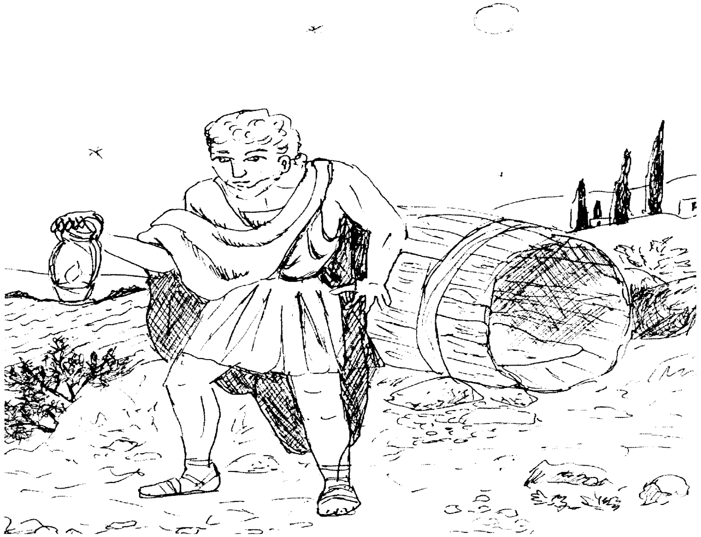
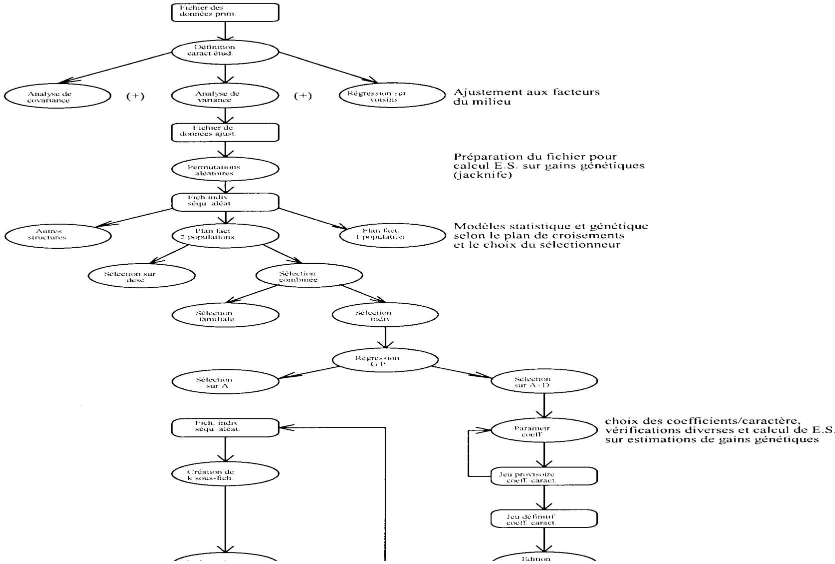



Fichier des données prim.
Définition caract.étud.
Analyse de covariance
(+)
Analyse de variance
(+)
Régression sur voisins
Ajustement aux facteurs du milieu
Fichier de données ajust.
Permutations aléatoires
Préparation du fichier pour calcul E.S. sur gains génétiques (jacknife)
Fich.indiv séqu. aléat.
Autres structures
Plan fact 2 populations
Plan fact. 1 population
Modèles statistique et génétique selon le plan de croisements et le choix du sélectionneur
Sélection sur desc.
Sélection combinée
Sélection familiale
Sélection indiv
Régression G P
Sélection sur A
Sélection sur A · D
Fich. indiv séqu. aléat.
Création de k sous-fich.
Parametr coeff
choix des coefficients/caractère, vérifications diverses et calcul de E.S. sur estimations de gains génétiques
Jeu provisoire coeff. caract.
Jeu définitif coeff. caract.
Edition

DIOGENE

A Plant Breeding Software

- Users
- **Students** (Master, Thesis)
- **Confirmed researchers** (INRA, CIRAD, Laval University)
- **Tree Breeding managers,** technicians and & engineers (INRA, CIRAD, CEMAGREF…)

- Present state
- Integration of **General Biometry**, **Quantitative** & **Population Genetics**
- **Modular Structure**
- **Original models** (Genotype x Environment interaction, Selection indices, Spatial statistics: Papadakis++…)
- Usable both in in **interactive** mode and by building complex '**processing sequences**' (automatic generation of **scripts**)
- **Multivariable** and **non-orthogonal** (MANOVA, Selection indices, Data Analysis…)
- Simultaneous processing of **quantitative** and **qualitative (0-1)** traits
- **Resampling** (Jackknife and Bootstrap) very fast and standardized

- Recent improvements (Ph. Baradat and Th. Perrier 2003-2009)
- **Porting in Fortran 95 and Linux**
- **Contextual input of parameters**

Specifications

- Integrated software (several programs chained)
- Great number of parameters, but most of them are ‘guessed’ (from context)
- Ability to process experiments even with strong non-orthogonality
- High speed (mandatory for resampling)

The original data file system is adapted to resampling. It is binary, with each data (identifier or observation) coded in single precision (4 bytes). A parameter file suffixed by ‘.p’ is associated. It gives all informations useful for data processing.

X vector

Identifier 1 …	**Identifier k**	X_{11} …	X_{1q} …	X_{zq}

Y vector

Identifier 1 …	**Identifier k**	Y_{11} …	$Y_{1q'}$ …	$Y_{zq'}$

A record (X vector), stored into memory at the processing time, is defined by three parameters:

- Number of identifiers (k)
- Maximum number of individuals (z)
- Number of traits observed per individual (q)

The traits are referenced by their relative rank within an individual. The parser (see next slide) generates a virtual record (Y vector) with the same structure where the q observed traits are replaced by q’ functions of these traits and/or already defined functions (recursivity).

Structure and use of data file record (1)

Schematic conception of the parser

Tetrad 1				Tetrad 2			
operator	address of result	operand 1	operand 2 or ‘0’	operator or ‘end stack’ code	address of result	operand 1	operand 2 or ‘0’

Generation of binary data (presence/absence) from the ‘y’ studied traits

The incidence matrix (0-1) of the binary data is managed by a specialised language or by internal routines (e.g. for molecular markers involving thousands of traits).

