Expanding the capabilities of ssGWAS with p-values for large genotyped populations

Natalia Leite, Matias Bermann, Shogo Tsuruta,
Ignacy Misztal, **Daniela Lourenco**

August 31, 2023

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

UNIVERSITY OF GEORGIA

Outline

1. Equivalence between BV and marker-effects models
2. Advantages of ssGWAS
3. The Algorithm for Proven and Young (APY)
4. Equivalence between BV and marker-effects models with APY
5. Theoretical and practical consequences
6. Application example
7. Concluding remarks

BV and marker-effects models

- Both models are equivalent under some assumptions
- Consequences:
  - Estimable functions are equal
  - SNP-effects can be back-solved from EBV
  - Perform GWAS from GBLUP
  - Choice of the most convenient model

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

$$\mathrm{E}[\mathbf{y}] = \mathbf{Xb}$$

$$\mathrm{Var}\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{G}+\mathbf{R} & \mathbf{G} & \mathbf{R}\\ \mathbf{G} & \mathbf{G} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{pmatrix}$$

$$\mathbf{G} = \mathbf{ZZ}'$$

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

$$\mathrm{E}[\mathbf{y}] = \mathbf{Xb}$$

$$\mathrm{Var}\begin{pmatrix}\mathbf{y}\\ \mathbf{a}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{ZZ}'+\mathbf{R} & \mathbf{Z} & \mathbf{R}\\ \mathbf{Z}' & \mathbf{I} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{pmatrix}$$

DOI: 10.1534/genetics.107.081190

The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values

D. Habier,[1] R. L. Fernando and J. C. M. Dekkers

J. Dairy Sci. 92:2971–2975
doi:10.3168/jds.2008-1929

Technical note: **Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit**

I. Strandén*[1] and D. J. Garrick†‡

BV and marker-effects models

- ssGBLUP ↔ ssBR ↔ ssSNP-BLUP
- Consequences:
  - Estimable functions are equal
  - SNP-effects can be back-solved from GEBV
  - Indirect predictions for young-animals
  - Perform GWAS from ssGBLUP

J. Dairy Sci. 92:4656–4663
doi:10.3168/jds.2009-2061

A relationship matrix including full pedigree and genomic information

A. Legarra,*[1] I. Aguilar,†‡ and I. Misztal†
*INRA, UR631 SAGA, BP 52627, 32326 Castanet-Tolosan, France
†Department of Animal and Dairy Science, University of Georgia, Athens 30602
‡Instituto Nacional de Investigación Agropecuaria, Las Brujas, Uruguay

Fernando *et al. Genetics Selection Evolution* 2014, **46**:50
http://www.gsejournal.org/content/46/50

Genetics Selection Evolution

RESEARCH Open Access

A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses

Rohan L Fernando[1*], Jack CM Dekkers[1] and Dorian J Garrick[1,2]

J. Dairy Sci. 97:5833–5850
http://dx.doi.org/10.3168/jds.2014-7924

A single-step genomic model with direct estimation of marker effects

Z. Liu,*[1] M. E. Goddard,† F. Reinhardt,* and R. Reents*
*Vereinigte Informationssysteme Tierhaltung w.V. (VIT), Heideweg 1, D-27283 Verden, Germany
†Melbourne School of Land and Environment, University of Melbourne, Parkville, Victoria 3010, Australia

Why ssGWAS?

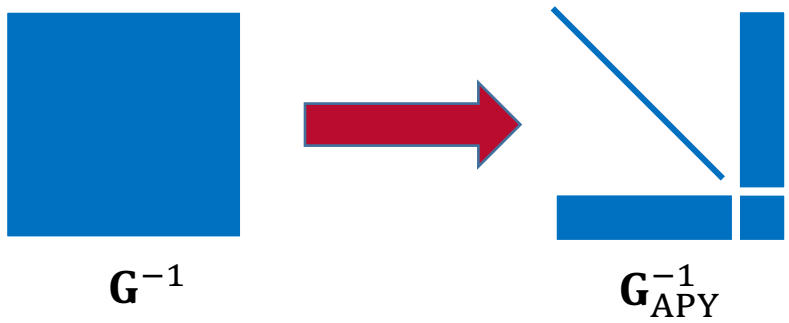
- Why ssGWAS?
- Assumption of single-marker GWAS: Genotyped individuals have phenotypes
- Animal populations: genotypes and phenotypes may not be on the same individuals
  - Deregressed EBV: are biased

- ssGWAS
  - All data on genotyped and non-genotyped individuals
  - Multi-trait models to accommodate correlations

- Negative aspect of ssGWAS
  - Heavy computations for p-values -> same limitation as REML

The Algorithm for Proven and Young

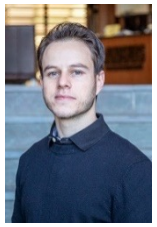
- Genomic information is redundant
- Most of the variation in GRM is explained by a set of core (proven) individuals
- BV of noncore individuals are a linear combination of BV of core individuals plus an error term
- The noncore (young) individuals are conditionally dependent on the core



$\mathbf{G}^{-1}$ $\quad$ $\mathbf{G}^{-1}_{\mathrm{APY}}$

$$\mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix} \Rightarrow \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\mathrm{APY}}^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix}$$

Equivalence APY ssGBLUP – ssSNPBLUP



JOURNAL ARTICLE

Indirect predictions with a large number of genotyped animals using the algorithm for proven and young

