
# Complex Trait Genomic Analysis

## Genomic Selection

# Genomic Selection (GS)

1. Fit vs. Prediction in modern genomics
2. The large p, small n paradigm
3. Variable selection vs. Ridge regression methods
4. Merging data: Single Step
5. GS using sequence data



## 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, University of Melbourne, Parkville 3052, Victoria, Australia*

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 overestimated and the accuracy of predicting genetic values of the offspring of the recorded animals was only 0.32. Best linear unbiased prediction of haplotype effects assumed equal variances associated to each 1-cM chromosomal segment, which yielded an accuracy of 0.73, although this assumption was far from true. Bayesian methods that assumed a prior distribution of the variance associated with each chromosome segment increased this accuracy to 0.85, even when the prior was not correct. It was concluded that selection on genetic values predicted from markers could substantially increase the rate of genetic gain in animals and plants, especially if combined with reproductive techniques to shorten the generation interval.

## Principles of Genome Selection

- It consists of using molecular data to predict phenotypic performance.
- It was formally proposed by T Meuwissen, B Hayes and M Goddard in 2001.
- It has revolutionized animal breeding, dairy cattle in particular.
- Resulted in increases in accuracy >30% in dairy cattle.

## Advantages of Genome Selection

- It increases accuracy for individuals without phenotypes (e.g., newborn).
- It decreases generation interval.
- Computes relationship (similarity) without the need for a pedigree, e.g., distinguishes between full sibs.
- Beneficial for low heritability traits.

# Limitations of Genome Selection

- Expensive, also in terms of logistics.
- It requires large datasets.
- Difficult to predict across breeds.
- Not clearly beneficial in all scenarios.

*But let's go back a little bit . . .*

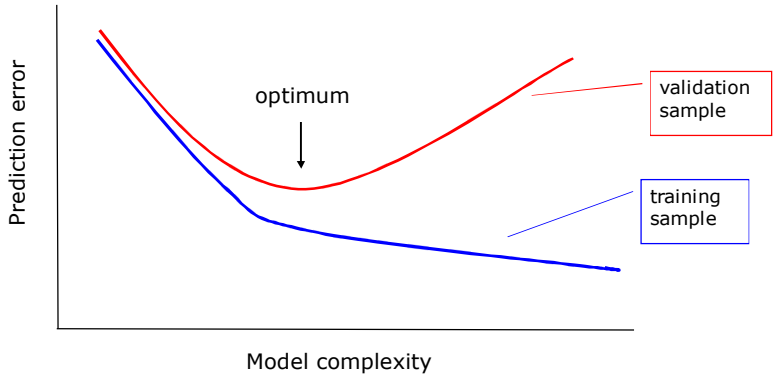
# Recall Statistics roles

1. Statistics was conceived as tools for measuring quantities of interest to the government (State), e.g., to collect taxes more efficiently.
2. For the largest part of 20th century though, Statistics predominant role has been inference, i.e., inferring unknown parameters, usually **model** based.
3. More recently, **prediction** has become central in Statistics, all the more with the enormous amounts of easily obtained data and powerful computer based methods. In the interim, and because of that, machine learning has evolved independently – occasionally converging with statistics.

## Two desirable characteristics: **parsimony** and **goodness of fit**.

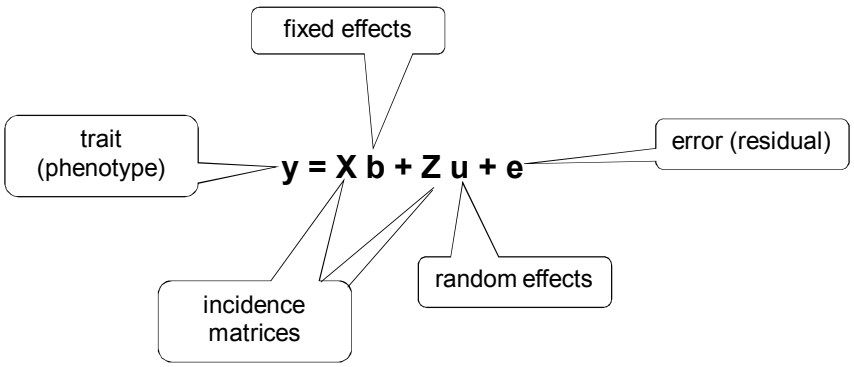
- ✓ For low p (# variables), goodness of fit is the most important issue.
- ✓ For high p, goodness of fit becomes critical (colinearity problems).
- ✓ For p > n (# observations), additional restrictions on the solutions must be posed. This is called prior information (Bayesian Statistics), shrinkage (Frequentist Statistics) or regularization (Machine Learning jargon)
- ✓ For n>>p, we are usually non longer interested in inference but only in prediction.

# A modeling tradeoff




Hastie et al. Elements of Statistical Learning

# Usual models: Mixed Linear Models



## Usual models: Mixed Linear Models

$$\mathbf{y = X\,b + Z\,u + e}$$

$$\begin{pmatrix} y \\ u \\ e \end{pmatrix} \sim N\left[\begin{pmatrix} Xb \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} ZGZ'\sigma_u^2 + I\sigma_e^2 & ZG\sigma_u^2 & I\sigma_e^2 \\ GZ'\sigma_u^2 & G\sigma_u^2 & 0 \\ I\sigma_e^2 & 0 & I\sigma_e^2 \end{pmatrix}\right]$$

Normality and additivity are two tightly linked phenomena

## Usual models: Mixed Linear Models

$$\mathbf{y = X\,b + Z\,u + e}$$

$$\begin{pmatrix} y \\ u \\ e \end{pmatrix} \sim N\left[\begin{pmatrix} Xb \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} ZGZ'\sigma_u^2 + I\sigma_e^2 & ZG\sigma_u^2 & I\sigma_e^2 \\ GZ'\sigma_u^2 & G\sigma_u^2 & 0 \\ I\sigma_e^2 & 0 & I\sigma_e^2 \end{pmatrix}\right]$$

- In classical breeding, Var(u) = G is computed from the pedigree.
- In GS, G is computed using marker information.

## large p small n paradigm

y = b1 x1 + b2 x2 + b3 x3 + b4 x4 + b5 x5 + … … … bn xn + e

Under the normal setting of breeding in which more variables than data are available, only two solutions exist (or a combination):

y = b1 x1 + b2 x2 + b3 x3 + b4 x4 + b5 x5 + … … … bn xn + e

Variable selection

y = b1 x1 + b2 x2 + b3 x3 + b4 x4 + b5 x5 + … … … bn xn + e

Shrinkage

## large p small n paradigm

Either variable selection or shrinkage are two kinds of regularization (in Machine Learning terminology).

Type 1 (L1) imposes a penalty on the sum of absolute values of coefficients

$$\min_w ||Xw - y||_2^2 + \alpha||w||_1$$

Lasso: results in variable selection

Type 2 (L2) imposes a penalty on the sum of squared coefficients

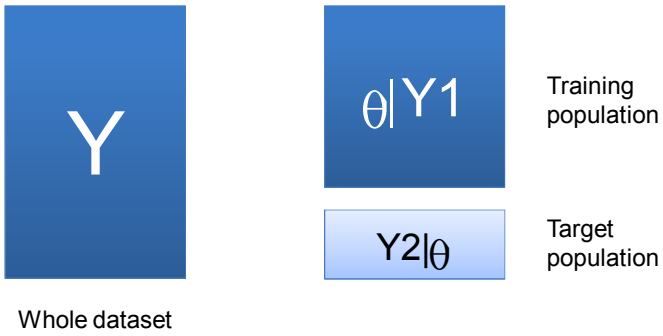
$$\min_w ||Xw - y||_2^2 + \alpha||w||_2^2$$

Ridge regression: results in shrinkage

Both regularization and combined methods have been used in GS

# Prediction is what matters in GS

Prediction, assessed by crossvalidation, is what matters. The population of reference is a critical choice here.




$$y = \mu + \sum_{k=1}^{m} \mathbf{x}_k \beta_k + \varepsilon$$

**Bayes A:** Each marker effect β follows, conditionally, a N distribution, variance is different for each marker, with prior inverted chi-2 distributed. Marginal of β is a t.

**Bayes B:** Same as Bayes A except that only a fraction of markers are assumed to have an effect.

**GBLUP:** Marker effects follow a normal distribution with equal variance for all markers.

$$y = \mu + \sum_{k=1}^{m} \mathbf{x}_k \beta_k + \varepsilon$$

Bayes A

$$p(\boldsymbol{\beta}_j, \sigma^2_{\beta_j}, S_\beta) = \left\{ \prod_k N(\beta_{jk}|0, \sigma^2_{\beta_{jk}}) \chi^{-2}(\sigma^2_{\beta_{jk}}|df_\beta, S_\beta) \right\} G(S_\beta|r, s)$$

Bayes B

$$p(\boldsymbol{\beta}_j, \sigma^2_\beta, \pi) = \left\{ \prod_k \left[ \pi N(\beta_{jk}|0, \sigma^2_\beta) + (1-\pi) 1(\beta_{jk} = 0) \right] \right\} \times \chi^{-2}(\sigma^2_\beta|df_\beta, S_\beta) B(\pi|p_0, \pi_0)$$

GBLUP

$$p(\boldsymbol{\beta}_i, \sigma^2_\beta) = \{ \prod_k N(\beta_{jk}|0, \sigma^2_\beta) \} \chi^{-2}(\sigma^2_\beta|df_\beta, S_\beta)$$

$$y = \mu + \sum_{k=1}^{m} \mathbf{x}_k \beta_k + \sum_{l=1}^{L} u_l + \varepsilon$$

