The 9th Workshop on QTL Mapping and Breeding Simulation
The University of Sydney, Cobbitty NSW, 7-9 March 2012

Principles of Modeling and Breeding Simulation

Jiankang Wang
CIMMYT China and CAAS

E-mail: wangjk@caas.net.cn or jkwang@cgiar.org
http://www.isbreeding.net

Outline of the presentation

- Quantitative genetics and plant breeding
- Why we need breeding simulation?
- Tools and principles of breeding simulation
- An illustrated example

Quantitative genetics and plant breeding

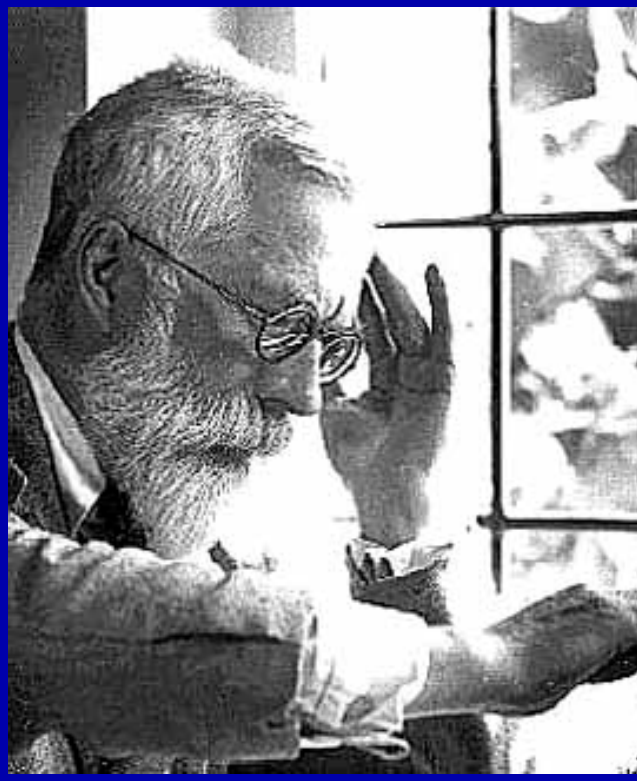
Polygene (or multi-factorial) hypothesis of the inheritance of quantitative traits

- R. A. Fisher (1918) "The correlation between relatives on the supposition of Mendelian inheritance"
- Multiple-factor hypothesis (polygene system)
  - A hypothesis to explain quantitative variation by assuming the interaction of a large number of genes (polygenes) each with a small additive effect on the character.
  - Number of genes, gene effects, environmental modifications

Classical quantitative genetics built on polygene hypothesis

(1920s-1940s)

- R. A. Fisher
- J. B. S. Haldane
- S. Wright
- Application in breeding (1940s-60s)
  - J. L. Lush
  - G. Malecot
  - G. F. Sprague and L. A. Tatum



From polygene hypothesis to QTL mapping

- $P = G + E + GE + e$
- $G = A + D + I$
- $V_G = V_A + V_D + V_I$
  - All genes contribute to A, D, I, V_A, V_D, and V_I

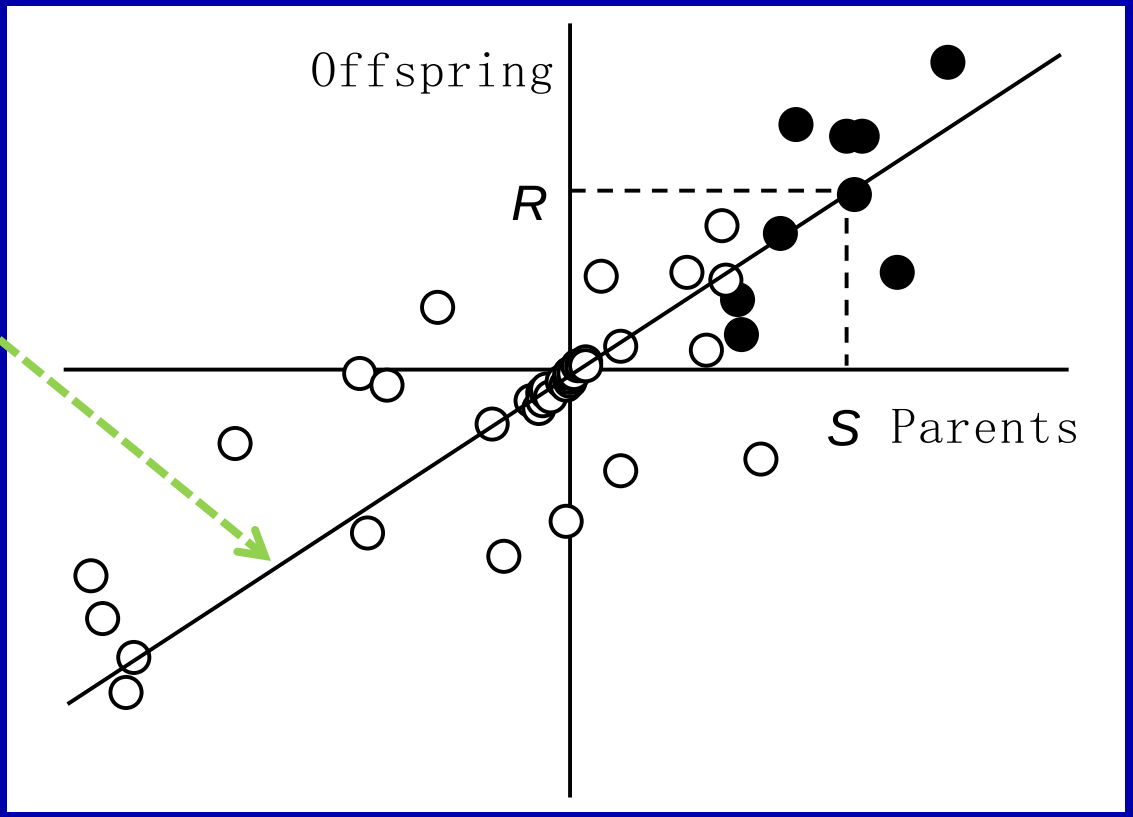
- QTL mapping: $G = \Sigma (wa+vd)$
  - Be able to study individual quantitative trait genes

Estimation of R based on heritability

- $y = b\,x = h^2\,x$

- Response to selection or Genetic gain:

$R \text{ or } \Delta G = \mu_1 - \mu_0 = h^2\,S$



Estimation of genetic gain

$$R = h^2 S = k_p h^2 \sqrt{V_P}$$

$$R = k_p h \sqrt{V_A}$$

$$R = \frac{k_p V_A}{\sqrt{V_P}}$$

Ways to increasing the response to selection

$$R = k_p h \sqrt{V_A}$$

- Increase selection intensity
  - p=10%, k_p=1.755; p=1% , k_p=2.665. So, R(p=1%) = 1.52 R(p=10%).
  - Limitation in increasing selection intensity
    - Strong selection needs much larger a population.
    - Certain amount of individuals is needed for retaining genetic variation for future genetic gain, and avoiding genetic drift
    - When 30 individuals are needed to form the next generation of breeding population, 300 individuals in the parental population are needed for p=10%; 3000 individuals are needed for p=1%.

