

# Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes

S. Maenhout<sup>1</sup> B. De Baets<sup>2</sup> G. Haesaert<sup>1</sup>

<sup>1</sup>Department of Plant Production  
University College Ghent, Belgium

<sup>2</sup>Department of Applied Mathematics, Biometrics and Process Control  
Ghent University, Belgium

XIV meeting of the Biometrics in Plant Breeding Section



# Outline

- 1 Introduction
- 2 WAIS estimator
  - Properties
  - Simulation study
  - Maize breeding data
- 3 CoCoa

# Outline

- 1 Introduction
- 2 WAIS estimator
  - Properties
  - Simulation study
  - Maize breeding data
- 3 CoCoa

## Coefficient of coancestry

The coefficient of coancestry (CoC) between two individuals  $i$  and  $j$  is defined as the probability that at an allele drawn from both  $i$  and  $j$  at the same locus is identical by descent (ibd) from a recent common ancestor.

- is often used to model the covariance between polygenic backgrounds in breeding value estimation or association studies
- can be estimated from pedigree information (e.g. tabular method)
  - requires pedigree knowledge tracing back to the natural populations
  - assumes equal contribution of both parents (absence of selection, genetic drift, ...)

## Molecular marker-based estimation

The CoC can be estimated from similarities in molecular marker fingerprints but:

- an allele identity does not necessarily reflect an identity by descent
- assumptions made by estimators from population genetics are often violated in breeding pools
  - knowledge of allele frequencies
  - linkage equilibrium between markers
  - Hardy-Weinberg equilibrium
- resulting estimators do not always lie within the unit interval  $[0, 1]$
- resulting coancestry matrices  $\mathbf{A}$  are not necessarily positive semi-definite

## Positive semi-definite

A matrix is psd if all its eigenvalues are greater or equal to zero or equivalently if

$$\mathbf{v}'\mathbf{A}\mathbf{v} \geq 0, \quad \forall \mathbf{v} \neq \mathbf{0}.$$

- every variance-covariance matrix is psd by definition !
- if  $\text{Var}(\mathbf{g}) = 2\sigma_g^2\mathbf{A}$  then CoC matrix  $\mathbf{A}$  must be psd as well
- linear mixed model packages rely on the psd property of user supplied coancestry matrices

# Outline

- 1 Introduction
- 2 **WAIS estimator**
  - Properties
  - Simulation study
  - Maize breeding data
- 3 CoCoa

# Calculation of WAIS

$$\mathbf{A}^{\text{WAIS}} = \frac{1}{4I} (\mathbf{XWX}' + \mathbf{Q}) ,$$

- $\mathbf{W}$  is a diagonal matrix containing a weight for each allele  $z$ :

$$\omega_z = P(\alpha_i \stackrel{\text{ibd}}{=} \alpha_j \mid \alpha_i = z, \alpha_j = z)$$

- $\mathbf{Q}$  diagonal matrix containing a correction factor for each genotype  $i$ :

$$q_i = \sum_{z=1}^p x_{(i,z)}^2 y_{i,z}$$

## Properties of the WAIS estimator

The WAIS estimator is ideally suited for use in hybrid breeding programs

- corrects the observed allele identities by means of allele-dependent weights
- weights are estimated from observed allele identities between unrelated genotypes (cf. Bernardo, 1993)
- allows for a mix of inbred and non-inbred genotypes
- all coancestry estimates are guaranteed to lie within the unit interval
- always produces psd coancestry matrices
- achieves a similar or better linear mixed model fit compared to other marker-based estimators

# Simulated breeding data

Studying the behaviour of WAIS in breeding pools requires simulation of the entire breeding history including:

- artificial phenotypical selection
- linkage disequilibrium between marker loci
- varying genotype relatedness

We simulated 8 breeding cycles (cf. Stich *et al.* 2007, Hohenheim University maize breeding program):

