



HAL
open science

Genomic selection; general principles & some French applications

Hervé Chapuis

► **To cite this version:**

Hervé Chapuis. Genomic selection; general principles & some French applications. École thématique. Tainan, Taiwan. 2022. hal-04163630

HAL Id: hal-04163630

<https://hal.inrae.fr/hal-04163630>

Submitted on 17 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



➤ Genomic selection: General principles & some French applications



➤ Outlines

I. Classical (pre-genomic) selection

II. Principles of genomic selection

III. Consequences on the selection process



➤ Genetic selection:

Organization of a breeding plan

Definition of selection objectives

Pedigree and performances recording

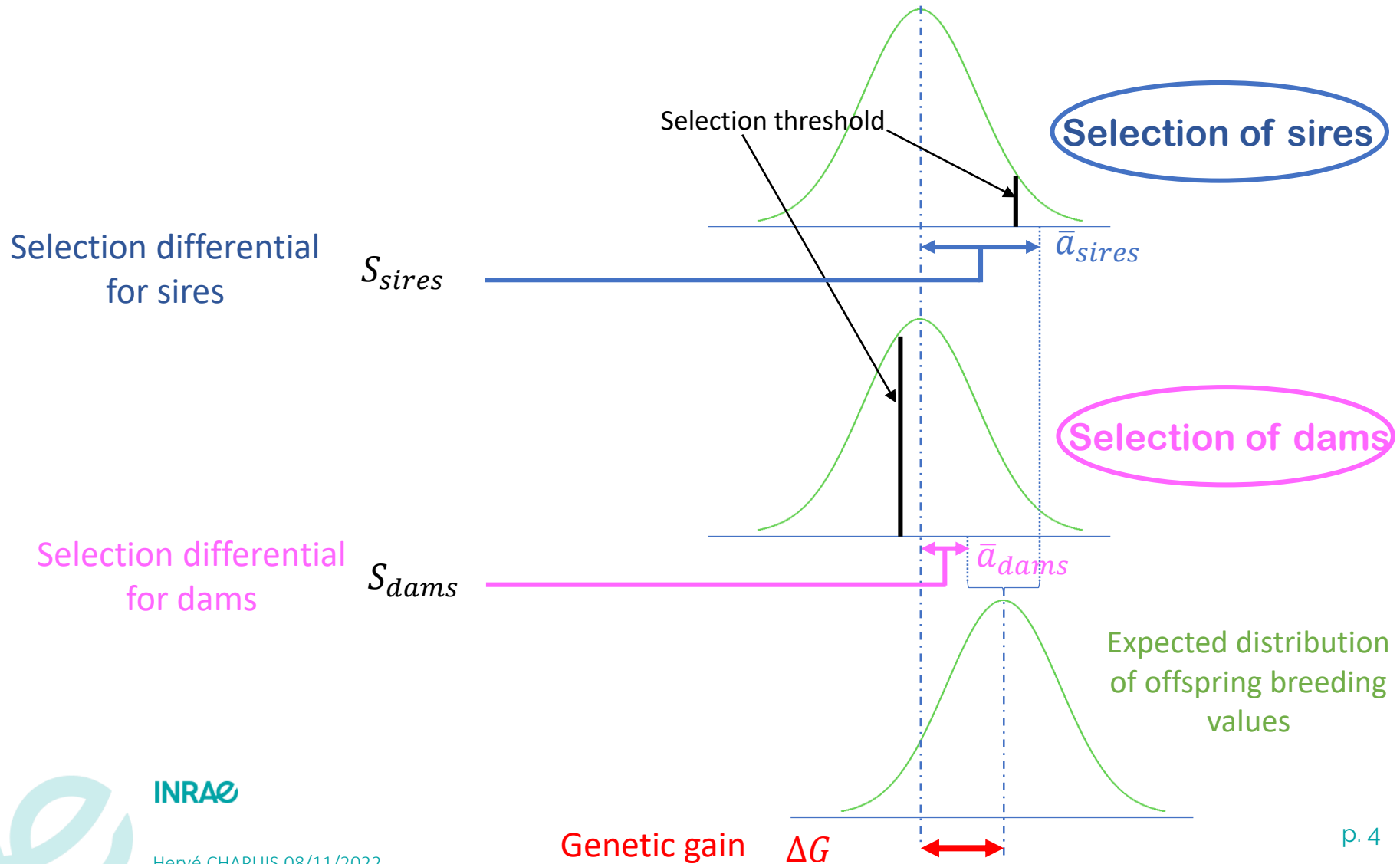
Genetic evaluation

Selection of sires & dams

Utilization of breeders



➤ Genetic selection: Creation of genetic gain



➤ Genetic selection:

Derivation of genetic gain

Annual genetic gain: $\Delta G = \frac{i \times \sqrt{CD} \times \sigma_g}{T}$

Selection intensity i (green circle) → Accuracy \sqrt{CD} (blue circle) → Genetic standard deviation (variability) σ_g (black arrow) → Generation interval T (orange circle)

Maximization of ΔG :

- ↗ i = ↘ % of selected animals
- ↗ \sqrt{CD} = ↗ quantity and quality of available information
- ↗ σ_g = ??
- ↘ T = culling of « old » breeders

Accounting for:

- *breeding goal, pop. size,*
- *heritability of economic traits,*
- *population structure,*
- *biology, available reproductive technologies...*



➤ Computations in genetic evaluation : requirements

Data Model Results

Performances **P**



Pedigree



Environment **M**
(hatch/herd/band, age,
season, sex...)

model to describe data:

$$\mathbf{P} = \mathbf{M} + \mathbf{a} + \mathbf{e}$$

model for transmission of genetic merit:

$$a_i = \frac{1}{2} a_s + \frac{1}{2} a_d + \phi_i$$

➔ Numerator relationship matrix **A**



{ Estimated breeding value $\hat{\mathbf{a}}$
Estimation of other effects $\hat{\mathbf{M}}$



C.R. Henderson

1973: "Sire evaluation and genetic trend"

1976: "A method to compute the inverse of the numerator relationship matrix"

BLUP

➤ Genetic evaluation = solving a system of equations

Henderson demonstrated that BLUE $\hat{\beta}$ and BLUP \hat{a} are solutions of Mixed Model Equations (MME)

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \frac{1}{\sigma_g^2}\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{a} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

Where :

- \mathbf{R}^{-1} = inverse of \mathbf{R} , (co)variance matrix of the residuals. Very simple. $\mathbf{R} = \sigma_e^2\mathbf{I} \Leftrightarrow \mathbf{R}^{-1} = \frac{1}{\sigma_e^2}\mathbf{I}$
- Setup of \mathbf{A}^{-1} is very easy (Henderson, 1976). $\sigma_g^2\mathbf{A}$ is the genetic variance
- σ_e^2 and σ_g^2 are supposed to be known.

Easily extended to more complicated models (multitrait,...)

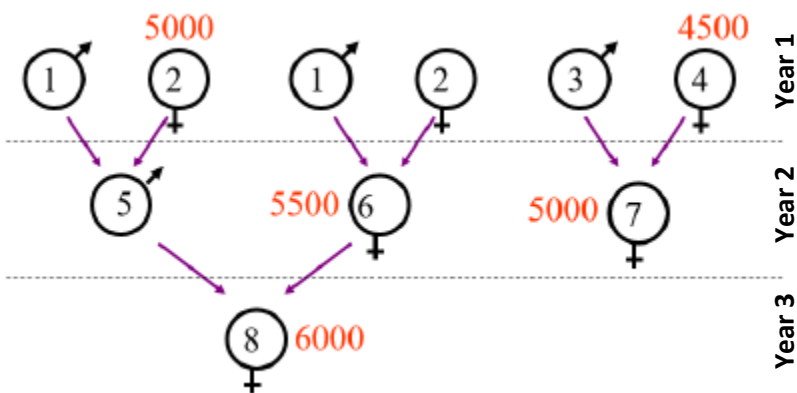


Size of the MME can be huge



Computations in genetic evaluation:

A Liliputian example



$$y_i = \beta_j + a_i + e_i$$

animal model

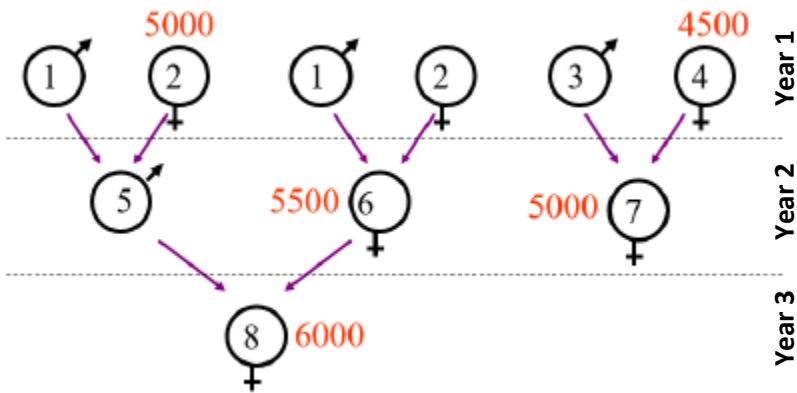
$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \alpha\mathbf{A}^{-1} \end{bmatrix}$$

where $\alpha = \frac{1-h^2}{h^2}$

Mixed Model Equations

$$\begin{pmatrix} 2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0+2\alpha & 0+\alpha & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0+\alpha & 1+2\alpha & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0+1.5\alpha & 0+0.5\alpha & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0+0.5\alpha & 1+1.5\alpha & 0 & 0 \\ 0 & 0 & 0 & 0-\alpha & 0-\alpha & 0 & 0 & 0+2.5\alpha & 0+0.5\alpha \\ 0 & 1 & 0 & 0-\alpha & 0-\alpha & 0 & 0 & 0+0.5\alpha & 1+2.5\alpha \\ 0 & 1 & 0 & 0 & 0 & 0-\alpha & 0-\alpha & 0 & 1+2\alpha \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0-\alpha & 0-\alpha \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \end{pmatrix} = \begin{pmatrix} 5000+4500 \\ 5500+5000 \\ 6000 \\ 0 \\ 5000 \\ 0 \\ 4500 \\ 0 \\ 5500 \\ 5000 \\ 6000 \end{pmatrix}$$

➤ Henderson rules : setup of \mathbf{A}^{-1}



1. Set $\mathbf{A}^{-1} = \mathbf{0}$
2. For each animal i with parents p & p' :
 - a) Add γ to (i,i)
 - b) Add $\frac{\gamma}{4}$ to $(p,p), (p',p'), (p,p'), (p',p)$
 - c) Add $-\frac{\gamma}{2}$ to $(i,p), (i,p'), (p,i), (p',i)$

where

- $\gamma = 2$ when both parents are known
- $\gamma = \frac{4}{3}$ when one parent is unknown
- $\gamma = 1$ when none of them is known

\mathbf{A}^{-1} is a very sparse matrix

➤ Models can be more or less sophisticated...

$$\underline{y}_i = \beta_j + \underline{a}_i + e_i \quad \text{same ID } i \rightarrow \text{animal model } a_i = \frac{1}{2}a_s + \frac{1}{2}a_d + \phi_i$$

Simpler models

Ignore the dam, which can be replaced by the maternal grand sire

More elaborate models

- More than one performance by animal
 - ❖ **Repeatability model**
- Two animals influence the same phenotype
 - ❖ **Maternal effect model**
- Many traits by animals or measured in more than one environment
 - ❖ **Multitrait animal model**

The derivation is quite straightforward but the computational cost can be a deterrent.



➤ Why bother ?

Because of desirable BLUP properties:

IF

- The base population is **neither inbred, nor related nor selected** and,
- All data previously used for **selection** and **mating** are accounted for in the genetic evaluation and,
- The A matrix encompasses **all relationships** and,
- **Genetic parameters** pertain to the **base population**

THEN

MME correctly account for all changes in additive genetic variance due to selection, genetic drift, inbreeding and preferential mating (Sorensen & Kennedy, 1982)

Using properly ALL available data leads to the best possible result

TAKE HOME MESSAGE

Ideal genetic evaluations

- ✓ Rely on an animal model
- ✓ Account for all available data (pedigree and performances) relative to directly or indirectly selected traits.