Ways to increasing the response to selection

$$R = k_p h \sqrt{V_A}$$

- Increase the coefficient of additive variance (V_A) in recurrent selection
  - By pollen control in selection
  - By recombination of S1 instead of half-sib families in half-sib family selection
  - By selection S2 instead of S1 families

$$Cov_{\mathrm{OP}} = Cov_{\mathrm{O\bar{P}}} = \tfrac{1}{2} V_A$$

$$Cov_{\mathrm{HS}} = \tfrac{1}{4} V_A$$

$$Cov_{\mathrm{FS}} = \tfrac{1}{2} V_A + \tfrac{1}{4} V_D$$

Ways to increasing the response to selection

$$R = k_p h \sqrt{V_A}$$

- Increase additive variance (V_A) itself
  - By introgressing other germplasm into the population
- Increase heritability
  - By reducing non-genetic effects

Why we need breeding simulation?

Breeding methods with self-pollinated crops

Allard, R.W. 1960. Principles of plant breeding. John Wiley & Sons, Inc.

Stoskopf, N.C. 1993. Plant Breeding — Theory and Practice. Westview Press.

- Mass and pure-line selection
- The pedigree system
- The bulk population method
- The backcross breeding method
- Single seed descent (SSD), a special case of pedigree system
- Recurrent selection breeding method
- Mutation breeding
- Haploid breeding system (doubled haploid)

Breeding methods in CIMMYT's wheat breeding program

- Pedigree system: before 1984.
  - "Pedigree selection" is used from F2 to F6.
- Modified pedigree/bulk (MODPED): in 1985-1989/94.
  - "Pedigree selection" is used in F2 and F6, and "bulk selection" is used in other generations.
- Selected bulk (SELBLK): after 1995.
  - "Pedigree selection" is used only in F6, and "bulk selection" is used in other generations.

Why do we need tools in breeding?

- To improve the efficiency of traditional phenotypic selection through exploring various options
- To better use the large amount of gene information available from
- To build a bridge between the biological data and breeders' requirements
- To combine all these sources of data into "knowledge" that breeders can use in their breeding programs
- To avoid the simplified assumptions made in classical quantitative genetic theory

Questions that can be studied by simulation

1. Comparison of breeding efficiencies from different selection strategies and their modifications. Which breeding method should be adopted?
2. Balance between the number of crosses and population size of segregating generations. What shall the breeder do if the available resources increase or reduce?
3. Evaluation of marker-assisted selection (MAS). When and how should MAS be used?
4. Comparative value of single, top, back, and double crosses in breeding?

Questions that can be studied by simulation

5. The correlation between parents and their offspring. Can F1 or F2 hybrids predict the performance of their advanced lines? For early generation selection, how early is too early?
6. When to use DH (doubled haploid)?
   - Early generation: many individuals need to be tested
   - Advanced line stage: lines will be good for other traits, but the desired genotype may be lost during selection due to population size, trait associations and genetic drift

Tools and principles of breeding simulation

Available breeding simulation tools

- **QuLine**, a computer software that simulates breeding programs for developing inbred lines
- **QuHybrid**, a computer software that simulates breeding programs for developing hybrids
- **QuMARS**, a computer software that simulates marker-assisted recurrent selection and genome-wide selection

QuLine: A simulation tool for genetics and breeding

- QU-GENE (QUantitative GENEtics)
  - A simulation platform for quantitative analysis of genetic models, developed by The University of Queensland, Australia
- QuCim (funded by GRDC 2000-2004)
  - A QU-GENE application breeding simulation module, specifically designed for CIMMYT's wheat breeding programs
  - Simulate most breeding programs for developing inbred lines
  - Version 1.1 released on July, 2003 (Workshop in Brisbane, Australia)
  - More than 100 global requests for QuCim 1.1
- Renamed as QuLine
- Version 2.0 available from http://www.uq.edu.au/lcafs/qugene/

What can QuLine do?

- Comparison of genetic gains from different selection methods
  - Change in population mean
  - Change in gene frequency
  - Change in Hamming distance (distance of a selected genotype to the target genotype)
- Comparison of cross performance
  - Selection history
  - Rogers' genetic distance
  - Number of lines retained from each cross
- Comparison of cost efficiency
  - Number of families
  - Individual plants per generation
- Validation of theories

In genetics
(implemented by the QU-GENE engine)

- Most genetic phenomena, if not all, can be defined in the QU-GENE engine input file (QUG).
- Among them are:
  - Multiple alleles (e.g. Glutenin genes in wheat)
  - Linkage (between gene and marker, between genes, between markers)
  - Additive, dominance and epistasis
  - Pleiotropy (one gene effects multiple traits)
  - Genotype by environment interaction
  - Molecular markers (dominant, or co-dominant)