- 250 QTL, allele number  $\sim \text{Pois}(\lambda+2)$ ,  $\lambda = 0 \dots 12$
- allele frequencies in subpopulations  $\sim \text{Dir}(\mathbf{a})$  where  $\mathbf{a} = \left\{ p \left( \frac{1-F_{st}}{F_{st}} \right) \right\}$  (Balding, 2003)
- 101 SSR markers (RAGT R2n linkage map)

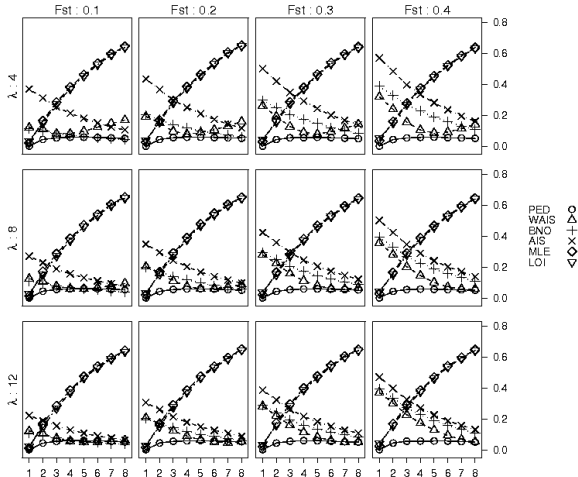
## The competition

After each breeding cycle the following CoC estimators were calculated:

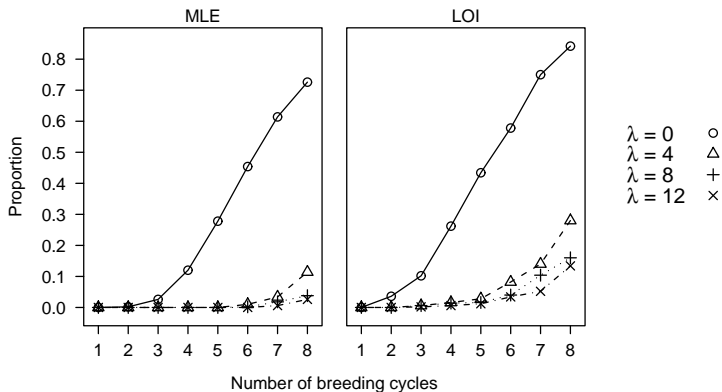
- the actual ibd CoC
- pedigree-based CoC
- uncorrected alikeness in state (AIS)
- WAIS
- BNO (Bernardo, 1993)
- MLE, maximum likelihood estimator (Thompson, 1975)
- LOI (Loiselle *et al.*, 1995)

# Bias and variance

Average root mean squared error over 100 iterations :



# Proportion of non-psd matrices

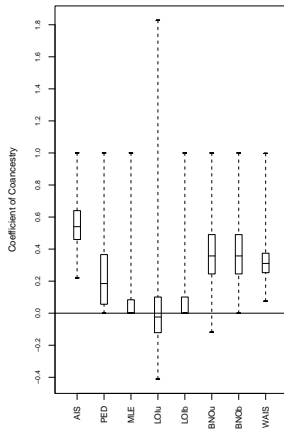


## Application to maize breeding data

The pedigree and five marker-based CoC measures were used to model the covariance of GCA and SCA components in a linear mixed model analysis of a subset of the phenotypic evaluation data of the maize breeding program of RAGT R2n:

- 2.367 hybrids
- 92 Iodent and 105 Iowa Stiff Stalk Synthetic lines genotypes with 75 SSR markers
- 40.432 plots
- 1280 multi-environment trials
- 110 locations spread over Europe
- 3 traits: grain yield, grain moisture content, days until flowering