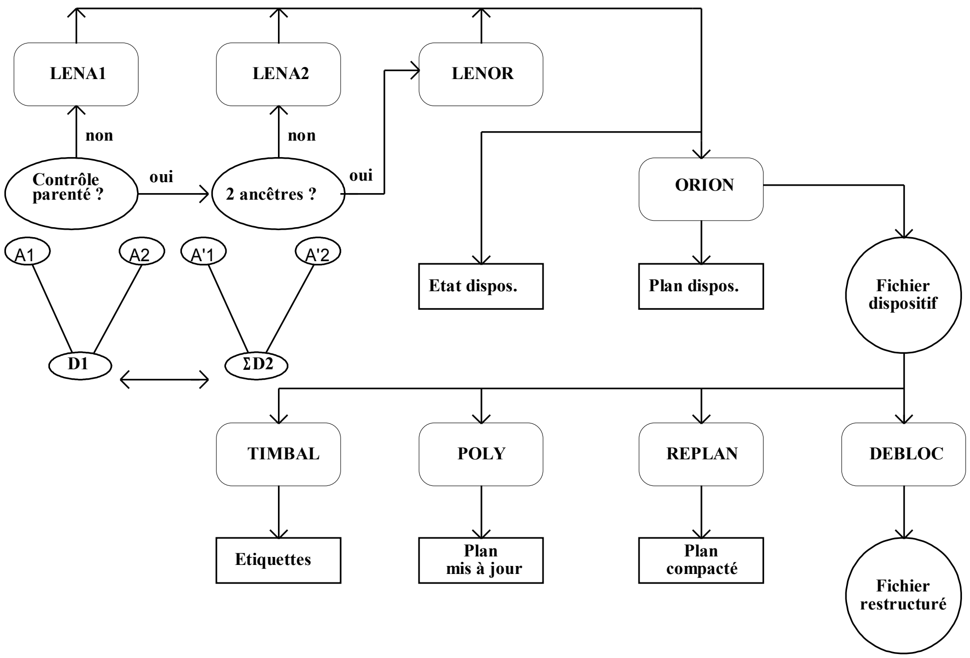
		Number of addressed column = y value						
	‘studied trait’ = line number	1	2	3	4	5	6	7
Number of the addressed line = rank of the trait	y1	1	0	1	0	0	1	0
	y2	0	0	1	1	0	0	1
	y3	1	0	0	0	0	0	0
	y4	0	0	0	1	1	0	0
	y5	1	1	1	0	0	0	0
	y6	0	0	0	0	0	1	1

Structure and use of data file records (2)

The ‘*y’* variables are defined in the form:
y(j)= F[x(1), x(2)...y(i), ctes]. According to this principle, the logarithm of the volume increment of a cone may be written:
log((x3**2*x4-x1**2*x2)*pi/3).
if (initial radius & height) and (final radius & height) are, in that order, the four ‘x’ variables.

Missing data are coded by ‘-9’ ou ‘-5’ according to the individual is dead or that the trait cannot be observed for another reason. Every individual whom at least one of the ‘*x’* variables which are required to define a ‘*y’* variable has one of these two values is excluded from the processing. Lastly, if n is the number of individuals per record, and $n < z$, a ‘logical end of record’ signal is coded by ‘9999’.

Structure and use of data file records (3)

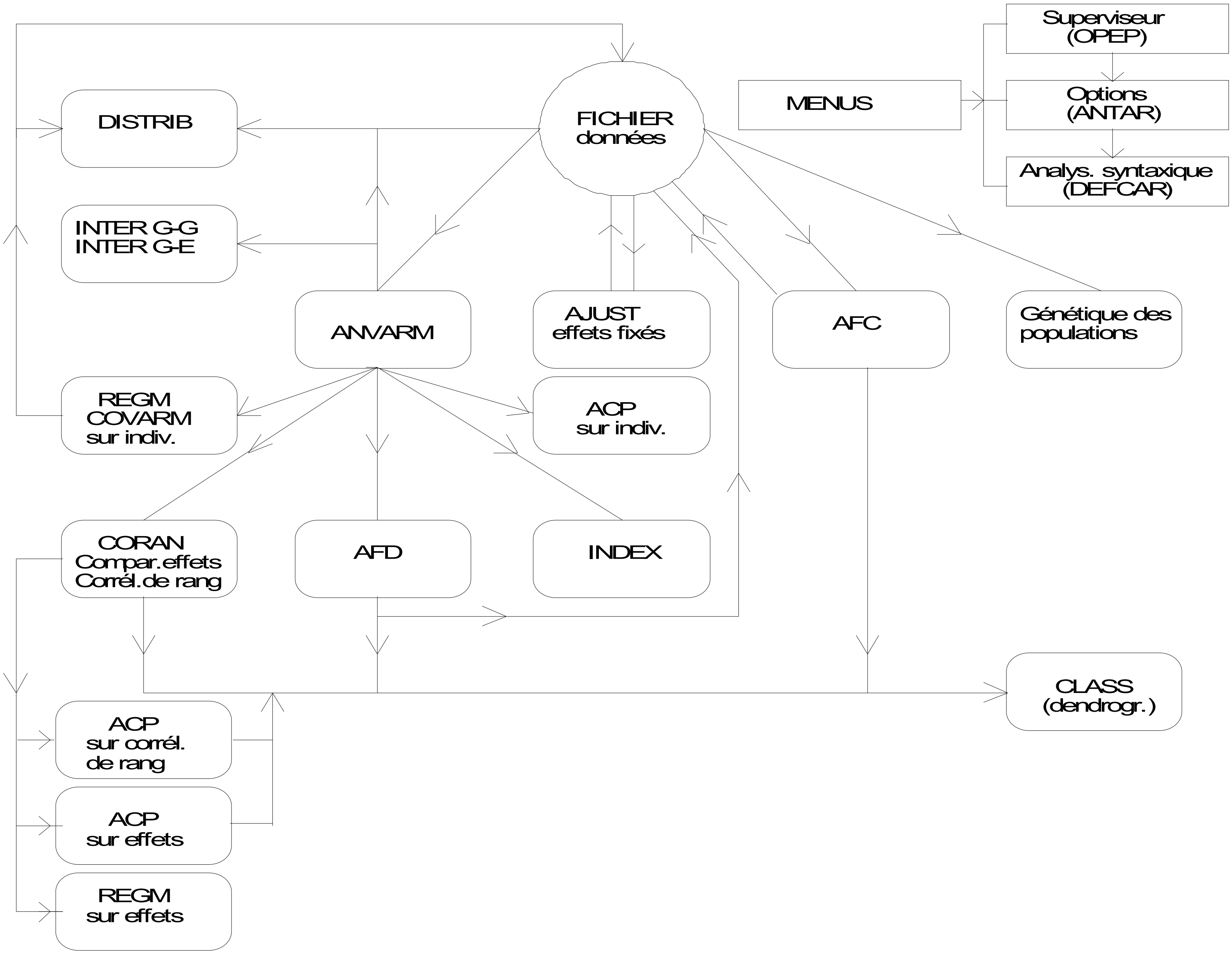


General flowchart of programs for creation/management of field trials (1)