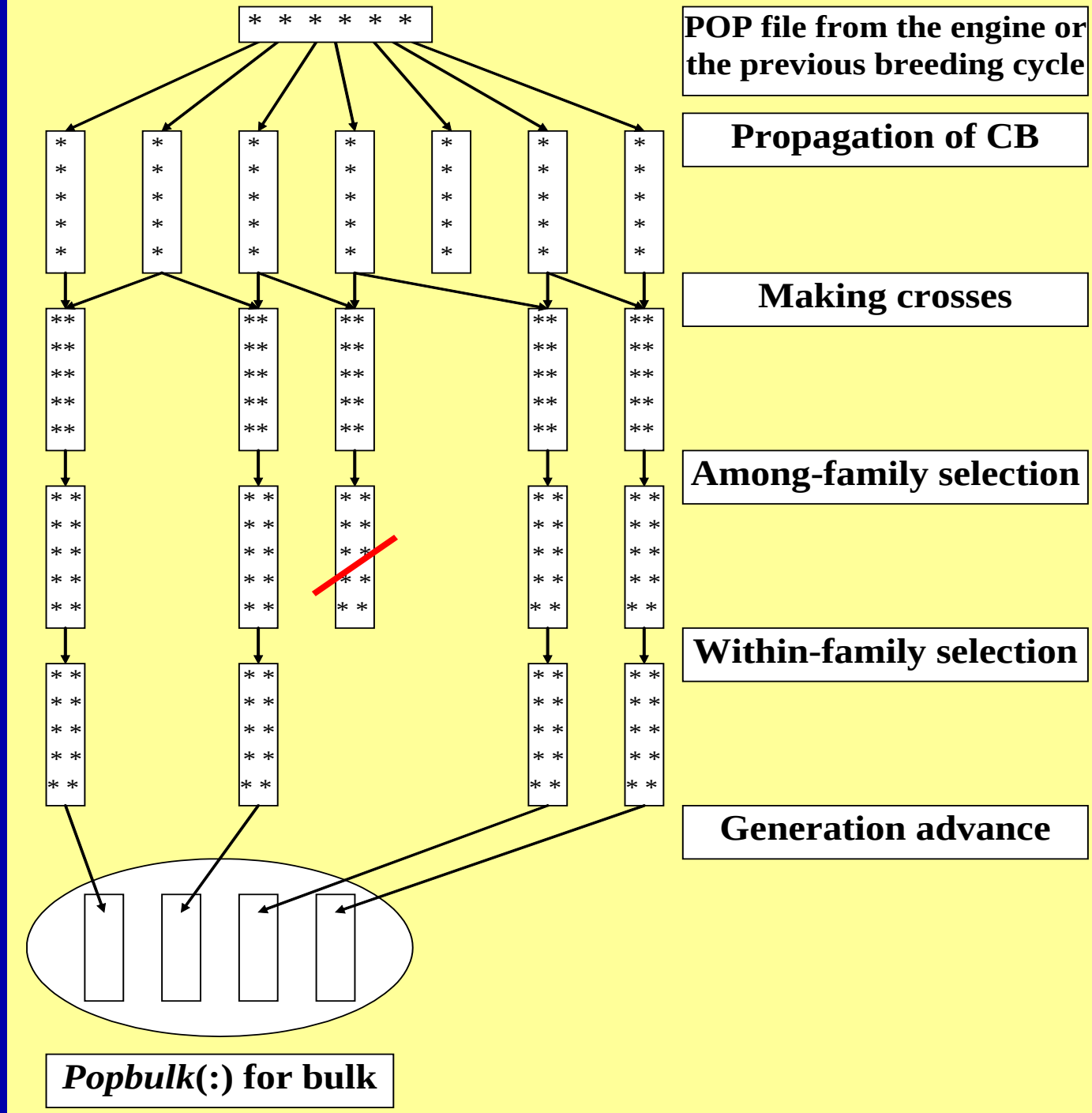
In breeding
(implemented by the QuLine module)

- Most, if not all, breeding methods for self-pollinated crops, can be defined and then simulated in QuLine.
- Among them are:
  - Pedigree system (including SSD)
  - Bulk-population
  - Doubled haploid
  - Marker-assisted selection (include marker-based selection)
  - Recurrent selection within one population
  - Many modifications and combinations

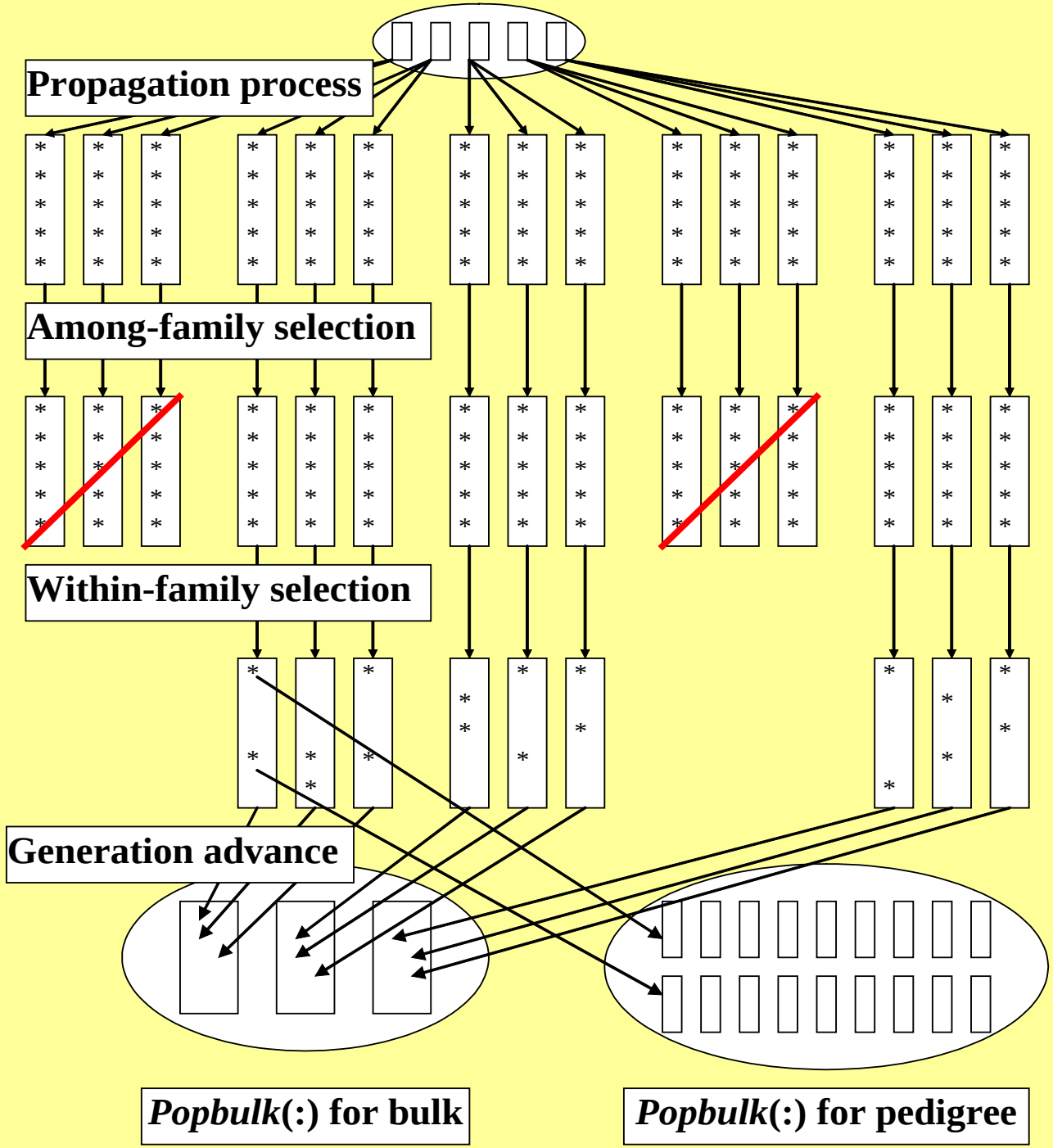
How does QuLine work?