MULTITRAIT BLUP ANIMAL MODEL



INRAE

➤ Ideal genetic evaluation

Categorical traits
(threshold models)

Continuous traits



Unfortunately no software can simultaneously
analyse all traits AND all models ☹️☹️



Longevity (censored data, Cox or Weibull models)

Longitudinal data
(random regression models)

Two-step process :

1. Each trait is analyzed with the most adequate model.
2. After computation, phenotypes are corrected for all non genetic effects.
 - A new phenotype (deregressed EBV) v
 - Associated weight ω

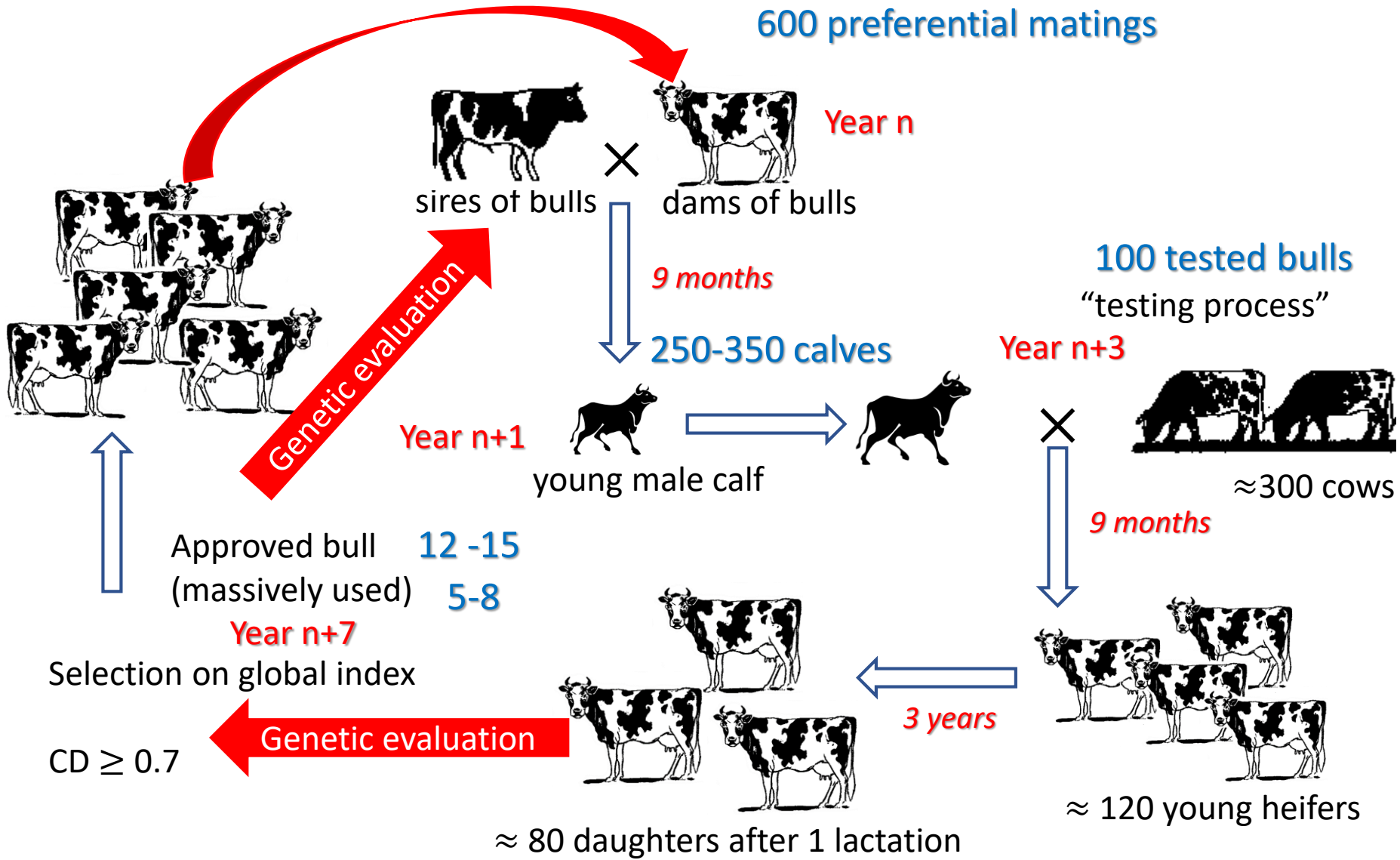
Then the (very simple) model $v = \mu + a + e$ gives the same result as the best model.

A multivariate evaluation of all v s with adequate weights ω is a good proxy for an ideal genetic evaluation (and is computationally more feasible).



➤ Pre-genomic selection in dairy cattle





➤ Pre-genomic selection in dairy cattle

- Long generation interval : ≈ 7 years
- Strong selection pressure.
1/4000 in Holstein (# tested bulls /# inseminations)
- High accuracy only for males : $CD \geq 0.7$ required for approval (not for all traits). Some traits with low heritability (fertility) were poorly evaluated.

➤ Pre-genomic selection in swine production



➤ Selective Breeding of Pig

In France :

- 14 000 farms
- 125 sows /farm
- 23 M slaughtered
- 2,2 M tons of meat

(2018 statistics)

Type femelle
Dam lines

♀

Axes de sélection :

- Prolificité
- Qualités maternelles
- Rusticité et facilités d'adaptation
- Prolificacy
- Maternal abilities
- Hardy and easy to adapt

Lignées Large White femelle
Large White dam lines

Lignées Landrace Français
Landrace French lines

Lignées sino-européennes
Chinese-european lines

Lignées Duroc femelle
Duroc dam lines

Type mâle
Sire lines

♂

Axes de sélection :

- Croissance
- Indice de consommation
- Composition des carcasses
- Qualité de viande
- Growth rate
- Feed conversion ratio
- Carcass muscle content
- Meat quality

Lignées Piétrain
Pietrain sire lines

Lignées Large White mâle
Large White sire lines

Lignées synthétiques
Composite lines

Lignées Duroc mâle
Duroc sire lines



INRAE

➤ Selective Breeding of Pig:

Breeding goals



- ✓ Increase muscle
- ✓ Reduce fat (backfat thickness)
- ✓ Improve growth (Average Daily Gain)
- ✓ Good feed efficiency (Daily Feed Intake)
- ✓ Meat quality (pH, meat quality index)

➔ Letal measure
➔ slaughtered sibs

- ✓ Higher prolificacy
- ✓ Maternal qualities



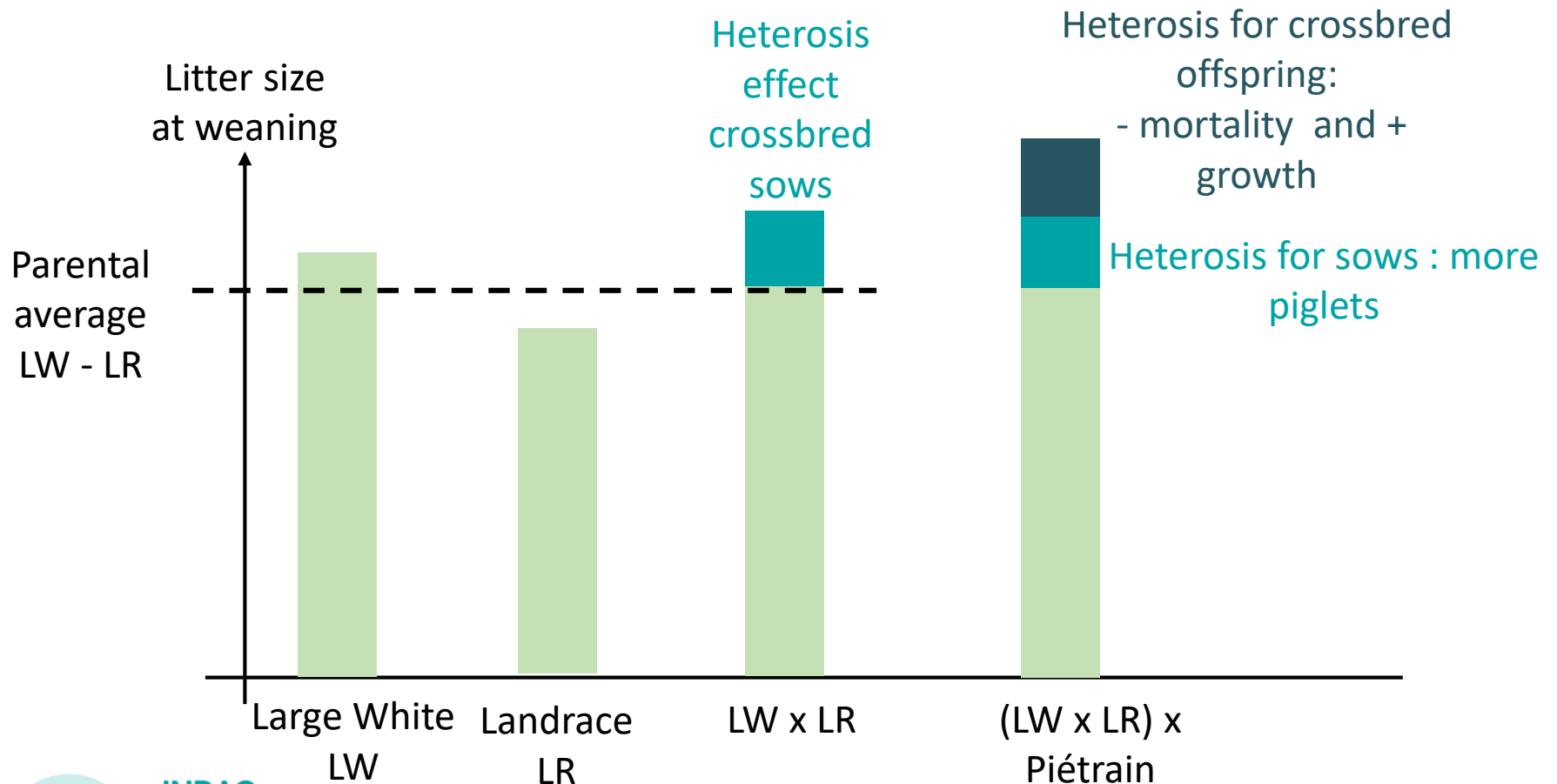
These traits are unfavorably correlated with the previous ones
➔ crossbreeding



➤ Selective Breeding of Pig:

advantages of crossbreeding 1

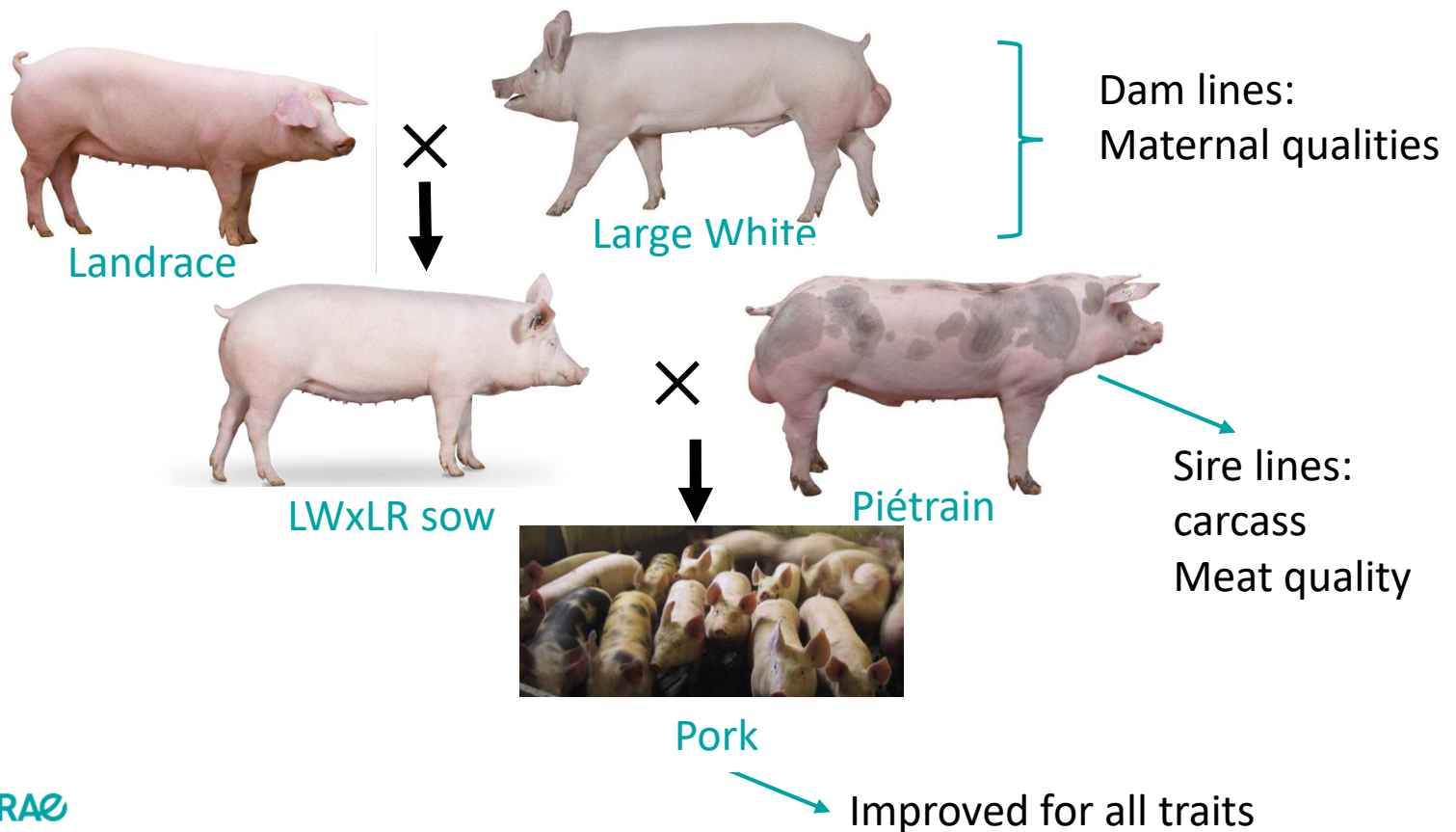
- heterosis: crossbred animals outperform parental average