The programs create random incomplete block trials which take into account environmental constraints met in the field, with a coordinate localization of individuals. Geometry of blocks and plots can be parametrized. Relativness between individuals of the same block may be controlled in the case of seedling seed orchards. Tn this case, the program checks for every new individual (D1) randomly drawn, that none individual among those already drawn in the block (D2) have in common one or two common ancestors using the constraint: $(A1'A'1)Ç(A1'A'2)Ç(A2'A'1)Ç(A2'A'2)$. The algorithm of random drawing of individuals from each genetic unit for allocation to blocs is deviced so that: $\Pr(D_{ij})=n_i/N$ where D_{ij} is an individual or a plot of the D_i genetic unit of size $n_{i.}$ during random drawing, if N individuals or plots are involved. This principle allows generation of trials optimized even with genetic units having very different sizes.

General flowchart of programs for creation/management of field trials (2)

General flowchart of programs for Biometry and Genetics



Some characteritics which make DIOGENE original and useful (1)

❑ Modular Structure (‘à la carte’ models)

- ***Complex adjustment to environment including multisite trials (Papadakis++)***
- ***MANOVA models including individual contribution to G x E Interaction***
- ***MANOVA + Discriminant Analyses corresponding to models model (eg. Diallel)***
- ***selection Indices including choice of predictors and target traits with easy weighting***

etc…

❑ Choice of standardized data file allowing:

- ***A selective processing of selected lines (records)***
- ***Great processing quickness (important for resampling)***

= ‘**ANTAR**’ which integrates:

- Data on a binary direct access file

- All informations on the data (associated parameter file)

Some characteristics… (2)

❑ A management of data processing by ‘scripts’

- ➢Easy to create to correct & to modify (usable in different context)
- ➢Allowing creation of scripts for complex computations

❑ Generalized resampling concerning chains of programs

- ➢Jackknife
- ➢Bootstrap
- ➢Each method may be used at individual or genetic entry levels

By choice of:

- ➢ The first and the last programs of the sequence
- ➢ Where is done the resampling (‘Upstream’ parameter)
- ➢ The level: individul or genetic entries (family, provenance…)

❑ Other kinds of reiterated computations (Papadakis++…)

Resampling (1)

- *The Jackknife method (1)*

One discards successively individuals of ranks 1 to u, u+1 to 2u,…(k-1)u+1 to ku.

It is possible to discard only one individual by subsample: k=N, u=1.

If u>1, the sub-sample must be représentative of the total population (all levels of factors). This may be realized by random permutation of the initial ranks of individuals. Each individual is associated to n variables : y_1, y_2 …y_n and one computes on the population a general function of these variables, $F(y_1,y_2,\ldots y_n)$.

Resampling (2)

- The Jackknife method

This function of observations is re-computed from each sub-sample. The positive autocorrelation between the sub-samples, with (k-2)u individuals in common, would lead to underestimate the error variance of the parameter estimate. An unbiased estimate of this error variance (Quenouille-Tukey's estimate) is given by:

$$\hat{S}^2 = \frac{1}{k(k-1)}\left(\sum_{i=1}^{k} F_i^2 - \frac{\sum_{i=1}^{k} F_i^2}{k}\right)$$

where:

$F_i = k\hat{F} - (k-1)F_i^*$ (Tukey's pseudo-value);

F_i^* is the value of the parameter computed on the subsample of rank i where individuals of ranks $u(i$-1)+1 to ui are removed;

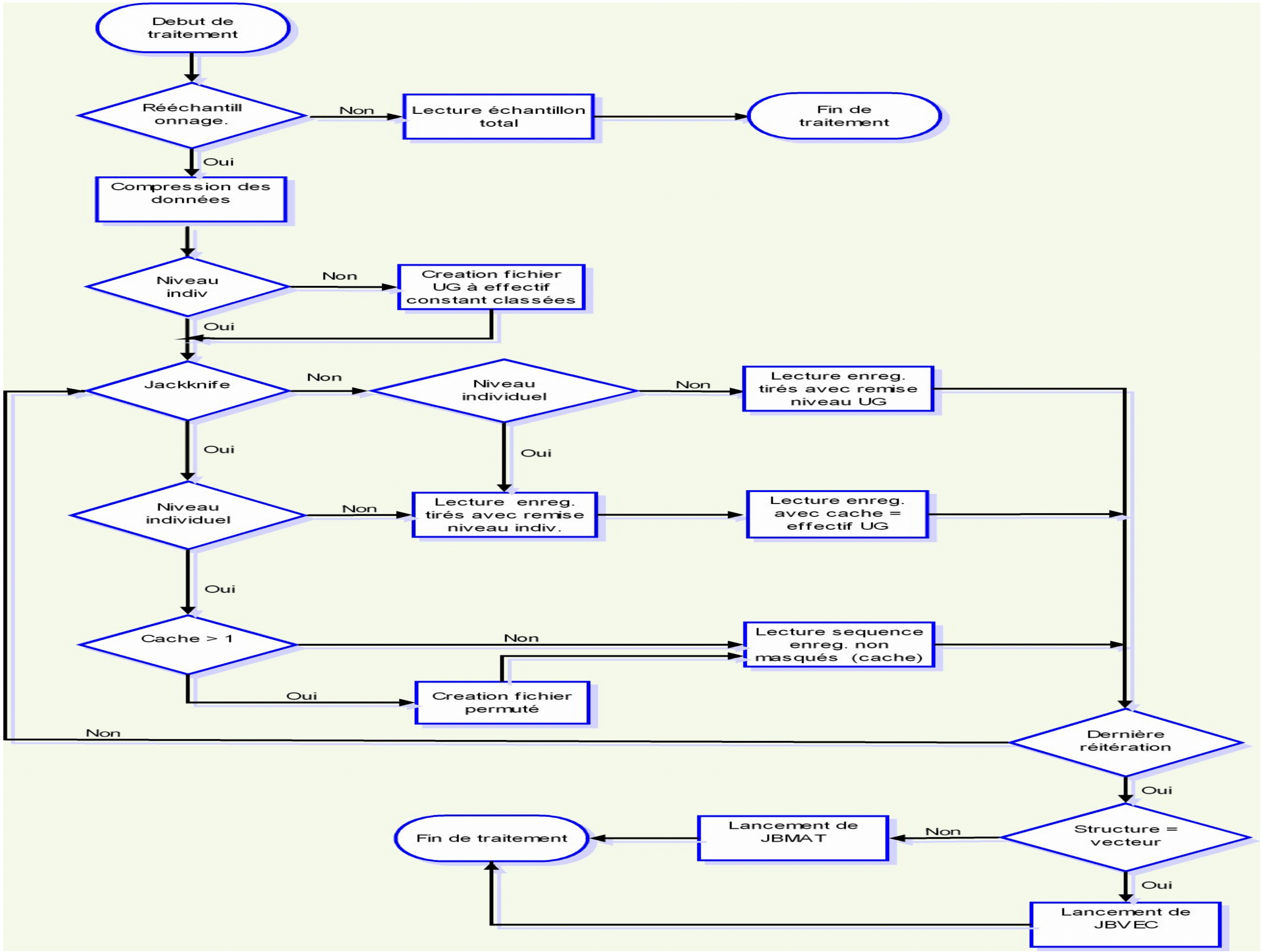
$\hat{F}$ is the parameter's value computed on the total sample (ku individuals). These pseudo-values are independent variables and the statistic: $\frac{\hat{F} - E(F)}{\hat{S}}$ follows the Student's ***t*** distribution with k-1 degrees of freedom.

