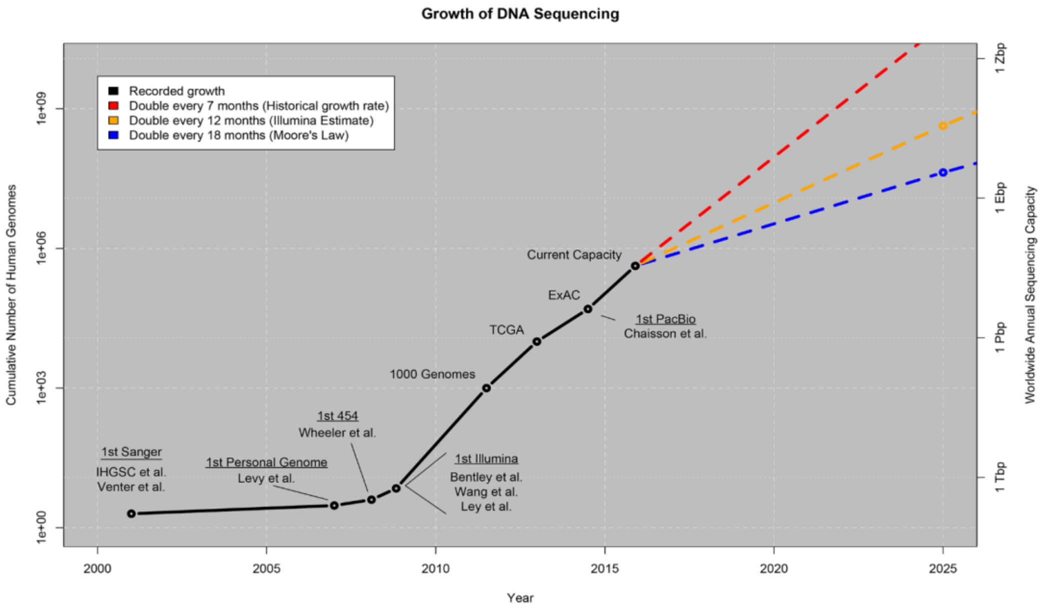
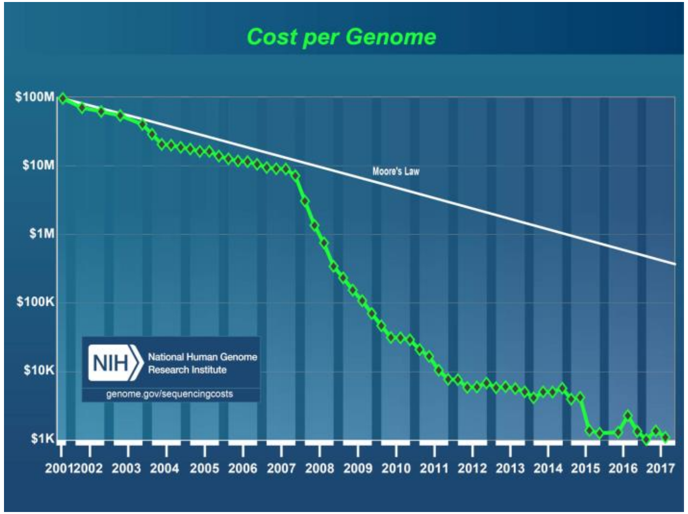
CORTEVA™
agriscience

Opportunities and challenges on the use of genomic information in plant breeding

Alencar Xavier
Corteva Biostatistics, alencar.xavier@corteva.com
Purdue University, xaviera@purdue.edu
https://alenxav.wixsite.com/home

