Selección Genómica

Luis Garreta

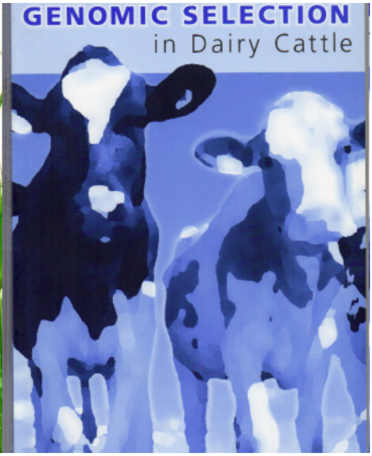
Electiva de Bioinformática
MAESTRÍA EN INFORMÁTICA BIOMÉDICA
Universidad del Bosque
Bogotá-Colombia

September 7, 2022

- Esta es una técnica que ha sido muy usada en mejoramiento de especies.
- Está muy avanzada en mejoramiento de animales y plantas.
- Sin embargo, en el campo humano todavía no hay “grandes” aplicaciones.
- Y no se ha dado, principalmente por cuestiones éticas:
  - mejorar la raza,
  - escoger el color de piel,
  - de ojos,
  - buscar seres libres de enfermedades, etc.
- Como vamos a ver, esta técnina no es más que inteligencia artificial aplicada
- Específicamente lo que se conoce como técnicas de aprendizaje de máquina.

Agrigenomics

The science of using genetics and computation (bioinformatics) for breeding both animals and crops, is the keystone of the global economy and health.



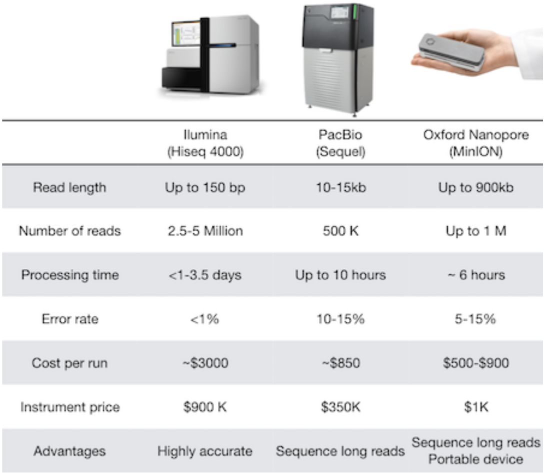
- Agrigenomics, se puede definir como la ciencia de acelerar el mejoramiento de especias utilizando información genómica y computacional (bioinformática).
- Y todo esto se da gracias a los diferentes desarrollos en:
  - genética,
  - computación,
  - bases de datos
- Advances in genetics, bioinformatics, and biotechnology present breeders with powerful tools to advance agriculture beyond the early days of these limited marker sets.
- Databases characterizing diversity within species are essential for driving breeding decisions.
- Sequence data and well-characterized marker sets can now be used to study phenotypes of interest.
- These data allow us to sequence new species, perform meta-analyses among large datasets, unravel complex traits, and empower our abilities in both marker-assisted selection (MAS) and GS.
- In the last few years, these technologies have revolutionized breeding of both livestock and crops in a field known as agrigenomics, the science of accelerating breeding decisions using whole genome information.
- Agrigenomics is enabling and revolutionizing how breeding decisions are made.

Algunos términos....

- Breeding: Mejoramiento, Crianza, Reproducción
- Breeding Value (BV): Valor de Cría, Mérito Genético,
- Estimated Breeding Value (EBV): Valor estimado de cría
- Genomic Estimated Breeding Value (gEBV): Valor genético, valor genético estimado de cría

Advances in DNA sequencing

Advances in DNA sequencing technologies now allow us to isolate DNA (or RNA) from multiple sample types, amplify and sequence regions of the genome, and sequence whole genomes.



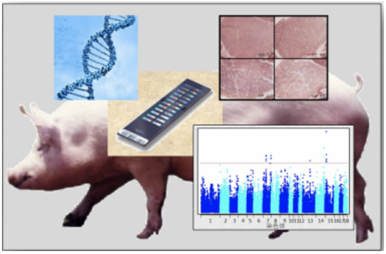
	Ilumina (Hiseq 4000)	PacBio (Sequel)	Oxford Nanopore (MinION)
Read length	Up to 150 bp	10-15kb	Up to 900kb
Number of reads	2.5-5 Million	500 K	Up to 1 M
Processing time	<1-3.5 days	Up to 10 hours	~ 6 hours
Error rate	<1%	10-15%	5-15%
Cost per run	~$3000	~$850	$500-$900
Instrument price	$900 K	$350K	$1K
Advantages	Highly accurate	Sequence long reads	Sequence long reads Portable device

Bueno y todo esto se da por los ultimos avances en genómica:

- Ahora es más fácil y barato sequenciar genomas
- Existen diversas technologias
- Aquí vemos algunas especies de plantas, junto con una aproximación del tamaño de sus genomas,y la tecnología usada para secuenciarla (los puntos de color)
- Vemos la gran diversidad de tecnologías de secuenciación usada, algunas de ellas:
  - Roche e Illumina, que son como las más tradicionales o conocidas.
  - Hasta las nuevas, más baratas y fáciles de usar, como las de nanoporo.
- Además, existen grandes repositorios o BD de información genómica, por ejemplo
  - BD de secuencias de genes, de proteínas
  - BD de marcadores mapeados o asociados a genes o rasgos (fenotipos) de interés:
    - Rasgos asociados a desempeño, como producción de leche o grosor del tallo
  - Rasgos asociados a enfermedades: gota, polilla.