➤ Selective Breeding of Pig: advantages of crossbreeding 2

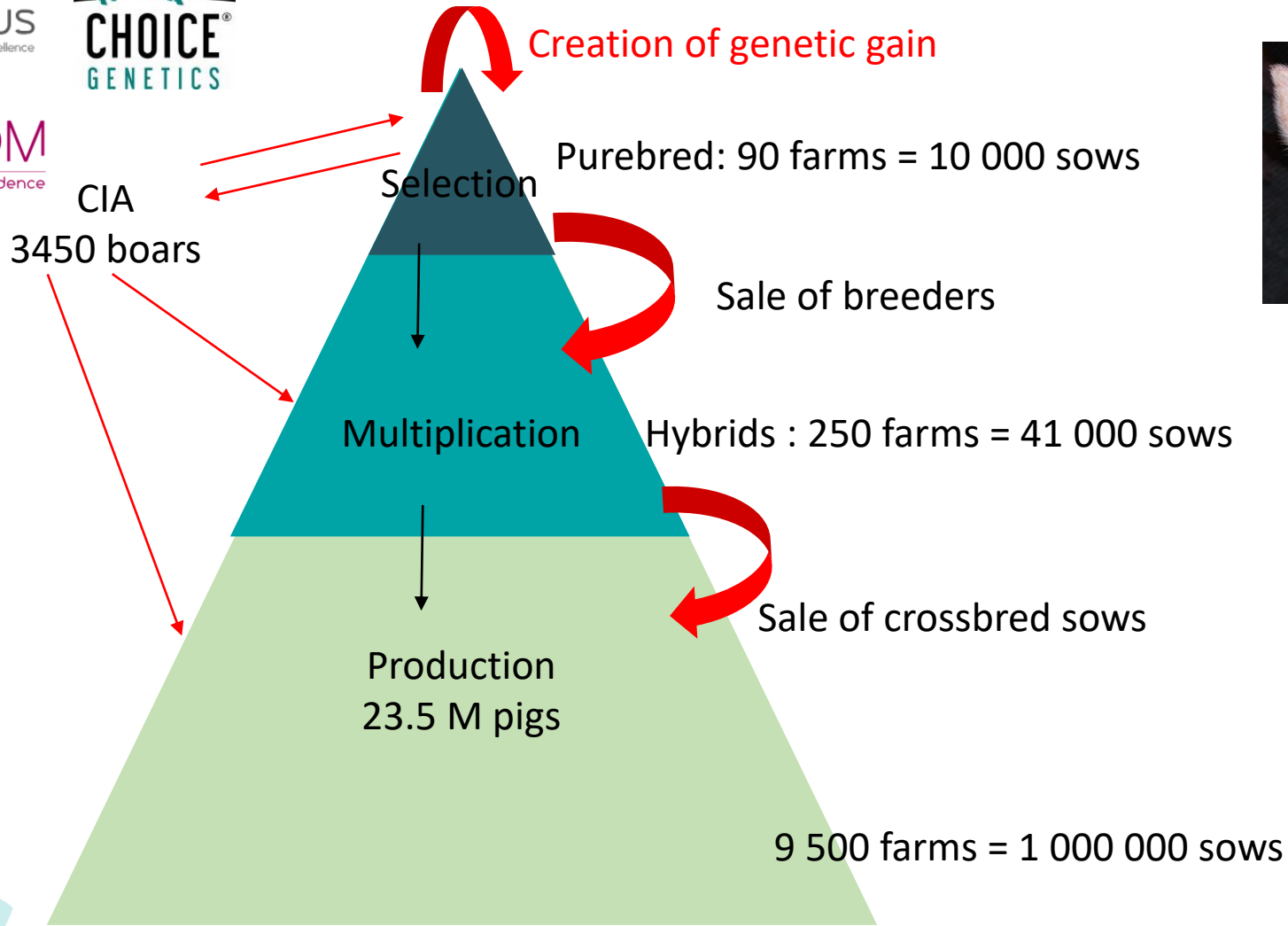
Pork is obtained from crossbreeding of multiple lines

Major cross in France :



➤ Selective Breeding of Pig:

pyramidal design



➤ Pre-genomic selection of pigs

performance monitoring



Purebred animals
Selection farm

weaning (≈ 1 month)

Selection candidate
(farm)

collateral sibs
performance monitoring
station



High number of
measured
animals
(80 000/an)

fattening

fattening

Costly measures
(2 500 indiv/an)

Evaluation around 100kg
($\approx 4-5$ monthes)

Evaluation around 100kg

selection
CIA

multiplication

slaughter ≈ 6 months

slaughter



- ➔ Generation intervals are short.
- ➔ Selection is based on imprecise EBVs

➤ Pre-genomic selection of pigs

- Short generation interval:

1 to 2 years

- High selection intensity:

**1/11 à 1/16 for female pathway
& 1/50 à 1/65 for male pathway**

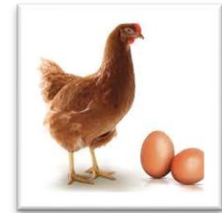
- Low accuracy of EBVs:

between 0.15 and 0.40

➤ Pre-genomic selection in poultry



➤ World leaders share the global market



Hy-Line®



LOHMANN BREEDERS



B.U.T.

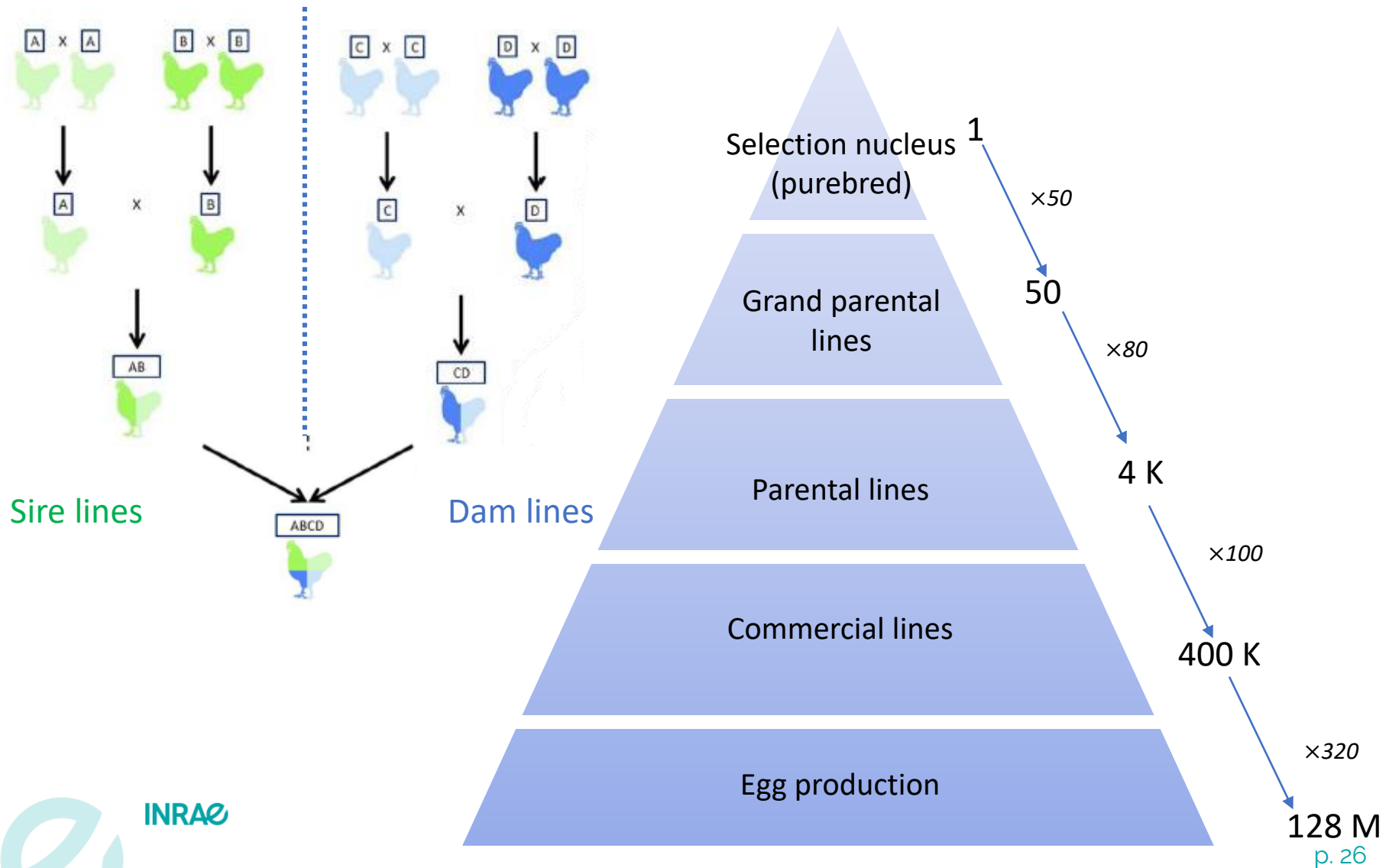


INRAE

Hervé CHAPUIS 08/11/2022



➤ Pyramidal design in layers

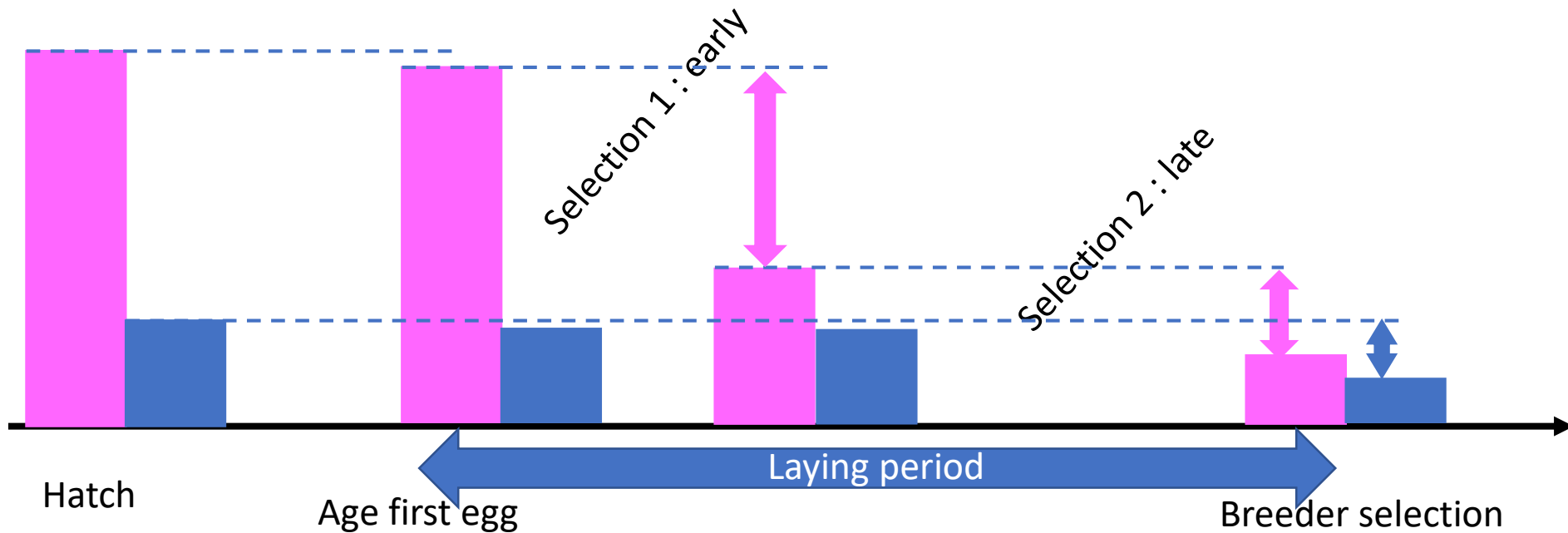


➤ Breeding goals in layers

Breeding goal	Selection criterion
Egg production	Laying intensity Persistency in lay Age at first egg
Egg quality	Egg weight Eggshell color and resistance Shape Yolk yield Albumen height
Feed efficiency	Feed intake FCR Number of meals
Viability	Feather pecking Disease resistance