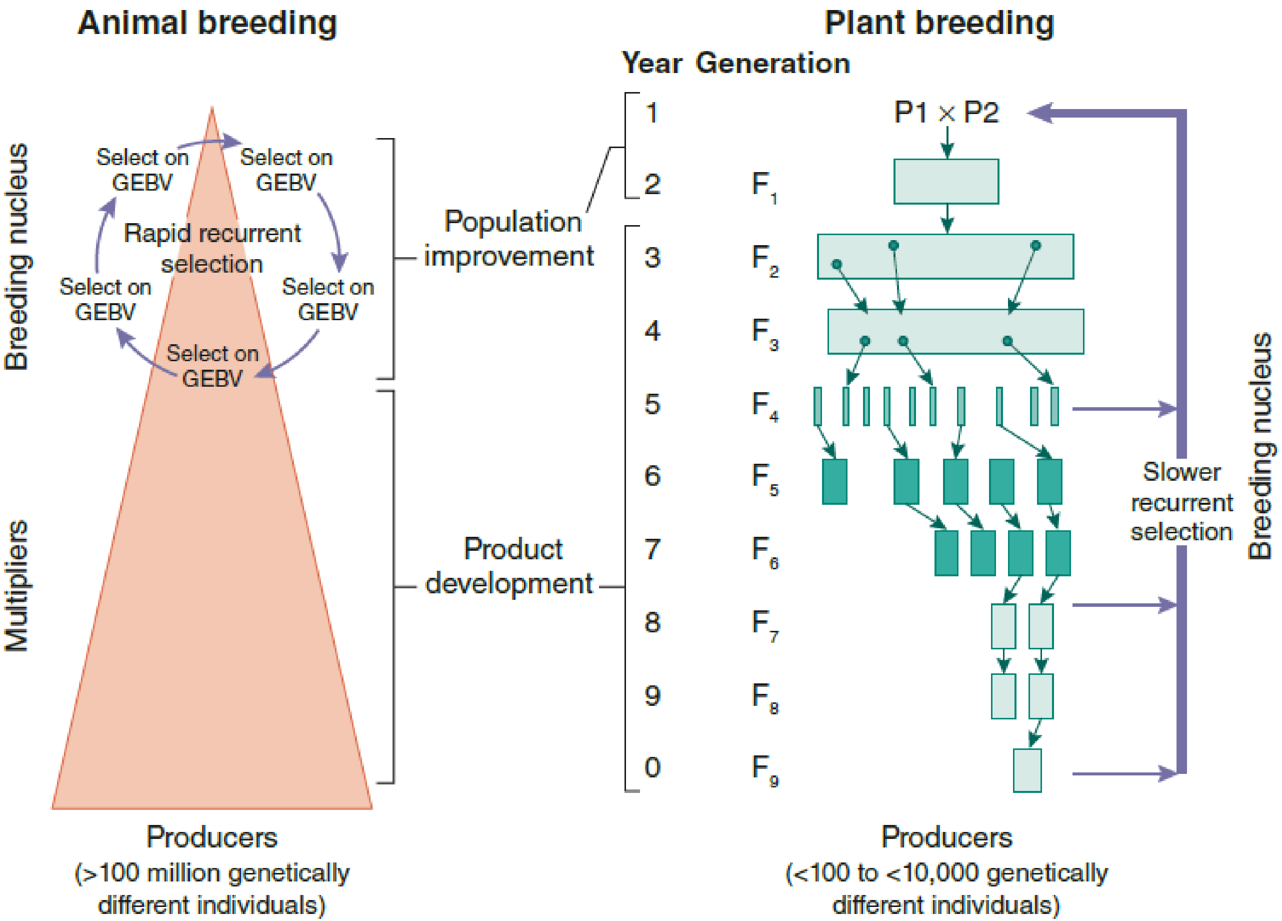


Stephens, Z. D.et al. (2015). Big data: astronomical or genomical? *PLoS biology*, *13*(7), e1002195.



The Cost of Sequencing a Human Genome. NIH. https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/