SNP chips

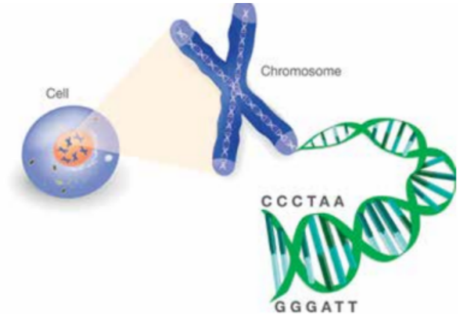
- Identification of the **genomic regions** or **markers**
- Markers responsible for economic traits such as:
  - meat quality
  - tuber shape
  - eye deep
- Type of studies:
  - GWAS: Genome Wide Association Studies
  - GWAS: Few genes for a trait
  - GS: Genomic Selection:
  - GS:Multiple genes for a trait



- E incluso ya no es necesario secuenciar todo un genoma.
- Sino que se consiguen ya arreglos o chips con información muy específica de los **puntos de interés** de ese genoma o lo que llamaremos de aquí en adelante como **marcadores**.
- Por ejemplo, se consiguen chips con las variaciones del genoma relacionados con un conjunto de características específicas, como:
  - la calidad de la carne
  - o, relacionados con una enfermedad, por ejemplo la gota en la papa.
- Con estos chips, es mucho más fácil y económico realzar estudios que permitan ya sea:
  - Asociar un único marcador o gene a una característica, por ejemplo el color de los ojos (características cualitativas)
  - O dar una medida (merito genético) de que un individuo vaya a desarrollar una característica en el futuro, por ejemplo vaya a ser propenso a una enfermedad o vaya a ser buen productor de leche o de muchos frutos.
- El primer caso, pocos genes asociados a una característica se analizan con el enfoque de GWAS, que ya hablamos de el.
- El segundo caso, muchos genes asociados una característica se analiza con predicción genómica, o lo que se conoce en plantas y animales con selección genómica o GS, que es lo que vamos a ver hoy.

DNA has the form of a double helix

- DNA has the form of a double helix.
- The 2 complementary strands of DNA are sequences of nucleotides that carry 1 of 4 possible nitrogen-containing bases.

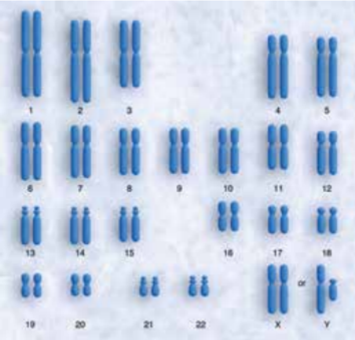


Algunos fundamentos:

- El material biológico principal de todos los organismos es el ADN,
- Que está compuesto de 4 base nucleótidas,
- Y está depositado en los cromosomas

Chromosomes

- The genome is packed and organized in structures called chromosomes.
- Most animals inherit 1 chromosome from each parent and are called **diploids** (i.e. they have 1 homologous pair of chromosomes).
- Some animals, and many plants, have multiple homologous pairs of chromosomes and are defined as **polyploid**. .

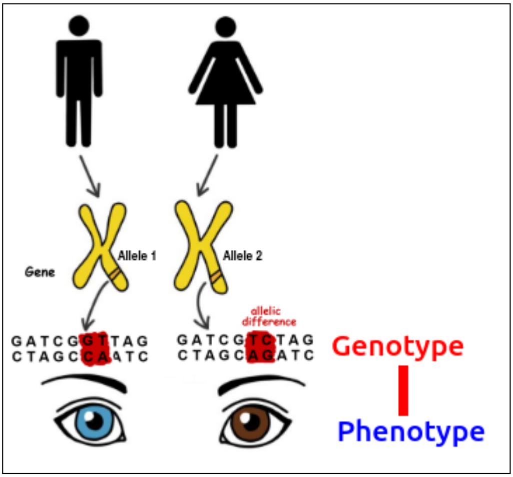


Organization of the diploid human genome in 23 condensed structures called chromosomes

- En los humanos tenemos 23 pares de cromosomas,
  - 1 heredado del padre
  - Otro heredado de la madre
- Es decir, somos seres diploides.
- Sin embargo, algunos animales y muchas plantas hereden más de dos cromosomas y se conocen como poliploides:
  - Ejemplo, la caña es hexaploide
  - Existen variedades de papa diploide y otras tetraploides (La variedan andigena es tetraploide)
  - Algunos peces son poliploides.

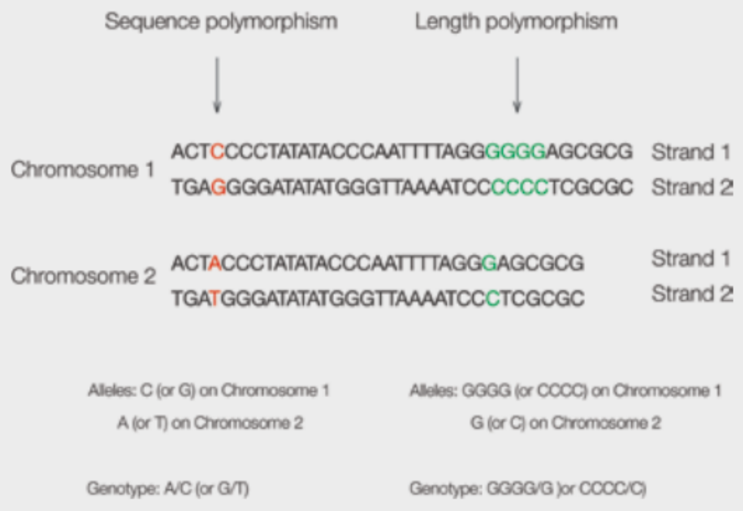
Alleles

- For each gene, diploid individuals will each have **two alleles**.
- Alleles are alternative **forms of the same gene** that can differ in 1 or more variations in the 4 bases A, G, C, or T.
- Alleles create **diversity**.
- The combination of these alleles is what defines a **genotype** for an individual.
- Genotype and other external variable determines largely the **phenotype** of an individual-

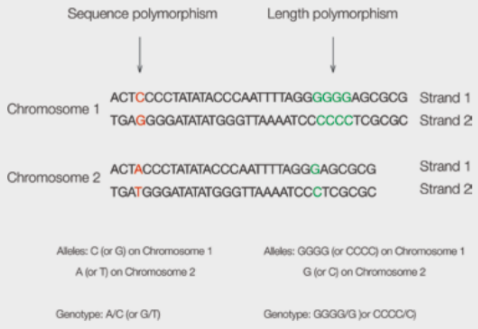


- Al ser organismos diploides,
- Por cada padre heredamos un gen, o mejor una variación del gen conocidos como alelos.
- Estos alelos pueden variar en una o más bases y esto es lo que brinda variabilidad y que no haya dos seres iguales
- Por ejemplo, un gen puede ser el determinate del color de los ojos,
- Y uno de los dos alelos de ese gen, puede determinar si tenemos el color de ojos de nuestro padre, o de nuestra madre.
- Todo este conjunto de alelos es lo que define el genotipo de un individuo
- Y este genotipo más un conjunto de evento externos (ambiente) es lo que produce lo observable, es decir su fenotipo.