- Two input files are needed
  - ***QUG file*** containing the necessary information for a genotype and environment (GE) system and initial population(s) of genotypes. It is the input for the QU-GENE engine. Two kinds of output files will be generated from the engine.
    - GES file for defining a GE system (= input for QuLine)
    - POP file for defining the initial population (= input for QuLine)
  - ***QMP file*** containing the necessary information for the breeding strategies to be simulated (e.g. pedigree, bulk, SSD, DH, etc.) (= input for QuLine)

General procedure for crossing and selection in F1



General procedure for propagation and selection in F2 and onwards



Parameters to describe a set of breeding strategies

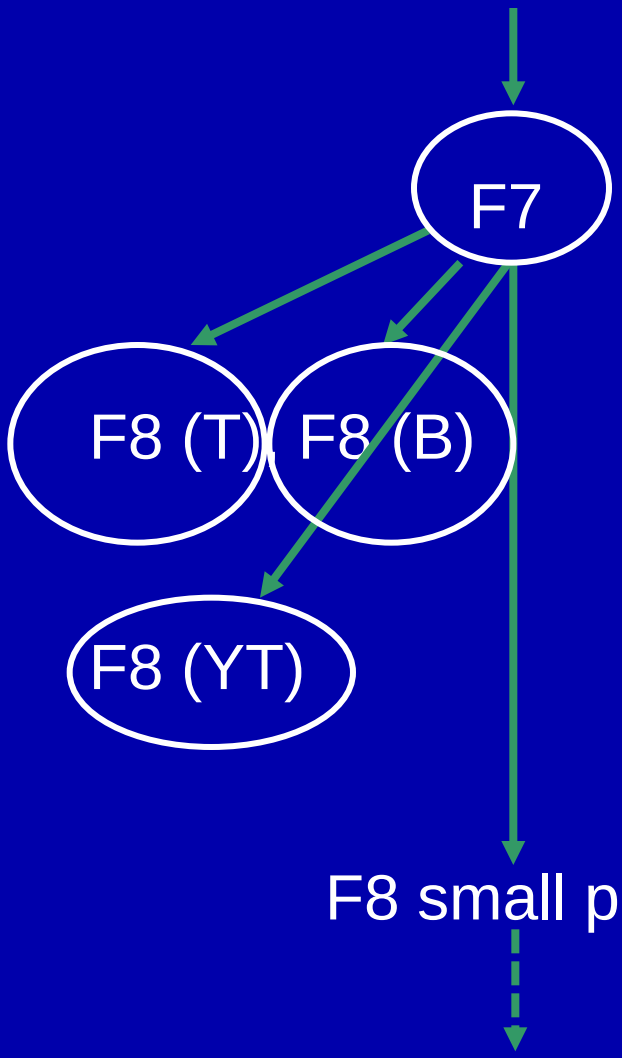
- Number of strategies
- For each strategy
  - Strategy name: any character
  - Number of generations in the strategy: any integer more than 0
  - Definition for each generation

Definition of a generation

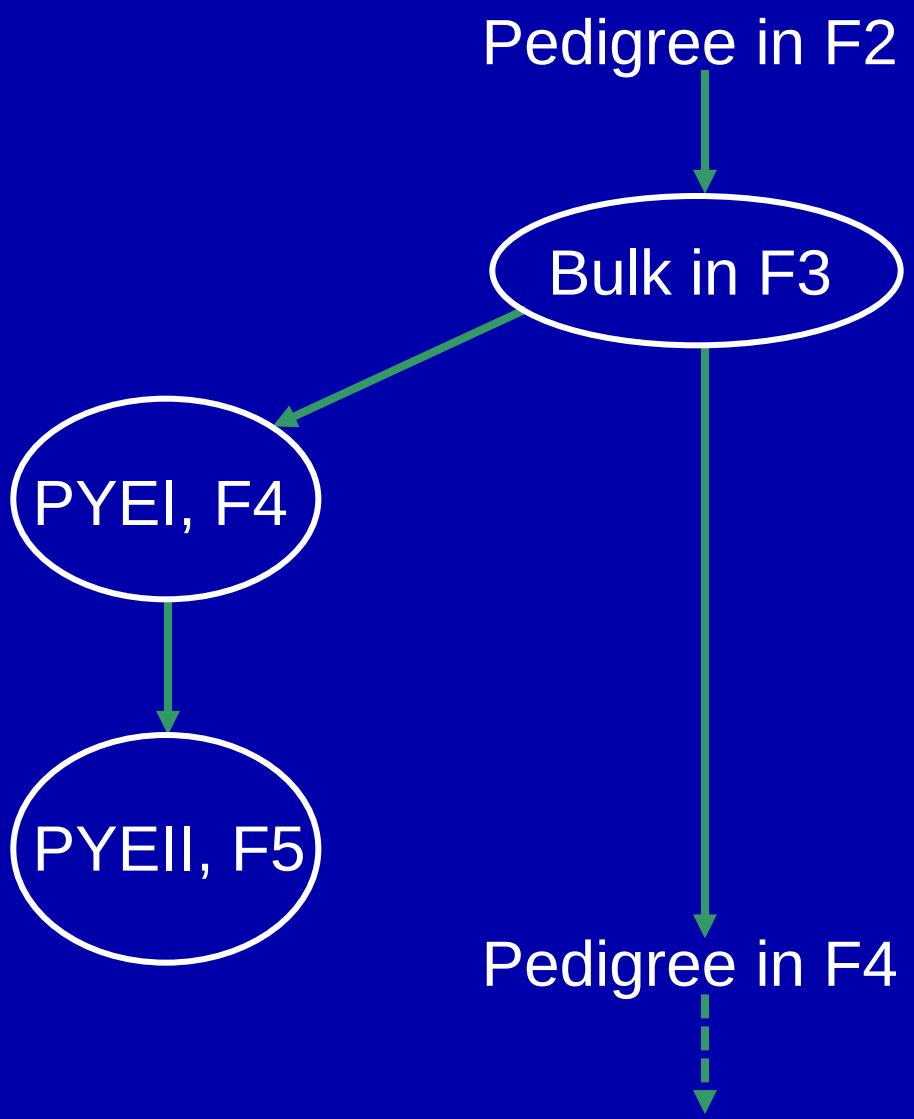
- Number of selection rounds in the generation: any integer more than 1
- Seed source indicator
  - 0: Seed for selection round *i* (*i* > 1) come from round 1
  - 1: Seed for selection round *i* (*i* > 1) come from round *i*-1
- Definition of each selection round