BREEDING PIPELINE

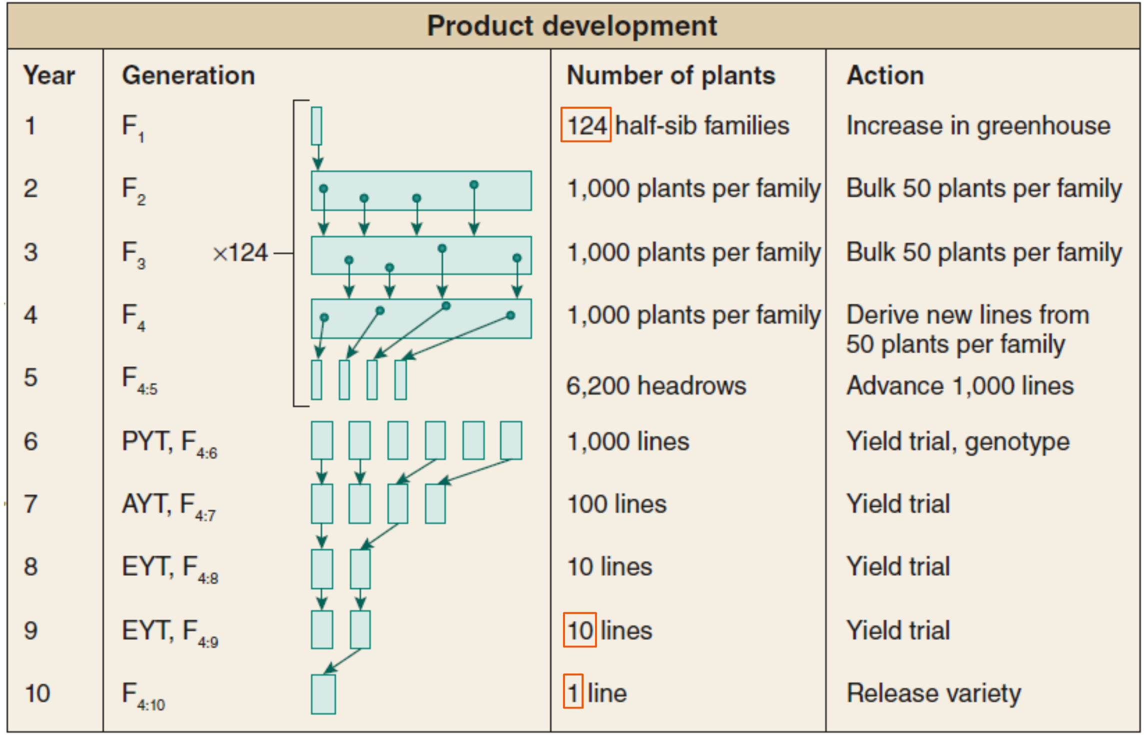


Hickey et al. (2017) *Nature genetics* 49(9):1297

Varietal Wheat

Example from one program in one geography

~120 crosses



Product development

Year	Generation	Number of plants	Action
1	F_1	124 half-sib families	Increase in greenhouse
2	F_2	1,000 plants per family	Bulk 50 plants per family
3	F_3 ×124	1,000 plants per family	Bulk 50 plants per family
4	F_4	1,000 plants per family	Derive new lines from 50 plants per family
5	$F_{4:5}$	6,200 headrows	Advance 1,000 lines
6	PYT, $F_{4:6}$	1,000 lines	Yield trial, genotype
7	AYT, $F_{4:7}$	100 lines	Yield trial
8	EYT, $F_{4:8}$	10 lines	Yield trial
9	EYT, $F_{4:9}$	10 lines	Yield trial
10	$F_{4:10}$	1 line	Release variety

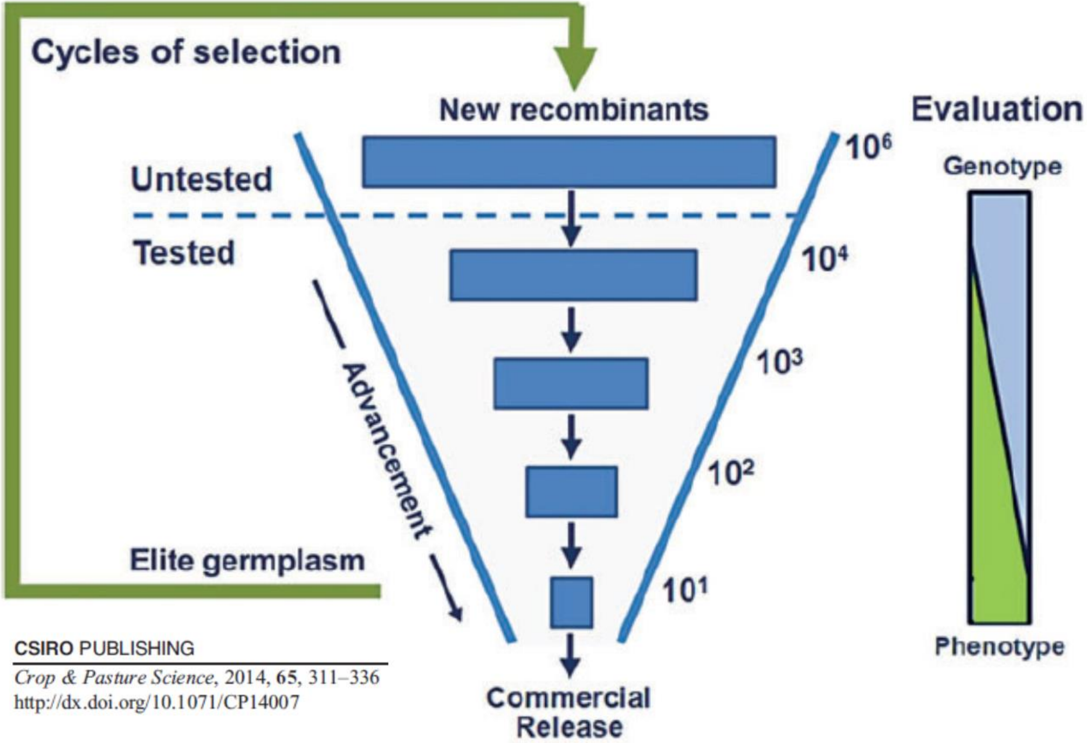
>120.000 F2s
(100% genotyped)

