Genomic evaluation

Vincent Ducrocq

GABI, Jouy-en-Josas

INRA
SCIENCE & IMPACT

Whole Genome Sequence

- After Human (2001) and mice, WGS of the chicken (2004), the dog (2005), bovine (2006), horse (2007), pig (2009), ...
- Entirely in the public domain
- Sequencing of different individuals => polymorphisms
  - 3.2 million bovine polymorphisms in dbSNP
  - >30 millions known today
- New technologies for genotyping and sequencing

SNP : Single Nucleotide Polymorphism

DNA Variation of one base

..GAATCTTATGCTATACATAATTATATACTAATCGGGTATTGTTCTTAT.

.

..GAATCTTATGCTATACATAATTATATACTAATAGGGTATTGTTCTTAT..

SNP

Genotyping chips

- Simultaneous genotyping of many SNP
- From few dozens up to several million SNP
- Two main technology providers, Illumina and Affymetrix
- Illumina products in cattle
  - 3000 (□7000□ 1000□ 20000=« LD »)
  - **54 000=« 50k »**
  - 777 000=« HD »

FIGURE 1: BOVINESNP50 BEADCHIP

The BovineSNP50 BeadChip features more than 54,000 evenly-spaced SNPs across the entire bovine genome.

Chips of different densities: IMPUTATION

- i_p_ta_io_ c_nsi_t_ i_ pr_di_t_n_ t_e m__s__g l_t_er_ _i_h__ w__d o_ a s__t__c_

+ a dictionary

- imputation consists in predicting the missing letters within a word or a sentence

7K

T				T				G					C				A			T			C
T				T				C					A				C			T			A

+ an imputation population

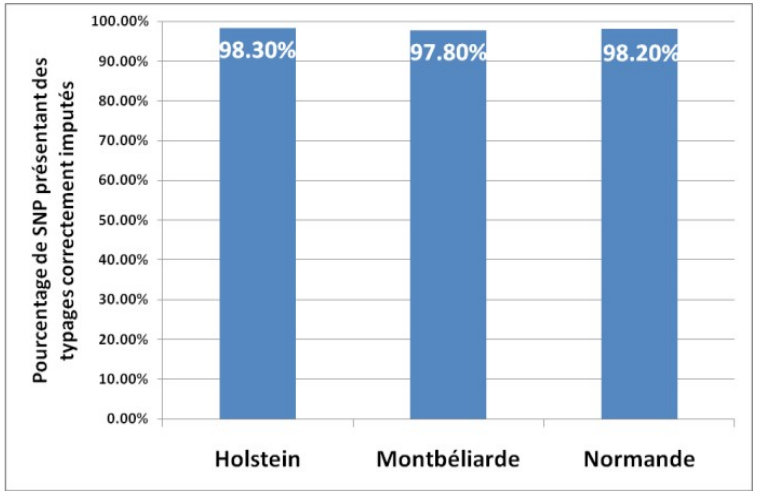
54K

T	C	A	G	T	T	G	T	G	A	A	T	G	A	C	T	G	G	A	C	G	T	G	C	C
T	A	A	G	T	G	G	T	C	A	C	T	G	A	A	T	G	G	C	C	G	T	T	C	A

Works well ! (~1% error in Holstein if sire genotyped with 54k)

Imputation

Beagle/DagPhase



(Dassonneville et al., 2011)

Fimpute

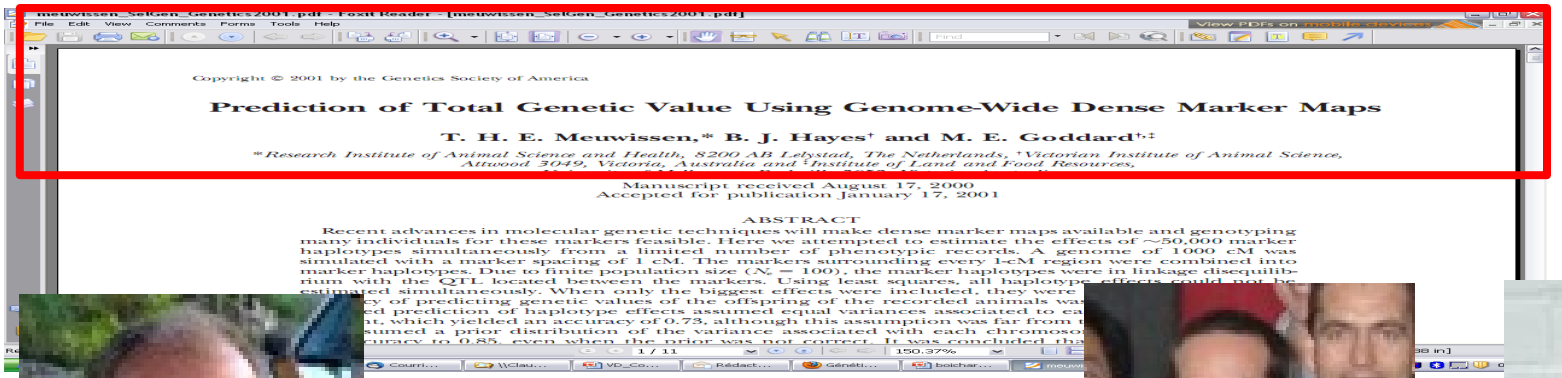
(Saintilan et al., 2014)

Computation time divided by 3 – 200-300 SNP/animal corrected

Genomic selection

selection based on the estimation of the genetic value of candidates using information on dense markers covering the whole genome

= selection based on results from a genomic evaluation



Copyright © 2001 by the Genetics Society of America

Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,* B. J. Hayes† and M. E. Goddard†,‡

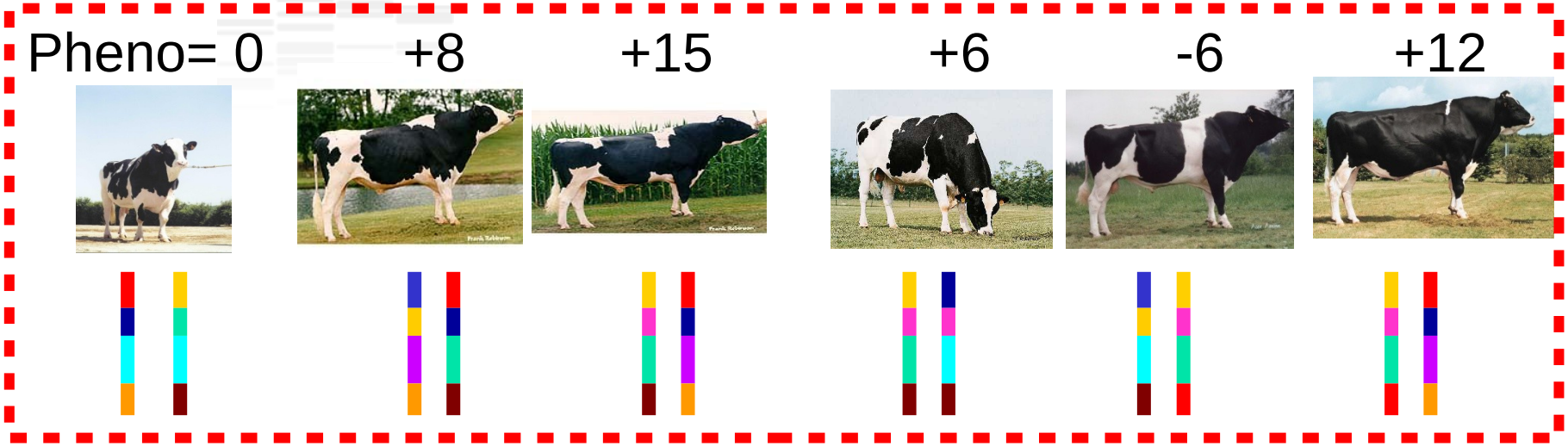
*Research Institute of Animal Science and Health, 8200 AB Lelystad, The Netherlands, †Victorian Institute of Animal Science, Attwood 3049, Victoria, Australia and ‡Institute of Land and Food Resources,