Genetic Variation and Polymorphisms

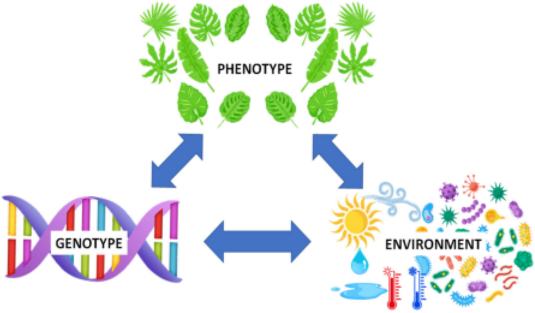

Sequence polymorphism
Length polymorphism
Chromosome 1
ACTCCCCTATATACCCAATTTTAGGGGGGAGCGCG
TGAGGGGATATATGGGTTAAAATCCCCCCTCGCGC
Strand 1
Strand 2
Chromosome 2
ACTACCCTATATACCCAATTTTAGGGAGCGCG
TGATGGGATATATGGGTTAAAATCCCTCGCGC
Strand 1
Strand 2
Alleles: C (or G) on Chromosome 1
A (or T) on Chromosome 2
Alleles: GGGG (or CCCC) on Chromosome 1
G (or C) on Chromosome 2
Genotype: A/C (or G/T)
Genotype: GGGG/G)or CCCC/C)

- Ahora, la mayor parte del genoma es el mismo entre seres de la misma especie (>90%)
- Pero existen algunas posiciones en las cuales difiere:
  - Y se conocen como polimorfismos,
- De los cuales los más importantes son los de un sólo nucleótido SNPs.
- La mayoría de estos son neutrales, es decir no afectan el fenotipo.
- Sin embargo, algunos si tienen un efecto observable:
  - Que pueder benefico, e.g. Mayor rendimiento (leche, carne, trabajo)
  - O a veces perjudicial, e.g. dar origen a una enfermedad


Sequence polymorphism
Length polymorphism
Chromosome 1
ACTCCCCTATATACCCAATTTTAGGGGGGAGCGCG
TGAGGGGATATATGGGTTAAAATCCCCCCTCGCGC
Strand 1
Strand 2
Chromosome 2
ACTACCCTATATACCCAATTTTAGGGAGCGCG
TGATGGGATATATGGGTTAAAATCCCTCGCGC
Strand 1
Strand 2
Alleles: C (or G) on Chromosome 1
A (or T) on Chromosome 2
Alleles: GGGG (or CCCC) on Chromosome 1
G (or C) on Chromosome 2
Genotype: A/C (or G/T)
Genotype: GGGG/G)or CCCC/C)

Phenotype = Gentype + Environment

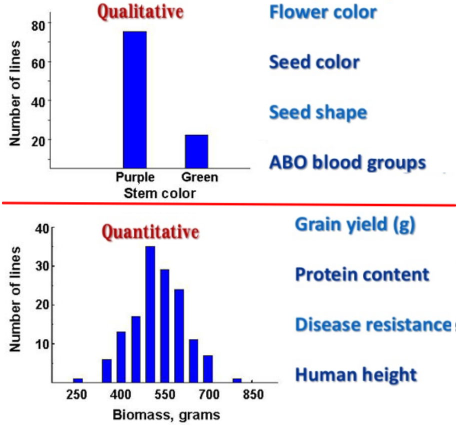
- Both animals and plants evolve in complex environments.
- Gradually acquiring the ability to cope with elements in that environment such as predators, adverse soil conditions, or adverse climates.

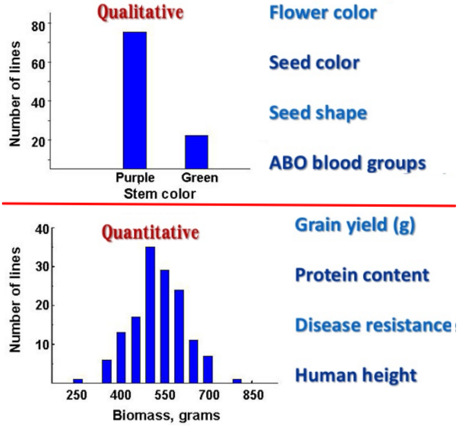


- Los seres son producto de sus genes y de los factores ambientales que los afecten:
- Se puede tener un gen que se asocie a una enfermedad, pero nunca se expresa.
- Sin embargo, si le brinda los factores ambientles (e.g. es fumador),
- Además, si la enfermedad se ha repetido siempre en su familia, heredabilidad, entonces hay mayor probabilidad de expresarse

Types of Phenotypes (traits)

- There are many **traits** where one change in a DNA location produce a qualitative change, e.g Eyes color
- Most (but not all) of the traits, or phenotypes, that are desirable in agriculture are defined as complex quantitative traits.

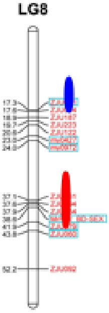




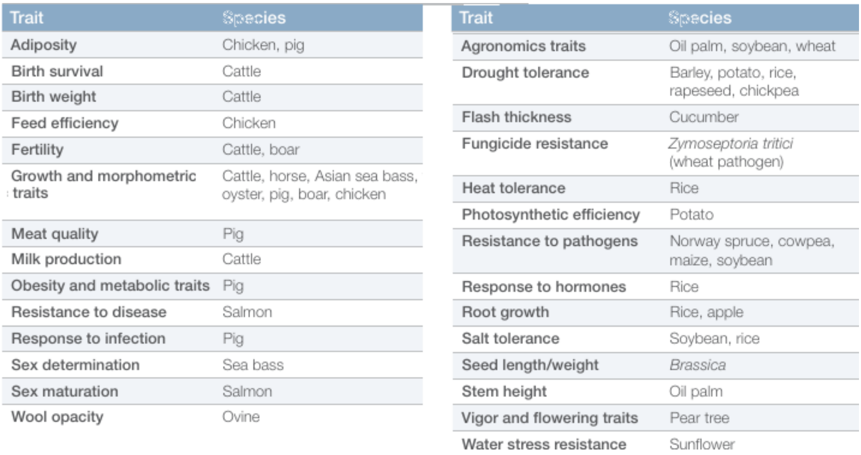
- Cuando hablamos de traits, nos referimos a fenotipos o caracteristicas observables.
- Estas características pueden ser de dos tipos:
  - Cualitativas (binarias)
  - Cuantitativas (valor)
- Generalmente las características cualitativas están relacionadas con pocos genes.
- Y las cuantitativas están relacionas con muchos genes.
- Cuando alguna región del genoma se encuentra correlacionada con una c

Quantitative Trait Loci (QTLs)

- When **locations in the genome** are found to be correlated to these traits, we call these **quantitative trait loc**i (QTL)
- To give an example, dairy traits were originally thought to be regulated by 50–100 genes, but they are now known to be regulated by 1000–2000 genes.