~10 parents
1-2 products

Hickey et al. (2017) *Nature genetics* 49(9):1297

Types of plant breeding

1. Varietal
2. Hybrid
3. Population
4. Clonal



CSIRO PUBLISHING

Crop & Pasture Science, 2014, **65**, 311–336
http://dx.doi.org/10.1071/CP14007

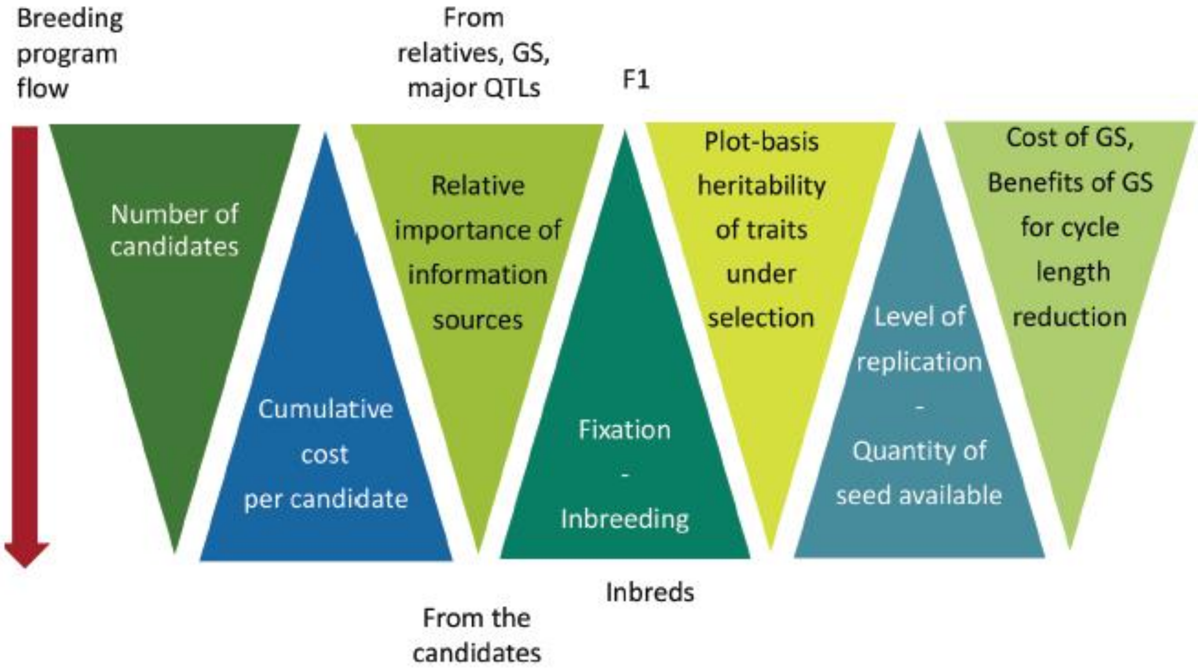
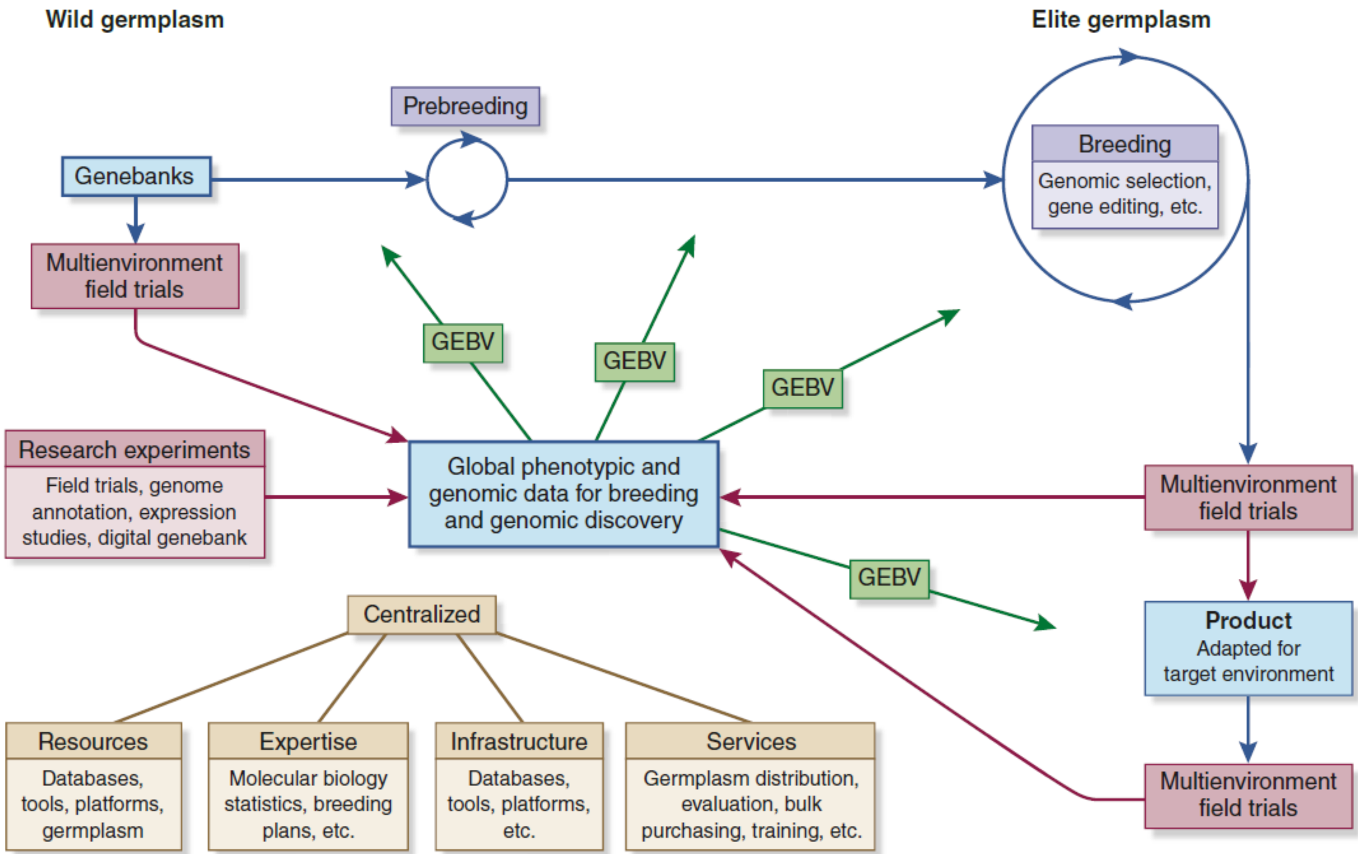

Breeding program flow
From relatives, GS, major QTLs
F1
Number of candidates
Cumulative cost per candidate
Relative importance of information sources
Fixation - Inbreeding
Plot-basis heritability of traits under selection
Level of replication - Quantity of seed available
Cost of GS, Benefits of GS for cycle length reduction
Inbreds
From the candidates

Figure 1. Key parameters and changes during a breeding cycle, to consider in implementing genomic selection (GS). The triangles indicate increase or decrease of the quantity considered. QTL, quantitative trait loci.

Heslot, N., Jannink, J. L., & Sorrells, M. E. (2015). Perspectives for genomic selection applications and research in plants. *Crop Science*, *55*(1), 1-12.



Wild germplasm
Elite germplasm
Prebreeding
Genebanks
Breeding
Genomic selection, gene editing, etc.
Multienvironment field trials
GEBV
GEBV
GEBV
Research experiments
Field trials, genome annotation, expression studies, digital genebank
Global phenotypic and genomic data for breeding and genomic discovery
Multienvironment field trials
GEBV
Centralized
Product
Adapted for target environment
Resources
Databases, tools, platforms, germplasm
Expertise
Molecular biology statistics, breeding plans, etc.
Infrastructure
Databases, tools, platforms, etc.
Services
Germplasm distribution, evaluation, bulk purchasing, training, etc.
Multienvironment field trials

Hickey et al. (2017) *Nature genetics* 49(9):1297

side note

Modeling genetic merit

Single-stage? Single-step?