An example for seed source indicator 0

Practical breeding	Virtual breeding	Families selected
F6 field test at Toluca	F6	452 → 408 → 14,760
F7 field test at Cd. Obregon	F7	*Round 1 for F7* 14,760 → 3,868
F8 field test at Toluca and El Batan	F8 (T), F8 (B)	*Rounds 2 and 3* 3,868 → 2,163 → 1,974
F8 yield trial and F8 small plot evaluation at Cd. Obregon	F8 (YT) F8 small plot	*Round 4* 1,974 → 779 **779**



An example (LRC, Toowoomba, Australia) for seed source indicator 1



Families selected

Round 1 for F3

4000 → 1200

Round 2

1200 → 600

Round 3

600 → 100

100 → 1000

Definition of each selection round

- Title for the generation
- Seed propagation type (within a family)
  - *clone*, asexual reproduction
  - *DH*, doubled haploid
  - *self*, self-pollination
  - *backcross*, backcrossed to one parent
  - *topcross*, crossed with a third parent (three-way cross)
  - *doublecross*, crossed with another F1
  - *random*, random mating
  - *noself*, random mating but self-pollination is eliminated

Definition of each selection round

- Generation advance method (or harvest method): management of the selected individual plants in a family
  - *pedigree*: the selected plants in each family are harvested individually, resulting a few families in the next generation
  - *bulk*: the selected plants in each family are harvested in bulk, resulting one family in the next generation
  - *superbulk*: Family structure will be broken down. All selected plants form one family in the next generation

Definition of each selection round

- Field experiment design
  - Number of replications for each family
  - Number of plants in each replication
  - Number of test locations
  - Environment type for each test location
    - defined in the GE system
    - if 0, randomly determined based on environment frequency

Definition of each selection round

- Among family and within family selection
  - Number of traits used for selection
  - Definition of each trait

Definition of each trait used in selection

- Trait number, for the trait in selection (0 when marker score is used in selection)
- Selection mode
  - **T for top, e.g. yield, tillering, grains per spike and 1000-kernel weight**
  - **B for bottom, e.g. lodging and rusts**
  - **M for middle, e.g. height and heading**
  - **R for random, for some special studies**
  - **TV for top value**
  - **BV for bottom value**
  - **TN for a number of individuals/families with top phenotypic values**
  - **BN for a number of individuals/families with bottom phenotypic values**
  - **RN for a number of individuals/families to be selected randomly**
- Selected proportion or value: the proportion or value of individual plants in a family (for within family selection) or of families (for among family selection) to be selected

