$\bar{x}_C$	5.64c	6.63c	5.88c	5.57b
Phenotypic variances (OC in%)				
σ_H^2	0.84	1.05	0.98	0.44
σ_C^2	0.61	0.26	0.22	0.30
Threshold for OC-based classification (OC in %)				
t [b]	4.50	5.50	4.75	4.40
Haploid induction rate (HIR)				
GS	8.90	9.82[c]	15.60[d]	9.14
OC	8.55	9.13	13.18	9.47
R1-nj	5.01	n.d.[e]	n.d.	n.d.
False-discovery rate (FDR)				
OC	0.18	0.21	0.30	0.10
R1-nj	0.21	n.d.	n.d.	n.d.
False-negative rate (FNR)				
OC	0.21	0.01	0.04	0.10
R1-nj	0.56	n.d.	n.d.	n.d.

Means ($\bar{x}_{♀}$, $\bar{x}_{♂}$) of source germplasm (♀) and inducer (♂) as well as means ($\bar{x}_H, \bar{x}_C$) and phenotypic variances (σ_H^2, σ_C^2) for oil content (OC) of haploid (H) or diploid crossing (C) seeds in four induction crosses in maize, threshold *t* for OC-based classification of H vs. C seeds, estimates of haploid induction rate (HIR), false discovery rate (*FDR*) and false negative rate (*FNR*) determined from the “gold standard” (GS) and classification based on the OC or the *R1-nj* marker.

[a] Numbers followed by the same letter in a column are not different from each other at P < 0.05 based on a t-test according to Snedecor and Cochran (1980, p. 97)

[b] Threshold *t* used for classification based on OC

[c] Extrapolated from the total number of seeds classified based on OC (T = C : N_T = 7914 and T = H : N_T = 795) and estimates of the *FDR* and *FNR* determined from the GS test results with a random subset of these seeds (T = C : N = 213 and T = H: N = 682)

[d] Extrapolated from the total number of seeds classified based on OC (T = C : N_T = 3426 and T = H : N_T = 474) and estimates of the *FDR* and *FNR* determined from the GS test results with a random subset of these seeds (T = C : N = 205 and T = H: N =318)

[e] n.d.= not determined

Supplementary Table S2 | Numerical examples

Induction cross	Gold standard	Result of test T		N[a]	N_T[b]
		OC < *t*	OC ≥ *t*		
(PHD3 x PDH8) x UH600	D = H	101	27	128	
t[c] = 4.50 %	D = C	22	1288	1310	
(F103 x F087) x UH600[d]	D = H	577	5	682	795
t = 5.50 %	D = C	105	208	213	7914
(P204 x P211) x UH600[c]	D = H	294	11	305	474
t = 4.75 %	D = C	24	194	218	3426
(S072 x P213) x UH601	D = H	76	8	84	
t = 4.40 %	D = C	11	824	919	
Classification based on R1-nj marker					
		T = W	T = P		
(PHD3 x PDH8) x UH600	D = H	57	71	128	
	D = C	15	1295	1310	

Classification of haploid (H) vs. diploid crossing (C) seeds in four induction crosses in maize based on (i) their oil content (OC < *t* or OC ≥ *t*) or (ii) expression of the *R1-nj* marker (W = white vs. P = purple) in their embryo and (iii) the "gold standard" results.

[a] N = number of germinated seeds in "gold standard" test

[b] N_T = Total number of seeds classified based on OC

[c] Threshold *t* used for classification based on OC

[d] Note that in theses crosses only a random subset of seeds was evaluated for the "gold standard" test

		T = test result		
		H	C	
D = true status	H	P [T = H , D = H] *True positive (TP)*	P [T = C, D = H] *False negative (FN)*	P [D = H]
	C	P [T = H , D = C] *False positive (FP)*	P [T = C , D = C] *True negative (TN)*	P [D = C]
		P [T = H]	P [T = C]	1

Figure S1 | Contingency table for seed classification.

Two-way contingency table of a binary classification test T showing the four possible outcomes (H = haploid seed; C = diploid crossing seed).