Single-step
(Animal Breeding)

$$Phe = Env + Gen$$
$$y = Xb + Zu + e$$
$$u \sim N(0, H\sigma_a^2)$$
$$e \sim N(0, I\sigma_e^2)$$

$$H = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & G \end{bmatrix}$$

Single-step is an alternative to two-step analysis where pedigree EBV are degressed then fed to the genomic model

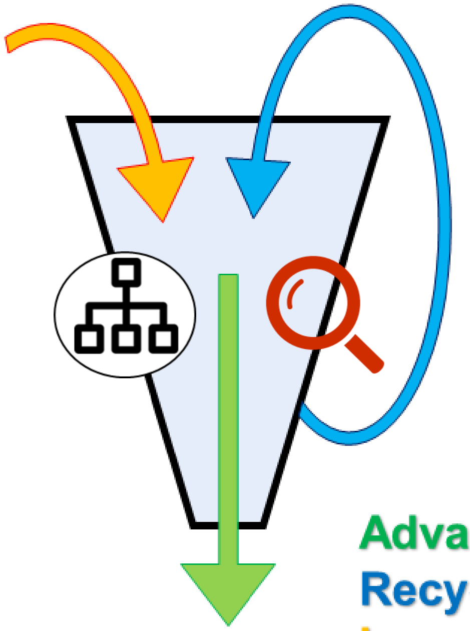
Single-stage
(Plant Breeding)

$$Phe = Env + Gen + G \times E$$
$$y = \mu + Xb + Zu + Wa + e$$
$$b \sim N\left(0, I\sigma_b^2\right)$$
$$u \sim N\left(0, G\sigma_a^2\right)$$
$$a \sim N\left(0, \begin{matrix} G\sigma_{a1}^2 & 0 & 0 \\ 0 & G\sigma_{a2}^2 & 0 \\ 0 & 0 & G\sigma_{a(\dots)}^2 \end{matrix}\right)$$
$$e \sim N\left(0, \begin{matrix} R_1\sigma_{e1}^2 & 0 & 0 \\ 0 & R_2\sigma_{e2}^2 & 0 \\ 0 & 0 & R_3\sigma_{e(\dots)}^2 \end{matrix}\right)$$
$$R_i = Z_{e_i}\left(\rho_{c_i} \otimes \rho_{r_i}\right)Z'_{e_i}$$

Single-stage is an alternative to two-stage analysis where spatially adjusted BLUEs are computed locally then fed to the (G)BLUP model

APPLICATIONS

Where is genomic information used for breeding?



Advancement

Recycling

Incorporation

Assessment

Classification

Where is genomic information used for breeding?

- Germplasm classification (***PCA, Clustering, Unsupervised ML, F_{ST}***)
- Incorporation (***GWAS, haplotype analysis***)
- Genomic selection (***BayesABC, Supervised ML, etc.***)
- Recycling (***Simulation and optimization***)
- Quantitative assessment (***Variance component analysis***)

Where is genomic information used for breeding?

- Germplasm classification (***PCA, Clustering, Unsupervised ML, F_{ST}***)
  - **Characterization** – Characterize diversity using unsupervised learning methods.
  - **Heterotic group** – Classify (if known) or infer (if unknown) heterotic groups on individuals and populations.
  - **Signatures of selection** – Use F_{ST} (or related methods) to identify signatures of selection, adaptation and domestication.
- Incorporation (***GWAS, haplotype analysis***)
- Genomic selection (***BayesABC, Supervised ML, etc.***)
- Recycling (***Simulation and optimization***)
- Quantitative assessment (***Variance component analysis***)

Where is genomic information used for breeding?

- Germplasm classification (***PCA, Clustering, Unsupervised ML, F_{ST}***)
- Incorporation (***GWAS, haplotype analysis***)
  - **Trait discovery** – Finding new QTLs via association analysis on breeding data and designed populations.
  - **Introduction of diversity** – Screening non-elite (or elite from elsewhere) germplasm for pre-breeding.
  - **Haplotype enrichment** – Assess genome of non-elite material to add diversity to regions where elite germplasm is fixed.
- Genomic selection (***BayesABC, Supervised ML, etc.***)
- Recycling (***Simulation and optimization***)
- Quantitative assessment (***Variance component analysis***)

Where is genomic information used for breeding?