Likelihood

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} N\left( y_i|\mu + \sum_{j=1}^{J} \sum_{k=1}^{K_j} x_{ijk} \boldsymbol{\beta}_{jk} + \sum_{l=1}^{L} u_{li}, \sigma^2_\varepsilon w_i^2 \right)$$

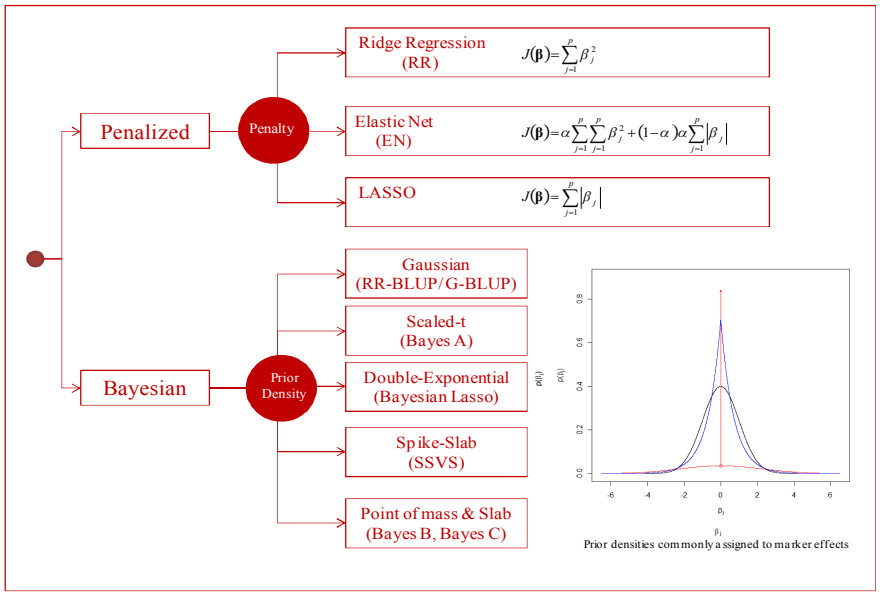
Prior

$$p(\boldsymbol{\theta}) = p(\mu) p(\sigma^2_\varepsilon) \prod_{j=1}^{J} p(\boldsymbol{\beta}_j) \prod_{l=1}^{L} p(\mathbf{u}_l).$$

Genome-Wide Regression and Prediction with the BGLR Statistical Package
Paulino Pérez and Gustavo de los Campos

| model= | Join distribution of effects and hyper-parameters |
|---|---|
| FIXED | $p(\boldsymbol{\beta}_j) \propto 1$ |
| BRR | $p(\boldsymbol{\beta}_j, \sigma^2_\beta) = \left\{\prod_k N(\beta_{jk}\|0, \sigma^2_\beta)\right\} \chi^{-2}(\sigma^2_\beta\|df_\beta, S_\beta)$ |
| BayesA | $p(\boldsymbol{\beta}_j, \sigma^2_{\beta_j}, S_\beta) = \left\{\prod_k N(\beta_{jk}\|0, \sigma^2_{\beta_{jk}})\chi^{-2}(\sigma^2_{\beta_{jk}}\|df_\beta, S_\beta)\right\} G(S_\beta\|r, s)$ |
| BL | $p(\boldsymbol{\beta}_j, \boldsymbol{\tau}^2_j, \lambda^2\|\sigma^2_\varepsilon) = \left\{\prod_k N(\beta_{jk}\|0, \tau^2_{jk} \times \sigma^2_\varepsilon) Exp\left\{\tau^2_{jk}\|\frac{\lambda^2}{2}\right\}\right\} \times G(\lambda^2\|r, s)$ , or<br>$p(\boldsymbol{\beta}_j, \boldsymbol{\tau}^2_j, \lambda\|\sigma^2_\varepsilon, \mathtt{max}) = \left\{\prod_k N(\beta_{jk}\|0, \tau^2_{jk} \times \sigma^2_\varepsilon) Exp\left\{\tau^2_{jk}\|\frac{\lambda^2}{2}\right\}\right\} \times B(\lambda/\max\|p_0, \pi_0)$, or<br>$p(\boldsymbol{\beta}_j, \boldsymbol{\tau}^2_j\|\sigma^2_\varepsilon, \lambda) = \left\{\prod_k N(\beta_{jk}\|0, \tau^2_{jk} \times \sigma^2_\varepsilon) Exp\left\{\tau^2_{jk}\|\frac{\lambda^2}{2}\right\}\right\}$ |
| BayesC | $p(\boldsymbol{\beta}_j, \sigma^2_\beta, \pi) = \left\{\prod_k \left[\pi N(\beta_{jk}\|0, \sigma^2_\beta) + (1-\pi)1(\beta_{jk}=0)\right]\right\} \times \chi^{-2}(\sigma^2_\beta\|df_\beta, S_\beta) B(\pi\|p_0, \pi_0)$ |
| BayesB | $p(\boldsymbol{\beta}_j, \sigma^2_\beta, \pi) = \left\{\prod_k \left[\pi N(\beta_{jk}\|0, \sigma^2_\beta) + (1-\pi)1(\beta_{jk}=0)\right] \chi^{-2}(\sigma^2_{\beta_{jk}}\|df_\beta, S_\beta)\right\} B(\pi\|p_0, \pi_0) \times G(S_\beta\|r, s)$ |
| RKHS | $p(\boldsymbol{u}_l, \sigma^2_{u_l}) = N(\boldsymbol{u}_l\|\mathbf{0}, \boldsymbol{K}_l \times \sigma^2_{u_l}) \chi^{-2}(\sigma^2_{u_l}\|df_l, S_l)$ |

Genome-Wide Regression and Prediction with the BGLR Statistical Package
Paulino Pérez and Gustavo de los Campos




Gustavo de Los Campos

## In general, it seems that

- Bayesian alphabet is quasi infinite.
- Most methods work similarly as size of data increases.
- Genetic architecture does influence the performance of the methods. In general, variable selection is better when the number of causal loci is small (Daetwyler et al., 2010).

## GBLUP (VanRaden)

G-BLUP is a widely used method for genomic selection. It consists in replacing the pedigree based matrix by a molecular marker based one (GRM).

**y = X b + Z u + e**

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} X'R^{-1}\boldsymbol{y} \\ Z'R^{-1}\boldsymbol{y} \end{pmatrix}$$

- There are several potential ways of calculating **G.**
- What is the relationship between **G** and **A** (pedigree NRM).
- **G** must be inverted brute force (good approx algorithms by Misztal).

## Genomic relationship matrix

**Not a a single metrics exists**. Depends on genotypes and on allele frequencies.

$$G = \left\{ \sum \frac{(z-2p)(z'-2p)}{2pq} \right\}$$

Correlation matrix

$$G = \left\{ \frac{\sum (z-2p)(z'-2p)}{\sum 2pq} \right\}$$

Covariance matrix (averaged)

## Genomic relationship matrix

**Not a a single metrics exists**. Depends on genotypes and on allele frequencies.

Low frequency variants have a very low variance and induce important numeric problems. Cov is recommended (see also Toro et al. 2011)

$$G = \left\{ \sum \frac{(z-2p)(z'-2p)}{2pq} \right\}$$

Correlation matrix

$$G = \left\{ \frac{\sum (z-2p)(z'-2p)}{\sum 2pq} \right\}$$

Covariance matrix (averaged)

## Relation between G and pedigree A
(Toro, García-Cortés, Legarra, GSE 2011)

A = $\{a_{ij}\}$ contains the additive relationship coefficients ≡ twice the probability of two alleles from individuals i and j being Identical by descent (IBD) | pedigree

G = $\{g_{ij}\}$ contains the covariance of molecular genotypes, which depends on the probability of two alleles from individuals i and j being Identical by State (IBS) | marker information

$$E(f_{M_{ij}}) = f_{ij} + (1 - f_{ij})(p^2 + q^2)$$

P (IBS) P (IBD) P (M=A) P (M=a)

In the base population!

## Merging Information: Single Step Approach
(Legarra, Aguilar, Misztal, JDS 2009;
Mogens, Lund, GSE, 2010)

- It addresses the issue of a large connected pedigree where only a subset of animals are genotyped.
- It models genotypes as multivariate normal variables, using **A** and **G** as covariance matrices.
- It derives an expression for **A** of ungenotyped individuals given **G**.

$$Cov(z_i, z_j) = A_{ij}\ 2\ p\ q$$

Marker genotype at individuals i & j

## Merging Information: Single Step Approach

Ungenotyped indivs

Genotyped indivs

$$\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix}=\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}+\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}+\varepsilon$$

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}|pedigree\Big)=\begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}\sigma_u^2$$

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}|\text{marker}\Big)=\begin{pmatrix}? & ?\\ ? & \mathbf{G}_{22}\end{pmatrix}\sigma_g^2$$

## Merging Information: Single Step Approach
(Mogens, Lund, GSE, 2010)

Ungenotyped indivs

Genotyped indivs

$$\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix}=\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}+\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}+\varepsilon$$