Proportion selection

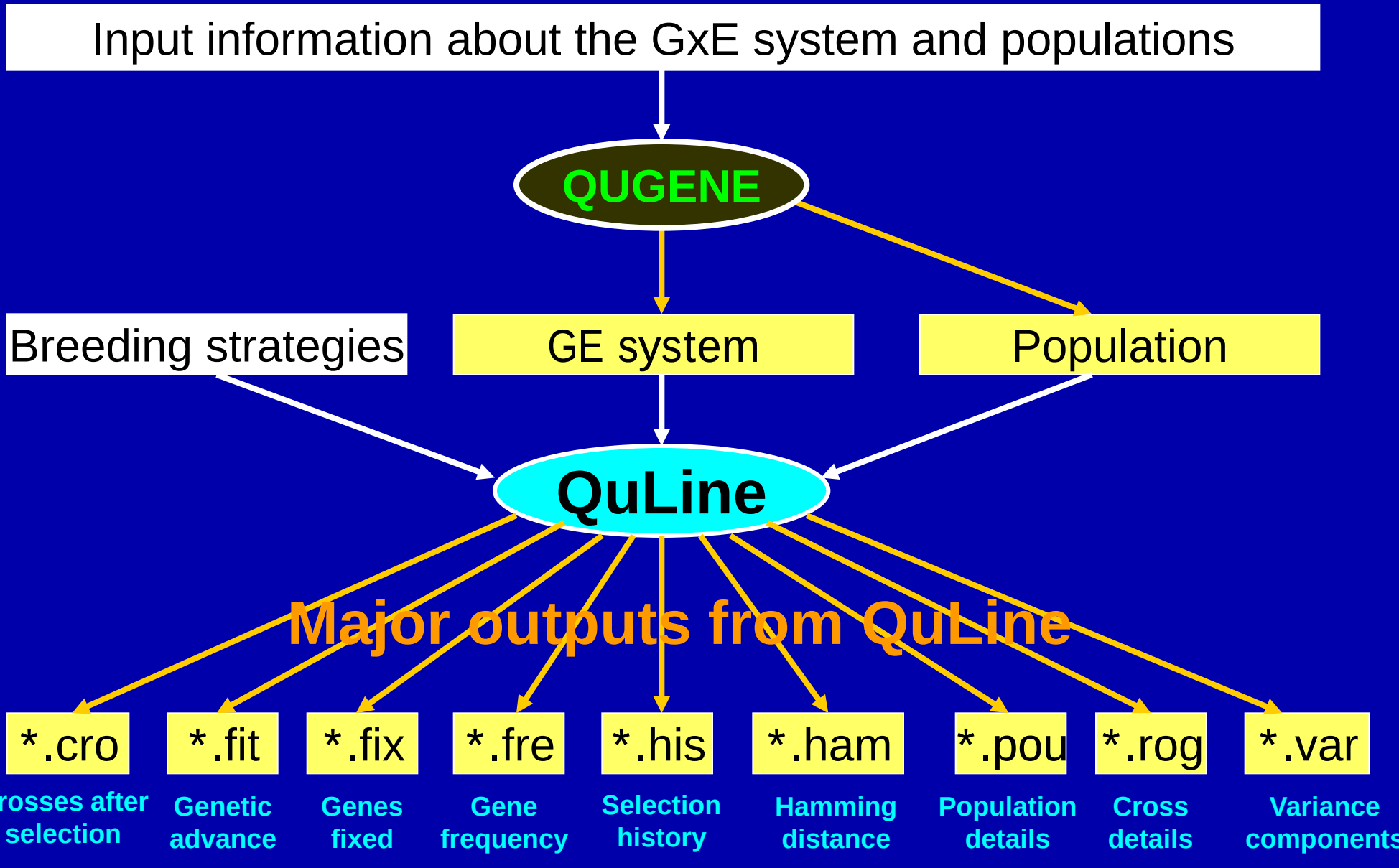
Threshold selection

Number selection

In QuLine, a breeding program looks like ...

```
!NR SS GT        PT         GA         RP PS    NL  ET...            Row 1
!                             AT (ID SP  SM)...                     Row 2
!                             WT (ID SP  SM)...                     Row 3
1   0  CB        self       bulk       1  10    1   2
                              0
                              0
1   0  F1     singlecross bulk        1  20    1   1
                              7   2 B 0.98  3 B 0.99  4 B 0.85  6 M 0.99  7 T 0.90  8 B 0.98  9 T 0.97
                              0
1   0  F2        self       pedigree   1  1000  1   2
                              7   2 B 0.99  3 B 0.99  5 B 0.90  6 M 0.99  7 T 0.99  8 B 0.99  9 T 0.99
                              8   2 B 0.95  4 B 0.99  5 B 0.40  6 M 0.85  7 T 0.60  8 B 0.90  9 T 0.50  10 T 0.60
1   0  F3        self       bulk       1  70    1   1
                              7   2 B 0.90  3 B 0.99  4 B 0.70  6 M 0.97  7 T 0.75  8 B 0.95  9 T 0.80
                              5   4 B 0.90  6 M 0.95  8 B 0.95  9 T 0.30 10 T 0.60
1   0  F4        self       bulk       1  70    1   2
                              6   2 B 0.90  5 B 0.65  6 M 0.97  7 T 0.85  8 B 0.97  9 T 0.85
                              5   5 B 0.90  6 M 0.95  8 B 0.95  9 T 0.30 10 T 0.60
1   0  F5        self       bulk       1  70    1   1
                              6   2 B 0.90  4 B 0.75  6 M 0.97  7 T 0.85  8 B 0.95  9 T 0.85
                              5   4 B 0.90  6 M 0.95  8 B 0.95  9 T 0.30 10 T 0.60
1   0  F6        self       pedigree   1  140   1   2
                              6   2 B 0.90  5 B 0.75  6 M 0.97  7 T 0.85  8 B 0.97  9 T 0.85
                              5   5 B 0.90  6 B 0.90  7 T 0.95  8 B 0.95  9 T 0.10
4   0  F7        self       bulk       1  70    1   1
                              7   2 B 0.90  4 B 0.75  6 M 0.97  7 T 0.90  8 B 0.95  9 T 0.85 10 T 0.75
                              0
       AL(T)     self       bulk       1  70    1   2
                              6   2 B 0.95  5 B 0.90  6 M 0.99  7 T 0.98  8 B 0.99  9 T 0.85
                              0
       AL(B)     self       bulk       1  70    1   3
                              1   4 B 0.90
                              0
       PYT       self       bulk       1  100   1   1
                              1   1 T 0.40
                              0
```