Resampling (3)

- **The Bootstrap method**

It is a resampling with remise, which generates samples of size *N* and then leads to the possibility to include several times the same data in different samples or within the same sample. This method can be applied when the autocorrelation between generated random samples is reduced and therefore the proportion of common data is low. These samples may be considered as independent. The variance between the estimates of the parameter is an estimate of its sampling variance. This method is much used population genetics because it uses simple et robust structure (generally, a single population or nested classifications with one or two levels). It is more difficult to use in the case of experimental designs with crossed or mixed classifications where some random sequences drawn with can generate disconnected levels of factors. Nevertheless, this method has an important interest: The number *E* of random samples that can be obtained from *N* individuals is practically infinite even is *N* represent some thenths: $E = N^N$.

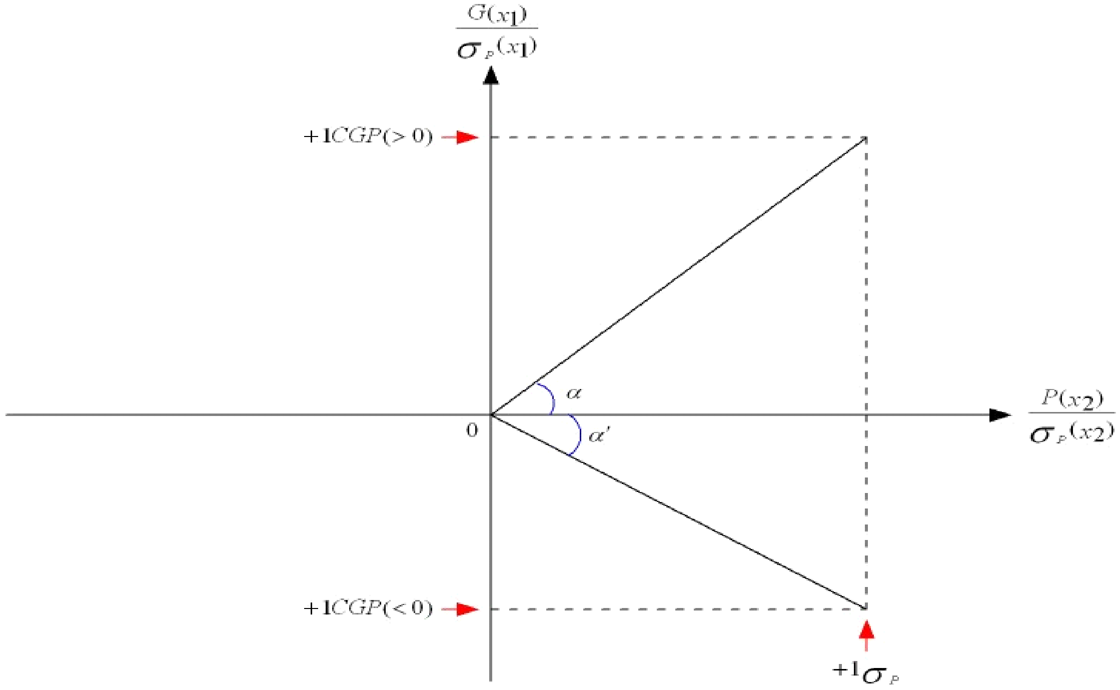
As the estimates of parameters are independent, the study of their distribution on several thousands of sequences allows to determine their confidence intervals without the hypothesis of a normal distribution.



Simplified organigram showing the realization of resampling in DIOGENE software.

DIOGENE gives of course the significance levels associated to statisticals tests

```
Mean squares & F tests assuming fixed effects

GCA mean square of genotype  Parent (11 d.f.)
       y  1        y  2        y  3        y  4        y  5
       ht84        pp85        ht85        pp86        ht86
   5.6699E+03  7.6234E+03  9.2083E+03  1.7853E+04  2.1153E+04

F tests (11 and 2551  degrees of freedom)

       y  1        y  2        y  3        y  4        y  5
       ht84        pp85        ht85        pp86        ht86
     13.164      13.431      12.893      15.791      13.941
      0.000%      0.000%      0.000%      0.000%      0.000%

Mean square of Specific Combining Ability, SCA (51 degrees of freedom)

       y  1        y  2        y  3        y  4        y  5
       ht84        pp85        ht85        pp86        ht86
   9.3257E+02  1.3669E+03  1.5766E+03  2.4063E+03  3.5983E+03

F tests (51 and  2551 d.f.)

       y  1        y  2        y  3        y  4        y  5
       ht84        pp85        ht85        pp86        ht86
      2.165       2.408       2.207       2.128       2.371
      0.000%      0.000%      0.000%      0.001%      0.000%
```