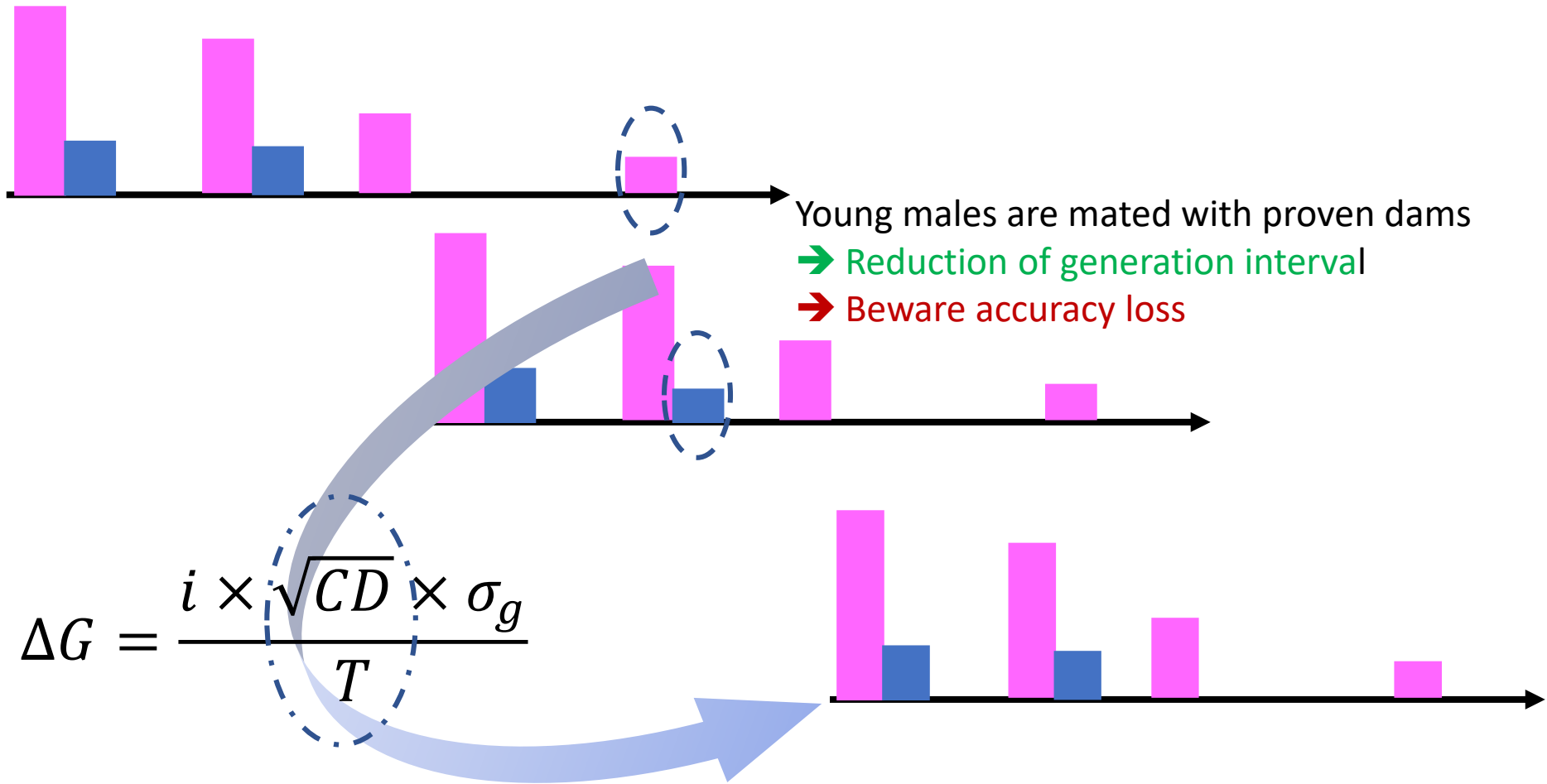


➤ Pre-genomic selection in layers: breeding scheme principles



Generation interval is high, as main phenotype of interest is persistency of lay
Selection intensity is low on the male pathway (no phenotype). Selected males are sibs of the “best” layers → low accuracy & low selection intensity (all sibs are considered equal)

➤ Pre-genomic selection in layers: alternative breeding scheme (utilization of young males)



Every variation from the original scheme should be carefully modelled before implementation.

➤ Outlines

I. Classical (pre-genomic) selection

II. Principles of genomic selection

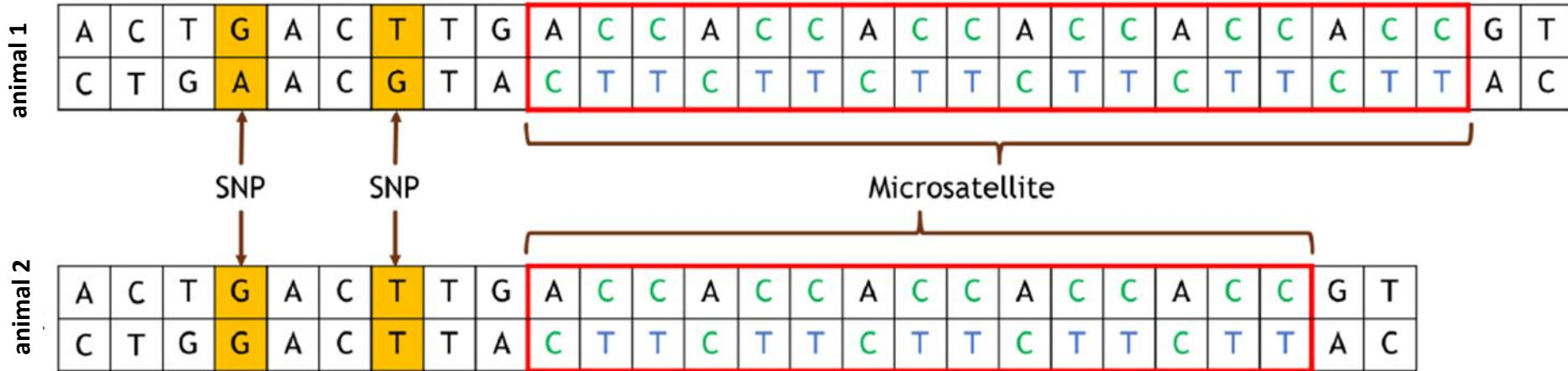
III. Consequences on the selection process



➤ Genomic selection

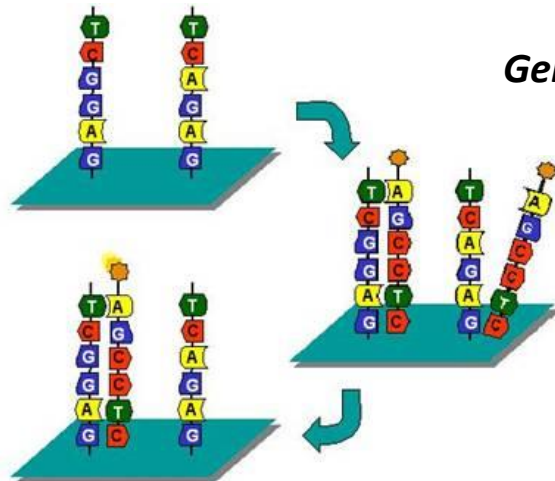
SNP markers

Single Nucleotide Polymorphisms are less informative (biallelic) than other markers but easier (and cheaper) to detect.



SNP chip

INRAE



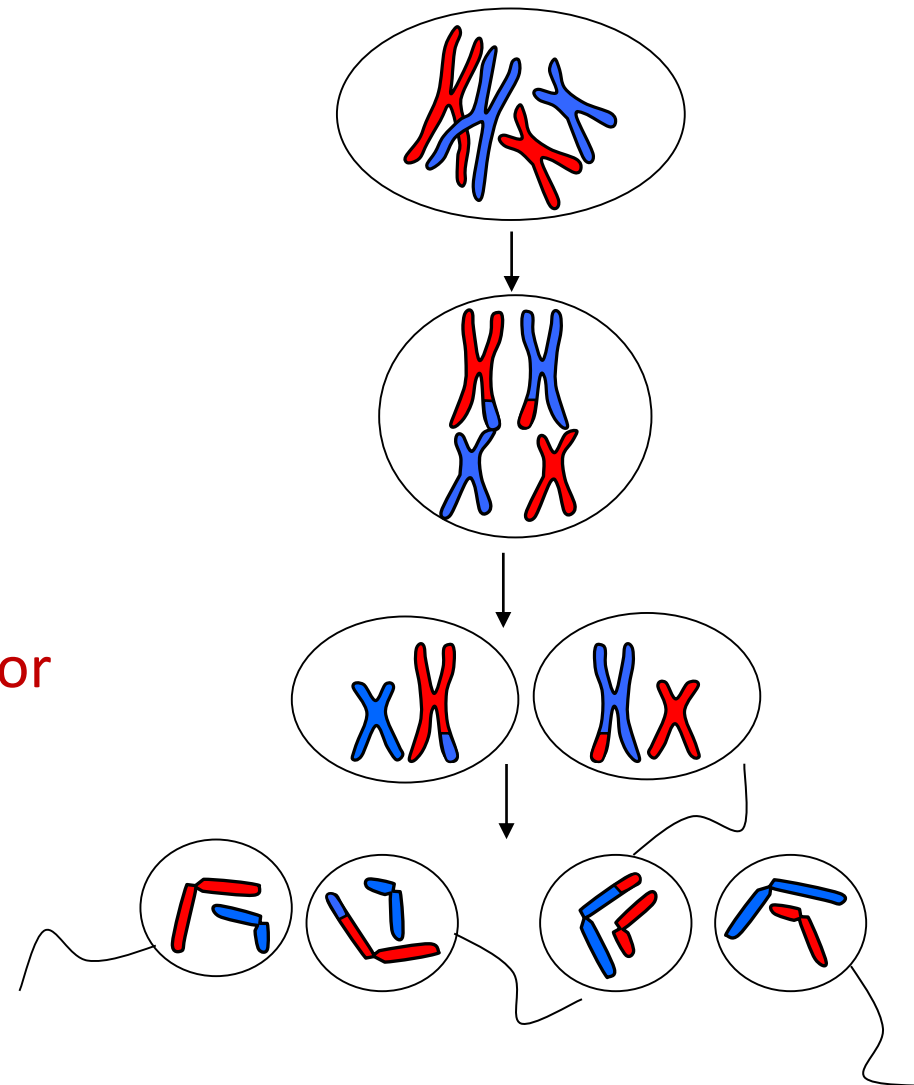
Genotyping through fixation of fluorescent nucleotides



➤ Genomic selection

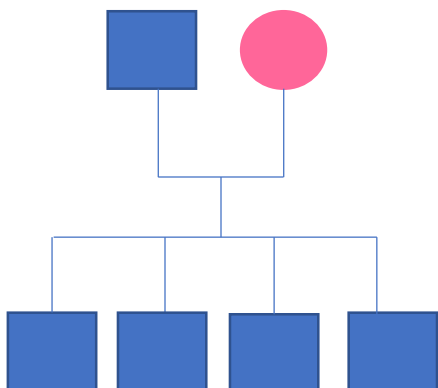
DNA transmission generates variability

- During meiosis:
 - Crossing-over
 - Homologous segments are shuffled
- Causes genetic variability
- Mendelian sampling accounts for 50% of σ_a^2



➤ Genomic selection

EBV are estimated earlier, with a better accuracy



Based on Pedigree BLUP (P-BLUP):

- Accuracy is low ($CD_i = 0,25(CD_s + CD_d)$)
- Can't distinguish the best among siblings
- Yet, due to genetic recombination, parents did not transmit the same qualities to all their offspring



Improve the knowledge of genetic potential through **assessed transmission** of chromosomal segments using tags (SNP chip).

Requires a **reference population** genotyped and whose genetic merit is known with high accuracy: allows for establishing relation between transmitted chromosomal segments and breeding value.