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}\Big|\begin{matrix}?\\ \mathbf{Z}_2\end{matrix}\Big)=\begin{pmatrix}? & ?\\ ? & \mathbf{G}_{22}\end{pmatrix}\sigma_g^2$$

$$Var(\mathbf{u}_1\,|\,\mathbf{Z}_2)=Var[E(\mathbf{u}_1\,|\,\mathbf{Z}_1,\mathbf{Z}_2)]+E[Var(\mathbf{u}_1\,|\,\mathbf{Z}_1,\mathbf{Z}_2)]$$

$$Var[E(\mathbf{u}_1\,|\,\mathbf{Z}_1,\mathbf{Z}_2)]=Var(\mathbf{A}_{21}\mathbf{A}_{22}^{-1}\mathbf{Z}_2)=\mathbf{A}_{21}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{12}\,\sigma_u^2$$

$$E[Var(\mathbf{u}_1\,|\,\mathbf{Z}_1,\mathbf{Z}_2)]=E[\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21}]=(\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21})\,\sigma_u^2$$

$$Cov(\mathbf{u}_1,\mathbf{u}_2\,|\,\mathbf{Z}_2)=\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}$$

$$Var(\mathbf{u}_2\,|\,\mathbf{Z}_2)=\mathbf{G}$$

Assumes normality between marker genotypes and equal p in Z1 and Z2

## Merging Information: Single Step Approach
(Legarra et al. 2009)

Ungenotyped indivs

Genotyped indivs

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} + \varepsilon$$

Assumes prior **A** has no influence on **u2** (ie, molecular information fully dominates over pedigree)

$$p\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = \mathrm{N}\left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \right)$$

prior

$$p(\boldsymbol{u}_1, \boldsymbol{u}_2 | markers) = p(\boldsymbol{u}_1 | \boldsymbol{u}_2) p(\boldsymbol{u}_2 | markers) = N(\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{u}_2, \sigma_u^2(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})).\ N(0,\ \sigma_u^2\boldsymbol{G}).$$

'posterior'

$$= [\mathbf{A}_{21}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{12} + \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21}]\sigma_u^2$$

## Merging Information: Single Step Approach
(Legarra et al 2009)

Putting all together

$$Var\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} | \text{markers} = \mathbf{H}\sigma_u^2 \quad = \sigma_u^2 \begin{pmatrix} \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{G} \end{pmatrix}$$

**H** can be interpreted as a projection (ie, multivariate regression) of **A** on **G**

Amazingly, **H** has a very simple inverse

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

## Single Step Approach: Advantages
(Legarra et al LPS 2014)

1. Automatic accounting of all relatives of genotyped individuals and their performances.
2. Simultaneous fit of genomic information and estimates of other effects (e.g., contemporary groups). Therefore not loss of information.
3. Feedback: the extra accuracy in genotyped individuals is transmitted to all their relatives (*e.g.* Christensen et al., 2012).
4. Simple extensions. Because this is a linear BLUP-like estimator, the extension to more complicated models (multiple trait, threshold traits, test day records) is immediate. Any model fit using relationship matrices can be fit using combined relationship matrices.
5. Analytical framework. The Single Step provides an analytical framework for further developments. This is notoriously difficult with pseudo-data.

## Single Step Approach: Limitations
(Legarra et al LPS 2014)

1. Programming complexity to fit complicated models for marker effects (Bayesian Regressions, machine learning algorithms, etc.).
2. Lack of experience on very large data sets.
3. Long computing times with current Single Step algorithms methods, for very large data sets.
4. Lack of an easy and elegant way of considering major genes in a multiple trait setting, this is a drawback of multiple step methods as well.
5. Lacks general theory to fit unequal marker density
6. Theory is asymptotic

# Genomic Selection in Dairy Cattle: The USDA Experience*

**Annual Review of Animal Biosciences**
Vol. 5:309-327 (Volume publication date February 2017)
First published online as a Review in Advance on November 16, 2016
https://doi.org/10.1146/annurev-animal-021815-111422

**George R. Wiggans,[1] John B. Cole,[1] Suzanne M. Hubbard,[1] and Tad S. Sonstegard[2]**
[1]Animal Genomics and Improvement Laboratory, Agricultural Research Service, US Department of Agriculture, Beltsville, Maryland 20705–2350; email: george.wiggans@ars.usda.gov, john.cole@ars.usda.gov, suzanne.hubbard@ars.usda.gov
[2]Acceligen of Recombinetics Inc., St. Paul, Minnesota 55104; email: tad@recombinetics.com



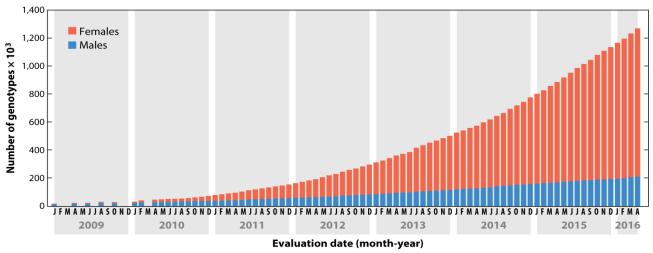
Sections



## Abstract

Genomic selection has revolutionized dairy cattle breeding. Since 2000, assays have been developed to genotype large numbers of single-nucleotide polymorphisms (SNPs) at relatively low cost. The first commercial SNP genotyping chip was released with a set of 54,001 SNPs in December 2007. Over 15,000 genotypes were used to determine which SNPs should be used in genomic evaluation of US dairy cattle. Official USDA genomic evaluations were first released in January 2009 for Holsteins and Jerseys, in August 2009 for Brown Swiss, in April 2013 for Ayrshires, and in April 2016 for Guernseys. Producers have accepted genomic evaluations as accurate indications of a bull's eventual daughter-based evaluation. The integration of DNA marker technology and genomics into the traditional evaluation system has doubled the rate of genetic progress for traits of economic importance, decreased generation interval, increased selection accuracy, reduced previous costs of progeny testing, and allowed identification of recessive lethals.

## Number of genotyped bulls and cows in USA



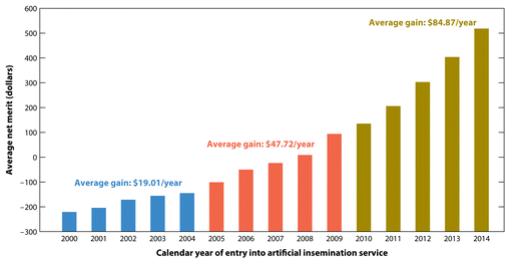

Wiggans GR, et al. 2017.
Annu. Rev. Anim. Biosci. 5:309–27

Figure 1 Number of genotyped animals included in US genomic evaluations for dairy cattle since January 2009. Official US genomic evaluations were first released to the dairy industry in January 2009 for Holsteins and Jerseys, in August 2009 for Brown Swiss, in April 2013 for Ayrshires, and in April 2016 for Guernseys. Data for figure generation were reported by the Council on Dairy Cattle Breeding (27). Months without data represent months in which official evaluations were not released.

**Table 4** Increases in reliability of US genetic evaluations of Holsteins from including genomic information[a]

| | August 2011 reliability (%) | | |
|---|---|---|---|
| **Trait** | **Parent average** | **Genomic evaluation** | **Gain[b]** |
| Milk (kg) | 38.5 | 72.5 | 34.0 |
| Fat (kg) | 38.5 | 72.2 | 33.8 |
| Protein (kg) | 38.4 | 63.3 | 24.9 |
| Fat (%) | 38.5 | 96.9 | 58.5 |
| Protein (%) | 38.4 | 87.4 | 49.0 |

# Evolution of net merit




Wiggans GR, et al. 2017.
Annu. Rev. Anim. Biosci. 5:309–27

Figure 5 Net merit in April 2015 of marketed US Holstein bulls that entered artificial-insemination service in 2000 and later. Net merit is a genetic-economic index that was developed as a lifetime profit function that uses actual incomes and expenses for traits of economic importance ( 37 ). The economic values and traits included in the net merit index are updated as needed to reflect changes in the dairy industry, and the latest revisions were made in 2014 ( 38 ). The data for figure generation were provided by the Council on Dairy Cattle Breeding (https://www.cdcb.us).

- ✓ Genomic selection has profoundly affected genetic improvement of dairy cattle.
- ✓ Producers have accepted genomic evaluations as accurate indications of a bull's eventual daughter-based evaluation.
- ✓ The AI organizations no longer rely on progeny-test herds to determine which bulls to market, and they purchase young bulls based on genomic evaluations.
- ✓ Both heifers and bulls are genotyped usually before they are one month old.
- ✓ Another benefit of genomics is the detection of carriers of undesirable recessive characteristics.
- ✓ The age of parents of marketed bulls has steadily decreased since the start of genomic selection, essentially halved the generation interval.
- ✓ The rate of improvement in average net merit has nearly doubled for Holstein bulls since the implementation of genomic evaluation in 2010.

How about sequence?

# Sequence (NGS) vs. Chip data

- About ~ 1000 € + 24h computing per sample for initial processing.
- Sophisticated, tricky bioinformatic procedures.
- 'Unbiased' estimates of SNP variability.
- Missing data is unavoidable
- You probably have heard that causal variants are in the data, do not trust it too much or so what?

REMEMBER: NGS DATA IS NOT SIMPLY HIGHER SNP DENSITY.

- About ~ 100 € per sample (20x less than NGS).
- Standard software already available.
- Heterozygosity is not interpretable due to SNP ascertainment.
- Missing data can be controlled.
- Bias is unavoidable, even in the highest density array.

Remember: RARE VARIANTS ARE THE MOST FREQUENT VARIANTS

# Claimed advantages of sequence

- Causal variants are in the data.
- No decrease in LD over generations.

Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

## Genomic relationships computed from either next-generation sequence or array SNP data

M. Pérez-Enciso[1,23,4]

**Summary**

The use of sequence data in genomic prediction models is a topic of high interest, given the decreasing prices of current 'next'-generation sequencing technologies (NGS) and the theoretical possibility of directly interrogating the genomes for all causal mutations. Here, we compare by simulation how well genetic relationships (G) could be estimated using either NGS or ascertained SNP arrays. DNA sequences were simulated using the coalescence according to two scenarios: a 'cattle' scenario that consisted of a bottleneck followed by a split in two breeds without migration, and a 'pig' model where Chinese introgression into international pig breeds was simulated. We found that introgression results in a large amount of variability across the genome and between individuals, both in differentiation and in diversity. In general, NGS data allowed the most accurate estimates of *G*, provided enough sequencing depth was available, because shallow NGS (4×) may result in highly distorted estimates of ***G*** elements, especially if not standardized by allele frequency. However, high-density genotyping can also result in accurate estimates of ***G***. Given that genotyping is much less noisy than NGS data, it is suggested that specific high-density arrays (~3M SNPs) that minimize the effects of ascertainment could be developed in the population of interest by sequencing the most influential animals and rely on those arrays for implementing genomic selection.