Coefficient of genetic prediction between traits 1 and 2: $CGP(x_1 \cdot x_2) = \frac{\mathrm{cov}_G(x_1 \cdot x_2)}{\sigma_P(x_1)\sigma_P(x_2)} = tg(\alpha)$ or $tg(\alpha')$

$G(x_1)$: genetic value of trait 1 $P(x_2)$: phenotypic value of trait 2

$\mathrm{cov}_G(x_1 \cdot x_2)$: genetic covariance between trait 1 and trait 2

$\sigma_P(x_1)$: phenotypic standard deviation of trait 1 $\sigma_P(x_2)$: phenotypic standard deviation of trait 2

Fig. 4. Graphical representation of the coefficient of genetic prediction (*CGP*)

On this figure is represented the correlated response of trait 1 (y-axis) selected via the trait 2 (x-axis). If one moves the phenotypic mean of the population of +1, for trait 2, in term of phenotypic standard deviation, the result is a correlated response (indirect selection) of 1 CGP for trait 1. This response may be positive or negative according to the sign of the genetic coefficient of prediction. The heritability of a trait is not other than the genetic coefficient of prediction of this trait by itself. In this case, the response is, by definition, positive or null. If one permutes traits 1 and trait 2 (by plotting them on x-axis and y-axis, respectively), the figure will be the same as the units are the identical on these two axes.

On this figure is represented the correlated response of trait 1 (y axis) selected using trait 2 as predictor (x axis). If one shifts the average phenotypic value of the population by +1 for the trait 2 in term of phenotypic standard deviation, the result is a positive correlated response (indirect selection) of 1 CGP for trait 1. This response may be positive or negative according to the sign of the CGP. The heritability of a trait is nothing else than the CGP of this trait by itself. In this case, the response is, by definition, positive or null. If one permutes the les traits 1 and 2 (by plotting them on x and y axes, respectively), the figure will be the same, as the units are identical on the axes (phenotypic standard deviations).

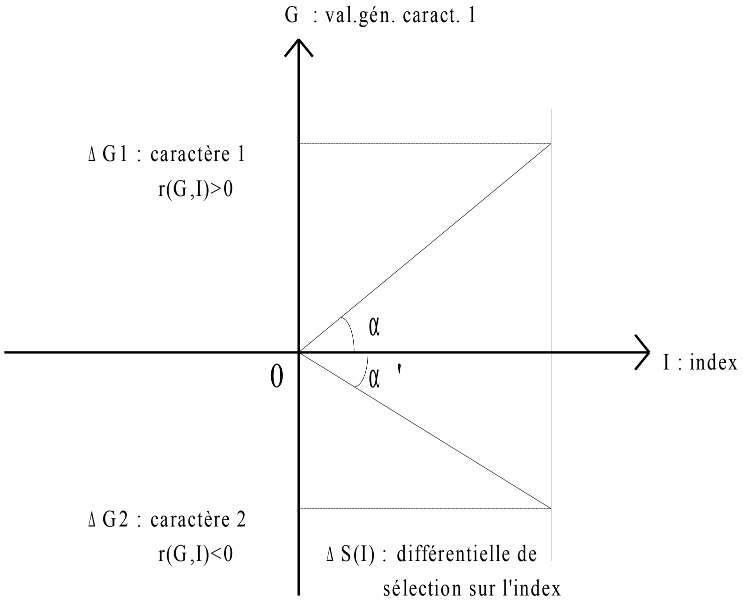
Graphical representation of the Coefficient of Genetic Prediction (2)

Genetic values predicted by par regression of genotype on phenotype

$$\left[\hat{G}\right] = \left[\Sigma\ GP\right]\left[\Sigma\ PP\right]^{-1}\left[\hat{p}\right]$$

Linear combination of predicted genotypic values using set of weights

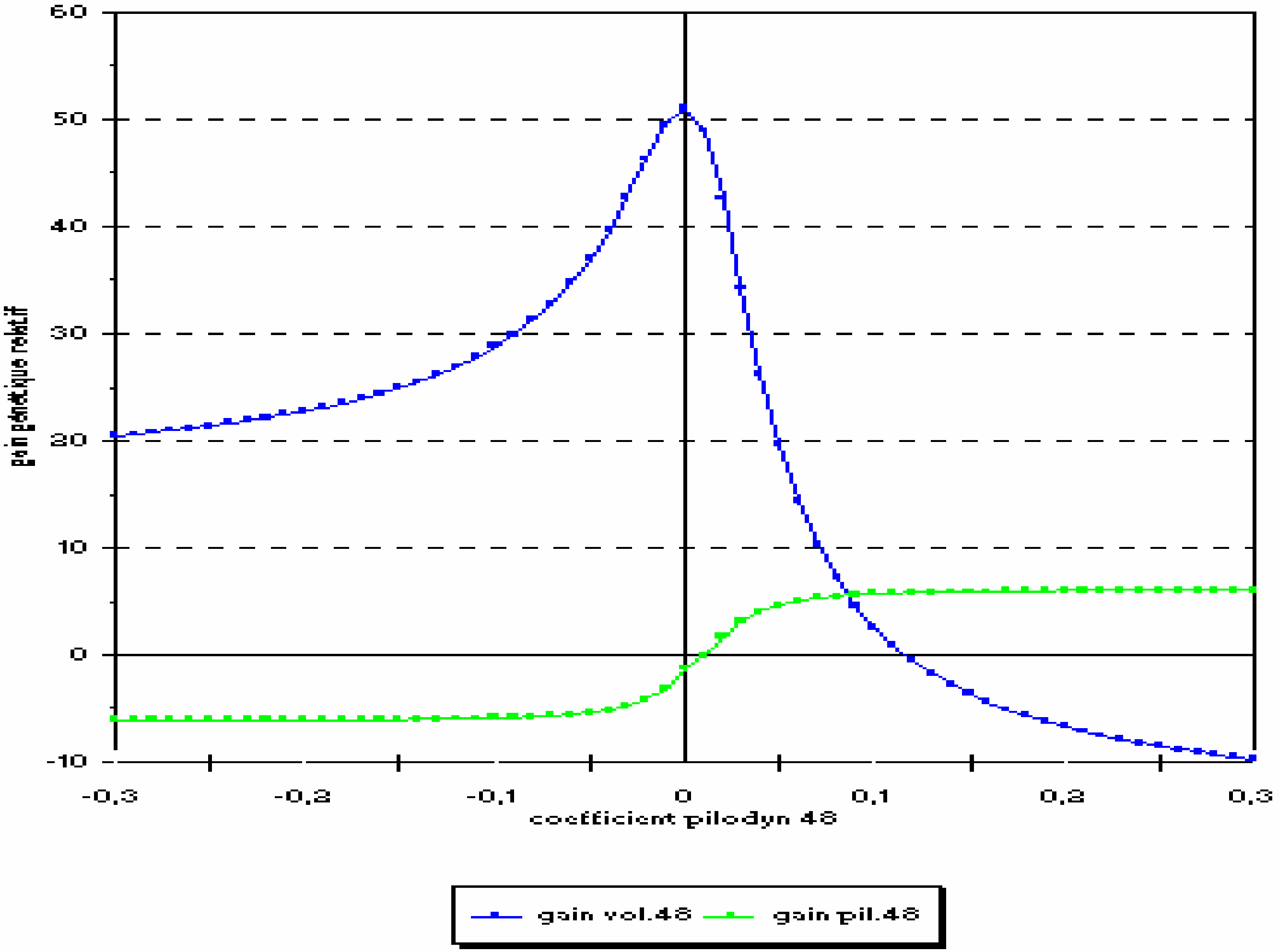
$$I = \left[b\right]'\left[\hat{G}\right]$$