Manuscript received August 17, 2000
Accepted for publication January 17, 2001

ABSTRACT

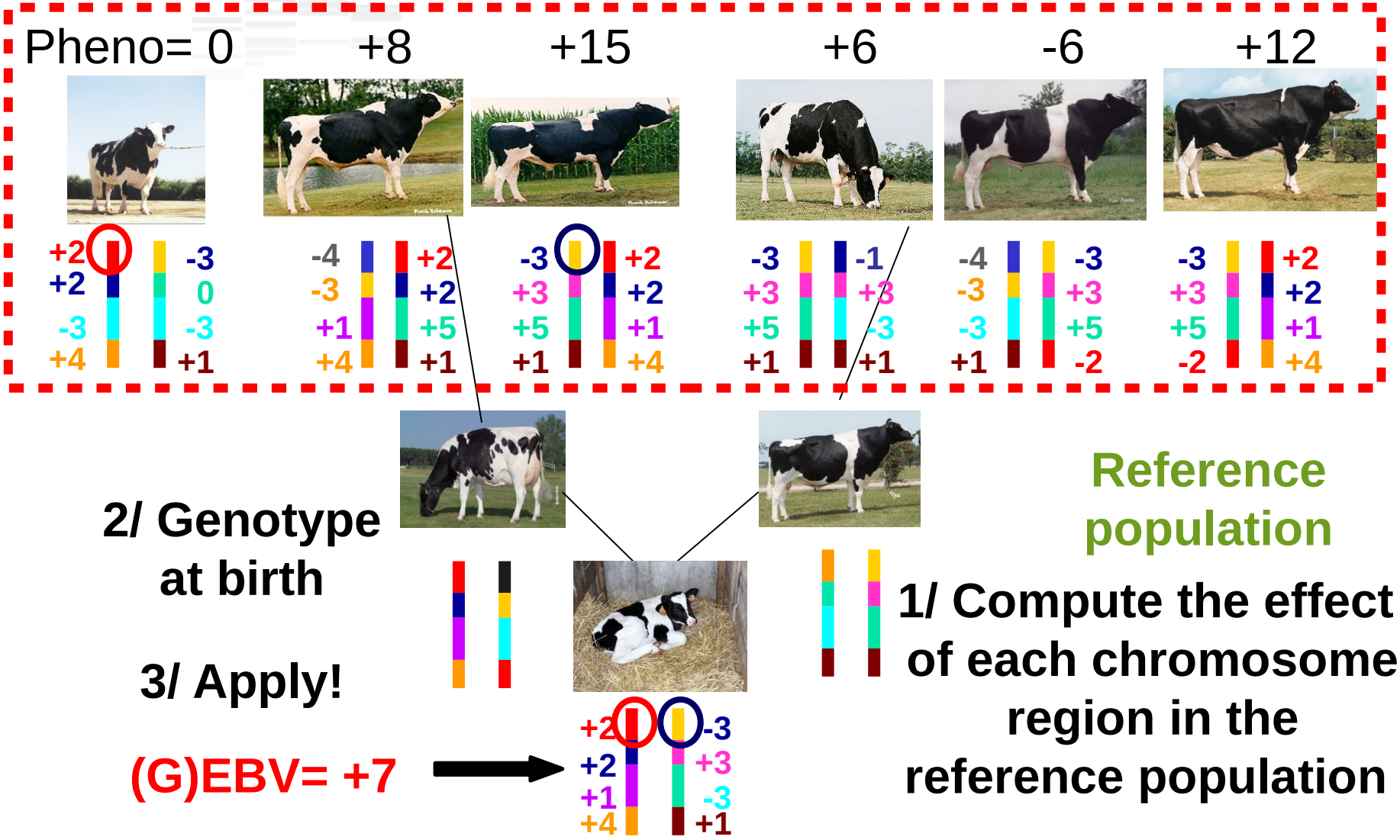
Recent advances in molecular genetic techniques will make dense marker maps available and genotyping many individuals for these markers feasible. Here we attempted to estimate the effects of ~50,000 marker haplotypes simultaneously from a limited number of phenotypic records. A genome of 1000 cM was simulated with a marker spacing of 1 cM. The markers surrounding every 1-cM region were combined into marker haplotypes. Due to finite population size ($N_e = 100$), the marker haplotypes were in linkage disequilibrium with the QTL located between the markers. Using least squares, all haplotype effects could not be estimated simultaneously. When only the biggest effects were included, they were

What is genomic selection?

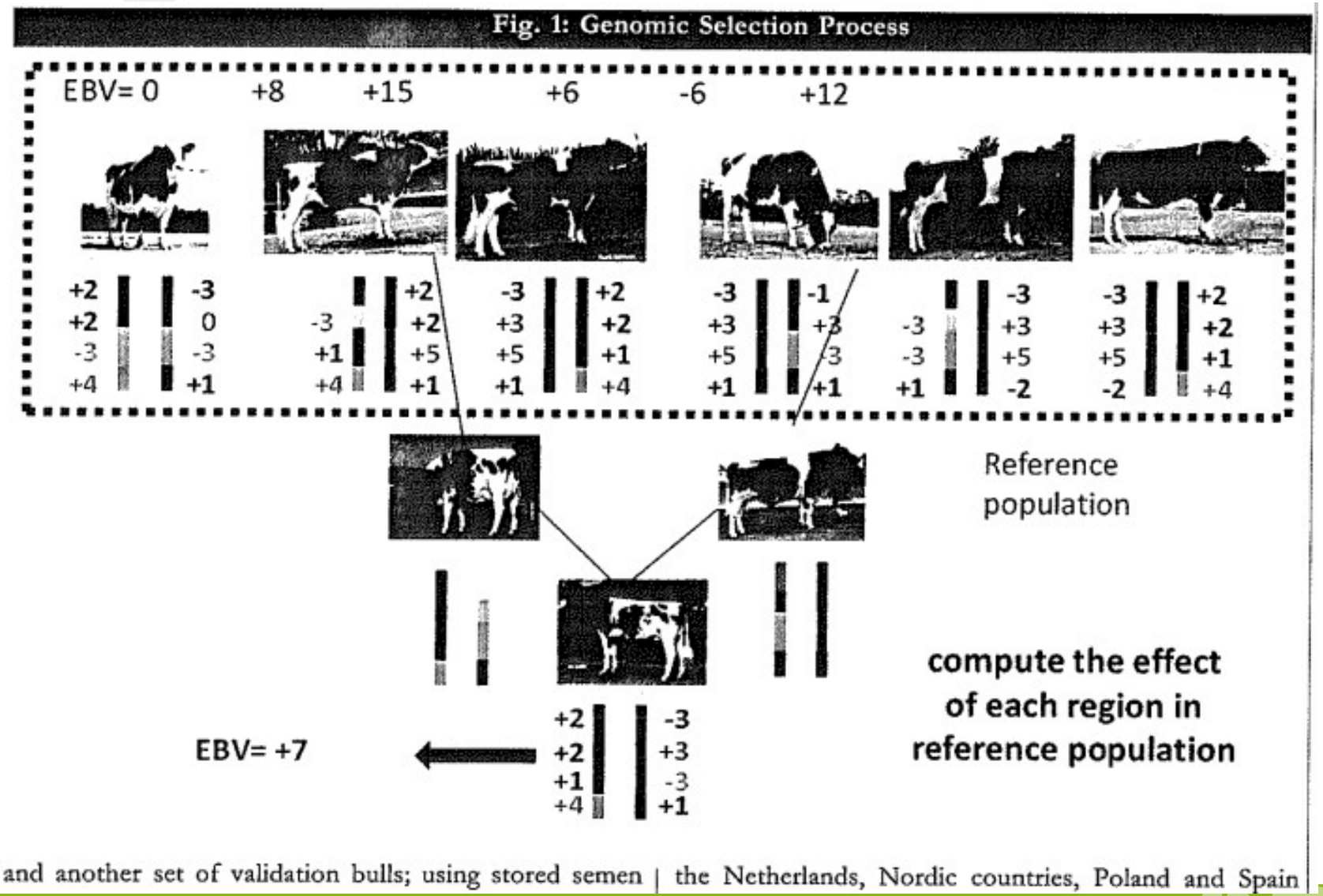


Genotyped and phenotyped animals ⟷ **Reference population**

What is genomic selection?