# Range of pairwise CoC estimates



# Restricted log-likelihoods

	yield	moisture %	flowering
PED	-222740.1 (4)	-194696.6 (1)	-55339.6 (1)
AIS	-222734.8 (2)	-194710.8 (2)	-55343.8 (2)
BNO	-222734.8 (1)	-194712.9 (3)	-55344.1 (3)
WAIS	-222739.2 (3)	-194715.3 (4)	-55347.7 (4)
MLE	-222743.2 (6)	-194716.2 (5)	-55357.0 (5)
LOI	-222741.0 (5)	-194725.6 (6)	-55361.9 (6)

- BNO and LOI values were bounded within the unit interval
- MLE and the bounded LOI matrices were bended towards the closest psd matrix

# Outline

- 1 Introduction
- 2 WAIS estimator
  - Properties
  - Simulation study
  - Maize breeding data
- 3 CoCoa

# CoCoa:

software for estimating the coefficient of coancestry from multilocus genotype data

- provides AIS, WAIS, BNO, MLE and LOI estimators
- uses the same input file format as Structure (Pritchard *et al.* 2000)
- allows to invert well-conditioned CoC matrices
- provides 2 bending routines for bending non-psd or (nearly) singular matrices
- allows to export (inverse) CoC matrices in file formats used by ASReml, SAS Proc Mixed, Wombat, Matlab, . . .
- available for Linux and Windows OS
- freely available, open source (GNU General Public License)

# CoCoo user interface

The screenshot displays the CoCoo software interface with the following components:

- Project Data Table:**

Label	Pop ID	Flag	Locus 1	Locus 2
			Z002	Z001
			-1	4.02
L11004	1	1	1	2
	1	1	1	2
- Project Information Panel:**

Project Name: maize-RAGT  
 Project Path: /home/smaenhout  
 Project Type: Structure

---

Number of Individuals: 190  
 Loci: 75  
 Ploidy: 2  
 Value represented as: -5
- LOI\_matrix.dat Table:**

Labels	L1004	L1015	L
L11004	0.77376846550...	0.33103191718...	0.00000...
L11015	0.33103191718...	0.00000...	
L11065	0.0000000		
L111088	0.0000000		
L111089	0.0000000	1.02483426056...	0.23061320700... -0.243
L111094	0.0000000	0.23061320700...	0.87250170524... -0.261
L111095	0.0000000	-0.2434317600...	-0.2616903944... 0.9551
L111096	0.0000000	-0.3250638101...	-0.2854148013... 0.6071
L111121	0.3339900	-0.2825194192...	-0.2776149963... 0.6181
L111128	0.2439955	-0.2887828990...	-0.2375523615... 0.2145
L111129	0.3326222	-0.1614451733...	-0.0754700500... 0.1337
L111132	0.3509700	-0.2538905894...	-0.2605676952... 0.3772
- Context Menu:**
  - Matrix Manipulation
    - Bound
    - Invert
    - End
  - Export
  - Delete
- CoCooCore version 1.0 Log:**

Parameter file: /home/smaenhout/maize-RAGT/maize-RAGT.cpf  
 Data file: /home/smaenhout/maize-RAGT/Project\_Data.dat  
 Output file: /home/smaenhout/maize-RAGT/results/WAS/WAS\_matrix.dat  
 Number of individual genotypes: 196  
 Number of loci: 75

Calculating WAS estimator...ok

# CoCoa homepage

<http://webs.hogent.be/cocoa>

provides:

- Windows installer
- 32 and 64 bit binaries for Linux
- manual
- C++ and Java source code
- ...

# Summary

## WAIS:

- is a new, marker-based CoC estimation procedure ideally suited for use in hybrid breeding programs
- estimators always lie within the unit interval
- resulting CoC matrices are guaranteed to be psd
- WAIS performs best in advanced cycle breeding pools

## CoCoa:

- is a free software tool for estimating the CoC from molecular marker data
- provides 5 estimation procedures: AIS, WAIS, BNO, MLE and LOI
- provides several useful matrix manipulation tools: bending, bounding, inversion, . . .