Some Phenotypic traits regulated by QTLs



Trait	Species
Adiposity	Chicken, pig
Birth survival	Cattle
Birth weight	Cattle
Feed efficiency	Chicken
Fertility	Cattle, boar
Growth and morphometric traits	Cattle, horse, Asian sea bass, oyster, pig, boar, chicken
Meat quality	Pig
Milk production	Cattle
Obesity and metabolic traits	Pig
Resistance to disease	Salmon
Response to infection	Pig
Sex determination	Sea bass
Sex maturation	Salmon
Wool opacity	Ovine

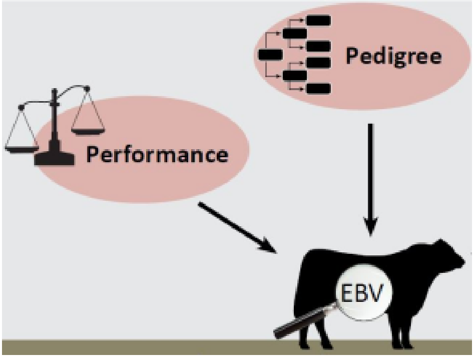
Trait	Species
Agronomics traits	Oil palm, soybean, wheat
Drought tolerance	Barley, potato, rice, rapeseed, chickpea
Flash thickness	Cucumber
Fungicide resistance	*Zymoseptoria tritici* (wheat pathogen)
Heat tolerance	Rice
Photosynthetic efficiency	Potato
Resistance to pathogens	Norway spruce, cowpea, maize, soybean
Response to hormones	Rice
Root growth	Rice, apple
Salt tolerance	Soybean, rice
Seed length/weight	*Brassica*
Stem height	Oil palm
Vigor and flowering traits	Pear tree
Water stress resistance	Sunflower

Genomic selection in agriculture: An overview of recent publications featuring Illumina® technology

Breeding

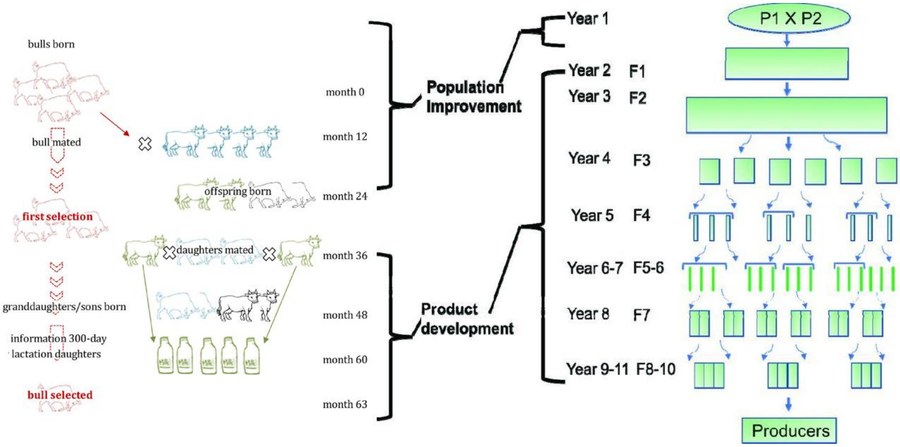
Estimated Breeding Value (EBV) (Traditional)

- Selection for breeding has historically been made using estimated breeding values (EBV), without identifying genes involved in phenotypes.
- EBVs were simply estimated from the study of pedigrees and phenotypic records with the knowledge of the heritability of each trait



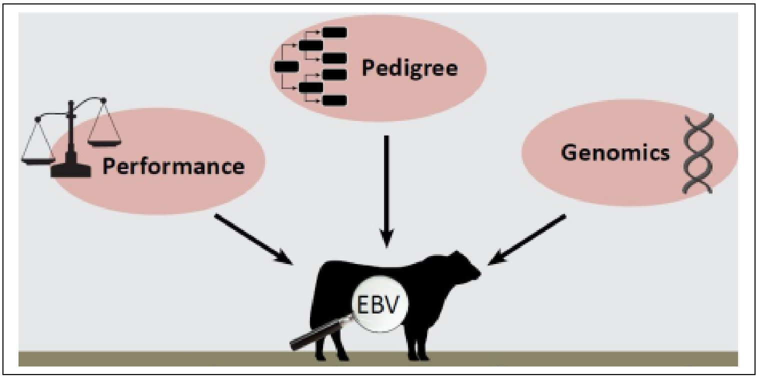
Estimated Breeding Value (EBV) (Traditional)

Traditional Breeding Problems



Genomic Estimated breeding values (EBV)

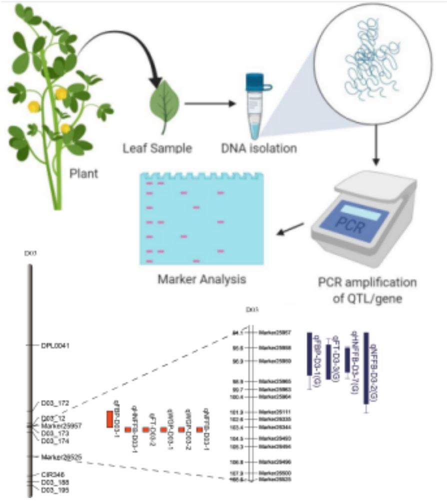
However, the efficiency of this method decreases as its use is expanded to traits that are difficult to measure, have low heritability, or can be measured only after several years and/or generations.



For this reason, the identification and knowledge of the genes underlying these traits in animals and plants is of great value in agrigenomics.

Marker-Assisted Selection (MAS)

- In MAS, breeders use a marker that has been correlated to a trait of interest to select the genetic determinant, or determinants, of a trait indirectly.
- Since the early 1990s, efforts to improve these methods have been intensive, but their implementation has been limited and, therefore, overall genetic improvement of bred species has been limited.