# G matrix with sequence vs chip data

(Pérez-Enciso 2013)

- Why molecular relationship matrices?
- Two definitions of molecular relationships (**G**).
- SNP ascertainment.
- Demography has an effect
- Definition of **G** has an effect
- Is NGS cost effective for selection?

# Why molecular relationship matrix?

1. Genetic improvement works by selecting those individuals that are predicted to have better performing offspring.
2. Speed of improvement depends on how accurately can we predict genotype (which is transmitted) from phenotype (which is observed).
3. This is done by regressing phenotype on relationship **G** (average allele sharing between individuals).
4. Therefore, the more accurate we can measure relationships, the higher genetic response will be.
5. Currently, **G** is computed using SNP arrays or simply the pedigree.

# Definitions of G

Broadly, a pairwise measure of similarity based on genotypes. It should result in a positive definite matrix if to be used in evaluation.

$G_L$: Average pairwise allele difference

GL

$$g_{ij} = \sum_{k=1,nsites} \delta(a_{ik}, a_{jk}) / L$$

| Genotypes | δ |
|---|---|
| AA,AA | 1 |
| AB,- | 0.5 |
| AA,BB | 0 |

$G_F$: Usual relationship matrix used in G-BLUP

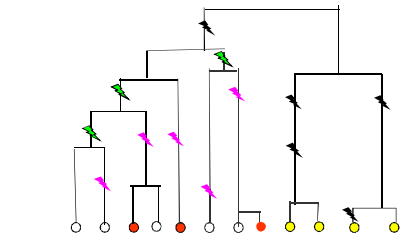
GF

$$g_{ij} = \frac{\sum_{k=1,nsp}(x_{ik} - \mu_k)(x_{jk} - \mu_k)}{\sum_{k=1,nsp} 2p_k(1-p_k)}$$

| Genotype | x |
|---|---|
| AA | -1 |
| AB | 0 |
| AA | 1 |

# SNPs in arrays are subject to ascertainment bias

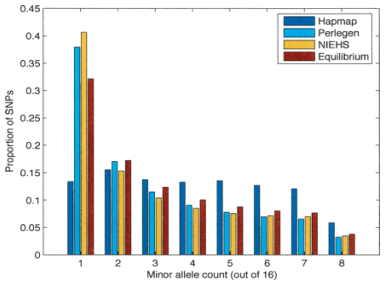
SNPs on a chip are chosen for technical reasons and to be as 'informative' as possible across breeds; they are used to genotype additional individuals from that or other breeds.