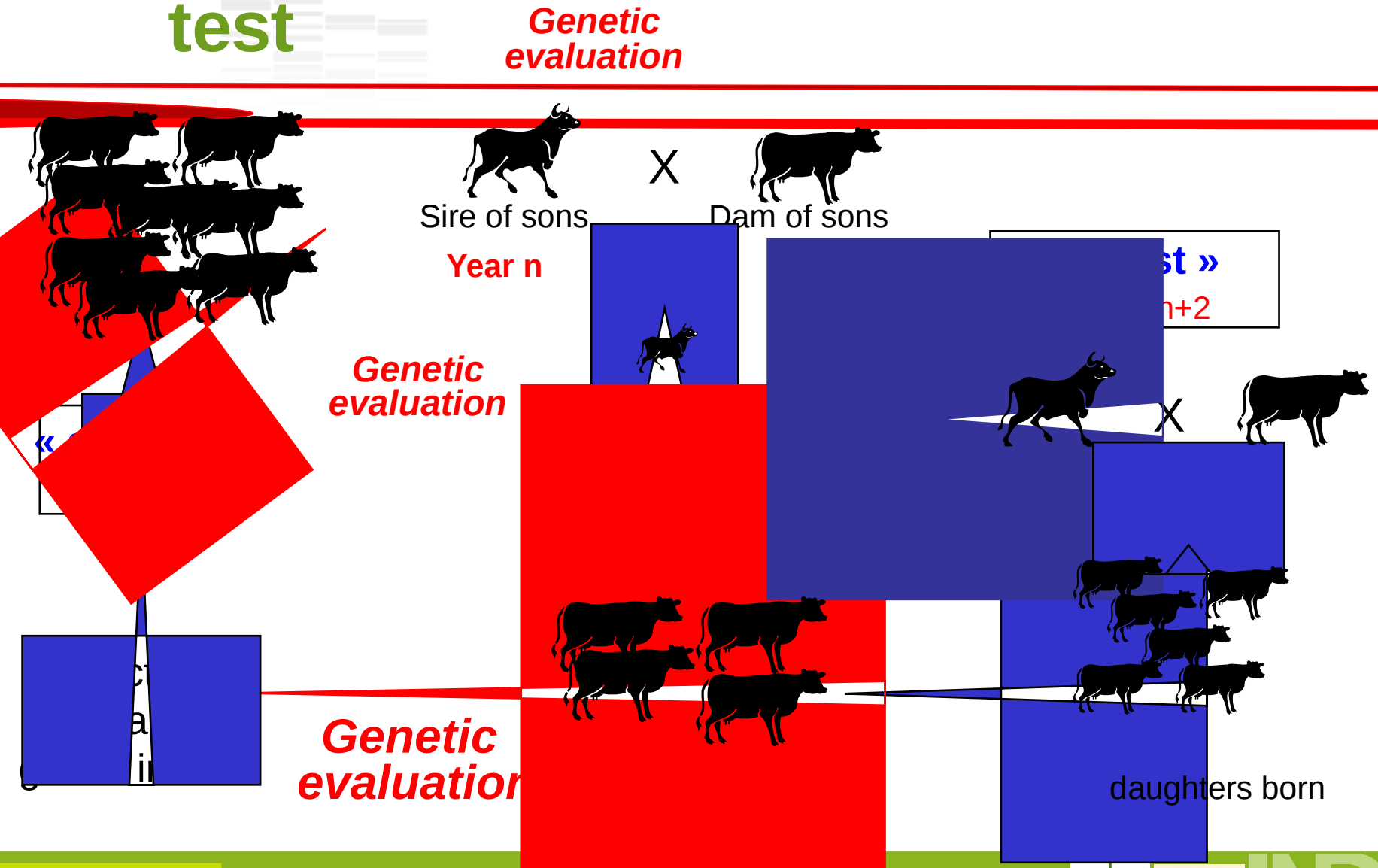


Digression: don't do this !

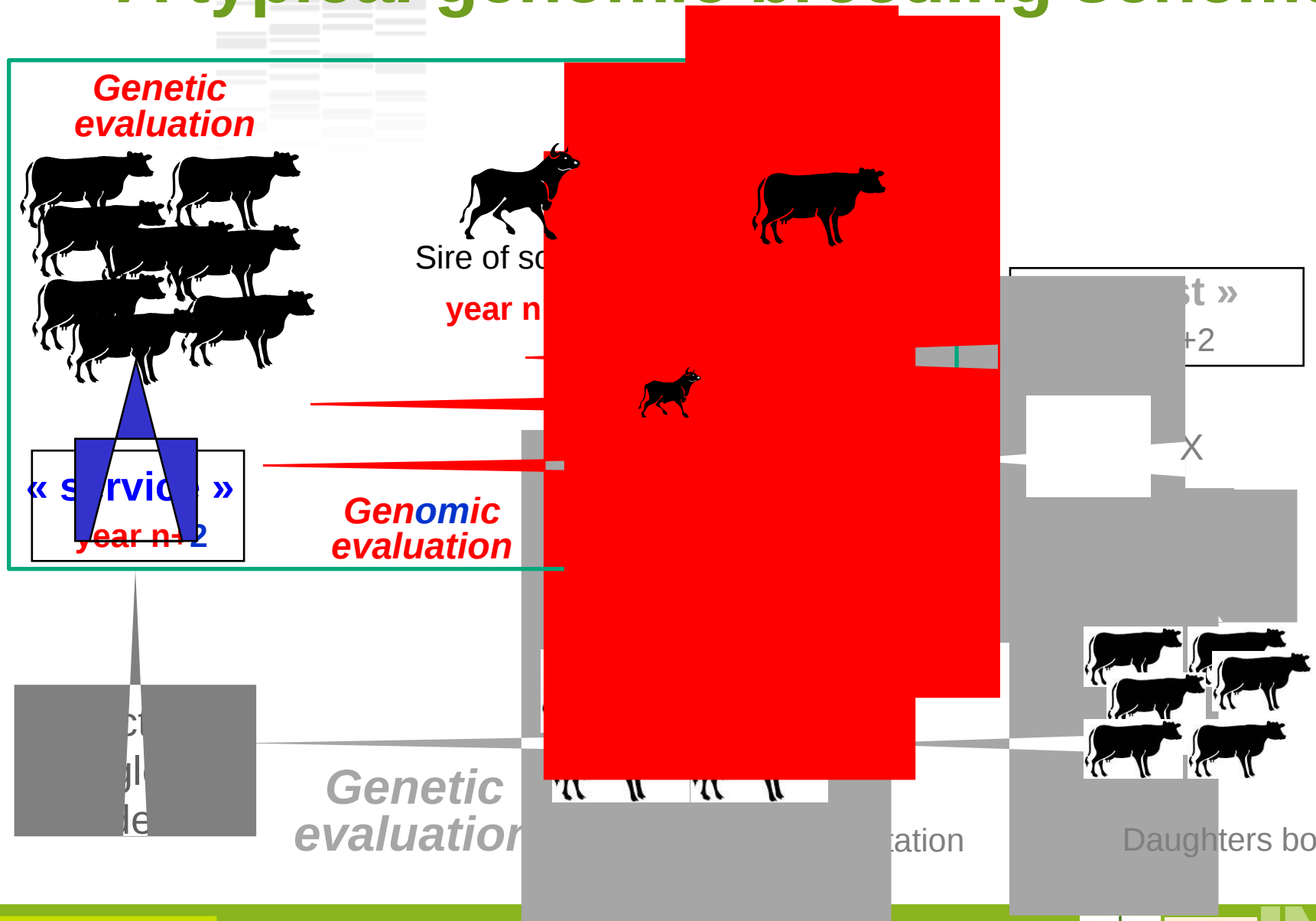


and another set of validation bulls; using stored semen | the Netherlands, Nordic countries, Poland and Spain