Genomic Selection

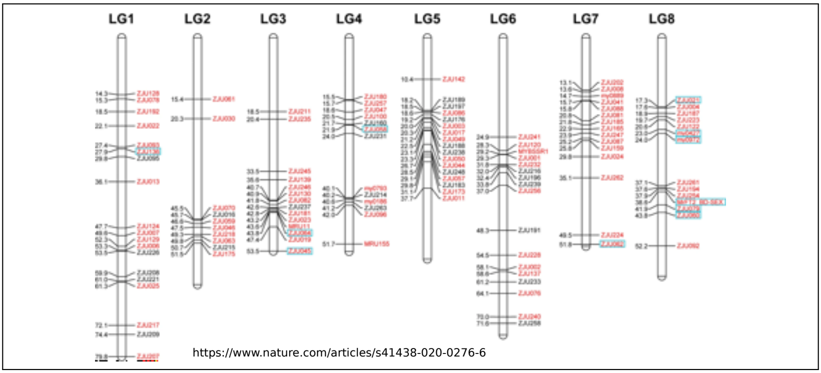
What is Genomic Selection?

Genomic Selection or Genomic Prediction is a new tool that can predict a plant/animal genetic merit based on scoring DNA markers such as single nucleotide polymorphisms (SNP).

GS can predict how an animal's progeny will perform before the traits are measured or even right after the birth without needing any other information (such as pedigree).

Genomic Selection (GS)

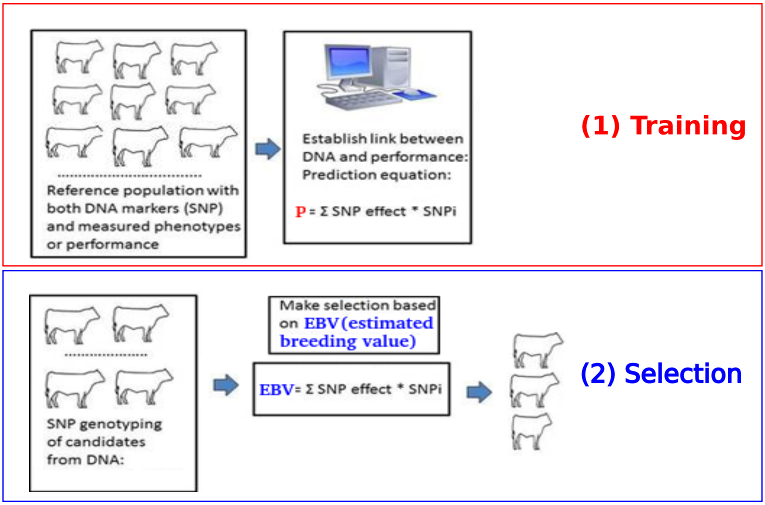
GS is based on the principle that information from a large number of markers, without knowledge of where genes are located, can be used to estimate breeding values



https://www.nature.com/articles/s41438-020-0276-6

It is similar to conventional MAS in that genetic information is used, but instead of explicitly introducing a single trait (as in MAS), multiple favorable traits are implicitly taken and selected from markers common to the training population.

How does GS work?



Reference population with both DNA markers (SNP) and measured phenotypes or performance

Establish link between DNA and performance: Prediction equation:

P = Σ SNP effect * SNPi

(1) Training

SNP genotyping of candidates from DNA:

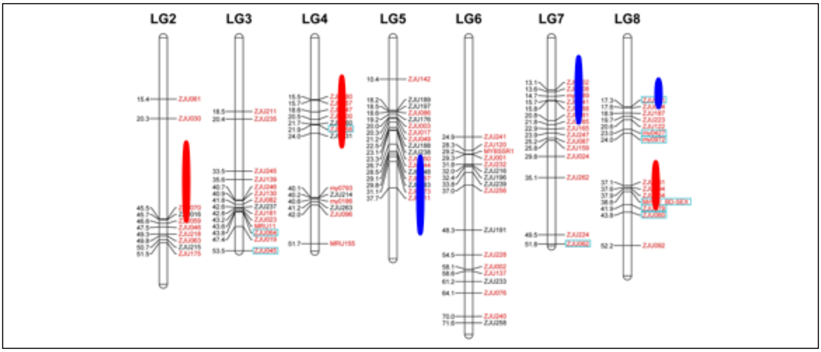
Make selection based on EBV (estimated breeding value)

EBV= Σ SNP effect * SNPi

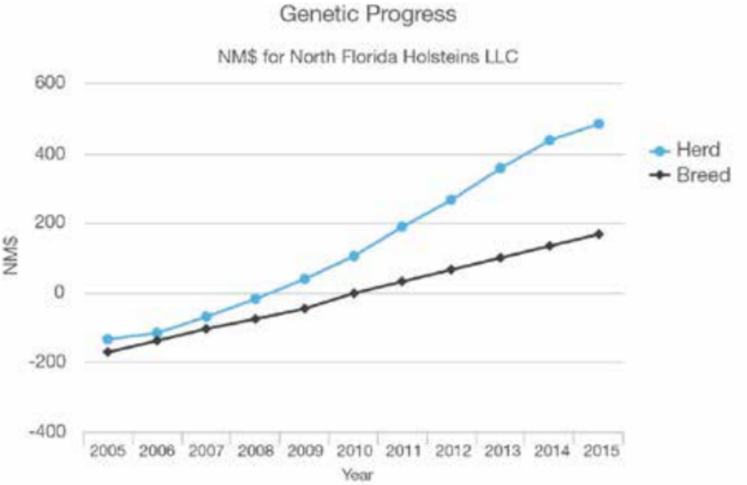
(2) Selection

QTLs are in Linkage Disequilibrium with Markers

- In GS, all QTLs are in linkage disequilibrium (LD) with at least 1 marker.
- In GS, EBVs (also called gEBVs) are calculated from the cumulative effect of large numbers of genetic markers covering the whole genome, and these values are used to score new potential breeding candidates.



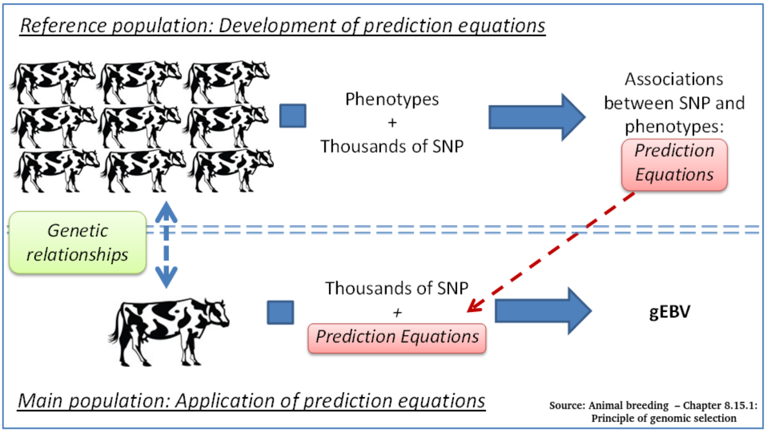
The Use of Genomics and the Return on Investment (ROI)



As evident from the trend lines, the appropriate implementation of the genomic technologies in 2008 resulted into a significantly faster genetic progress of the North Florida Holsteins LLC compared to the breed's average.