The genetic value of trait 1 (expected genetic gain = Δ G1) is positively correlated with the index ; the genetic value of trait 2 (expected genetic gain = Δ G2) is negatively correlated. The two expected genetic gains, Δ G1 et Δ G2, are determined by the selection differential on the index : $\Delta S(I) = i\sigma_I$, where **i** is the selection intensity, and by the coefficient of regression of each genetic value on the index: $b = \mathrm{cov}(G, I) / \sigma_I^2$, with: b1= $\mathrm{tg}(\alpha)$ et b2= $\mathrm{tg}(\alpha')$.

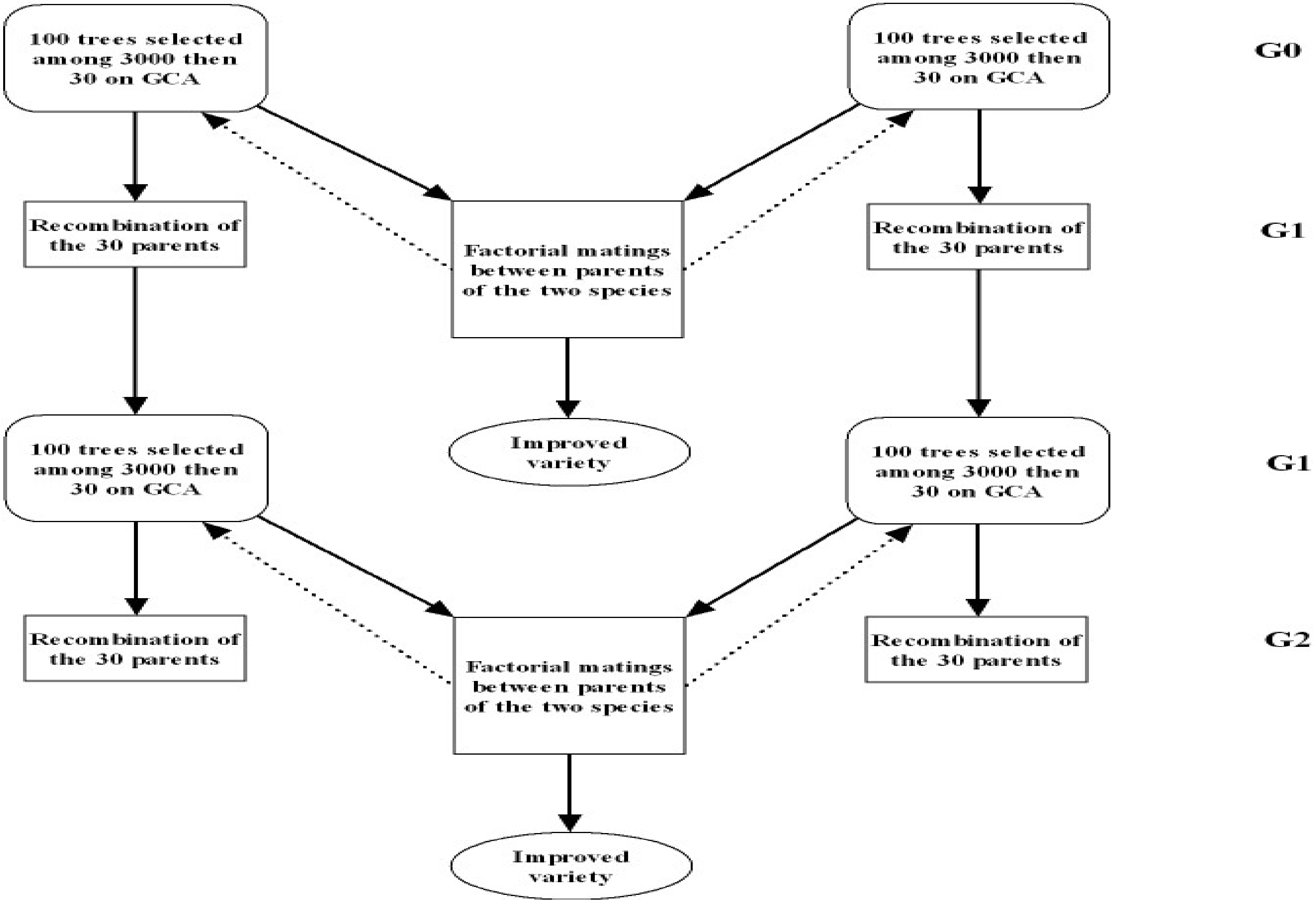
Realization of partial genetic gains on two traits by population truncation for an index correlated with their genetic values

DIOGENE computes Selection Indices for all the f practical situations met in Tree Breeding..