- Germplasm classification (***PCA, Clustering, Unsupervised ML, F_{ST}***)
- Incorporation (***GWAS, haplotype analysis***)
- Genomic selection (***BayesABC, Supervised ML, etc.***)
  - **F2 enrichment (WF)** – Entire population is genotyped with few markers and selected for specific QTL (e.g. disease resistance)
  - **Pre-selection (WF/AF)** – Entire population is genotyped and 0% is phenotyped. Selection is based on the genomic merit estimated a predefined estimation set that is either made by design or using breeding data.
  - **Test-and-shelf (WF/AF)** – Entire population is genotyped and X% is phenotyped. Within-season selection is based on the genomic merit estimated with a genomic model from phenotyped individuals.
  - **Advancement (WF/AF)** – Entire population is genotyped and phenotyped. Selection is based on the genetic merit of the individuals using one or more seasons of data from those individuals.
  - **Product placement (AF)** – Similar to advancement but GxE takes the spotlight from G.
- Recycling (***Simulation and optimization***)
- Quantitative assessment (***Variance component analysis***)

Where is genomic information used for breeding?

- Germplasm classification (***PCA, Clustering, Unsupervised ML, F_{ST}***)
- Incorporation (***GWAS, haplotype analysis***)
- Genomic selection (***BayesABC, Supervised ML, etc.***)
- Recycling (***Simulation and optimization***)
  - **Selection of parents** – Selection of high BV individuals with complementary polygene or traits.
  - **Select combinations** – Providing a set of candidate parents (100% genotyped), combinations are based on clustering, simulate crosses or predefined criterium (OHV or OPV).
- Quantitative assessment (***Variance component analysis***)

Where is genomic information used for breeding?

- Germplasm classification (***PCA, Clustering, Unsupervised ML, F_{ST}***)
- Incorporation (***GWAS, haplotype analysis***)
- Genomic selection (***BayesABC, Supervised ML, etc.***)
- Recycling (***Simulation and optimization***)
- Quantitative assessment (***Variance component analysis***)
    - **Heritability** – Narrow-sense and GxE (e.g. compound symmetry)
    - **Genetic variance decomposition** – Classic (Vg = Va + Vd + Vi) and hybrid (Vg = V_{GCA1} + V_{GCA2} + V_{SCA})
    - **Genetic correlations** – Across traits or within-trait across environments
    - **Effective population size** – Eigen analysis of the G matrix
    - **Genetic progress and rate of genetic gains** – Assess multiple years
    - **Evaluate breeding strategies** – Simulations and retrospective studies to ask ***what if*** questions

CHALLENGES

KEY CHALLENGES

- Improve accuracy with modeling + pop design + experimental design
- Better use of GxE and better understanding TPEs
- Use environment data (soil, weather, management) in genomic models
- Handle multi-parental crosses
- Collaborate effectively and breed consistently across programs
- Educate breeders on how to use genomic data
- Data management – easy access to any type of data & visualization tools

ITERATIVELY TUNING BREEDING DESIGN

- Define the number of reps and locations for each breeding stage
- From which breeding stage to select parents to recycle
- At which stage to GxE outweighs G
- Strategies to increase heritability and optimize GS models

References in optimization procedures

- Rincent et al. (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals. Genetics, 192(2), 715-728.
- Isidro et al. (2015). Training set optimization under population structure in genomic selection. TAG 128(1), 145-158.
- Habier (2016). Improved molecular breeding methods. US20160321396A1.
- Ou and Liao (2019). Training set determination for genomic selection. TAG 132(10), 2781-2792.
- Brauner et al. (2019). Genomic prediction with multiple biparental families. TAG

FOR THE BREEDERS

- Understand your germplasm
- Know your target environments
- Have clear breeding objectives

Concluding Remarks

1. GS is utilized differently for advancement, recycling and incorporations
2. Experimental settings and breeding design play key role in GS
3. Breeding pipeline is dynamic and constantly improved

Thank you for your attention!

Questions??

Alencar Xavier

alencar.xavier@corteva.com
alenxav.wixsite.com/home