Figure 7. Comparison of genetic progress between the North Florida Holstein LLC herd and the breed average (kindly provided by Don Bennink, North Florida Holsteins LLC).

Process



Reference population: Development of prediction equations

Phenotypes + Thousands of SNP

Associations between SNP and phenotypes: Prediction Equations

Genetic relationships

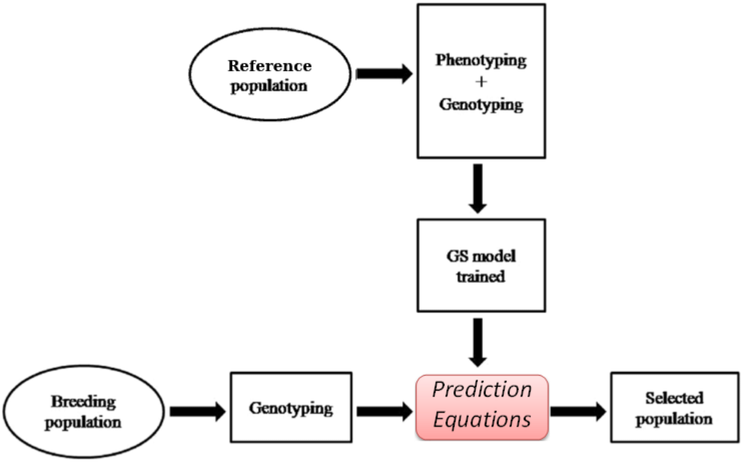
Thousands of SNP + Prediction Equations

gEBV

Main population: Application of prediction equations

Source: Animal breeding – Chapter 8.15.1: Principle of genomic selection

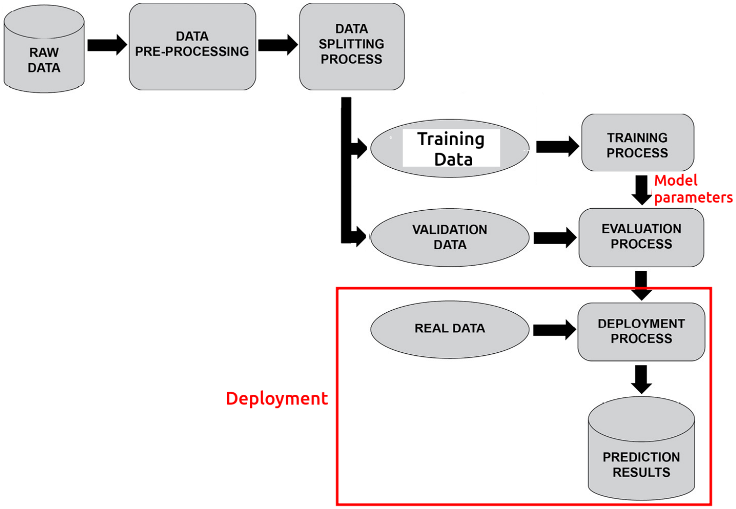
Workflow of GS



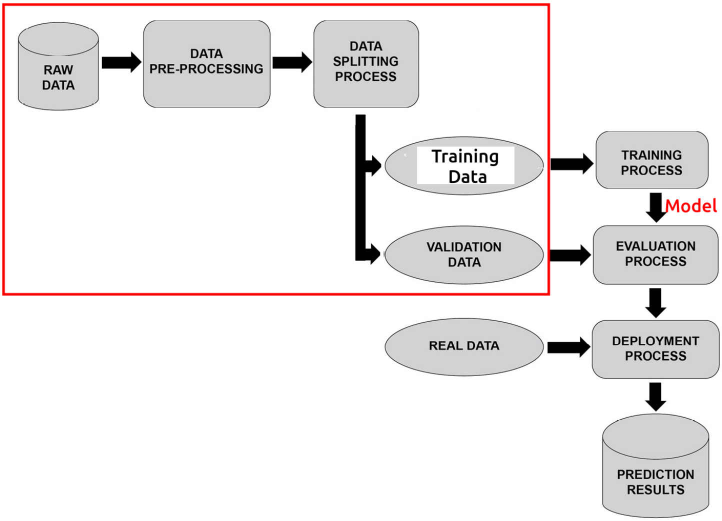
The process of GS includes the following steps, also

1. Collection and documentation of phenotype and genotype data for each marker of interest in the reference sample (or discovery dataset)
2. Representation of each genotype by a variable, x, that can have 3 values:
   - 0 (homozygote for one allele),
   - 1 (heterozygote), and
   - 2 (homozygote for the second allele)
3. Statistical analysis on a reference population to estimate the effect of each marker (w) on the phenotype
4. Generation of a prediction equation for the gEBV that combines all the marker genotypes with their effects on the predictive value of each animal (see below)
5. Application of the prediction equation to a group of animals for which genotypes (but not phenotypes) are available.
6. Breeding values are estimated and the best animals are selected for breeding

Machine Learning Process

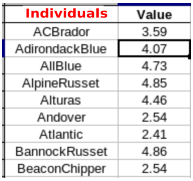


Data


RAW DATA
DATA PRE-PROCESSING
DATA SPLITTING PROCESS
Training Data
VALIDATION DATA
TRAINING PROCESS
Model
EVALUATION PROCESS
REAL DATA
DEPLOYMENT PROCESS
PREDICTION RESULTS

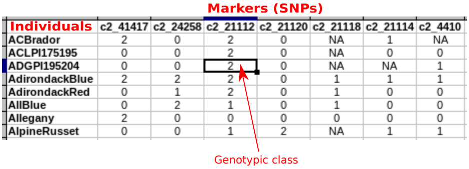
Data: Genotypes and Phenotypes (traits)

- Documentation of phenotypes



Individuals	Value
ACBrador	3.59
AdirondackBlue	4.07
AllBlue	4.73
AlpineRusset	4.85
Alturas	4.46
Andover	2.54
Atlantic	2.41
BannockRusset	4.86
BeaconChipper	2.54