➤ Genomic selection

Focus on reference population

Reference population (genotyped AND phenotyped)

➔ establish prediction equations

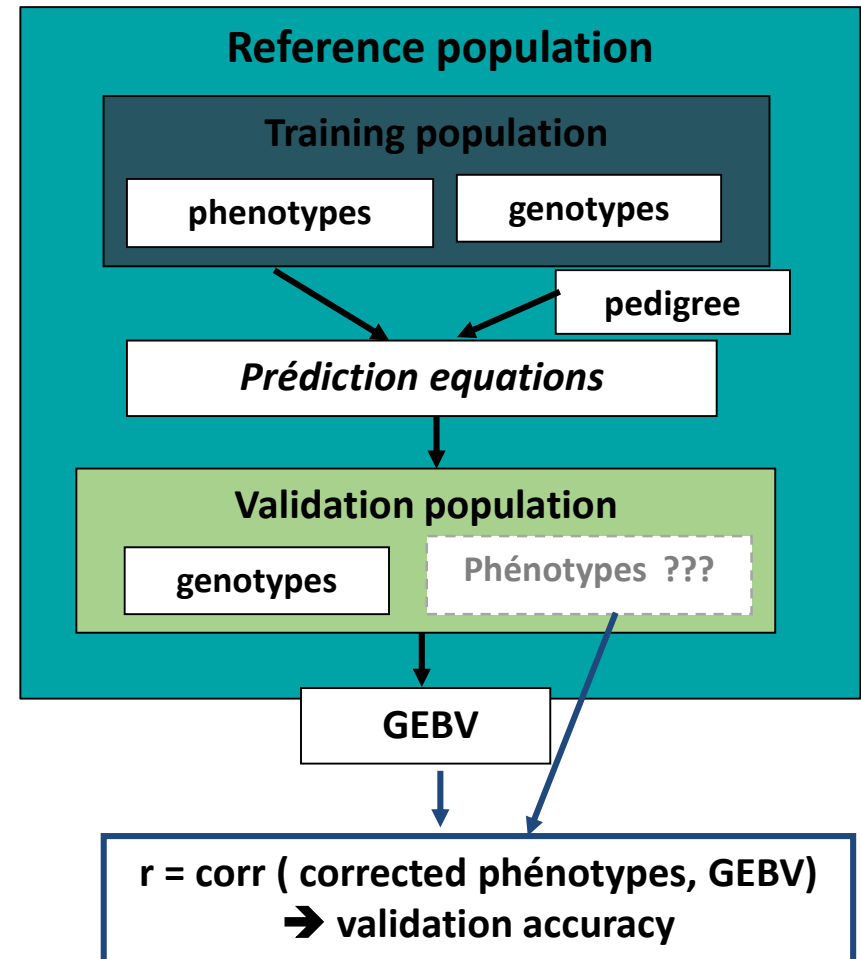
Setup of reference population is crucial BEFORE using GEBV in candidate population.

The reference population should be large enough to catch all possible haplotypes in the candidate population (depends on N_e).

Better sample in different families (maximization of genetic diversity, e.g. off-diagonal terms of A matrix).

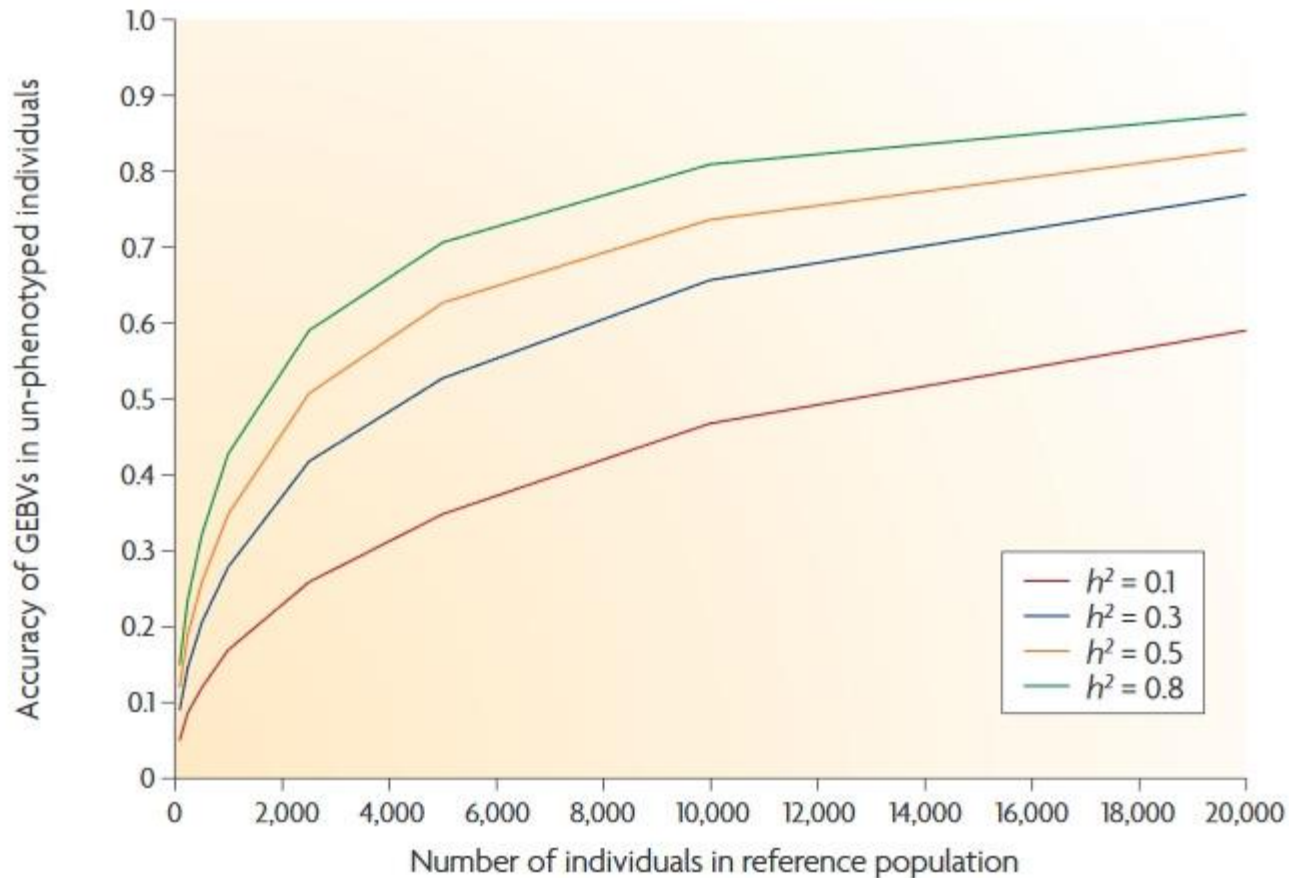
The reference population should be close enough to the candidate population (inclusion of dams should be considered).

Sufficient number of SNP (50k)



➤ Genomic selection

Influence of reference population size on accuracy



Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics. 2009;10:381-391.

➤ Genomic selection

Genomic evaluation

$$y_i = \beta_j + a_i + e_i$$

Estimated using **pedigree** & **genetic markers**

SNP chip ➔ several biological samples can be used to extract DNA: blood, hair, biopsies.

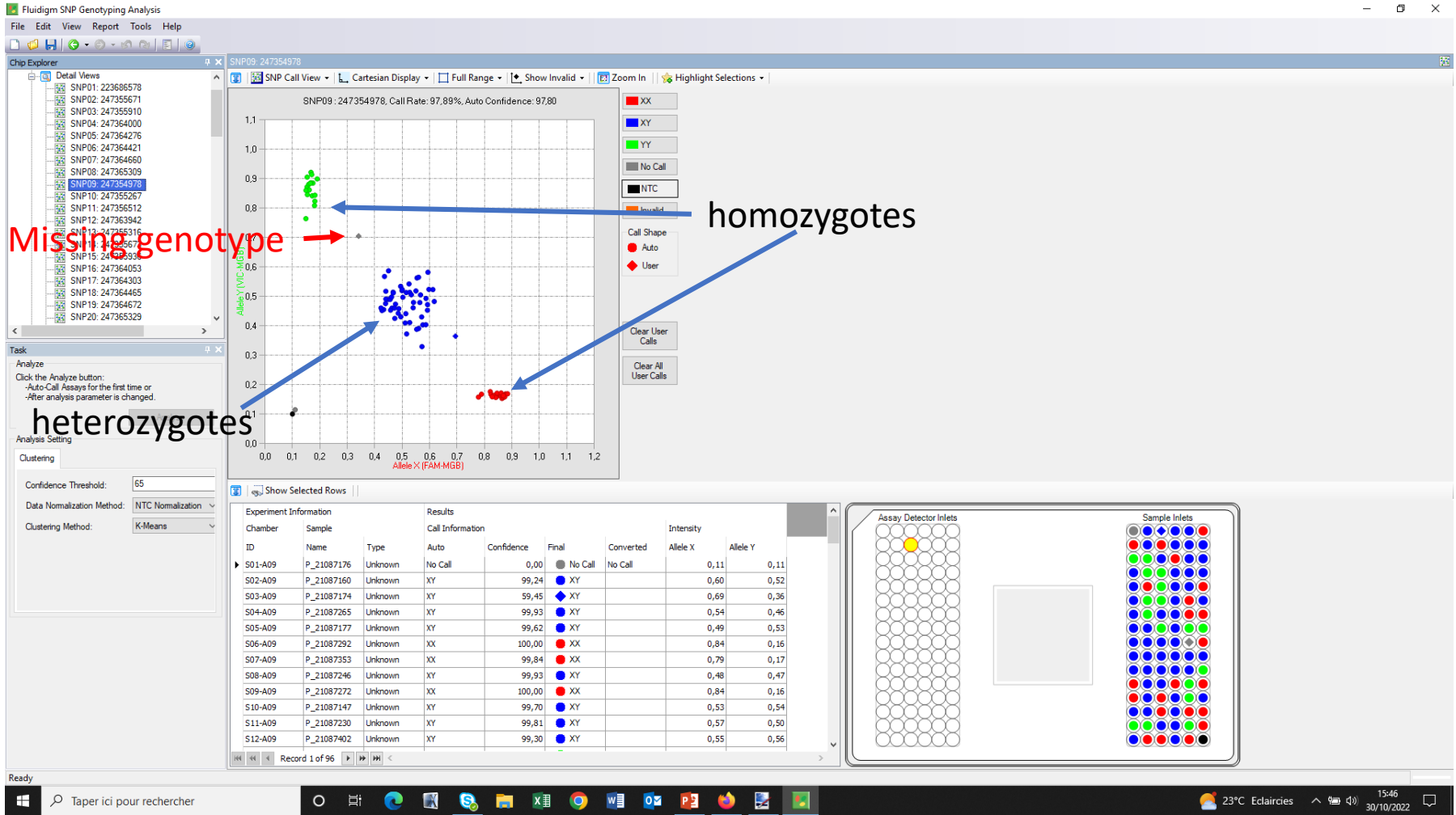
From 100K SNP (LD chip) to 600K SNP (HD chip)
Illumina & Thermo Fisher are two big chip providers

For each locus we obtain the nucleotide info : A T C G
Illumina uses another notation : A/B for each locus. X/Y can also be found.

Considering a SNP is biallelic, the information can be condensed, using an integer code to count the number of copies of a reference allele.

Genomic selection

SNP genotyping



➤ Genomic selection

“ped” file **Data for genomic evaluation**

“map” file

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8
ID1254	AA	BB	AB	AB	BB	AA	..	AB
ID1869	AB	BB	..	AA	AB	AA	BB	AB
...	...							

SNP_name	chromosome	Position (bp)
SNP1	1	1,245
SNP2	1	458,796
SNP3	1	586,987
SNP4	1	796,874
SNP5	1	1,200,687
SNP6	1	1,265,973
SNP6	1	1,364,789
SNP8	1	1,400,278
...		
SNP _i	5	560,785

genotype	code
AA	0
AB or BA	1
BB	2
Missing	5

ID1254	02112051
ID1869	12501021

➤ Genomic selection

Quality control of genomic data

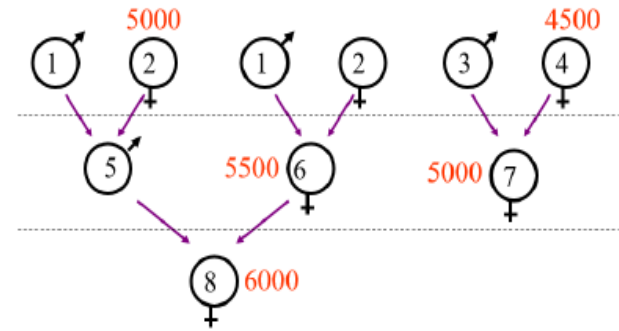
Mandatory !!! Can be achieved using PLINK or within blupf90 suite.

1. **Animal call rate.** N genotyped SNP / N SNP on the chip. Depends on sample quality. Below 0.95 the genotype is not reliable.
2. **SNP call rate.** N animals with genotype at a SNP/ N animals. Can eliminate SNPs with poor technical quality. Usual threshold =0.95
3. **Minor Allele Frequency.** Chip was designed using mixed populations and some alleles may be non informative in some lines. They are also more error-prone. Usual threshold = 0.05
4. **Deviation from Hardy-Weinberg Equilibrium.** Used as a signal for genotyping errors. If $P_value < 10^{-4}$, markers are generally discarded.
5. **Heterozygosity.** Too low H_e can signal a poor DNA quality or inbreeding issue. Too high H_e may be due to pollution. Discard animal i if $H_{e_i} - \overline{H_e} > 3\sigma$
6. **Pedigree check.** Detection of Mendelian conflicts. If parents are opposite homozygotes, offspring is heterozygote. If the number of mismatches exceeds a given threshold, the pedigree is false.