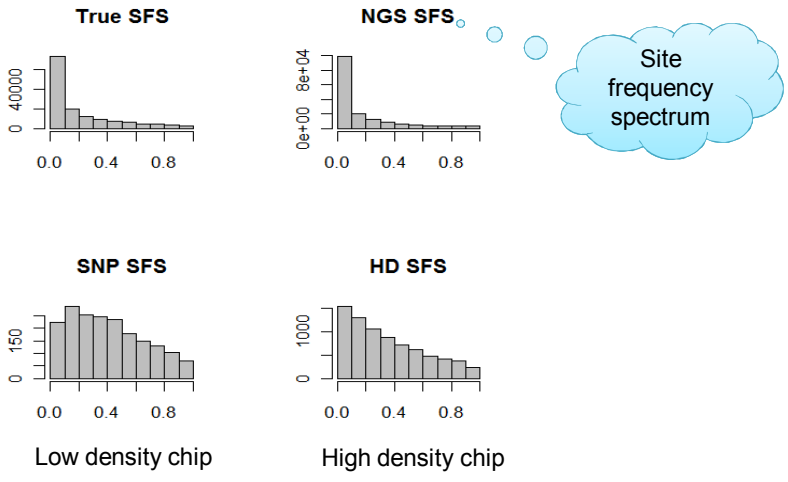
SNPs in panel

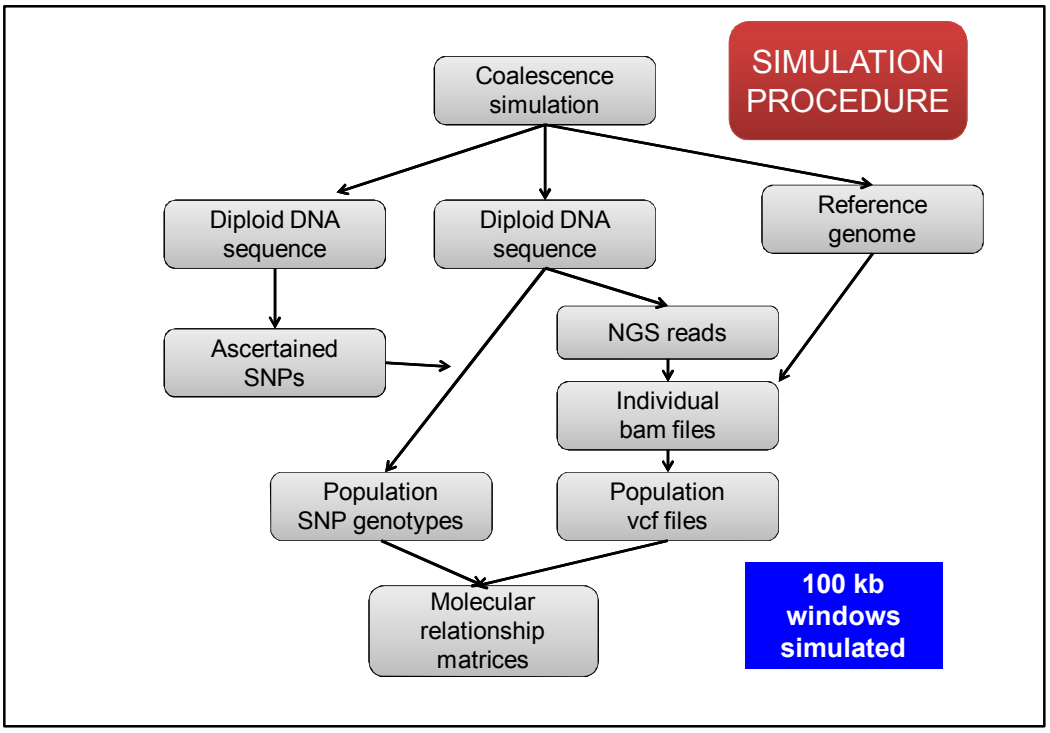
singletons in panel

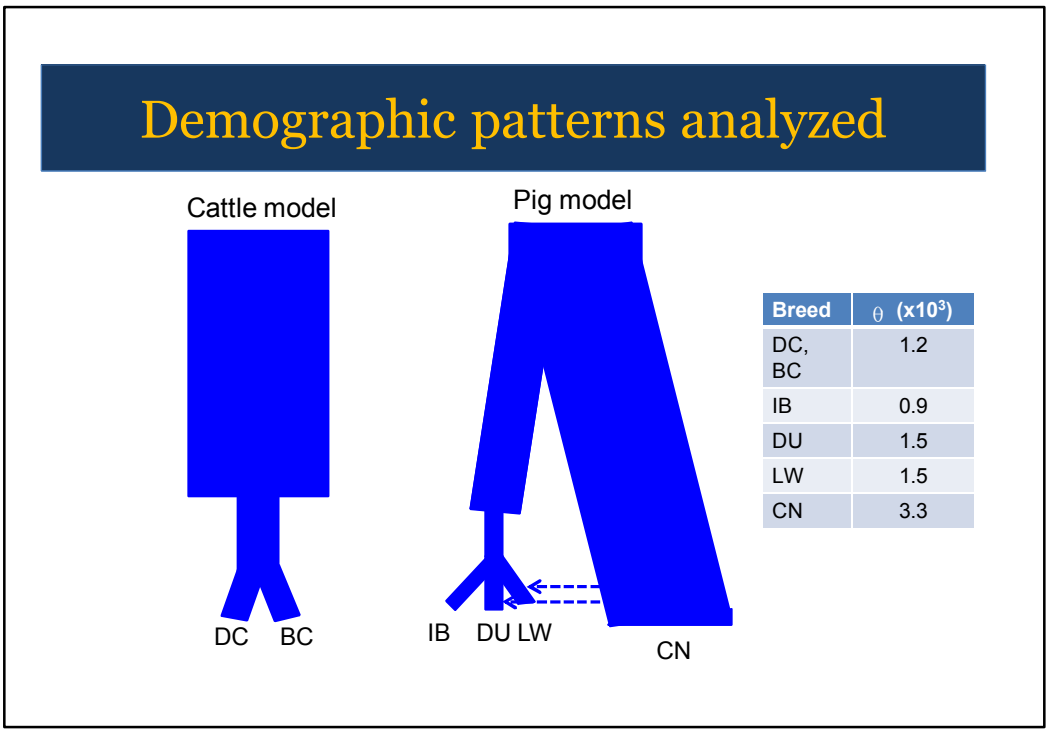
undiscovered SNPs



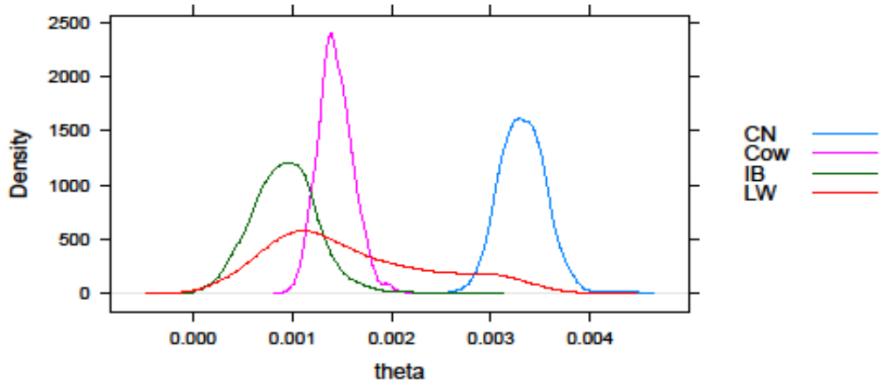

# SNP ascertainment effect on allele frequencies



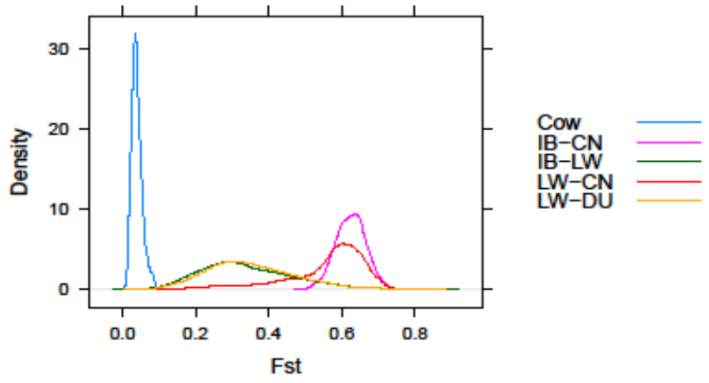

SIMULATION PROCEDURE
Coalescence simulation
Diploid DNA sequence
Diploid DNA sequence
Reference genome
Ascertained SNPs
NGS reads
Individual bam files
Population SNP genotypes
Population vcf files
Molecular relationship matrices
100 kb windows simulated


Demographic patterns analyzed
Cattle model
Pig model
DC
BC
IB
DU LW
CN
Breed
θ (x10³)
DC, BC
1.2
IB
0.9
DU
1.5
LW
1.5
CN
3.3

# Results: Variability



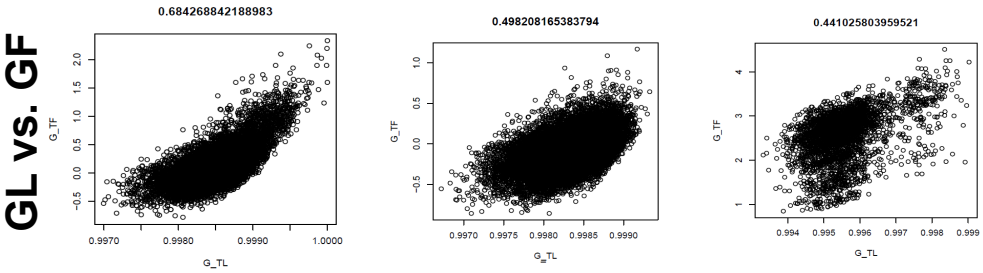

# Results: Differentiation



# GF and GL are not the same metrics

Cattle: within breed | Cattle: between breeds | Pig: within China

GL vs. GF

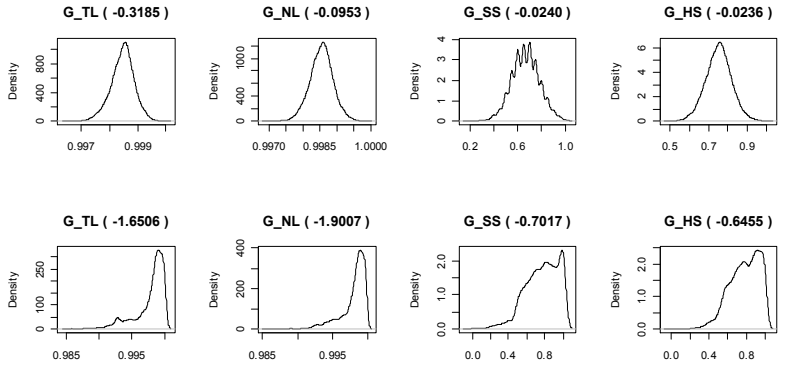



Each dot represents the relationship in a 100 kb window computed according to GL and GF between two random individuals, including self-relations, from a random replicate. In GF, freqs are computed across all breeds and individuals.

# The distribution of GL - values

cattle

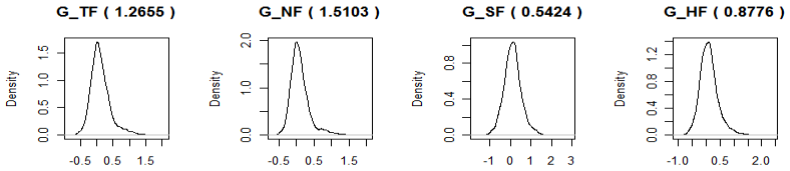
Pig (admixed)



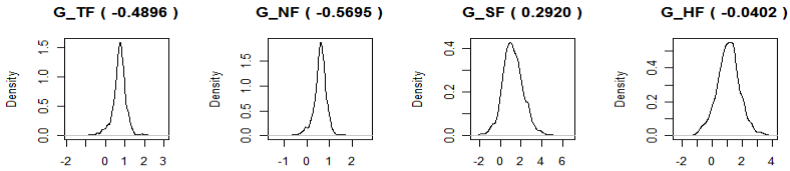

True | NGS | SNP | HD SNP

# The distribution of GF - values

cattle



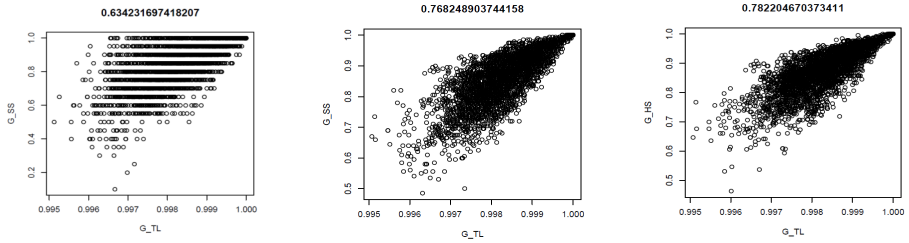

Pig (admixed)




True | NGS | SNP | HD SNP

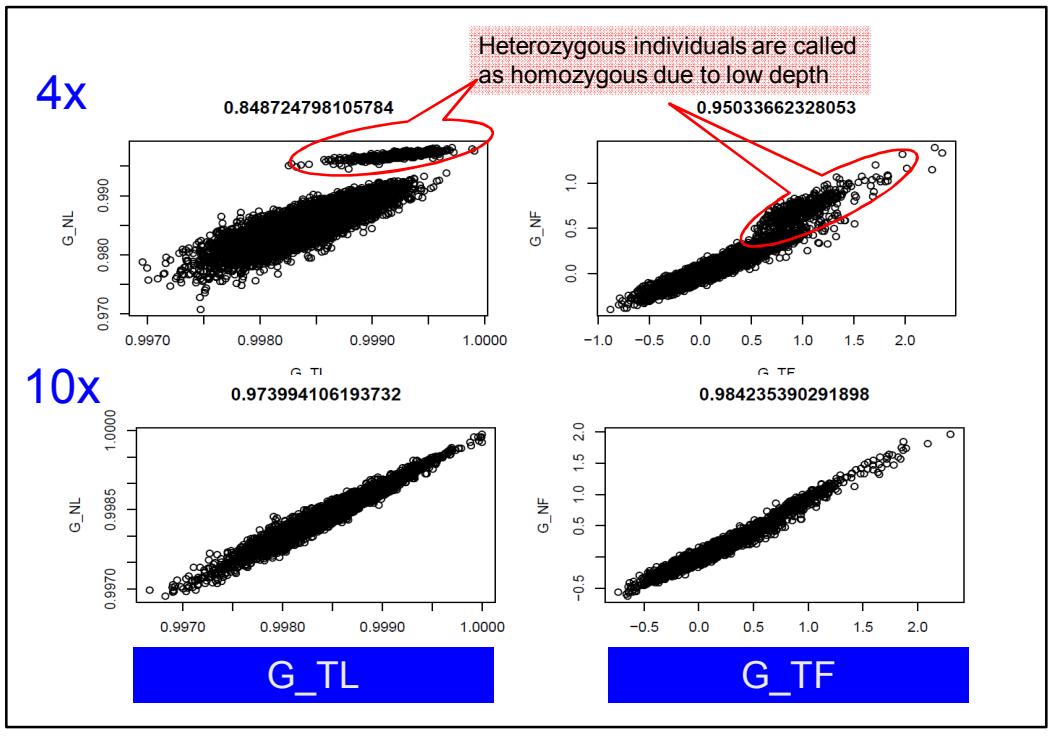
# The number of SNPs chosen is also important

Low density, MAF > 0.15 | High density, MAF > 0.15 | High density, MAF > 0.05




Within Iberian pig population.