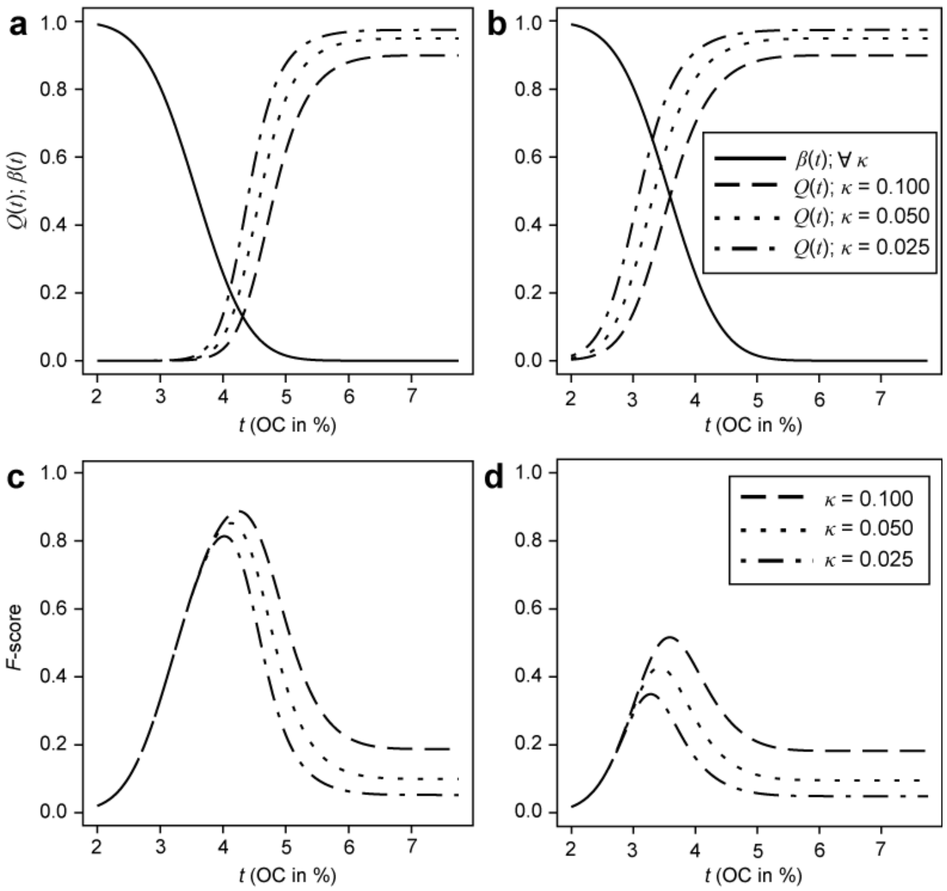

Figure S2 | Test performance in dependence on the selected threshold *t*.

False discovery rate $Q(t)$ and false negative rate $\beta(t)$ (**a**,**b**) as well as *F*-score (**c**,**d**) for classification of seeds into haploid (H) and diploid crossing (C) seeds based on their oil content (OC in %) as a function of the threshold t (H: OC < t; C: OC ≥ t). Assumptions are: (i) seeds were produced by pollination with a high oil (HO) inducer having a haploid induction rate (HIR) κ = 0.100, 0.050, 0.025, (ii) H seeds have a mean μ_H = 3.57% and standard deviation σ_H = 0.55% for OC, and (iii) C seeds have a mean μ_C = 5.51% (**a**,**c**) or 4.48% (**b**,**d**) and a standard deviation σ_C = 0.66%.

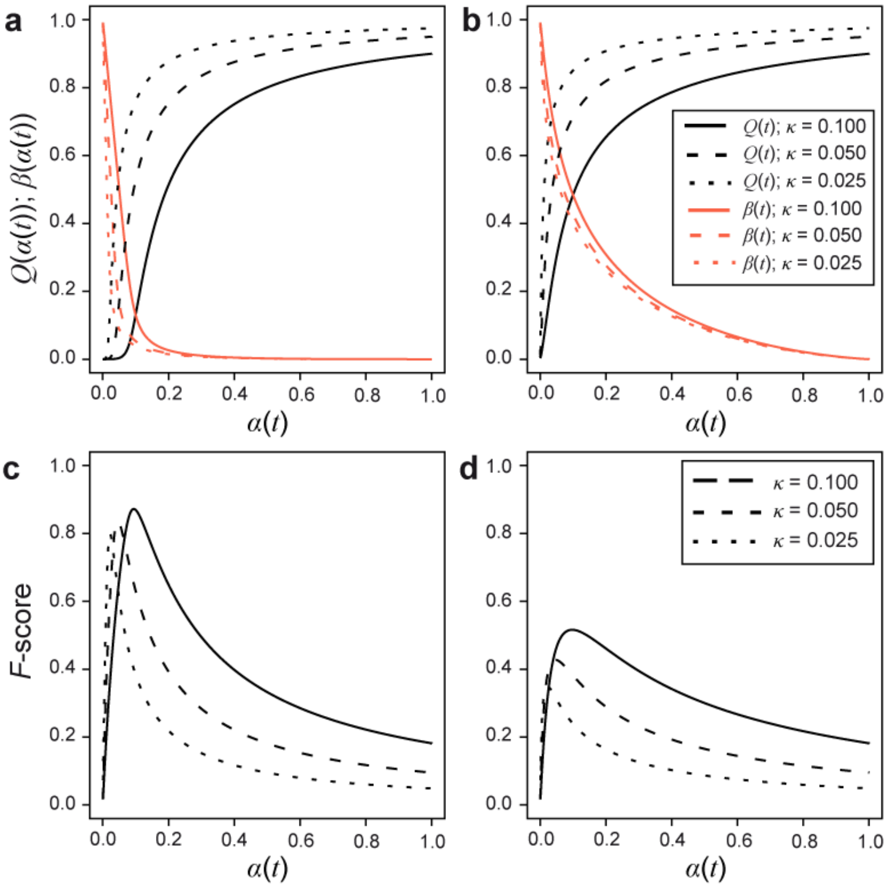

Figure S3 | Test performance in dependence on α.

False discovery rate $Q(t)$ and false negative rate $\beta(t)$ (**a**,**b**) as well as *F*-score (**c**,**d**) for classification of seeds into haploid (H) and diploid crossing (C) seeds based on their oil content (OC in %) as a function of the proportion α of selected seeds. Assumptions are: (i) seeds were produced by pollination with a high oil (HO) inducer having a haploid induction rate (HIR) κ = 0.100, 0.050, 0.025, (ii) H seeds have a mean μ_H = 3.57% and standard deviation σ_H = 0.55% for OC, and (iii) C seeds have a mean μ_C = 5.51% (**a**,**c**) or 4.48% (**b**,**d**) and a standard deviation σ_C = 0.66% for oil content.