Steps to run QuLine

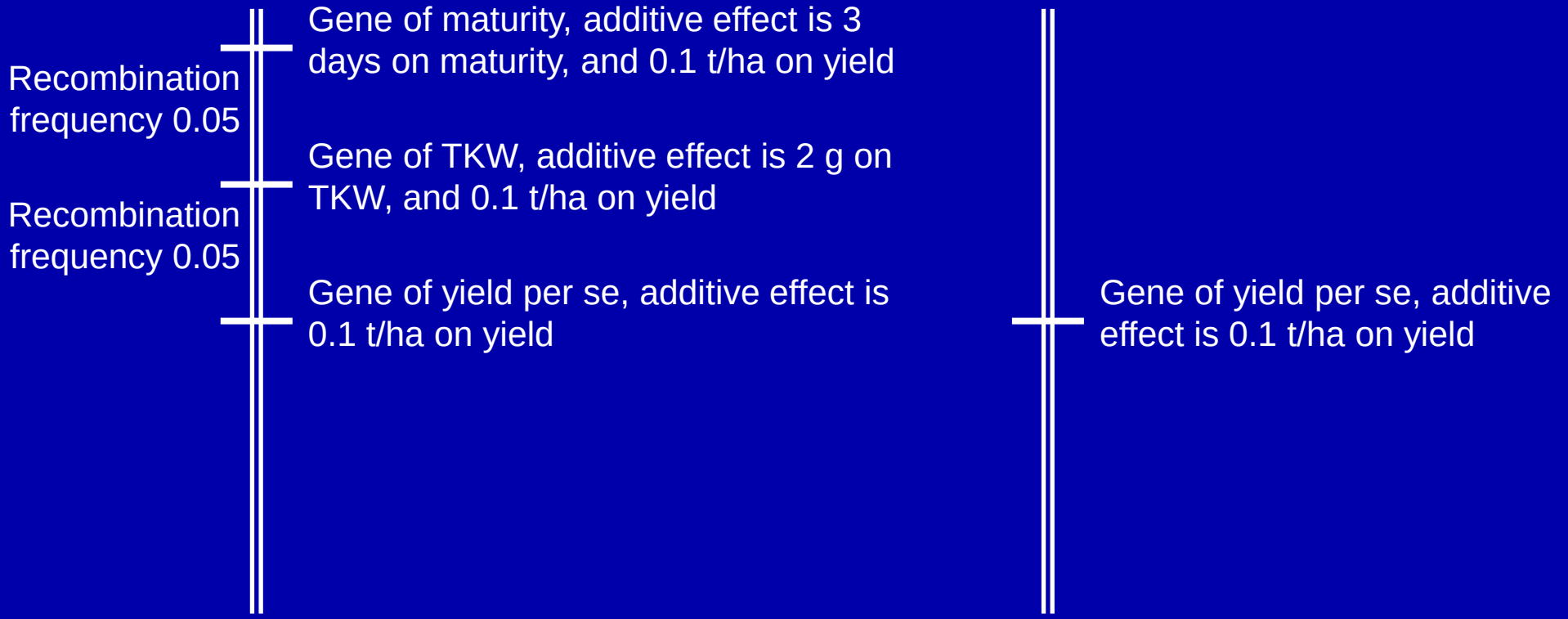


An illustrated example

A putative genetic model on maturity, TKW (thousand kernel weight) and yield

Genes distributed on chromosomes 1 to 5

Genes distributed on chromosomes 6 to 20



General information about a gene and environment system in QU-GENE

```
! ************************************************************************
! *  QUGENE engine input file
! *
! ************************************************************************
! *** General information on the G-E system ***

! Engine G-E output filename prefix (*.ges)
WheatModel

1                ! Number of models
0                ! Random seed of random gene effects
30               ! Number of genes (includes markers and qtls)
1                ! Number of environment types
3                ! Number of traits (not including markers)
1 1 1 0 0 0 0    ! Specify names (ETs, Trts, Genes, Alls, EPN, GPM, pop)
```

Environment and trait definitions in QU-GENE

```
! ************************************************************************
! *** Environment Type Information ***
!     Row 1: Number
!     Row 2: Name (if defined)
!     Row 3: Frequency of occurrence in TPE
! ************************************************************************

1
Obregon
1.000

! ************************************************************************
! *** Trait Information ***
!     Row 1: Number
!     Row 2: Name (if defined)
!     Row 3: Error Specification Type (for within,among,mixture)
!               1=heritability (spb); 2=error
!     Row 4+: Within, Among, Mixture error [each ET]
! ************************************************************************

1
Maturity
1      1      2
0.400  1.000  0.000

2
TKW
1      1      2
0.300  1.000  0.000

3
Yield
1      1      2
0.200  1.000  0.000
```

Gene definition in QU-GENE

```
!
*************************************************************************
!                    Columns
! CH     RF    NA NT      WT   ET   GP   EF   Gene effects
! 1      2     3  4       5    6    7    8    9+
!
*************************************************************************
   13
   Mat5
   5  0.5000   2  2
                        1    1    1   -1    0.0 3.000 0.000
                        3    1    1   -1    0.0 0.100 0.000
   14
   TKW5
   5  0.0500   2  2
                        2    1    1   -1    0.0 2.000 0.000
                        3    1    1   -1    0.0 0.100 0.000
   15
   Yld5
   5  0.2300   2  1
                        3    1    1   -1    0.0 0.100 0.000
   16
   Yld6
   6  0.5000   2  1
                        3    1    1   -1    0.0 0.100 0.000
```

Locus ID number

Locus name

Chromosome ID number, recombination with previous locus, number of alleles at the locus, and number of traits affecting

Genetic effects for all affected traits in all defined environments

Four populations defined in QU-GENE

```
4    ! Number of populations to create
1    ! Which population to use for error
estimates

1
Poperror
100
1
   0   1   1   2   1   0    0.5000

2
Pop02
20
1
   0   1   1   2   1   0    0.2000

3
Pop05
20
1
   0   1   1   2   1   0    0.5000

4
Pop08
20
1
   0   1   1   2   1   0    0.8000
```

Flowchart of breeding strategies I-M and II-M

Breeding strategy I-M	Generation flow A x B	Breeding strategy II-M
100 single crosses made from 50 parental lines		100 single crosses made from 50 parental lines
10 plants for each F_1; no selection; each F_1 population is harvested in bulk	F_1	10 plants for each F_1; no selection; each F_1 population is harvested in bulk
500 plants for each F_2 population; select for 20% with medium maturity, and 10% for TKW; **selected F2 plants are harvested individually**	F_2	500 plants for each F_2 population; select for 20% with medium maturity, and 10% for TKW; **selected F_2 plants are harvested in bulk**
30 plants in each $F_{2:3}$ family; select for 50% families with medium maturity; each selected family is harvestled in bulk	F_3	**100 plants in each F_3 family; select for 50% individuals with medium maturity in each family; each family is harvestled in bulk**
40 plants in each $F_{2:4}$ family; select for 50% families with high TKW; each family is harvestled in bulk	F_4	**150 plants in each F_4 family; select for 50% individuals with high TKW in each family; each family is harvestled in bulk**
50 plants in each $F_{2:5}$ family; in each family, select for 20% individuals with medium maturity, and 20% with high TKW; selected plants are harvestled individually	F_5	**200 plants in each F_5 family; in each family, select for 5 individuals with medium maturity and high TKW;** selected plants are harvestled individually
500 F_6 families are grown in one location, each having 50 plants with 2 replications; select for 20% families with high yield; each selected family is harvested in bulk	F_6	500 F_6 families are grown in one location, each having 50 plants with 2 replications; select for 20% families with high yield; each selected family is harvested in bulk
100 F_7 families are grown in three locations, each having 50 plants with 2 replications; select for 20% families with high yield; each selected family is harvested in bulk	F_7	100 F_7 families are grown in three locations, each having 50 plants with 2 replications; select for 20% families with high yield; each selected family is harvested in bulk