1. NGS results in highly accurate of G estimates, but
2. Shallow depth may cause strong biases in SNP calling.
3. Joint SNP calling distorted if mixing highly divergent individuals



# Conclusions

- ✓ Demography has an important influence on genetic relationships.
- ✓ SNP ascertainment biases can be remedied either by increasing SNP number or by relaxing frequency restriction.
- ✓ NGS can lead to misleading estimates if depth too low (<5x) or if joint SNP calling of multiple divergent populations.
- ✓ Otherwise, the accuracy of NGS based relationships is very high although at a price >> 10x larger.
- ✓ Is in practice NGS to be cost-effective for genomic selection? An alternative would be to develop high-density population-specific SNP arrays

# How about genomic selection itself?

Pérez-Enciso *et al. Genetics Selection Evolution* (2015) 47:43
DOI 10.1186/s12711-015-0117-5

GSE Genetics Selection Evolution

RESEARCH Open Access

## Sequence- vs. chip-assisted genomic selection: accurate biological information is advised

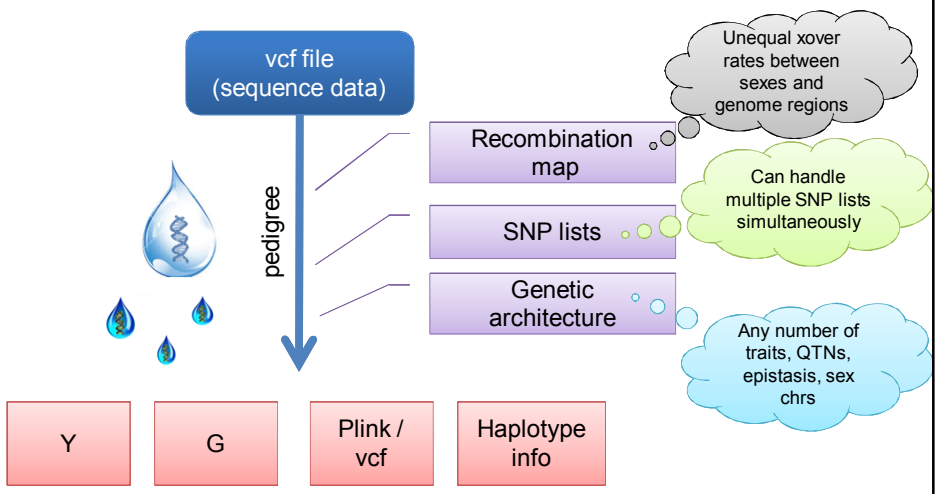
Miguel Pérez-Enciso[1,2,3*], Juan C Rincón[1,4] and Andrés Legarra[5]

GENETICS | GENOMIC SELECTION

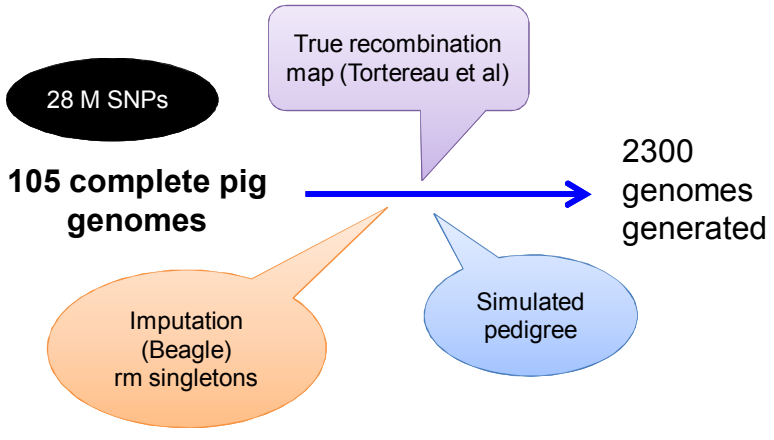
## Evaluating Sequence-Based Genomic Prediction with an Efficient New Simulator

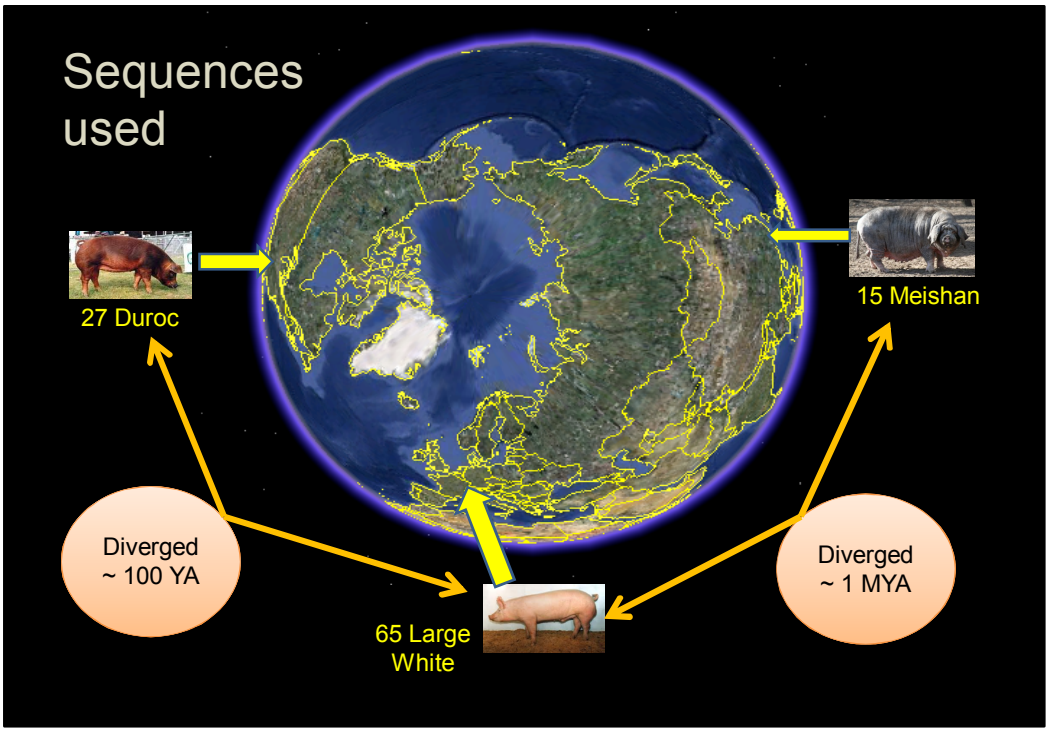
**Miguel Pérez-Enciso,*,†,‡,§,1 Natalia Forneris,* Gustavo de los Campos,‖,** and Andrés Legarra††**
*Centre for Research in Agricultural Genomics (CRAG), Consejo Superior de Investigaciones Científicas - Institut de Recerca i Tegnologia Agroalimentàrias - Universitat Autònoma de Barcelona - Universitat de Barcelona (CSIC-IRTA-UAB-UB) Consortium and †Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain, ‡Institut Català de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain, §Department of Epidemiology and Biostatistics and **Department of Statistics, Michigan State University, East Lansing, Michigan 48824, and ††Institut National de la Recherche Agronomique (INRA), Unité Mixte de Recherche 1388 GENPHYSE, Castanet-Tolosan 31326, France
ORCID IDs: 0000-0003-3524-995X (M.P.-E.); 0000-0001-8893-7620 (A.L.)

## Sequence Based Virtual Breeding (SBVB)




## Sequence Based Virtual Breeding

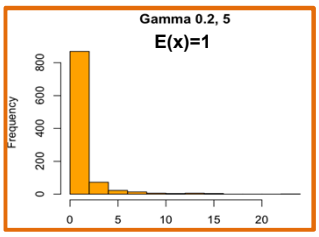





## Genetic architectures compared

‘NEUTRAL’

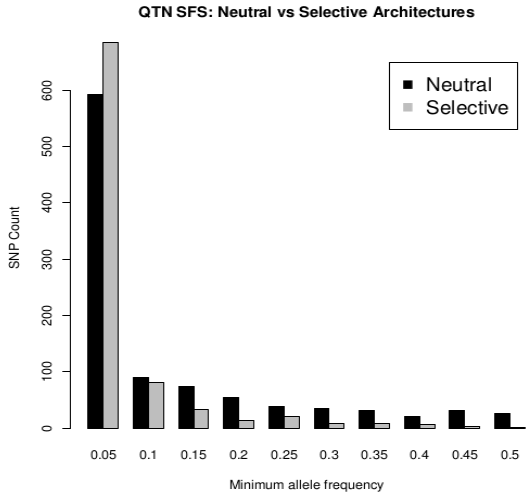
- ✓ 1000 SNPs randomly chosen as QTNs
- ✓ QTN effects ~Γ(0.2,5)
- ✓ No correlation btw effect and frequency



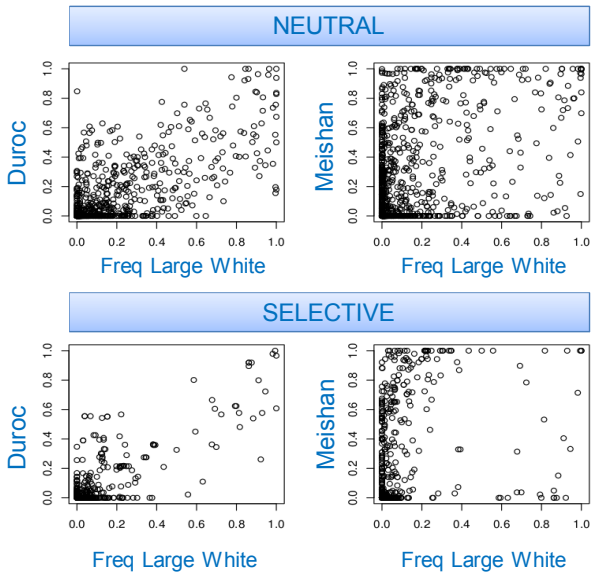

‘SELECTIVE’

- ✓ 200 genes with highest Fst (wild – domestic) and lowest Tajima’s D
- ✓ 1000 SNPs with ‘moderate’ or ‘high’ impact + UTRs
- ✓ QTN effects ~ Γ(0.2,5)
- ✓ Negative correlation btw effect and frequency