Genomic selection

Genomic evaluation



```
ID001 21220201011002211211112121001121111122120001010122111110012012111120221002012201100222002200201111120
ID002 1201121111012200221100012110202021210101000220020202020002002222000222000011221101112200120020012220
ID003 1212020111001211212000102001202022201112000110020102011001012212110122100001210101112200210010022220
ID004 2122020210000221111111112010011100011112000202002112120112201221121022100111220110111200210010111220
ID005 211112110111121012121112210202021211211000110021211111001101211101122100111221000022200120020101120
ID006 22111212001012112202111121011212111201000110011202020002102222001122000101221000021200110010101220
ID007 222202020000022221100021200111211112212000101011102011001112212121022100102210110022200210010022220
ID008 2122020101210121121222101200112122102212000010012211111001101211101122100110220000021200210010200120
```

A matrix (pedigree)

	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8
ID1	1.00	0.00	0.0	0.0	0.50	0.50	0.0	0.25
ID2	0.00	1.00	0.0	0.0	0.50	0.50	0.0	0.25
ID3	0.00	0.00	1.0	0.0	0.00	0.00	0.5	0.00
ID4	0.00	0.00	0.0	1.0	0.00	0.00	0.5	0.00
ID5	0.50	0.50	0.0	0.0	1.00	0.50	0.0	0.75
ID6	0.50	0.50	0.0	0.0	0.50	1.00	0.0	0.75
ID7	0.00	0.00	0.5	0.5	0.00	0.00	1.0	0.00
ID8	0.25	0.25	0.0	0.0	0.75	0.75	0.0	1.25

G matrix (markers)

	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8
ID1	0.549	-0.589	-0.323	0.196	0.007	-0.207	0.143	0.224
ID2	-0.589	1.169	0.372	-0.428	0.187	0.200	-0.384	-0.528
ID3	-0.323	0.372	0.863	-0.291	-0.223	-0.307	0.268	-0.359
ID4	0.196	-0.428	-0.291	1.000	-0.379	-0.174	0.208	-0.130
ID5	0.007	0.187	-0.223	-0.379	0.557	0.087	-0.432	0.196
ID6	-0.207	0.200	-0.307	-0.174	0.087	0.582	-0.227	0.047
ID7	0.143	-0.384	0.268	0.208	-0.432	-0.227	0.767	-0.343
ID8	0.224	-0.528	-0.359	-0.130	0.196	0.047	-0.343	0.895

➤ Genomic selection: 2 approaches

2 step vs. single step

SNP-based 2 step approaches:
SNP-BLUP, Bayes A, Bayes B, Bayes $C\pi$,...

Reference population

Performances +
Pedigree +
"best model"

Pseudo performances + weight
(milk traits for dairy bulls)

Genotypes

Prediction equations

Genotypes of candidates

Genomic EBVs of candidates



➤ Genomic selection: 2 approaches

2 step vs. single step

Reference population

Performances +
Pedigree +
“best model”

Pseudo performances + weight
(milk traits for dairy bulls)

Genotypes

Genotypes of candidates

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \alpha G^{-1} \end{bmatrix}$$

Where **G** is the matrix of genomic relationships.
(The system is no longer sparse)

GBLUP and SNP-BLUP were found to be equivalent

Genomic EBVs of candidates



➤ Genomic selection: 2 approaches

2 step vs. single step

Only a small portion of the animal in a given population are genotyped.

To avoid multi step approach, the idea is combine pedigree and genomic relationships and use this matrix in MME.

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix} \quad \mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

Setup of MME to get **ssGLUP** is straightforward :

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wu} + \mathbf{e}$$

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \mathbf{H}^{-1} \otimes \mathbf{G}_0 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

➤ Genomic selection:

A word about G matrix

G is singular if **Z** uses centered coding with observed allele frequencies (last row can be predicted from the other ones).

➔ Inclusion of a small part of pedigree matrix **A**₂₂: $\mathbf{G} = (1 - \alpha)\mathbf{G} + \alpha\mathbf{A}_{22}$

G was computed using current (available) allele frequencies.

To account our ignorance as to the difference between pedigree and genomic bases, a correction is proposed to have same mean diagonal and off-diagonal as **A**₂₂:

$$G^* = (G + 11'a) \text{ where } a = \overline{A_{22}} - \bar{G}$$

➤ Genomic selection:

Interest of ssGBLUP

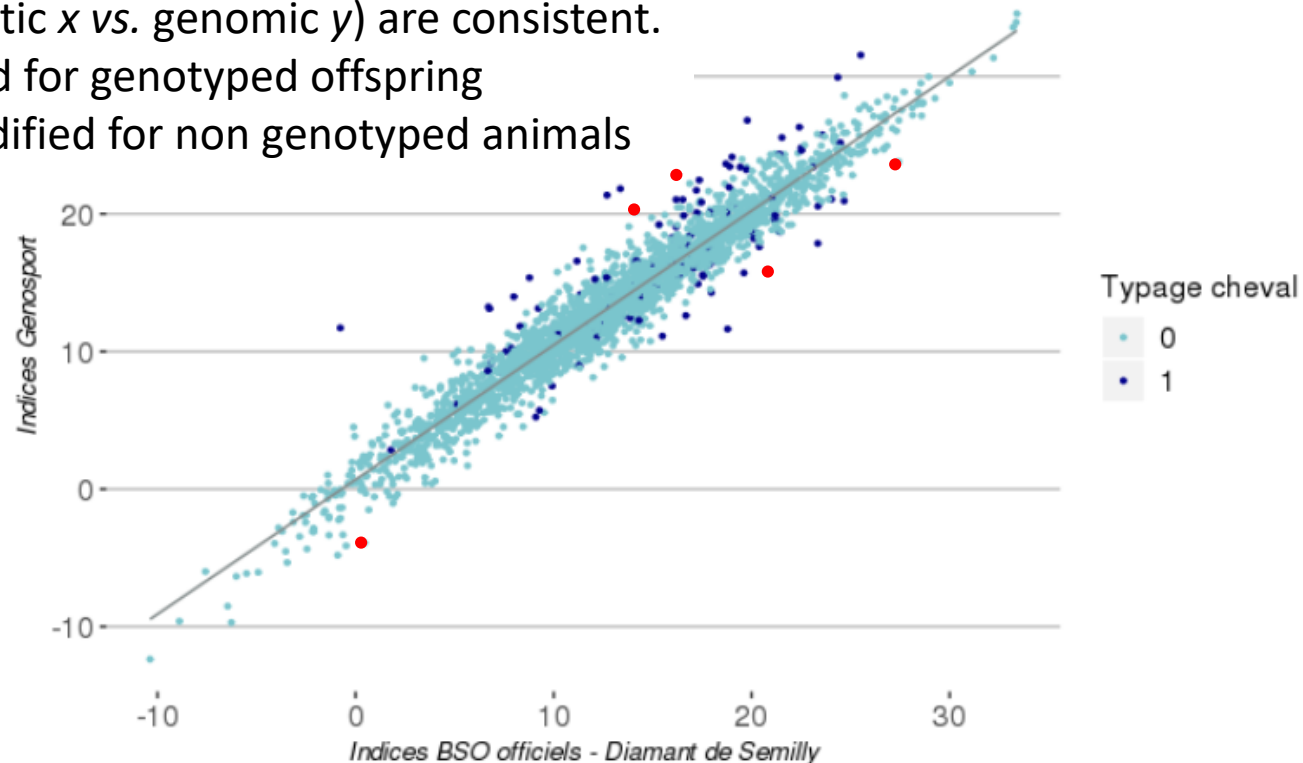
As u_1 and u_2 are jointly estimated, **EBVs of non genotyped animals also benefit from genotypes**



Diamant de Semilly was long considered the best French stallion for show jumping.

He sired a very large number of offspring with performances. ➔ CD =1

On average EBVs (plain genetic x vs. genomic y) are consistent.
Largest changes are observed for genotyped offspring
but some EBVs are quite modified for non genotyped animals



➤ Genomic selection

What are the costs ?



In 2019, the cost for a HD 600K genotype was around 150€
(TWD 4815.00)

In 2019, the cost for a LD 1K genotype was around 35€
(TWD 1125.00)



A cost-saving strategy is:

1. genotype selection candidates with a LD or MD chip
2. **impute genotypes of candidate to the MD or HD chip**
3. Compute GEBVs
4. Select candidates
5. Genotype selected breeders with MD or HD chip so that they are included in the reference population

The advantages of a HD (or sequence) over a MD genotype should be carefully evaluated.

➤ Genomic selection

What is imputation ?

IMPUTATION IS T_E PR__ICT_N _F MI_SI_GL__T__S _N W_R_S O_ S__T_NC_S
_H__HR_I__O__I_K_G_DI_EQ__LI_R__M

IMPUTATION IS THE PREDICTION OF MISSING LETTERS IN WORDS OR SENTENCES
WHICH RELIES ON LINKAGE DISEQUILIBRIUM

What is linkage disequilibrium ?

How can it be used for imputation ?



➤ Genomic selection

Linkage disequilibrium LD

Non-random association of alleles between two loci.

	Observed frequency	
A B	42	50
A b	28	20
a B	18	10
a b	12	20

Alleles frequencies :

$$p_A = 70/100 = 0.7$$

$$p_a = 30/100 = 0.3$$

$$p_B = 60/100 = 0.6$$

$$p_b = 40/100 = 0.4$$

Expected haplotypes frequencies :

$$\text{Freq}(AB) = p_A \times p_B = 0.42$$

$$\text{Freq}(Ab) = p_A \times p_b = 0.28$$

$$\text{Freq}(aB) = p_a \times p_B = 0.18$$

$$\text{Freq}(ab) = p_a \times p_b = 0.12$$

Deviation from expected haplotype frequency ➔ Linkage disequilibrium

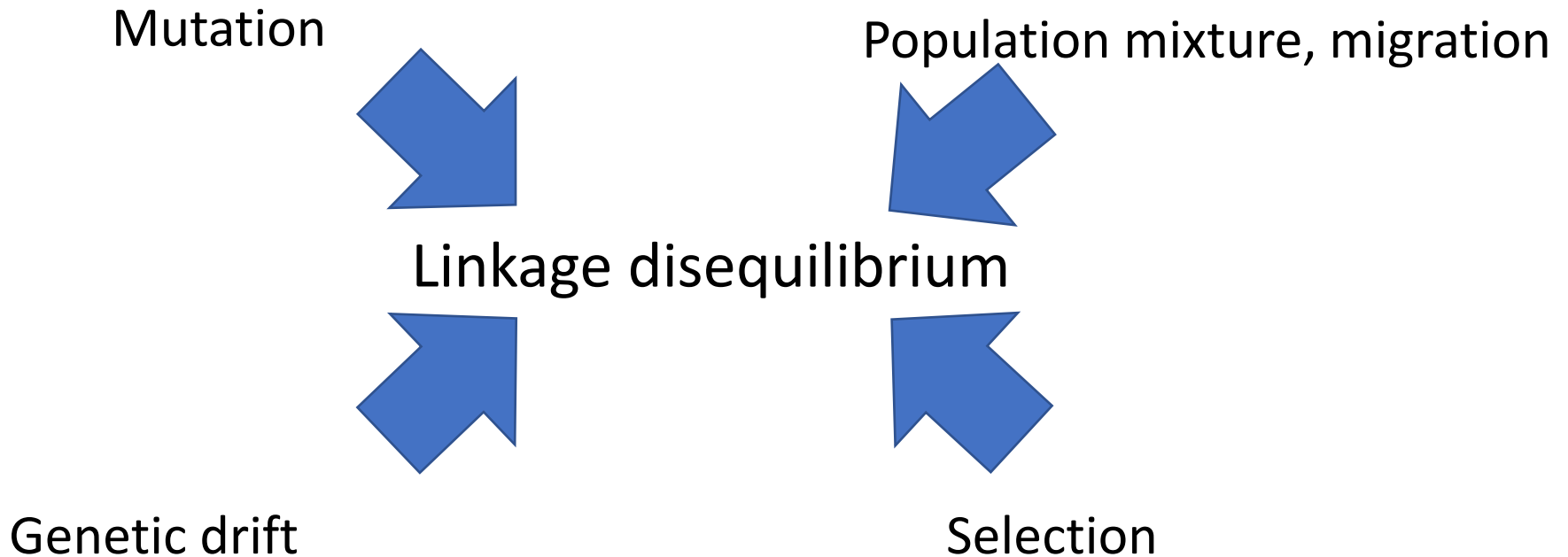
There is no single best statistic that quantifies the extent of DL

$$D = \text{Freq}(AB) \times \text{Freq}(ab) - \text{Freq}(Ab) \times \text{Freq}(aB)$$

$$r^2 = \frac{D^2}{p_A \times p_a \times p_B \times p_b} \quad (\text{preferred, less dependent on allele frequencies})$$

> Genomic selection

Linkage disequilibrium LD



➤ Genomic selection

Linkage disequilibrium LD

An example on how mixture of populations creates DL

	Pop. 1			Pop. 2			Mixture pop. 1 & pop. 2	
	A	a		A	a		A	a
B	81	9		42	18		123	27
b	9	1		28	12		37	13
	D=	0		D=	0		D=	0,015
	r²=	0		r²=	0		r²=	0,225

NO DL

NO DL

DL



➤ Genomic selection

Linkage disequilibrium LD

Average DL = f(#SNP, distance between markers)

	Duroc	New Hampshire
#SNP on chip	34K	37K
100 kb	0.36	0.27
1Mb	0.19	0.12

An average DL above 0.3 is considered favorable for implementation of genomic selection.