Curves for choice of coefficients of predicted traits in Selection Indices

The coefficient of volume, b1, is constant (b1=1) and the coefficient of pilodyn, b2, varies from -0.3 to + 0.3. Note the very large variation induced on the relative expected gain for the volume by a small variation of the pilodyn's coefficient around the value b2 = 0. On the other hand, the curve of expected genetic gains for pilodyn present a slightly negative value for b2 = 0. This is the effect of the (small) genetic negative correlation between volume and pilodyn (-0.08).



100 trees selected among 3000 then 30 on GCA

100 trees selected among 3000 then 30 on GCA

G0

Recombination of the 30 parents

Factorial matings between parents of the two species

Recombination of the 30 parents

G1

Improved variety

100 trees selected among 3000 then 30 on GCA

100 trees selected among 3000 then 30 on GCA

G1

Recombination of the 30 parents

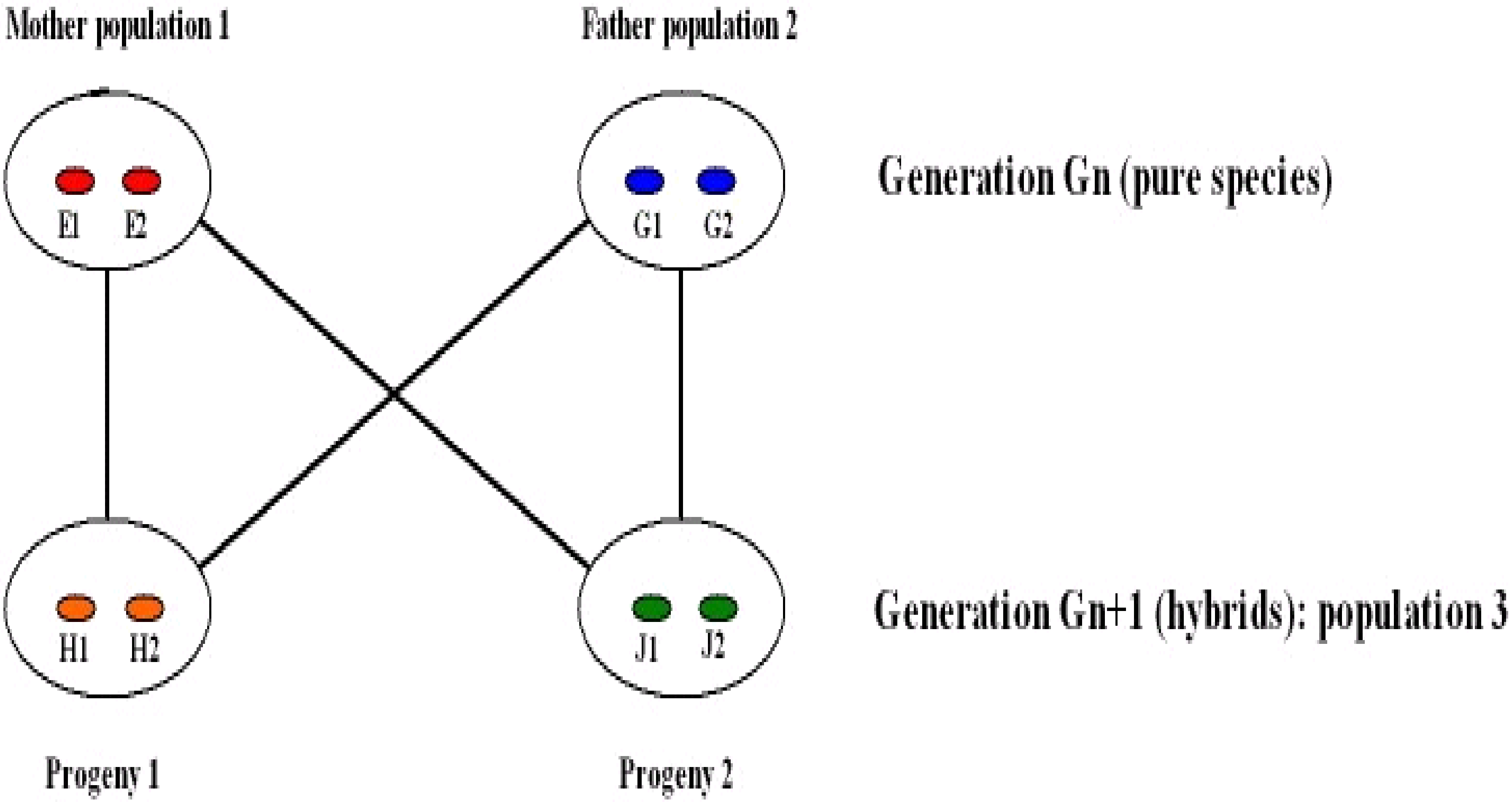
Factorial matings between parents of the two species

Recombination of the 30 parents

G2

Improved variety

Example of Selection indices for Reciprocal Recurrent Selection e.g. *Eucalypts grandis* and *E. urophylla in Congo*



SIm:single identity of maternal allele

SIp:single identity of paternal allele

DI: double identity of maternal and paternal alleles

Identity relationship		Probability	Variance components
Mother-progeny:	$(H1 \equiv E1) \cup (H1 \equiv E2)$	1	$\frac{1}{2}\text{cov}\,A_m = \text{cov}\,A_1$
	or $(J1 \equiv E1) \cup (J1 \equiv E2)$	--	
Father-progeny:	$(H2 \equiv G1) \cup (H2 \equiv G2)$	1	$\frac{1}{2}\text{cov}\,A_p = \text{cov}\,A_2$
	or $(J2 \equiv G1) \cup (J2 \equiv G2)$	--	
Between descendants:			
SIm	$(H1 \equiv J1) \mid (H2 \not\equiv J2)$	$\frac{1+F_1}{2}$	$\frac{1}{2}\text{cov}\,A_m = \text{cov}\,A_1$
SIp	$(H2 \equiv J2) \mid (H1 \not\equiv J1)$	$\frac{1+F_2}{2}$	$\frac{1}{2}\text{cov}\,A_p = \text{cov}\,A_2$
DI	$(H1 \equiv J1) \cap (H2 \equiv J2)$	$\frac{(1+F_1)(1+F_2)}{4}$	$\text{cov}\,A_1 + \text{cov}\,A_2 + \text{cov}\,D$

Genetic model for computation of covariances between relatives.

SIm = single identity from common mother, SIp = single identity from common father, DI = double identity.