Remember: “Classical” progeny test



A typical genomic breeding scheme



In practice …

- Genomic selection
  - Requires **phenotypes** and **genotypes**
  - To minimize costs, genotype bulls in countries with large existing progeny test programs
  - But bulls don't have phenotypes (of interest)

➔ summarize daughters' phenotypes at bull level

✓ *Advantage* : more precise phenotypes

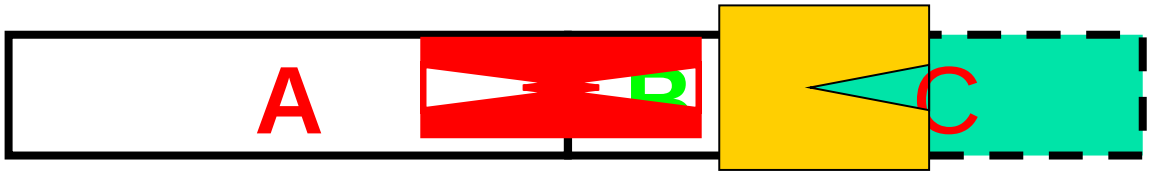
Basic principle for implementation

A reference population A+B : several thousands of genotyped animals with phenotype / performances

- From population A (training population), find the relationship between the phenotypes and the markers (=a prediction equation)
- purely statistical approach (no/very few genetic assumptions)

Basic principle for implementation

- then validate this prediction equation in population B (validation population)
- If satisfactory, redo the analysis on A+B and apply the resulting prediction equation to obtain a « genomic breeding value » for selection candidates C, for example at birth

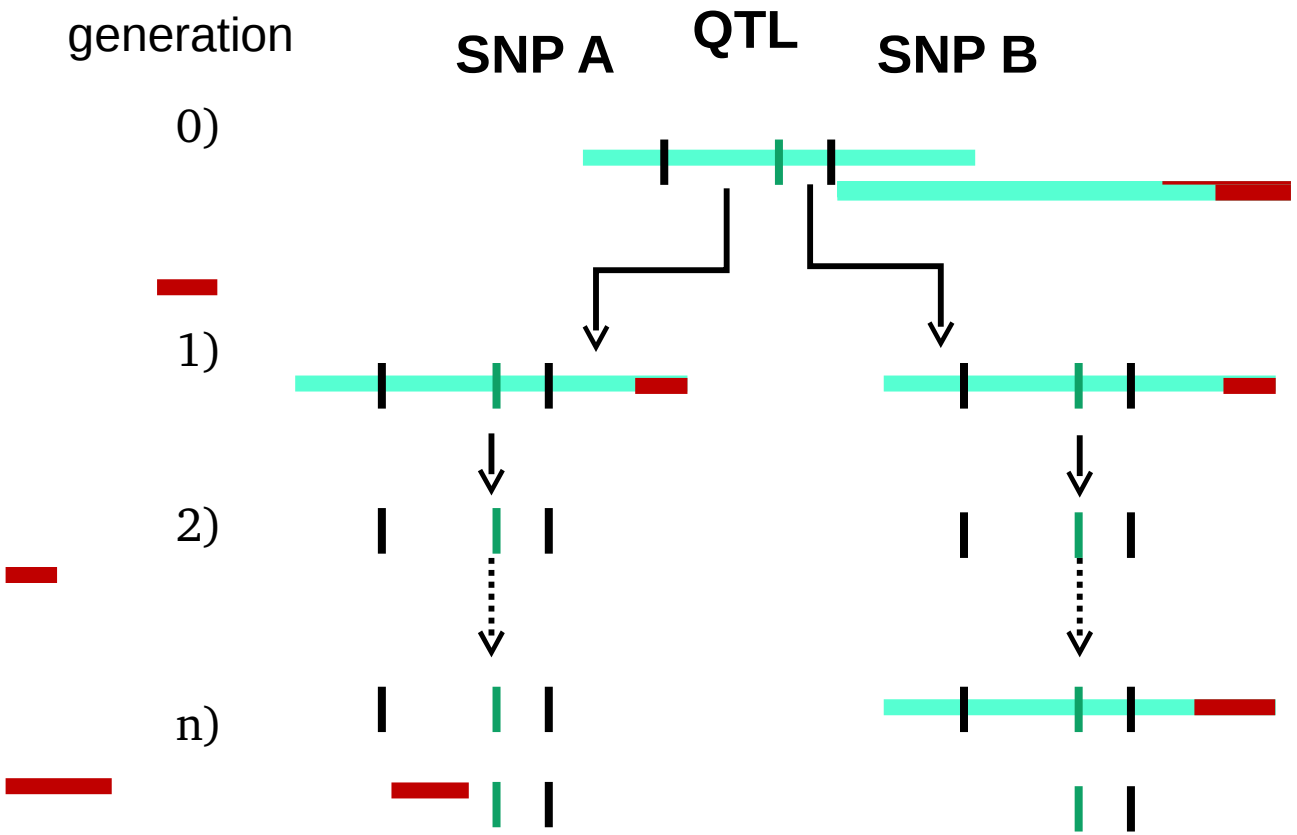


Genomic Selection: efficiency

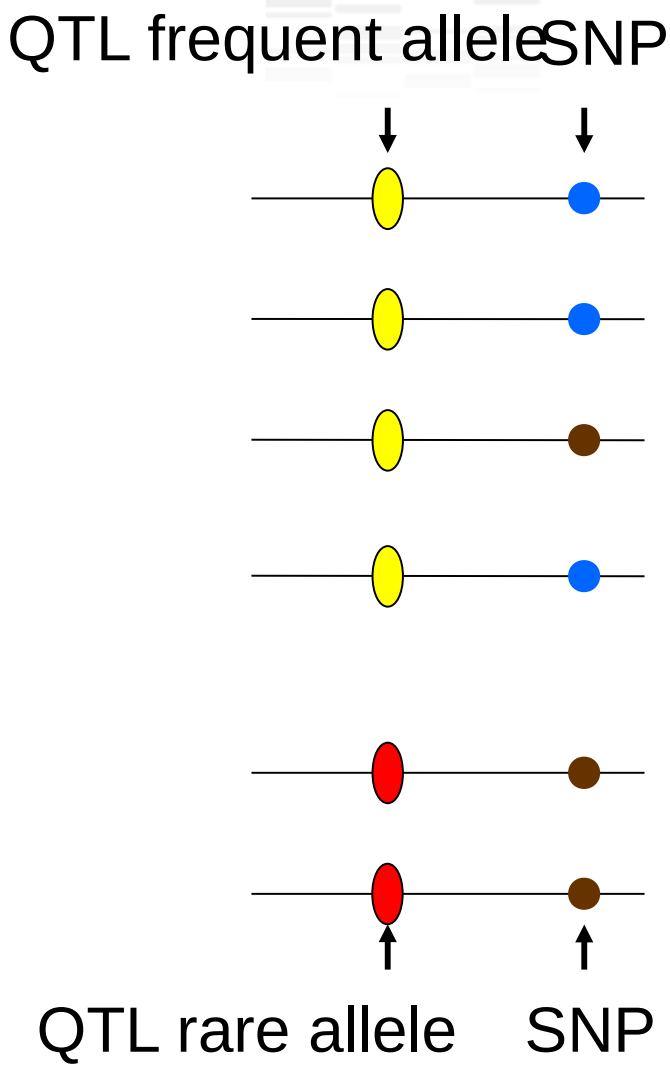
- Two main factors :
  - Accuracy of SNP effect estimation
    - size of reference population
    - heritability of the trait
    - statistical methodology used
  - Linkage Disequilibrium (LD) between markers and QTL
    - marker density
    - effective size of the population => number of « independent » segments
    - Relationship between candidates and reference population