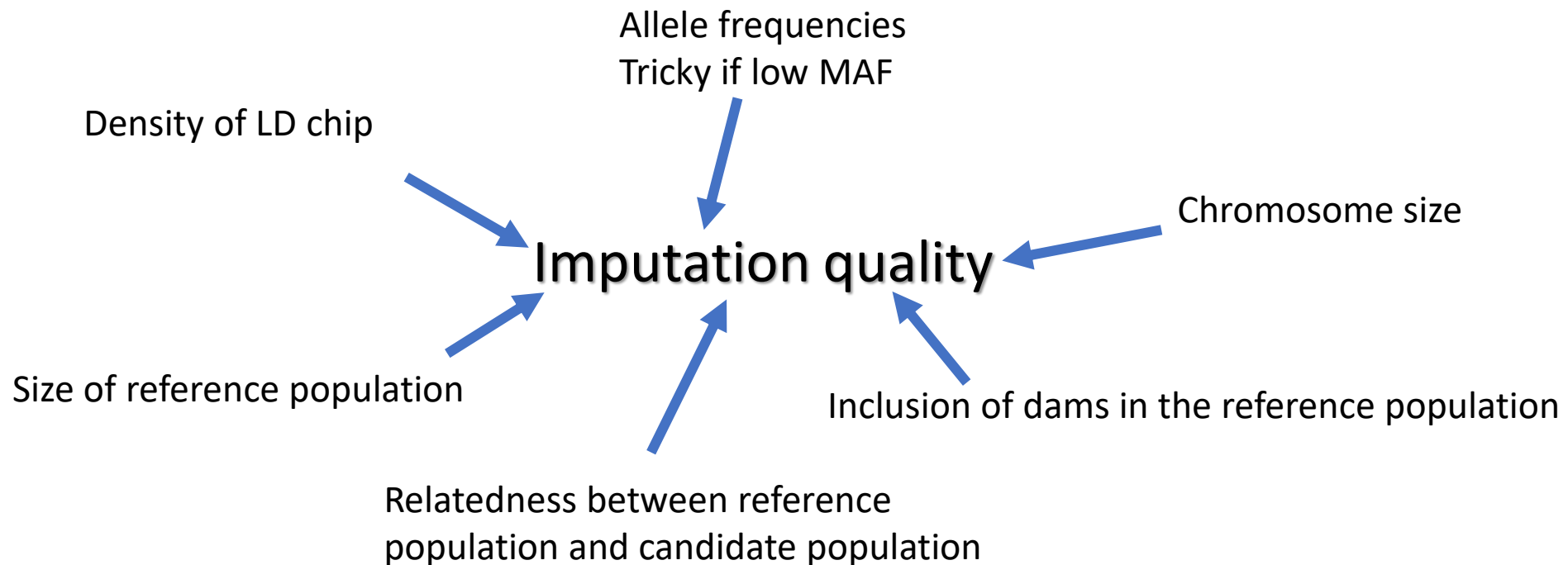
In poultry species, DL was found lower in micro chromosomes compared to macro chromosomes.



➤ Genomic selection

Imputation quality

Imputation using pedigree information is faster, as parents can be phased (not implemented in first software, that were used in human studies)



➤ Genomic selection

Imputation strategy

Given the genotyping costs, how to design a LD chip to perform genomic selection ?

1. Start from the reference population with HD genotypes.
2. Split in two populations. One will be used to test scenarios (LD chip design), the rest will be considered as a reference population.
3. *in silico* discard SNPs on the test population following a given scenario.
4. Impute the HD genotypes on the test population using the reference population.
5. Compare the imputed and true genotypes (mean correlation, one SNP at a time for all candidates)



➤ Genomic selection

Expected outcome

- Get an EBV for non phenotyped animals or animals without offspring or sib
 - Select new traits
 - Select sooner (at birth !)
 - ➔ decrease generation interval
- Increase EBV accuracy
- Reduce/abolish progeny testing costs
 - ➔ increase selection intensity.
- Choose from litter siblings

$$\Delta G = \frac{i \times \sqrt{CD} \times \sigma_g}{T}$$

Increased genetic gain !!

INRAE



➤ Genomic selection

Pedigree testing and parentage assignment

During quality control phase, Mendelian errors are tracked.

➔ Control of supposed pedigree

➔ In case of pedigree errors, the true parental pair can be found.

➔ Use a sample of SNP (N=100 to 200) adequately covering the autosomes and with desirable high MAF.

➔ Many software can be used.

R package APIS (Griot et al. *Mol Ecol Resour.* 2020;20:579–590.)

➔ **Make sure all potential parents are genotyped ! 💣**

➤ Outlines

I. Classical (pre-genomic) selection

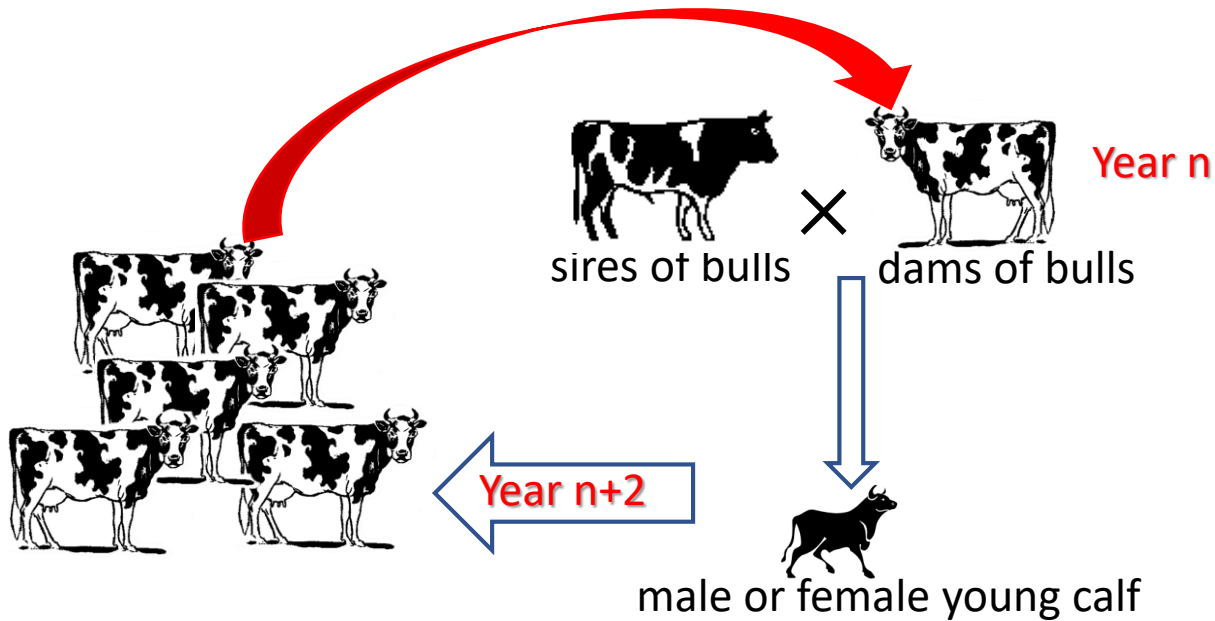
II. Principles of genomic selection

III. Consequences on the selection process



Consequences of the genomic selection

Dairy cattle



- Large reduction of selection interval
 - Large increase of accuracy (especially for low heritable traits)
 - Costs are dramatically reduced
- ➔ Testing of young bulls has been discontinued as soon as genomic indexes have been published

End of the “star system” era.

Each year new well indexed young bulls replace older ones

2.2 M cows, 9200 genotyped bulls, 140 approved bulls (Holstein, 2020)

Introduction of new traits in the selection index

Consequences of the genomic selection

Pig selection

Genomic selection is supposed to:

- Increase accuracy of EBVs

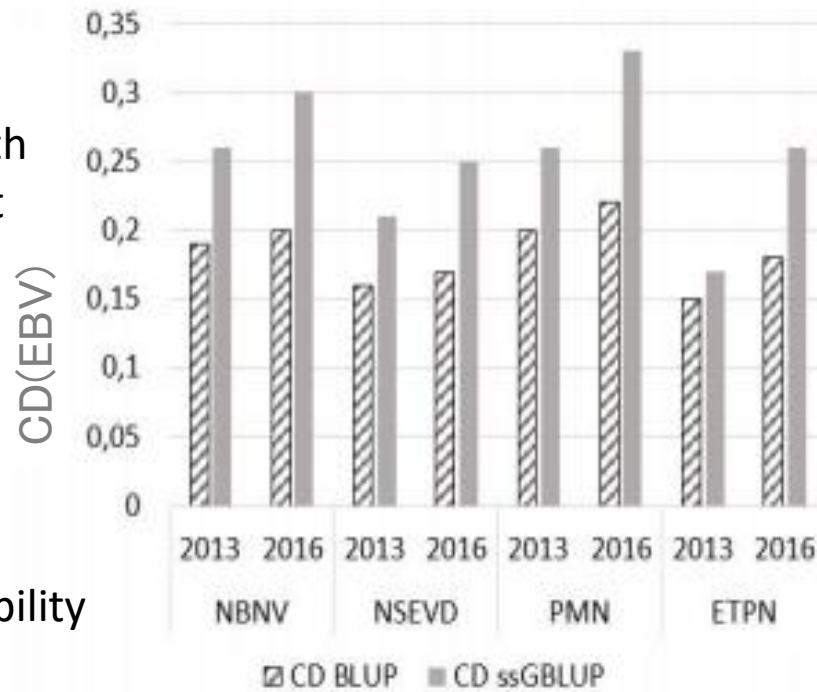


NBNV= # piglets born alive

NSEVD= #weaned piglets

PMN= Average piglet weight at birth

ETPN = standard deviation of piglet weight at birth



A. Bouquet, JRP 2017

Reproductive traits → low heritability



INRAE

➤ Consequences of the genomic selection

Pig selection

Genomic selection is supposed to:

- Allow for an increased selection intensity



Not possible, as almost all piglets are measured at 100 kg.

➔ The number of candidates depends on the number of farms.

➤ Consequences of the genomic selection

Pig selection

Genomic selection is supposed to:

- Decrease interval between generations



For reproductive traits (with low heritability) the generation interval actually decreased from 1 year to 6 months

➤ Consequences of the genomic selection

Pig selection

Genomic selection is supposed to:

- Allow for selection of new traits



➔ Crossbred selection

➔ Traits measured on a tiny part of the population (boar taint, microbiota/ digestibility, health, ...)

➤ Consequences of the genomic selection

Layers

Genomic selection is supposed to:

❖ Improve accuracy of EBVs 

❖ Reduce generation interval 

- Males EBVs are better estimated at young age. No need to wait for late performances of female sibs.

❖ Increase selection intensity. 

- Full-sibs can be ranked, based on the alleles they received. It is worth hatching a large number of male candidates.

➤ Consequences of the genomic selection

Layers

Incidentally, the introduction of genomic selection in poultry breeding schemes revealed that 5 to 10% of pedigrees were wrong 😞

Wrong pedigree → lower heritabilities.

Simple correction of pedigree errors, which is a by-product of genomic evaluation, also resulted in improved genetic gain. 😊

➤ Consequences of the genomic selection

Poultry breeders

Each breeding company developed its own strategy, depending on genotyping costs. Some used LD chip + imputation, others use MD (50K) chips on all candidates.

TIP1 : Genotype dams. If you don't genotype female candidates, at least genotype selected dams.