- Documentation of genotypes for each marker of interest

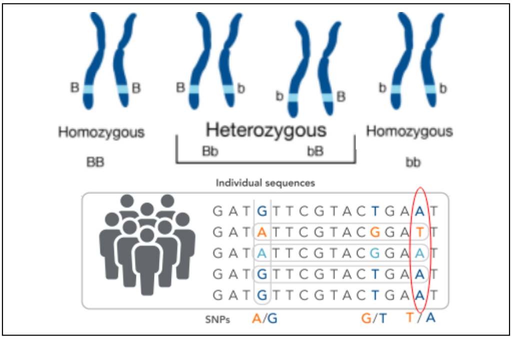


Markers (SNPs)

Individuals	c2_41417	c2_24258	c2_21112	c2_21120	c2_21118	c2_21114	c2_4410
ACBrador	2	0	2	0	NA	1	NA
ACLPI175195	0	0	2	0	NA	0	0
ADGPI195204	0	0	2	0	NA	NA	1
AdirondackBlue	2	2	2	0	1	1	1
AdirondackRed	0	1	2	0	1	0	0
AllBlue	0	2	1	0	1	0	0
Allegany	2	0	0	0	0	0	0
AlpineRusset	0	0	1	2	NA	1	1

Genotypic class

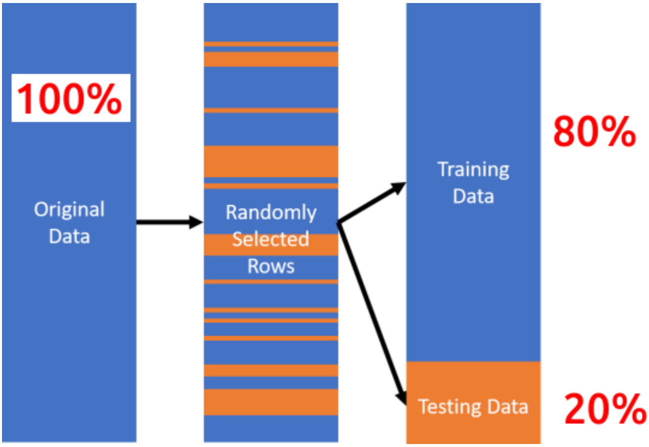
Data: Genotype coding schemes



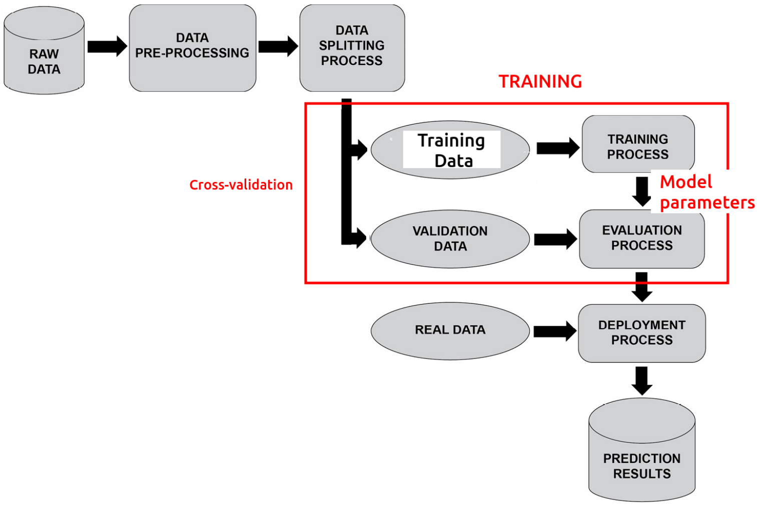
G A T G T T C G T A C T G A A T
G A T A T T C G T A C G G A T T
G A T A T T C G T A C G G A A T
G A T G T T C G T A C T G A A T
G A T G T T C G T A C T G A A T

Minor Allele	BB	Bb	bB	bb
Binary	00	01	10	11
Numeric	0	1		2
ACGT	AA	AT	TA	TT

Data: Splitting into Training and Testing datasets



Training


RAW DATA
DATA PRE-PROCESSING
DATA SPLITTING PROCESS
TRAINING
Cross-validation
Training Data
TRAINING PROCESS
Model parameters
VALIDATION DATA
EVALUATION PROCESS
REAL DATA
DEPLOYMENT PROCESS
PREDICTION RESULTS

Training: Simple model

A simple but frequently used genetic model is that the phenotypic value of an individual (P) is expressed as the summation of the genetic value (G) and the residual environmental effect (E):

$$P = G + E$$

- P : Phenotype
- G: Genotype
- E: Environment

Training: Linear model for a single marker

$$y_i = \beta_0 + x_{1i}\beta_1 + \varepsilon_i$$

- Y_i: Observed phenotype individual i
- β_o: Fixed effect
- X_{1i}: Genotype of the marker 1 of the ith individual
- ε_i: Error term

Training: Linear model for *M* marker

- General equation for M markers

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 \ldots + x_{Mi}\beta_M + \varepsilon_i$$
$$= \Sigma_{j=0}^{M} x_{ji}\beta_j + \varepsilon_i$$

- Minimization term:

$$E = \Sigma_i \,(y_i - \beta_0 - x_{1i}\,\beta_1)^2$$

Training: ‘large p, small n problem’

- p >> n
  - n: Number of individuals (observations)
  - p: Number of markers (variables)

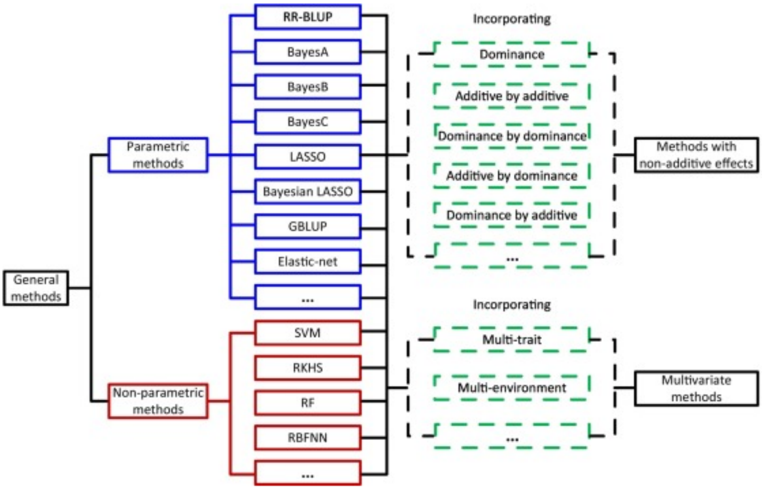
- Over-fitting:
  - The linear model only works well in the training population

- A penalty term is introduced:

$$E = \Sigma_{i=1}^{N} (y_i - \Sigma_{j=0}^{M} x_{ji} \beta_j)^2 + \lambda \Sigma_{j=0}^{M} |\beta_j|^q$$

λ: is a parameter that controls the effects of the penalty term.