# Site frequency spectrum (LW)

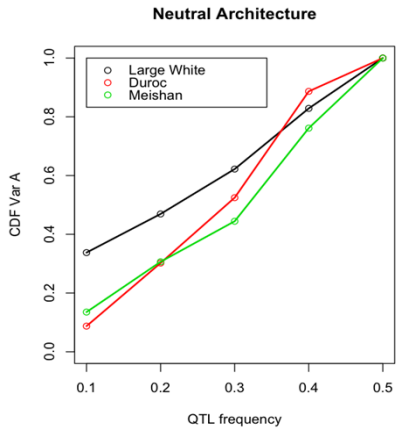


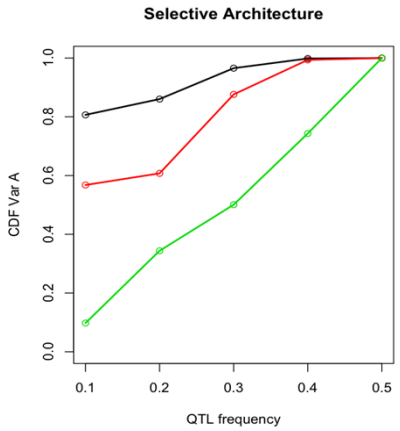

# Joint frequency spectra




These are 'real' (observed) frequencies

# Genetic Variance CDF



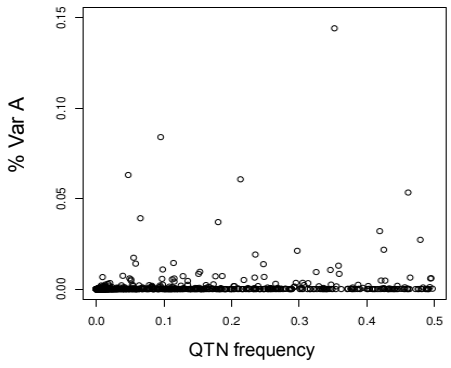




In a neutral architecture, the variance CDF explained is proportional to frequency

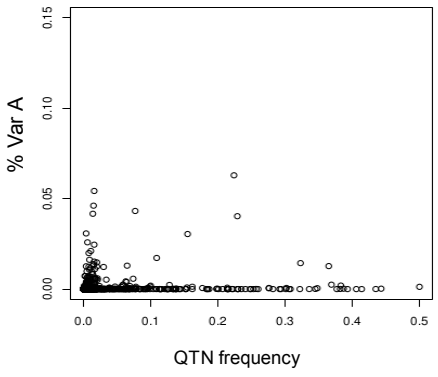
In a selective architecture, there is an excess of contribution from low frequency variants

# QTN frequency vs. % variance

Neutral architecture (LW)



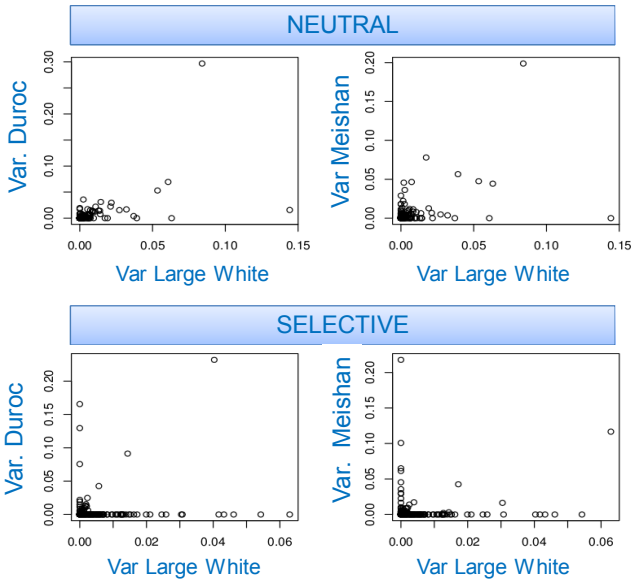

Selective architecture (LW)




Large effect QTNs at moderate frequencies may appear

Large effect QTNs tend to be rare variants

# Joint additive variances (%)




ρ ~ 0.50

ρ ~ 0.18

# Impact of architecture on genomic selection accuracy

- Computed as ρ2 btw true and predicted breeding value.
- GBLUP solved with BGLR (de los Campos & Pérez, 2015: https://github.com/gdlc/bglr-r ).
- h2=0.50, 10 replicates per case.

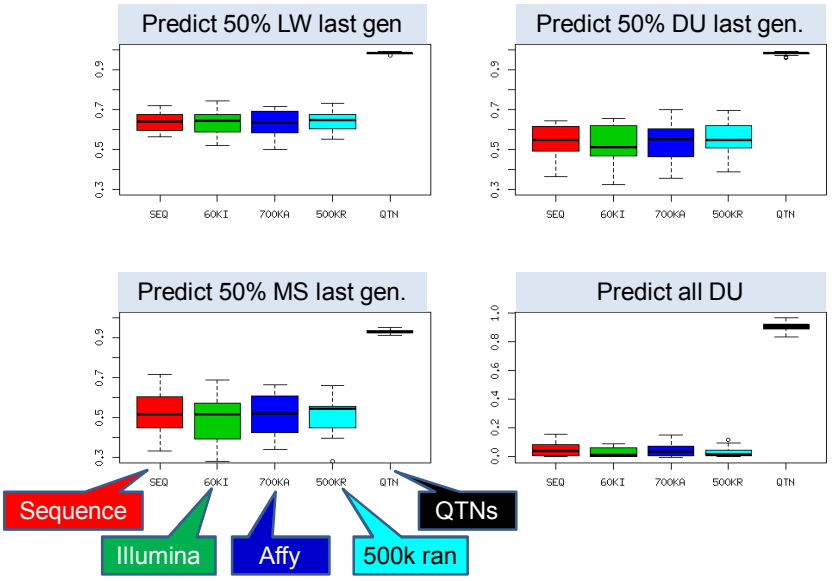
Scenarios
- Predict 50% of Large White from last generation
- Predict 50% of Duroc in last generation
- Predict 50% of Meishan
- Predict all Duroc

# Molecular information used

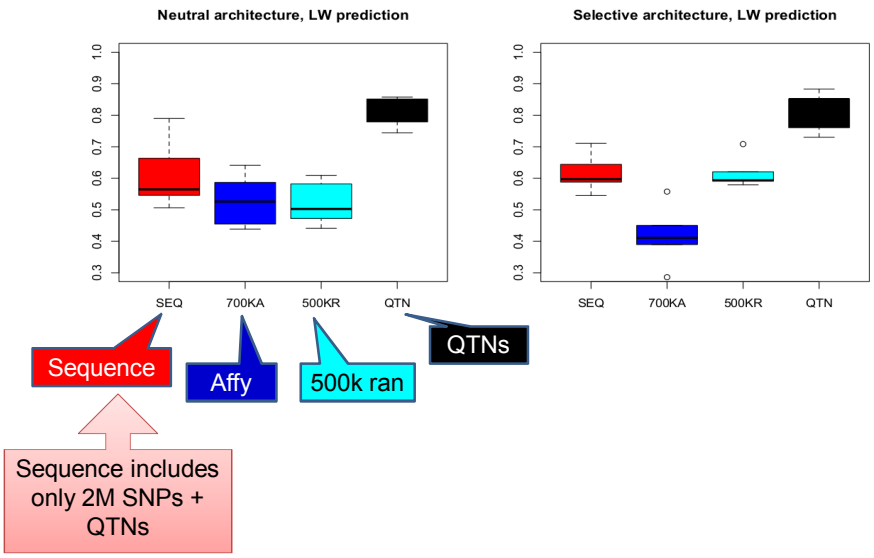
- Sequence (28M)
- Illumina 60k chip (~50k)
- Affy 700k chip (~500k)
- 500k random SNPs
- ~1k causal SNPs
- 60k SNPs in 200 causal genes

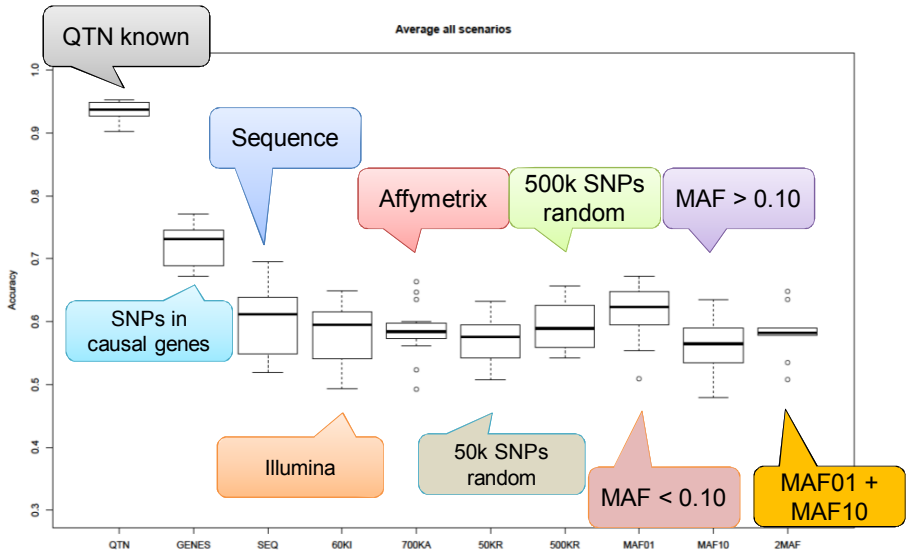
selective architecture only

# GS accuracy: Neutral architecture



# GS accuracy: Bayes R method



# What do real data say?

## Genomic prediction from whole genome sequence in livestock: the 1000 Bull Genomes Project

**Conference Paper** · August 2014 *with* 218 Reads

Cite this publication

Conference: 10th World Congress on Genetics Applied to Livestock Production

1st **Ben J Hayes**

2nd **Iona Macleod**
26.73 · University of Melbourne

3rd **Hans D. Daetwyler**
33.74 · Department of Economic...