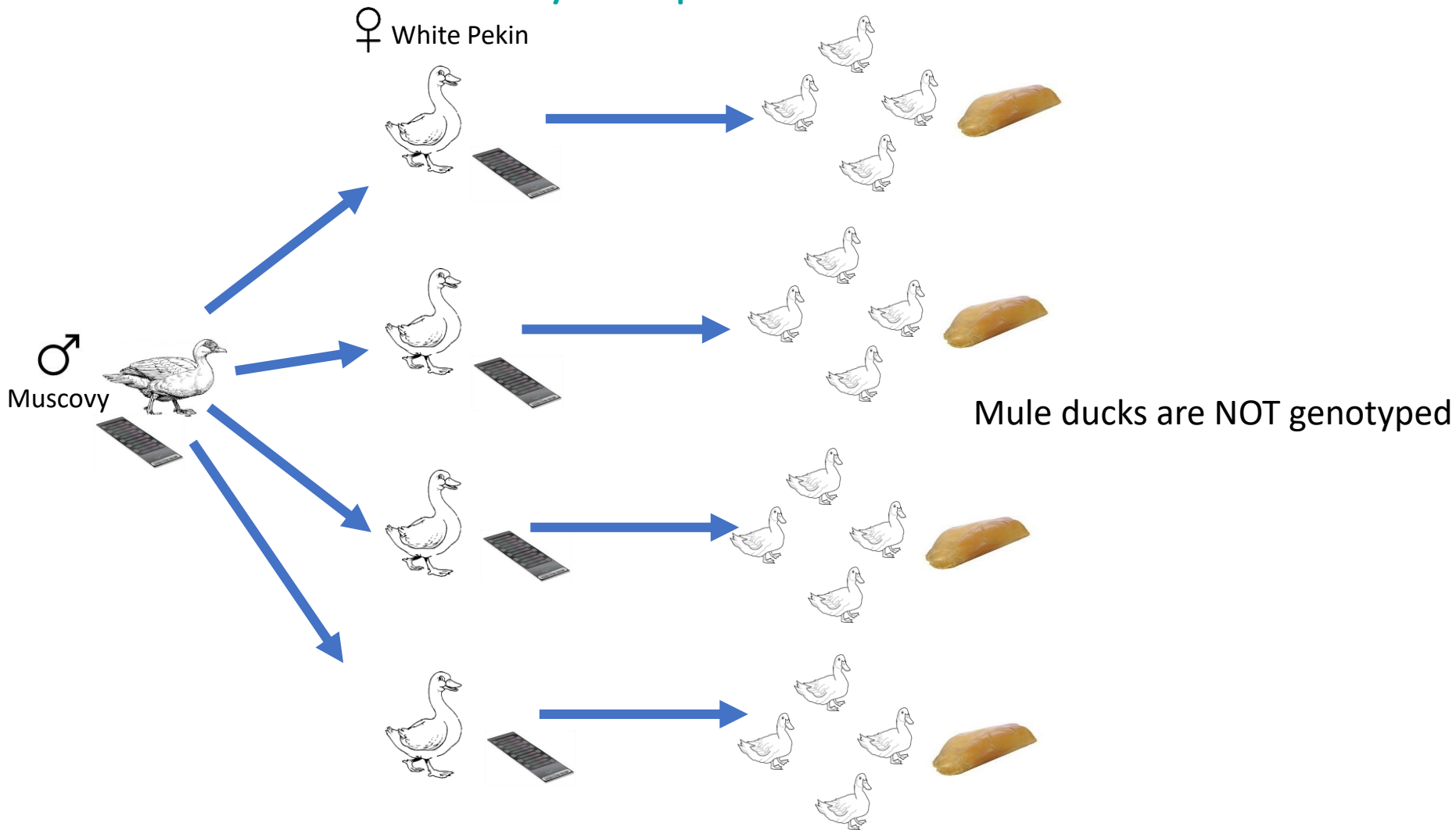
This will improve imputation results. And help maintain a connection between reference population and candidate population.

TIP2 : Discard genotypes of unselected animals without phenotypes. Typically unselected male candidates in layers. They increase computational cost without carrying any useful information.



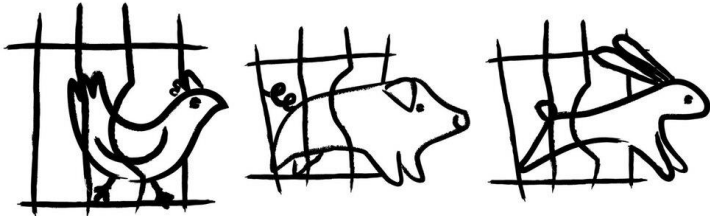
➤ Consequences of the genomic selection

Ducks selected for fatty liver production



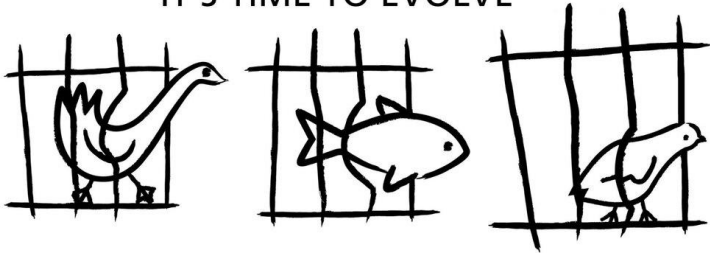
➤ Consequences of the genomic selection

Ducks selected for fatty liver production (in Europe)



END THE CAGE AGE

IT'S TIME TO EVOLVE



In Europe cages are likely to be banned very shortly

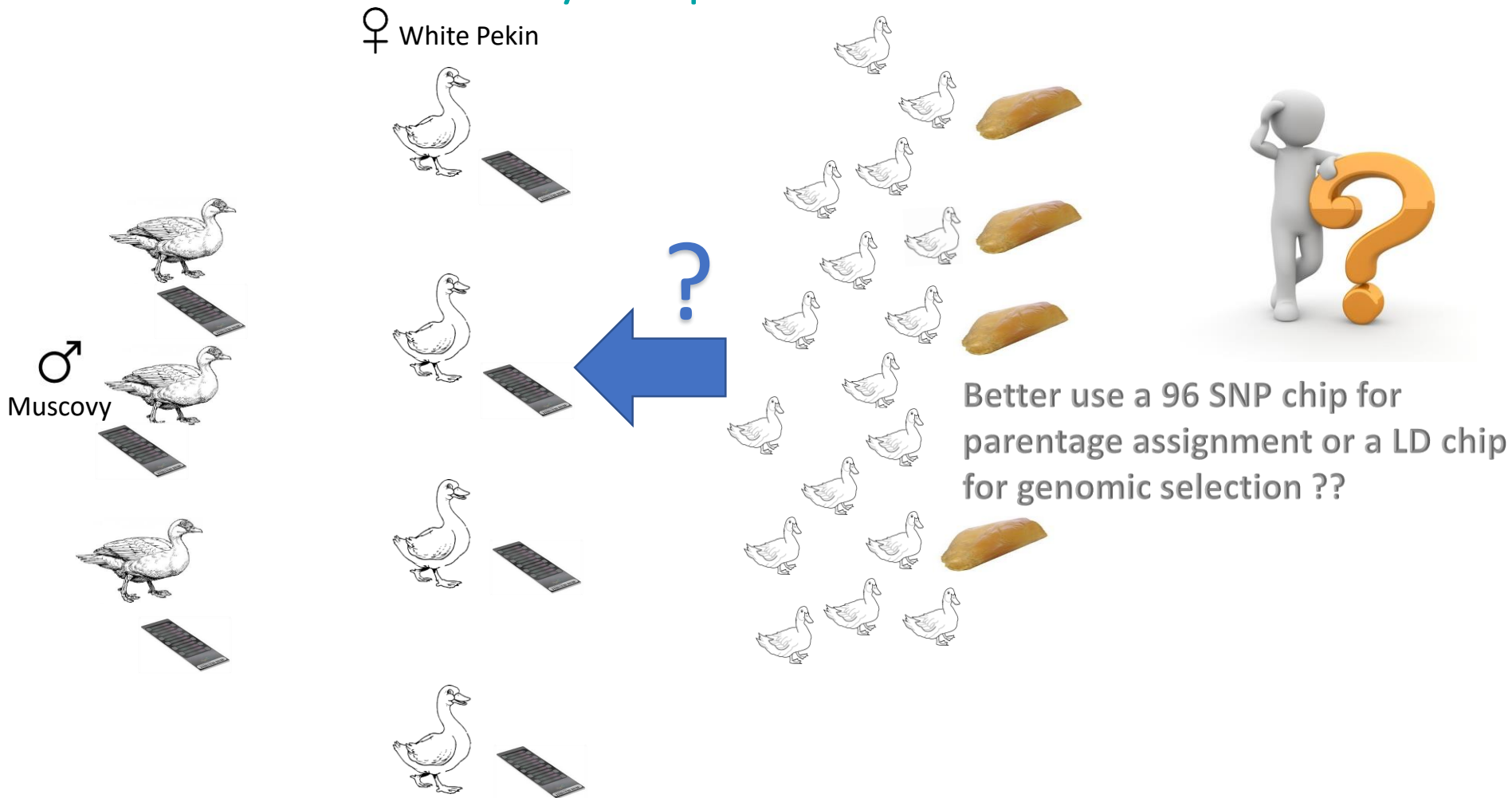
Need to rethink the way birds are bred in the poultry selection schemes.

One solution could be to house females on floor (large pens) and inseminate* them with pooled semen of several drakes.

➔ Offspring pedigree is obtained through parentage assignment using molecular markers.

➤ Consequences of the genomic selection

Ducks selected for fatty liver production



➤ For those who want more on the subject...





Journal of Animal Science, 2020, Vol. 98, No. 4, 1–14

doi:10.1093/jas/skaa101
Advance Access publication April 8, 2020
Received: 29 January 2020 and Accepted: 7 April 2020
Board Invited Review

Review

Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90

Daniela Lourenco ^{1,*}, Andres Legarra ², Shogo Tsuruta ¹ , Yutaka Masuda ¹, Ignacio Aguilar ³  and Ignacy Misztal ¹

¹ Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA; shogo@uga.edu (S.T.); yutaka@uga.edu (Y.M.); ignacy@uga.edu (I.M.)

² Institut National de la Recherche Agronomique, UMR1388 GenPhySE, 31326 Castanet Tolosan, France; andres.legarra@inra.fr

³ Instituto Nacional de Investigación Agropecuaria (INIA), 11500 Montevideo, Uruguay; iaguilar@inia.org.uy

* Correspondence: danilino@uga.edu

Received: 19 June 2020; Accepted: 6 July 2020; Published: 14 July 2020



BOARD INVITED REVIEW

Current status of genomic evaluation

Ignacy Misztal,^{†,1} Daniela Lourenco,[†] and Andres Legarra[‡]

[†]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, [‡]Department of Animal Genetics, Institut National de la Recherche Agronomique, Castanet-Tolosan, France

[†]Corresponding author: ignacy@uga.edu

ORCID number: 0000-0002-0382-1897 (I. Misztal).

Abstract

Early application of genomic selection relied on SNP estimation with phenotypes or de-regressed proofs (DRP). Chips of 50k SNP seemed sufficient for an accurate estimation of SNP effects. Genomic estimated breeding values (GEBV) were composed of an index with parent average, direct genomic value, and deduction of a parental index to eliminate double counting. Use of SNP selection or weighting increased accuracy with small data sets but had minimal to no impact with large data sets. Efforts to include potentially causative SNP derived from sequence data or high-density chips showed limited or no gain in accuracy. After the implementation of genomic selection, EBV by BLUP became biased because of genomic preselection and DRP computed based on EBV required adjustments, and the creation of DRP for females is hard and subject to double counting. Genomic selection was greatly simplified by single-step genomic BLUP (ssGBLUP). This method based on combining genomic and pedigree relationships automatically creates an index with all sources of information, can use any combination of male and female genotypes, and accounts for preselection. To avoid biases, especially under strong selection, ssGBLUP requires that pedigree and genomic relationships are compatible. Because the inversion of the genomic relationship matrix (G) becomes costly with more than 100k genotyped animals, large data computations in ssGBLUP were solved by exploiting limited dimensionality of genomic data due to limited effective population size. With such dimensionality ranging from 4k in chickens to about 15k in cattle, the inverse of G can be created directly (e.g., by the algorithm for proven and young) at a linear cost. Due to its simplicity and accuracy, ssGBLUP is routinely used for genomic selection by the major chicken, pig, and beef industries. Single step can be used to derive SNP effects for indirect prediction and for genome-wide association studies, including computations of the P -values. Alternative single-step formulations exist that use SNP effects for genotyped or for all animals. Although genomics is the new standard in breeding and genetics, there are still some problems that need to be solved. This involves new validation procedures that are unaffected by selection, parameter estimation that accounts for all the genomic data used in selection, and strategies to address reduction in genetic variances after genomic selection was implemented.

Key words: genomic evaluation, genomic selection, large data, single-step GBLUP



INRAE

Hervé CHAPUIS 08/11/2022



Special thanks to Vincent Ducrocq, Celine Carillier-Jacquín and Helene Larroque (INRAE) for their contribution to these slides.