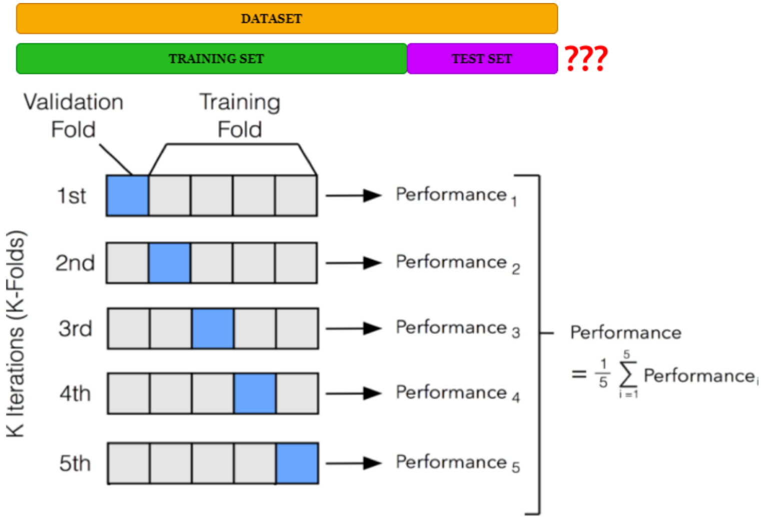
Training: Genomic Selection Methods



Source: Genomic selection methods for crop improvement: Current status and prospects

In recent years, many methods have advanced GS, including general methods and their extensions. General GS methods are based on additive models, and their accuracies may be different because they vary in their assumptions and algorithms with respect to the variances of complex traits. Incorporating non-additive effects or multiple variates, the general methods can be extended. The principles and characteristics of current popular GS methods are presented in this section.

Training: Evaluating a Model by Cross-Validation



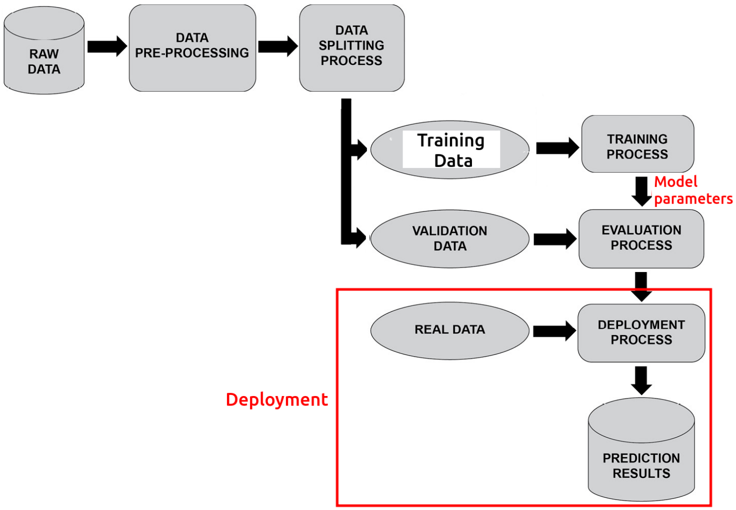
$$= \frac{1}{5}\sum_{i=1}^{5} \text{Performance}_i$$

Cross Validation:
It is a statistical method to estimate the sill of machine learning models. It is primarily used in applied machine learning to estimate the skill of the model on unseen data.

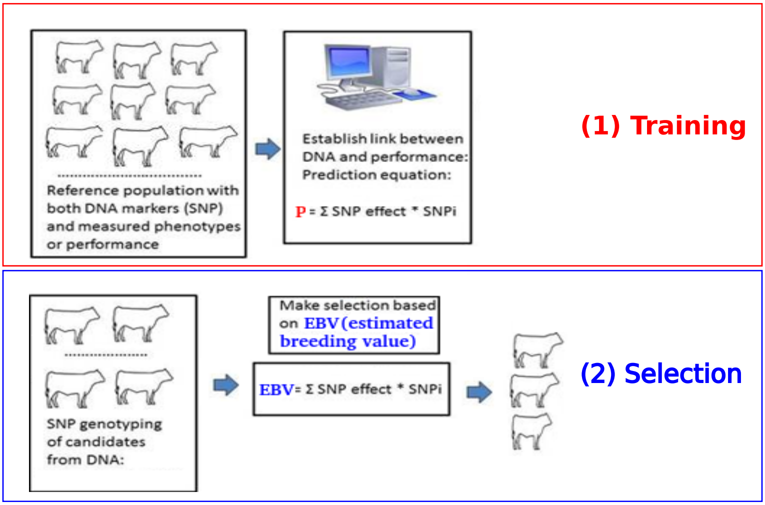
The general procedure is as follows:

- Shuffle the dataset randomly.
- Split the dataset into k groups
- For each unique group:
  - Take the group as a hold out or test data set
  - Take the remaining groups as a training data set
  - Fit a model on the training set and evaluate it on the test set
  - Retain the evaluation score and discard the model
- Summarize the skill of the model using the sample of model evaluation scores

Deployment


RAW DATA
DATA PRE-PROCESSING
DATA SPLITTING PROCESS
Training Data
TRAINING PROCESS
Model parameters
VALIDATION DATA
EVALUATION PROCESS
Deployment
REAL DATA
DEPLOYMENT PROCESS
PREDICTION RESULTS

Selection



Reference population with both DNA markers (SNP) and measured phenotypes or performance

Establish link between DNA and performance: Prediction equation:

P = Σ SNP effect * SNPi

(1) Training

SNP genotyping of candidates from DNA:

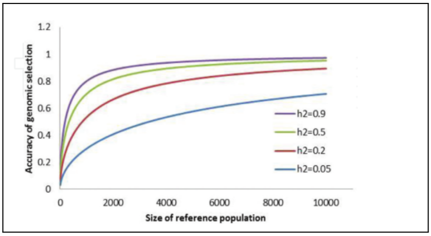
Make selection based on EBV (estimated breeding value)

EBV= Σ SNP effect * SNPi

(2) Selection

Factors responsible for the estimation accuracy of GS

- Population size
- Heritability
- Linkage Disequilibrium



- Thus, establishing a GS model based on a training population does not work in a breeding population if the genetic structures of both populations are different.

- Indeed, in most reported GS studies, the training populations were assumed to consist of ancestors or randomly selected individuals in a breeding population