Linkage Disequilibrium

QTL-SNP association transmitted over generations



Linkage Disequilibrium



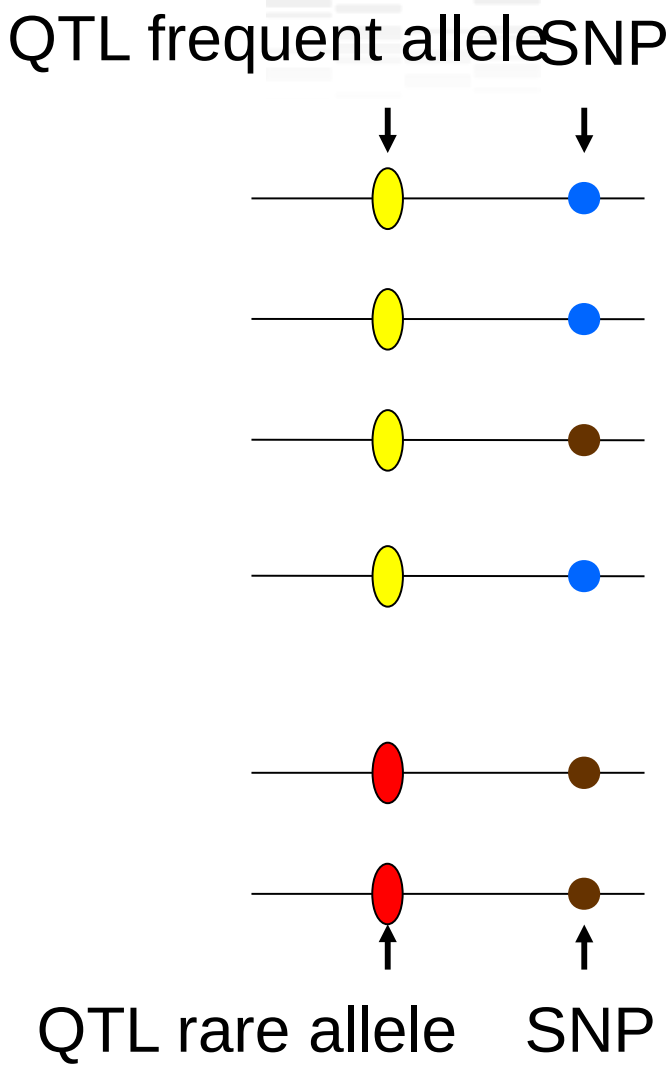
SNP ● more often associated with QTL allele