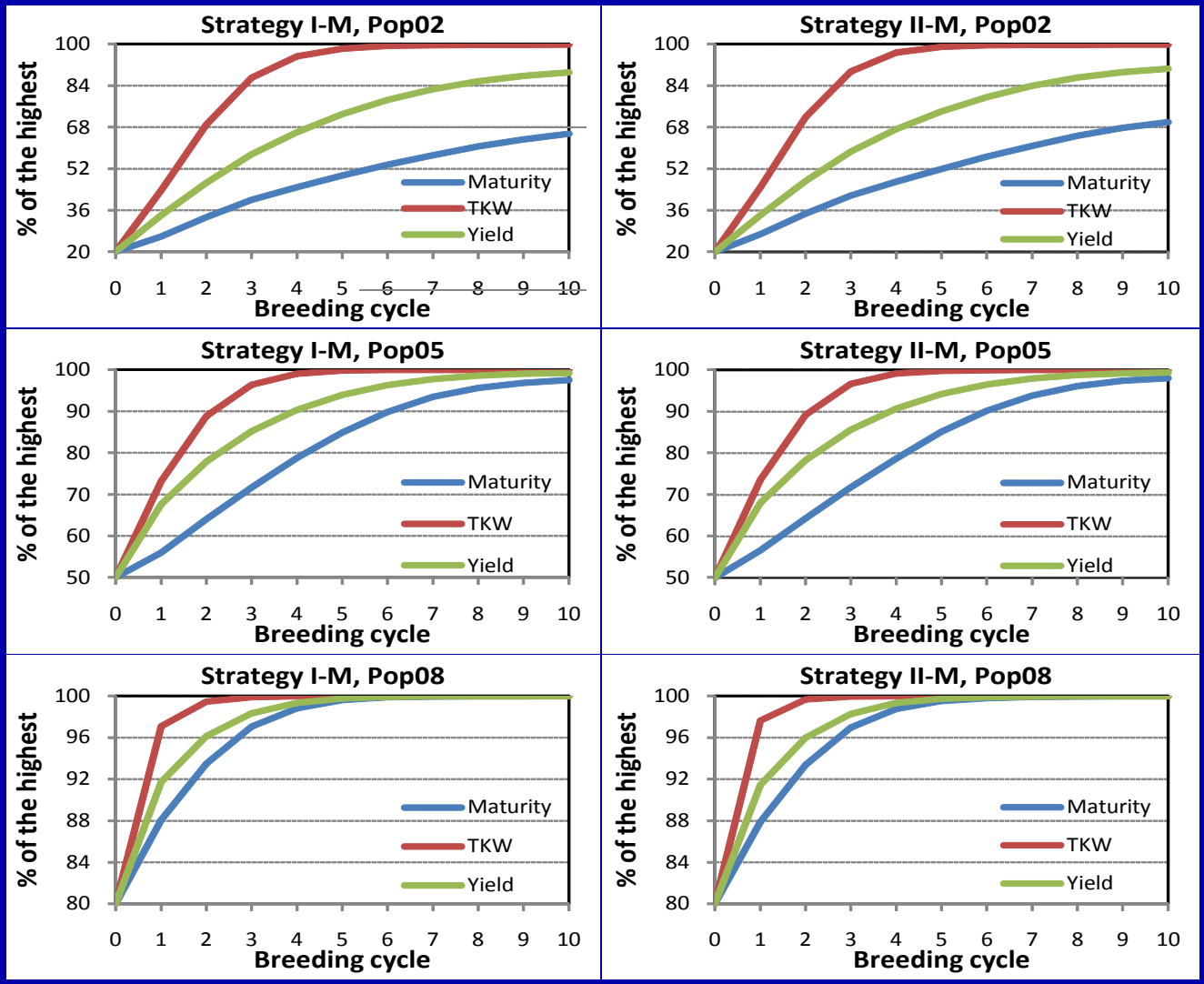
Definition of general simulation information and strategy I-M in QuLine

```
!********************General information for the simulation experiment*******************************
!NumStr NumRun NumCyc NumCro CBUpdate OutGES OutPOP OutHIS OutROG OutCOE OutVar Cross  RMtimes PopSize
 2      5      10     100    0        0      0      0      0      0      0      random 0       0


!********************Information for selection strategies to be simulated*****************************
!StrategyNumber StrategyName  NumGenerations
 1              StrategyI-M   7

!NR SS  GT      PT           GA        RP  PS  NL  ET...                                  Row 1
!                                          AT (ID  SP  SM)...                            Row 2
!                                          WT (ID  SP  SM)...                            Row 3
 1  0   CB      clone        bulk      1   1   1   1
                                              0
                                              0
 1  0   F1      singlecross  bulk      1  10   1   1
                                              0
                                              0
 1  0   F2      self         pedigree  1  500  1   1
                                              0
                                              2   1  M  0.20    2  T  0.10
 1  0   F3      self         bulk      1  30   1   1
                                              1   1  M  0.50
                                              0
 1  0   F4      self         bulk      1  30   1   1
                                              1   2  T  0.50
                                              0
 1  0   F5      self         pedigree  1  50   1   1
                                              0
                                              2   1  M  0.20    2  T 0.20
 1  0   F6      self         bulk      2  50   1   1
                                              1   3  T  0.20
                                              0
 1  0   F7      self         bulk      2  50   3   1  1  1
                                              1   3  T  0.20
                                              0
```

Definition strategy II-M in QuLine

```
!******************Information for selection strategies to be
simulated*****************************
!StrategyNumber StrategyName  NumGenerations
 2              StrategyII-M     7

!NR SS  GT      PT           GA        RP  PS  NL  ET...                          Row 1
!                                          AT (ID  SP  SM)...                    Row 2
!                                          WT (ID  SP  SM)...                    Row 3
 1  0   CB      clone        bulk      1   1   1   1
                                              0
                                              0
 1  0   F1      singlecross  bulk      1   10  1   1
                                              0
                                              0
 1  0   F2      self         bulk      1   500 1   1
                                              0
                                              2   1  M  0.20    2  T  0.10
 1  0   F3      self         bulk      1   50  1   1
                                              0
                                              1   1  M  0.50
 1  0   F4      self         bulk      1   50  1   1
                                              0
                                              1   2  T  0.50
 1  0   F5      self         pedigree  1   200 1   1
                                              0
                                              2   1  M  0.20    2  T 0.125
 1  0   F6      self         bulk      2   50  1   1
                                              1   3  T  0.20
                                              0
 1  0   F7      self         bulk      2   50  3   1  1  1
                                              1   3  T  0.20
                                              0
```

Adjusted ggenetic gains from breeding strategies I-M and II-M



Adjusted ggenetic gains from breeding strategies I-B and II-B