+ 21

Last **Michael E. Goddard**

In a dairy data set, predictions using BayesRC and imputed sequence data from 1000 Bull Genomes were 2% more accurate than with 800k data.

# What do real data say?

## Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture

Guiyan Ni[1*], David Cavero[2], Anna Fangmann[1], Malena Erbe[1,3] and Henner Simianer[1]

Our results show that little or no benefit was gained when using all imputed WGS data to perform genomic prediction compared to using HD array data regardless of the weighting factors tested. However, using only genic SNPs from WGS data had a positive effect on prediction ability.

## Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection

Mario P. L. Calus, Aniek C. Bouwman, Chris Schrooten and Roel F. Veerkamp

*Genetics Selection Evolution* 2016 48:49 | https://doi.org/10.1186/s12711-016-0225-x | © The Author(s) 2016

Predictions computed as the average of the predictions computed for each subset achieved the highest accuracies, i.e. 0.5 to 1.1 % higher than the accuracies obtained with the 50k-SNP chip, and yielded the least biased predictions.

## Accuracy of genomic prediction using imputed whole-genome sequence data in white layers


M. Heidaritabar, M.P.L. Calus, H-J. Megens, A. Vereijken, M.A.M. Groenen, J.W.M. Bastiaansen

With sequence data, there was a small increase (~1%) in prediction accuracy over the 60 K genotypes.

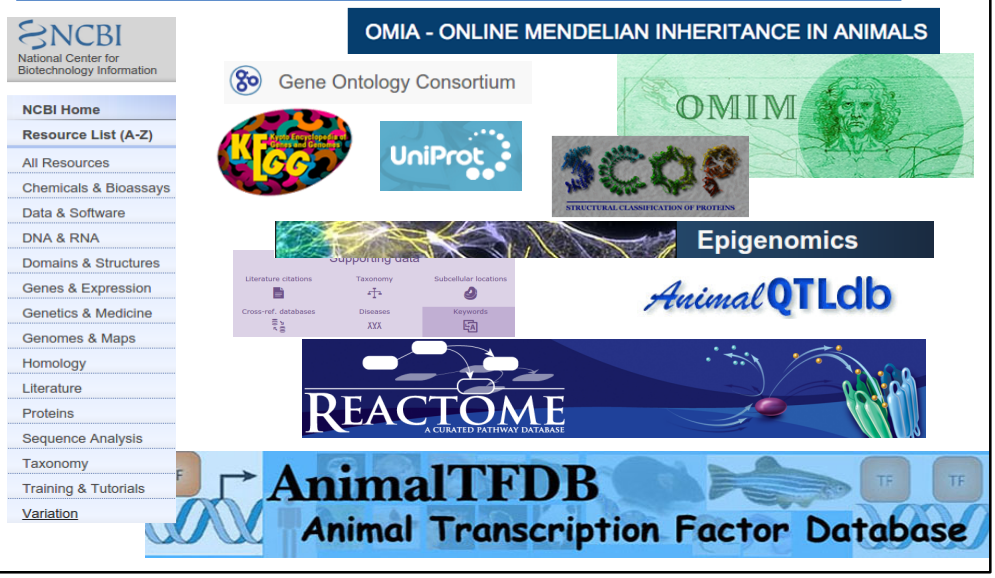
## In summary so far…

- ✓ Our simulations show that increasing SNP number only does not necessarily improve extant tools.
- ✓ Big data require not only efficient tools but also incorporation of several sources of information.
- ✓ **Conjecture:** It is by judicious use of biology that we are to make the most of sequence for prediction of genetic merit.

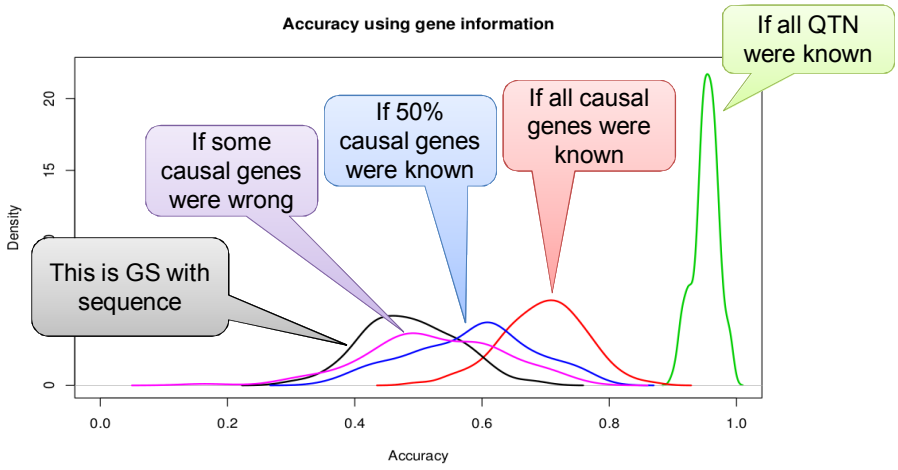
## ARE WE HEADING TOWARDS A BIOLOGICALLY INFORMED BREEDING?

- ❖ Seems to be a widespread feeling: Goddard, Hayes, Petersen, Simianer, Buckler, Groenen …
- ❖ An advantage of biology is that (some) information can be transferred across species.
- ❖ But recall highly heterogeneous sources: reactome, omim, …
- ❖ Multiple options: machine learning or parametric methods that combine multiple sources of information (lots of adhoclogy).
- ❖ So far results are mixed.

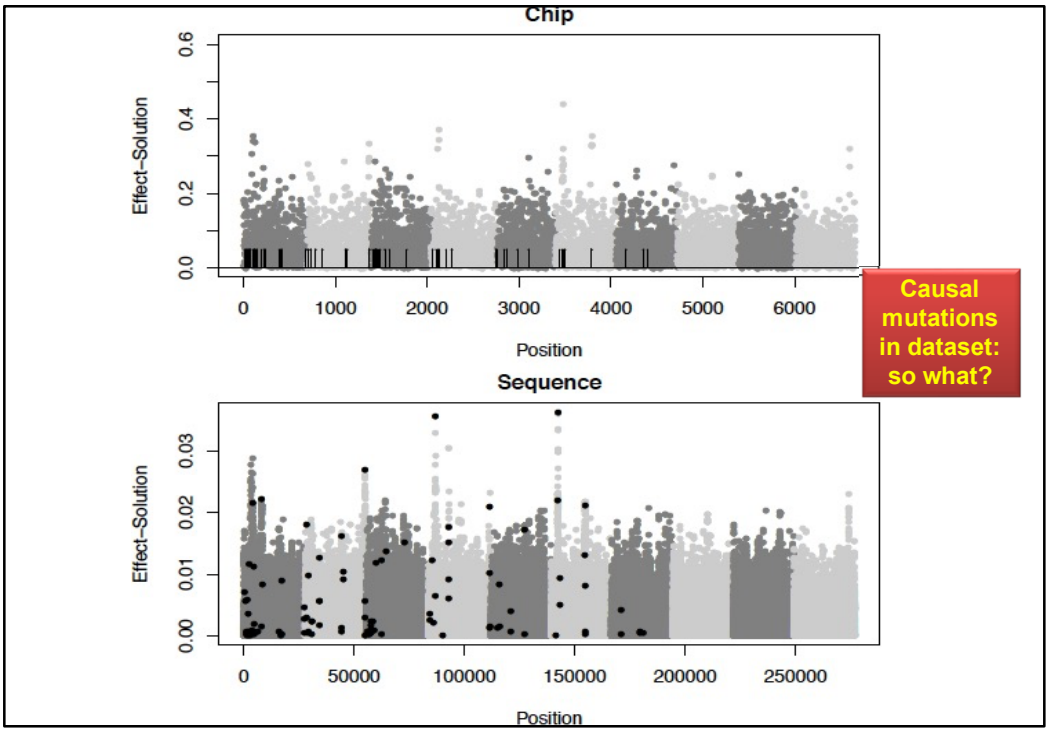
HOW MANY FUNCTIONAL KINDS?




# Biology can make a difference… if true

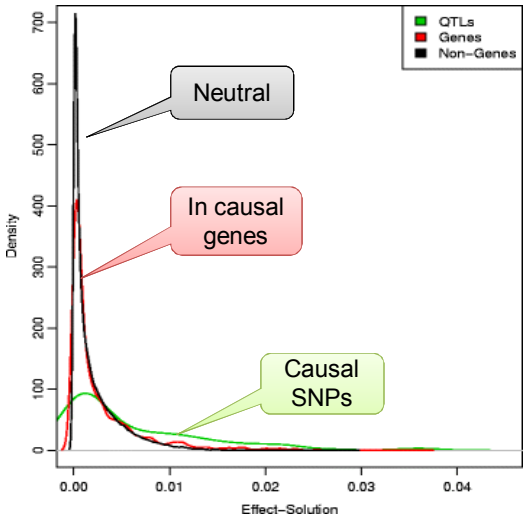



Pérez-Enciso et al (2015), Genet Sel Evol




**Causal mutations in dataset: so what?**

## Distribution of SNP estimated effects according to annotation



# In summary

- ✓ At least some GS methods such as GBLUP are robust to extreme genetic architectures of complex traits.
- ✓ There is a quick law of diminishing returns with increasing SNP density. Sequence and array data are operationally equivalent for genomic selection purposes.
- ✓ Unless accurate biological information is provided, it is unlikely an increase in accuracy over 4% with sequence (in agreement with experimental results).
- ✓ It is by using 'true' biological information that we will move forward, but data at hand may not help to determine what is this: use strong (and correct!) priors.
- ✓ You probably have heard that causal variants are in the data, do not trust it too much or so what?