SNP ● more often associated with QTL allele

use effect of ● as proxy of effect of

and effect of ● as proxy of effect of

Linkage Disequilibrium



SNP ● more often associated with QTL allele

SNP ● more often associated with QTL allele

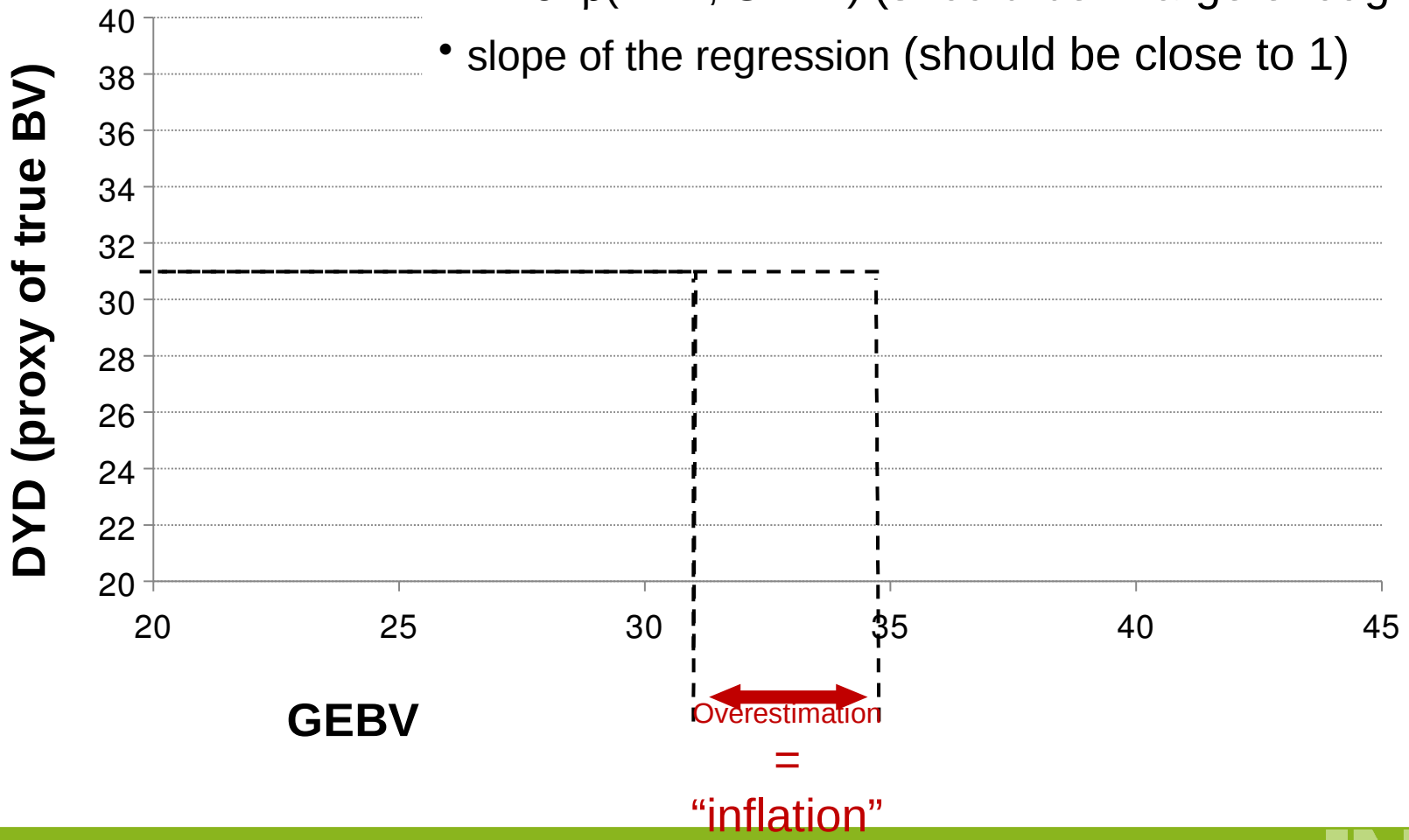
use effect of ● as proxy of effect of

and effect of ● as proxy of effect of

Validation test

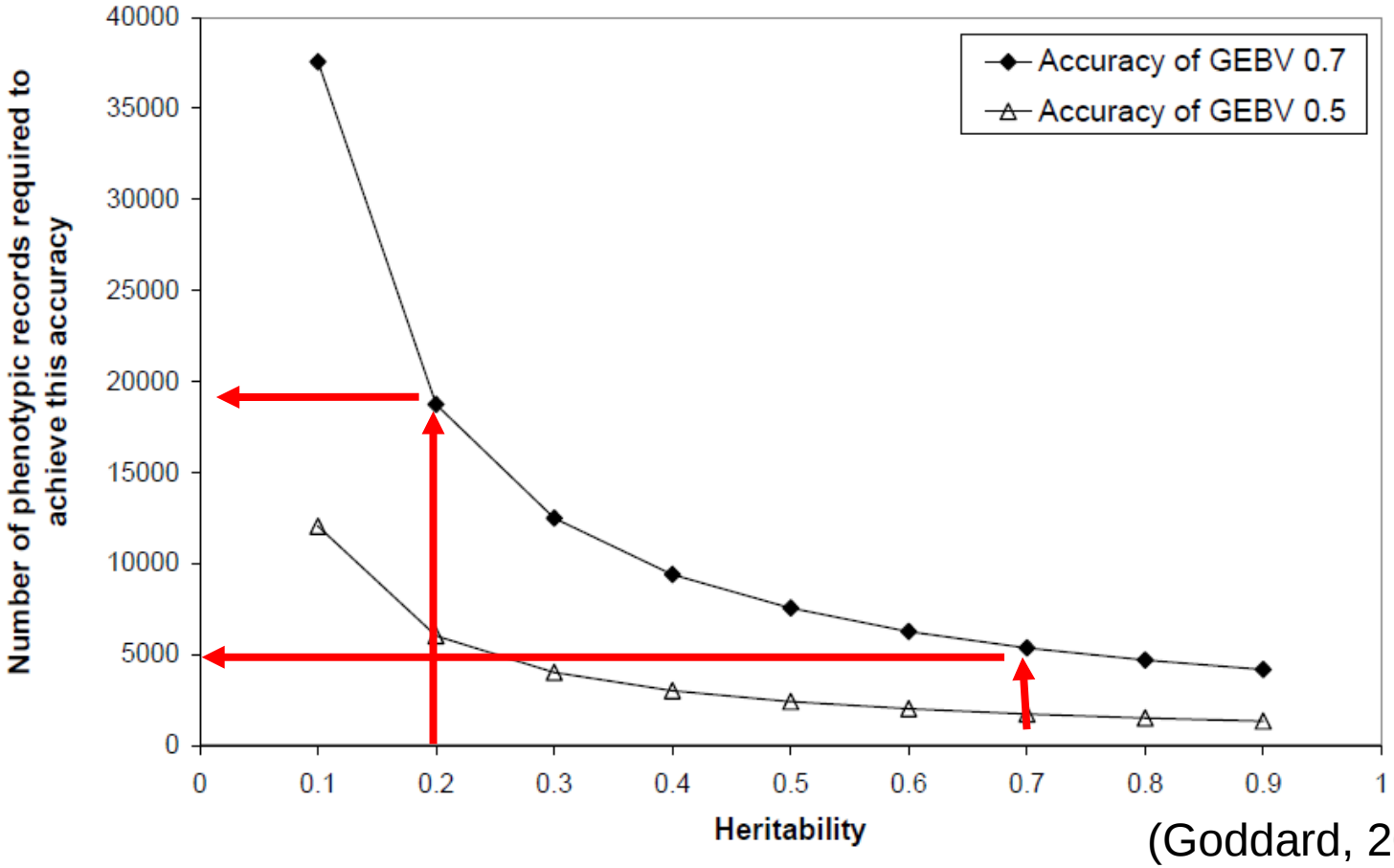
Two important parameters :

- R2 or ρ(DYD, GEBV) (should be « large enough »)
- slope of the regression (should be close to 1)



Size of the reference population

- Greatly influences the accuracy of genomic evaluations



(Goddard, 2008)

Increase size of reference populations

- Genotype as many progeny tested bulls as possible
- International collaborations
  - Holstein: 2 big consortia
    - **USA + Canada** ~>25000 bulls (?) + Italy + GB
    - **Eurogenomics** France, The Netherlands, (Germany), Nordic countries, Spain,Poland ~34000 buls

  - For small breeds: not enough AI bulls
    - combine with many genotyped cows

Information from bulls and cows

Number of cows with 1 own record
equivalent to a bull with performances on daughters

	h2				
CD*	0.1	0.2	0.3	0.4	0.5
0.40	6.0	2.7	1.6	1.0	0.7
0.50	9.0	4.0	2.3	1.5	1.0
0.60	13.5	6.0	3.5	2.3	1.5
0.70	21.0	9.3	5.4	3.5	2.3
0.80	36.0	16.0	9.3	6.0	4.0
0.90	81.0	36.0	21.0	13.5	9.0

* CD : coefficent of determination
= reliability of EBV based on progeny performance only (Boichard, 2015)

Basic data for genomic evaluation

- **y** = YD (Yield deviation) = Individual record corrected for all fixed and non genetic random effects

or

- **y** = DYD (Daughter yield deviation) = (2 x) average for a bull of all YD of their daughters corrected for ½ genetic effect of their dam (with associated weight = EDC (Equivalent Daughter Contribution)

or

· **y** = deregressed proofs (= DYD reconstructed from EBV, EDC and relationships: a genetic evaluation based on deregressed proofs will give back the bulls' EBV))

or

· ~~**y = EBV** (~« proxy » of DYD, valid only if reliabilities are high)~~

Different genomic evaluation methods

Can be grouped into 4 families:

1. Derived from BLUP: GBLUP

2. General multidimension methods when the number of unknowns (p) is much larger than the number of observations (n) : the « p >>n problem » (PLS, LASSO, Elastic Net…)

3. Bayesian methods A, B, C, Cπ, D, R …

4. ~~« Machine learning »~~ methods (big black box)

To be applied on individual SNP or haplotypes

GBLUP: Basic « genomic » model

Hypotheses :

1) QTLs (quantitative Trait loci) have a purely additive effect
2) All the variability at QTLs is explained the markers (!)

N biallelic markers

2 alleles A/B but only one effect defined = substitution effect mi

	xi mi =		xi mi =
	wi= (xi – 2pi) mi		
example: AA	-1mi	*or* 0	*or* – 2pi mi
AB	0	+mi	(1 – 2pi) mi
BB	+1mi	+2mi	(2 – 2pi) mi

where pi is the allelic frequency of allele B of marker i

Distribution of marker effects

$$g = BV = \sum_{i=1}^{N} x_i m_i \qquad \text{GEBV} = \sum_{i=1}^{N} x_i \hat{m}_i$$

With the third form, and assuming that each marker contributes the same way to the variance σ_g^2 of the trait, we have (Van Raden, 2009):

σ_m^2

mi~ N (0,) with $\sigma_m^2 = \dfrac{\sigma_g^2}{2 \sum_j p_j (1 - p_j)}$

Why this formula?: for consistency with the additive genetic variance. But what allelic frequency should be used? Quite a few were proposed (observed ones/ in the base population..)

Evaluation model = BLUP on markers

If all animals with records are genotyped, and if we take allelic frequencies into account by centering xi:

wi = xi – 2 pi (grouped in a matrix **W**) , we can write:

The corresponding Mixed Model Equations are:

$$\begin{bmatrix} X'X & X'W \\ W'X & W'W + \frac{\sigma_e^2}{\sigma_m^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ W'y \end{bmatrix}$$

For any new animal * without performance: $GEBV = \sum_{i=1}^{N} w_i^* \hat{m}_i$

An equivalent model

- We can also write: **g = Wm**

But, since we have mi~ N(0, σ^2_m), we also have:

g ~N(**0**, σ^2_g **G**)

with **G** = var(**Wm**) = **W** var(**m**) **W'**

= **WW'**/k

where k is a standardisation factor connecting the marker effects variance to the additive genetic variance

$$k = 2 \sum_j p_j \left(1 - p_j\right)$$

An equivalent model

- **G** plays the same role as the classical relationship matrix **A**

  G is called « genomic relationship matrix »

- So, reparameterizing the model

  if all recorded animals are genotyped:

or more generally (adding columns of 0 to **Z** if no phenotype):

GBLUP

The Mixed Model Equations are the same as in the classical case except that the relationship matrix **A** (=« expectation » of relationship coefficients) is replaced by the genomic relationship matrix **G:**

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \frac{\sigma_e^2}{\sigma_g^2} G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

(some) limits of GBLUP

- Assuming that **G is** correct, how to compute it efficiently?

  G= **WW'**/k **W** is dense!

  Very long to compute, can be too big to store and to invert

  → iterative solutions

- How to compute reliabilities → approximations
- A marker model seems more attractive…. (at least, its complexity does not increase too much when the number of genotyped animals increases!)

(some) limits of GBLUP

- It is not realistic to think that all the genetic variation at QTLs is captured by markers. This leads to genomic values which show «inflation » □ requires an additional « residual » polygenic effect:

Depending on the trait, ω = 5 to 30%

(some) limits of GBLUP

- How to combine genomic and classical (e.g., phenotypes from progeny) information?