- The values ‘on diagonal’ are the classical heritabilities
- *DIOGENE* computes and displays the CGP lower-triangular matrices
- The user thus obtains synthetic informations on the compared efficiencies of direct & indirect selection.

```
Matrices of the Coefficients of Genetic Prediction (heritabilities on the diagonal)


Narrow-sense Coefficients of Genetic Prediction

                    y  1        y  2        y  3        y  4        y  5
                    ht84        pp85        ht85        pp86        ht86
y  1:   ht84       0.102
y  2:   pp85       0.098       0.101
y  3:   ht85       0.097       0.100       0.099
y  4:   pp86       0.100       0.110       0.108       0.125
y  5:   ht86       0.096       0.102       0.100       0.112       0.106

Broad-sense Coefficients of Genetic Prediction

                    y  1        y  2        y  3        y  4        y  5
                    ht84        pp85        ht85        pp86        ht86
y  1:   ht84       0.208
y  2:   pp85       0.215       0.229
y  3:   ht85       0.205       0.218       0.209
y  4:   pp86       0.192       0.215       0.206       0.227
y  5:   ht86       0.192       0.211       0.203       0.223       0.231
```

DIOGENE **also computes and displays these estimates after re-sampling (significance tests for estimated parameters):**

```
Parameters and tests of the matrix number 9
Narrow-sense Coefficients of Genetic Prediction

                     y  1      y  2      y  3      y  4      y  5
                     ht84      pp85      ht85      pp86      ht86
y  1 :    ht84       0.102
  Standard E.:       0.021
      t test :       4.878
 Signif. (%) :       0.000
y  2 :    pp85       0.098     0.101
  Standard E.:       0.021     0.021
      Test t :       4.722     4.770
 Signif. (%) :       0.001     0.001
y  3 :    ht85       0.097     0.100     0.099
  Standard E.:       0.020     0.021     0.021
      Test t :       4.718     4.747     4.711
 Signif. (%) :       0.001     0.001     0.001
y  4 :    pp86       0.100     0.110     0.108     0.125
  Standard E.:       0.020     0.021     0.021     0.022
      Test t :       4.945     5.230     5.113     5.615
 Signif. (%) :       0.000     0.000     0.000     0.000
y  5 :    ht86       0.096     0.102     0.100     0.112     0.106
  Standard E.:       0.019     0.020     0.020     0.021     0.021
      Test t :       5.032     5.179     5.028     5.283     5.148
 Signif. (%) :       0.000     0.000     0.000     0.000     0.000
```

…and confidence intervals at the level chosen by the user

```
            Confidence intervals of the matrix    9
            Narrow sense of Coefficients of Genetic Prediction


                        y  1       y  2       y  3       y  4       y  5
                        ht84       pp85       ht85       pp86       ht86
y  1 :     ht84         0.143
                        0.061
y  2 :     pp85         0.139      0.143
                        0.057      0.060
y  3 :     ht85         0.137      0.141      0.140
                        0.056      0.058      0.058
y  4 :     pp86         0.140      0.151      0.149      0.169
                        0.060      0.069      0.066      0.082
y  5 :     ht86         0.133      0.141      0.138      0.153      0.146
                        0.059      0.064      0.061      0.070      0.066
```

Example of modular treatment (data processing sequence)

Sequence of programs: ENVIR - DIAL

Mixed model MANOVA of a half-diallel with random genetic effects and incomplete block design (fixed block effect).

Mean square and variance of effect	Deg.of Fr.	Expectation of mean square : $E(CM)$	F tests
bloc, CM_b	B-1	$\sigma_e^2+[1/(B-1)]\sum_{k=1}^{B} n_{..k.}\beta_k^2$	CM_b/CM_e unbiased
General Comb. Ability: *GCA* CM_a, s_a^2	P-1	$\sigma_e^2+k_1\sigma_a^2+k_2\sigma_s^2$	CMa/CM_s biased
Specific Comb. Ability: *ASC* CM_s, s_s^2	C-P	$\sigma_e^2+k_3\sigma_s^2$	CM_s/CM_e unbiased
Within family: CM_e, s_e^2	N-D-B+1	s_e^2	

B: number of blocks, *P*: number of parents, *C*: number of crosses, reciprocal confounded, *N*: number of individuals. F tests for significance of *AGC* variance is realized using *ASC* mean square. It is biased if the half-diallel is non-orthogonal and unbalanced. F test for *ASC* variance est uses the within-family mean square. It is unbiased in all cases.

To estimate the components of the variance, the system to be solved is:

$$\hat{\sigma}_e^2=CM_e \quad \text{et} \quad \begin{bmatrix}\hat{\sigma}_a^2\\ \hat{\sigma}_s^2\end{bmatrix}=\begin{bmatrix}k_1 & k_2\\ k_3 & 0\end{bmatrix}^{-1}\begin{bmatrix}CM_a-CM_e\\ CM_s-CM_e\end{bmatrix}$$

For the components of covariance, mean products replace mean squares by the mean for any couple of traits.

When tests for nullity of variance components are biased, **use of resampling is mandatory**. More generally, resampling allows check of significance for all parameters derived from variances-covariances (for instance heritabilities and genetic correlations).

Combination of genetic models and Spatial Statistics, allowed by *DIOGENE*, due to its modularity, leads to much more complex (and powerful) models.

Models of Population Genetics for help to Selection

Modelization of inbreeding of seed orchard’s progeny :

- Management of actual seed orchards (genetic thinning)
- Optimization of new seed orchards (compromise between expected genetic gain and consanguinity)

Measurement of selfing rate and contamination by wild pollen

DIOGENE places thus at the disposal of the user

❑ Powerful methods of trial generation/management and adjustment to Environment.

❑ Possibility to measure Genotype x Environment interaction of each genetic unit.

❑ Processing of all mating designs

❑ Aptitude to process highly non-orthogonal experiments

❑ Complete set of selection indices

❑ Very flexible and fast system to compute confidence intervals using re-sampling

CONCLUSION

DIOGENE = development platform

- ❑ *Unified Architecture*
  - ➢ Generic tools
  - ➢ Inter-compatibles modules
  - ➢ Normalised data file structure

- ❑ *Cell of development*
  - ➢ Maintenance of a permanent competence
  - ➢ College of users (preferably international: required internationalization of the software using modern methods)
  - ➢ Task sharing of design/development
  - ➢ Regular update of the notices

- ❑ ***DIOGENE*** may be downloaded from http://amap.cirad.fr