Van Raden's (2009) approach:

1/ run a classical genetic evaluation with all animals

2/ run a genomic evaluation with all genotyped animals

3/ rerun the same (classical) evaluation with genotyped animals

…(note: BLUP hypotheses are not fulfilled!)

4/ combine the 3 sources of information (more or less: 1 + 2 -3)

with selection index theory, for each animal

- How to make non genotyped animals benefit from genomic information of relatives?

An example

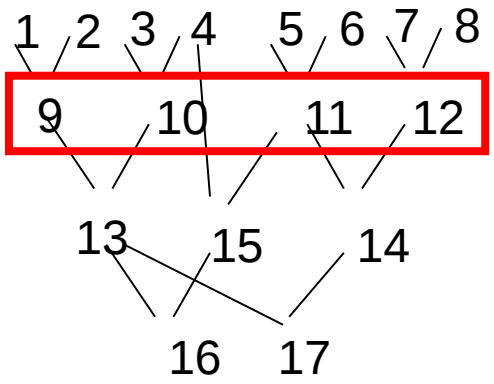
```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17]
 [1,] 1.00                                                0.50                 0.25                 0.12  0.12
 [2,]      1.00                                           0.50                 0.25                 0.12  0.12
 [3,]           1.00                                            0.50           0.25                 0.12  0.12
 [4,]                1.00                                       0.50           0.25        0.50     0.38  0.12
 [5,]                     1.00                                            0.50           0.25 0.25 0.12  0.12
 [6,]                          1.00                               0.50           0.25 0.25 0.12  0.12
 [7,]                               1.00                                    0.50        0.25           0.12
 [8,]                                    1.00                               0.50        0.25           0.12
 [9,] 0.50 0.50                                           1.00                 0.50                 0.25  0.25
[10,]           0.50 0.50                                       1.00           0.50        0.25     0.38  0.25
[11,]                     0.50 0.50                                   1.00           0.50 0.50 0.25  0.25
[12,]                               0.50 0.50                               1.00        0.50           0.25
[13,] 0.25 0.25 0.25 0.25                                0.50 0.50                 1.00        0.12  0.56  0.50
[14,]                     0.25 0.25 0.25 0.25                         0.50 0.50        1.00  0.25  0.12  0.50
[15,]                0.50 0.25 0.25                           0.25           0.50        0.12  0.25  1.00  0.56  0.19
[16,] 0.12 0.12 0.12 0.38 0.12 0.12                0.25 0.38           0.25           0.56  0.12  0.56  1.06  0.34
[17,] 0.12 0.12 0.12 0.12 0.12 0.12 0.12 0.12 0.25 0.25           0.25 0.25 0.50 0.50  0.19  0.34  1.00
>
```

A

?

?

G



Transmission of information (to non genotyped anima

Let **g2** be the genetic (genomic) values of genotyped animals and **g1** the genetic values of non genotyped animals

Let $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$ $\begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix}^{-1}$ and $p(\mathbf{g}_2) \sim N(\mathbf{0}, \mathbf{G}\sigma^2_{\mathbf{g}})$

By regression of **g1** on **g2:**

$$p(\mathbf{g}_1 | \mathbf{g}_2) \sim N(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{g}_2, \mathbf{V}\sigma^2_{\mathbf{g}})$$

with $\mathbf{V} = (\mathbf{A}^{11})^{-1} = \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$

Transmission of information (to non genotyped anima

Let **g2** be the genetic (genomic) values of genotyped animals and

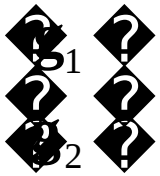
g1 the genetic values of non genotyped animals

An estimate of **g1** based on genomic information is obtained by regression of **g1** on **g2**

This information is added to regular BLUP equations

«Single-Step» approach

Legarra, Misztal, Aguilar; Christensen & Lund; D. Johnson

- **H =** Variance of 

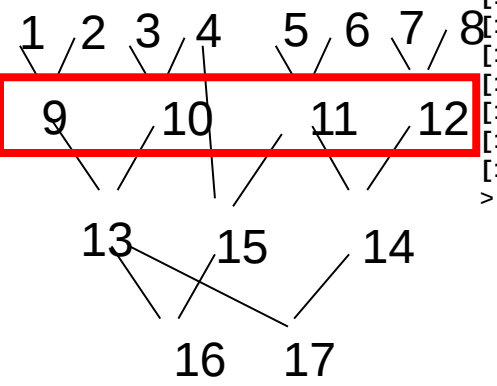
«Single-step» approach

$$\mathbf{H}=\begin{matrix} & \textit{non genotyped} & \textit{genotyped} \\ \\ \\ \end{matrix}\begin{matrix} \\ \\ \end{matrix}\begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix}=\begin{bmatrix} \mathbf{A}_{11}+\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G}-\mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}$$

- Incredible but true ! **H-1** has a relatively simple form:

$$\mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1} \end{bmatrix}$$

Back to the example




```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17]
 [1,] 1.00                                            0.50                  0.25                 0.12  0.12
 [2,]      1.00                                       0.50                  0.25                 0.12  0.12
 [3,]           1.00                                        0.50             0.25                 0.12  0.12
 [4,]                1.00                                   0.50             0.25        0.50     0.38  0.12
 [5,]                     1.00                                    0.50              0.25 0.25     0.12  0.12
 [6,]                          1.00                               0.50              0.25 0.25     0.12  0.12
 [7,]                               1.00                                    0.50        0.25                0.12
 [8,]                                    1.00                               0.50        0.25                0.12
 [9,] 0.50 0.50                                       1.00                        0.50                 0.25  0.25
[10,]           0.50 0.50                                   1.00                  0.50        0.25     0.38  0.25
[11,]                     0.50 0.50                                   1.00              0.50 0.50     0.25  0.25
[12,]                               0.50 0.50                                   1.00        0.50                0.25
[13,] 0.25 0.25 0.25 0.25                                  0.50  0.50                  1.00        0.12  0.56  0.50
[14,]                     0.25 0.25 0.25 0.25                        0.50  0.50              1.00 0.25  0.12  0.50
[15,]                0.50 0.25 0.25                              0.25  0.50        0.12  0.25 1.00 0.56  0.19
[16,] 0.12 0.12 0.12 0.38 0.12 0.12                        0.25  0.38  0.25        0.56  0.12 0.56 1.06  0.34
[17,] 0.12 0.12 0.12 0.12 0.12 0.12 0.12 0.12 0.25  0.25  0.25  0.25  0.50  0.50  0.19 0.34 1.00
>
```

A

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17]
 [1,] 1.00      0.17 0.17 0.17 0.17 0.17 0.17 0.50  0.35  0.35  0.35  0.42  0.35  0.26  0.34  0.39
 [2,]      1.00 0.17 0.17 0.17 0.17 0.17 0.17 0.50  0.35  0.35  0.35  0.42  0.35  0.26  0.34  0.39
 [3,] 0.17 0.17 1.00      0.17 0.17 0.17 0.17 0.35  0.50  0.35  0.35  0.42  0.35  0.17  0.30  0.39
 [4,] 0.17 0.17      1.00 0.17 0.17 0.17 0.17 0.35  0.50  0.35  0.35  0.42  0.35  0.68  0.55  0.39
 [5,] 0.17 0.17 0.17 0.17 1.00      0.17 0.17 0.35  0.35  0.50  0.35  0.35  0.42  0.34  0.34  0.39
 [6,] 0.17 0.17 0.17 0.17      1.00 0.17 0.17 0.35  0.35  0.50  0.35  0.35  0.42  0.34  0.34  0.39
 [7,] 0.17 0.17 0.17 0.17 0.17 0.17 1.00      0.35  0.35  0.35  0.50  0.35  0.42  0.26  0.31  0.39
 [8,] 0.17 0.17 0.17 0.17 0.17 0.17      1.00 0.35  0.35  0.35  0.50  0.35  0.42  0.26  0.31  0.39
 [9,] 0.50 0.50 0.35 0.35 0.35 0.35 0.35 0.35 1.00  0.70  0.70  0.70  0.85  0.70  0.52  0.69  0.77
[10,] 0.35 0.35 0.50 0.50 0.35 0.35 0.35 0.35 0.70  1.00  0.70  0.70  0.85  0.70  0.60  0.72  0.77
[11,] 0.35 0.35 0.35 0.35 0.50 0.50 0.35 0.35 0.70  0.70  1.00  0.70  0.70  0.85  0.68  0.69  0.77
[12,] 0.35 0.35 0.35 0.35 0.35 0.35 0.50 0.50 0.70  0.70  0.70  1.00  0.70  0.85  0.52  0.61  0.77
[13,] 0.42 0.42 0.42 0.42 0.35 0.35 0.35 0.35 0.85  0.85  0.70  0.70  1.35  0.70  0.56  0.96  1.02
[14,] 0.35 0.35 0.35 0.35 0.42 0.42 0.42 0.42 0.70  0.70  0.85  0.85  0.70  1.35  0.60  0.65  1.02
[15,] 0.26 0.26 0.17 0.68 0.34 0.34 0.26 0.26 0.52  0.60  0.68  0.52  0.56  0.60  1.18  0.87  0.58
[16,] 0.34 0.34 0.30 0.55 0.34 0.34 0.31 0.31 0.69  0.72  0.69  0.61  0.96  0.65  0.87  1.41  0.80
[17,] 0.39 0.39 0.39 0.39 0.39 0.39 0.39 0.39 0.77  0.77  0.77  0.77  1.02  1.02  0.58  0.80  1.52
```

H

«Single-step» approach

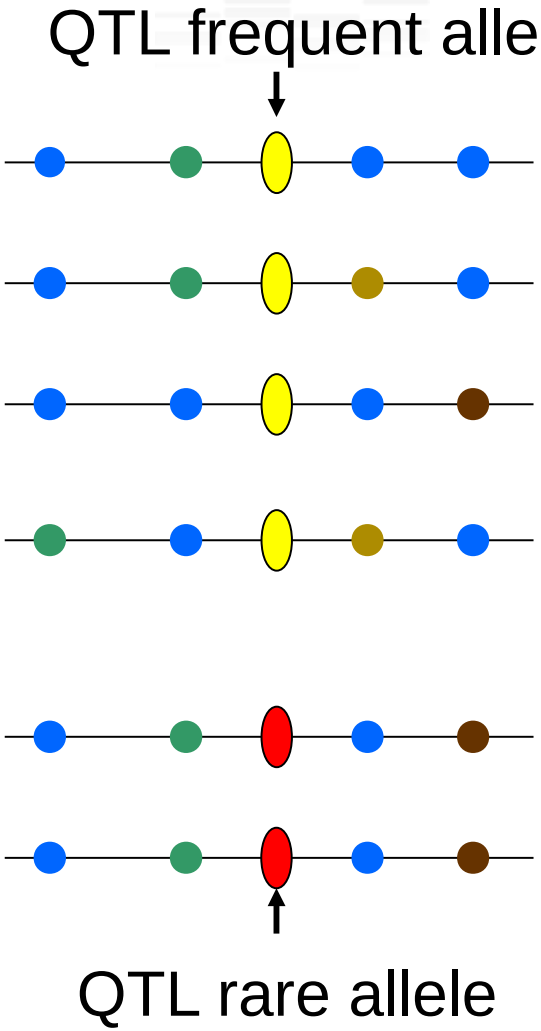
$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{a} + \mathbf{e}$ □ **(G)BLUP** on all data

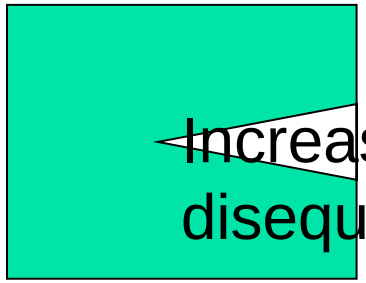
$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \alpha\mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

- Huge (and relatively dense) system but with some tricks, can be applied to large data sets (Legarra)
- Avantages:
  - Conceptually very simple (=BLUP)
  - quasi-optimal use of information from genotyped animals
  - automatically accounts for (genomic) selection

«Single-step» approach: current issues

Haplotypes vs SNP



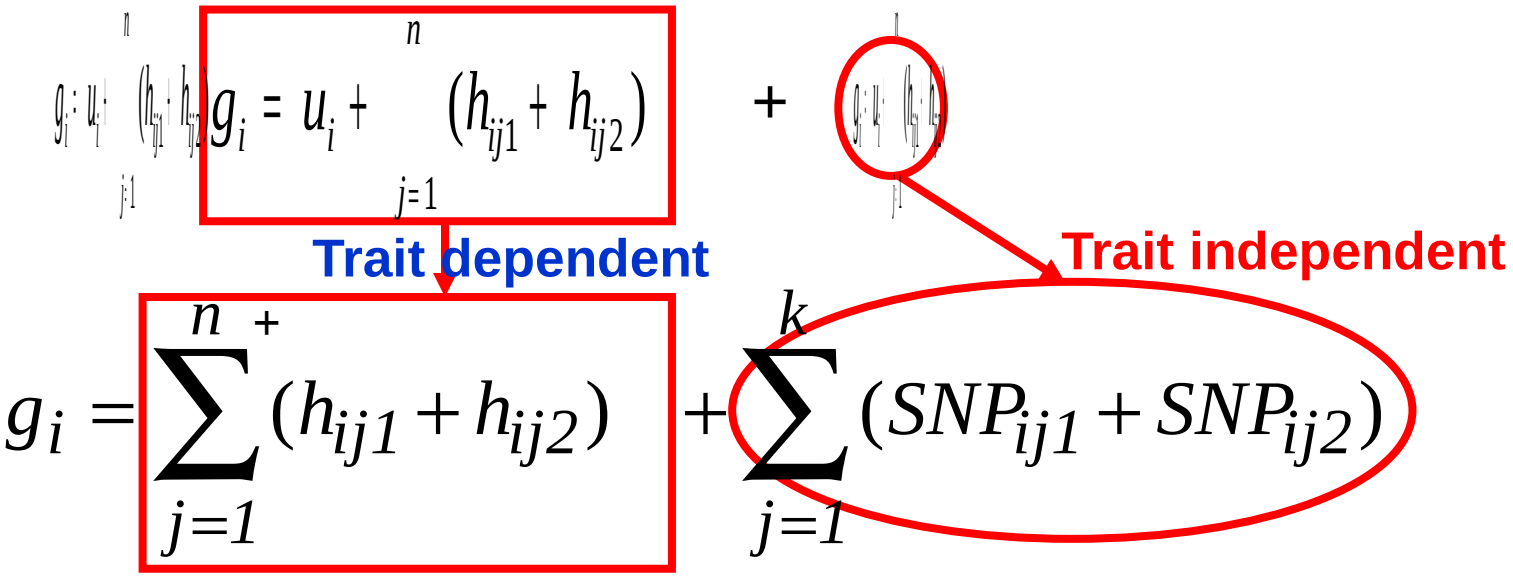


- Requires marker phasing
  (but imputation)
- more effects to estimate

The French genomic model (since April 2015)

QTL size	Genomic evaluation
Large Moderate Small	traced with markers □ haplotype effects □ DGV = $\sum_{j=1,..K}^{+} \hat{h}_j$
Tiny	Consider their sum only: $\hat{u} = \sum_{j'} \hat{m}_{j'} \sim N(0, \textbf{\textit{pedigree}}\text{ relationship matrix})$ $\sim N(0, \textbf{\textit{genomic}}\text{ relationship matrix})$

SNP & Haplotypes



$$g_i = \sum_{j=1}^{n} (h_{ij1} + h_{ij2}) + \sum_{j=1}^{k} (SNP_{ij1} + SNP_{ij2})$$

Comparison of genomic evaluation methods

- correlation between GEBV and DYD in validation population