Andre L S Garcia, Yutaka Masuda, Shogo Tsuruta, Stephen Miller, Ignacy Misztal, Daniela Lourenco

Journal of Animal Science, Volume 98, Issue 6, June 2020, skaa154,
https://doi.org/10.1093/jas/skaa154

Garcia *et al. Genetics Selection Evolution* (2022) 54:66
https://doi.org/10.1186/s12711-022-00752-4

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Theoretical accuracy for indirect predictions based on SNP effects from single-step GBLUP

Andre Garcia[1*], Ignacio Aguilar[2], Andres Legarra[3], Shogo Tsuruta[1], Ignacy Misztal[1] and Daniela Lourenco[1]

Bermann *et al. Genetics Selection Evolution* (2022) 54:52
https://doi.org/10.1186/s12711-022-00741-7

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

On the equivalence between marker effect models and breeding value models and direct genomic values with the Algorithm for Proven and Young

Matias Bermann[1*], Daniela Lourenco[1], Natalia S. Forneris[2,3], Andres Legarra[4] and Ignacy Misztal[1]

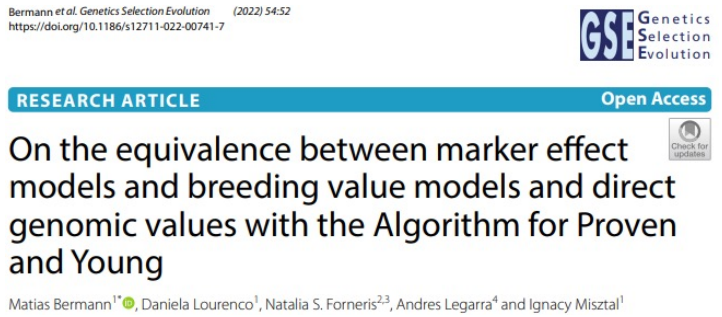
- If using APY in ssGBLUP
  - Numerical equivalence
  - $\hat{\mathbf{u}} = \mathbf{Z}\hat{\boldsymbol{a}}$
  - $\hat{\boldsymbol{a}}|\hat{\mathbf{u}} = k\mathbf{Z}'\mathbf{G}_{\text{APY}}^{-1}\hat{\mathbf{u}}$
  - $\text{Var}(\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}}) = k\mathbf{Z}'\mathbf{G}_{\text{APY}}^{-1}(\mathbf{G} - \mathbf{C}^{\mathbf{u_2}\mathbf{u_2}})\mathbf{G}_{\text{APY}}^{-1}\mathbf{Z}k$

↳ Function of all genotyped animals

- If using APY in ssGBLUP
  - Equivalent APY ssSNPBLUP model
  - $\hat{\mathbf{u}} = \mathbf{Z}^{\dagger}\hat{\boldsymbol{a}}$
  - $\hat{\boldsymbol{a}}|\hat{\mathbf{u}} = k\mathbf{Z}^{\dagger'}\mathbf{G}_{\text{APY}}^{-1}\hat{\mathbf{u}} = k\mathbf{Z}_{\text{c}}'\mathbf{G}_{\text{cc}}^{-1}\hat{\mathbf{u}}_{\text{c}}$
  - $\text{Var}(\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}}) = k\mathbf{Z}_{\text{c}}'\mathbf{G}_{\text{cc}}^{-1}\,(\mathbf{G}_{\text{cc}} - \mathbf{C}^{\mathbf{u}_{2\text{c}}\mathbf{u}_{2\text{c}}})\mathbf{G}_{\text{cc}}^{-1}\mathbf{Z}_{\text{c}}k$

↳ Function of CORE animals

Equivalence APY ssGBLUP – ssSNPBLUP


Bermann et al. Genetics Selection Evolution (2022) 54:52
https://doi.org/10.1186/s12711-022-00741-7

Genetics Selection Evolution

RESEARCH ARTICLE Open Access

On the equivalence between marker effect models and breeding value models and direct genomic values with the Algorithm for Proven and Young

Matias Bermann[1*], Daniela Lourenco[1], Natalia S. Forneris[2,3], Andres Legarra[4] and Ignacy Misztal[1]

- If using APY in ssGBLUP
  - Equivalent APY ssSNPBLUP model
  - $\hat{\mathbf{u}} = \mathbf{Z}^{\dagger}\hat{\boldsymbol{a}}$
  - $\hat{\boldsymbol{a}}|\hat{\mathbf{u}} = k\mathbf{Z}^{\dagger'}\mathbf{G}_{\mathrm{APY}}^{-1}\hat{\mathbf{u}} = k\mathbf{Z}_{\mathrm{c}}'\mathbf{G}_{\mathrm{cc}}^{-1}\hat{\mathbf{u}}_{\mathrm{c}}$
  - $\mathrm{Var}(\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}}) = k\mathbf{Z}_{\mathrm{c}}'\mathbf{G}_{\mathrm{cc}}^{-1}\left(\mathbf{G}_{\mathrm{cc}} - \mathbf{C}^{\mathbf{u}_{2\mathrm{c}}\mathbf{u}_{2\mathrm{c}}}\right)\mathbf{G}_{\mathrm{cc}}^{-1}\mathbf{Z}_{\mathrm{c}}k$

?

- Exact - Inverse of the LHS of MME
- Approximating reliabilities of GEBV

GEBV are published with reliability

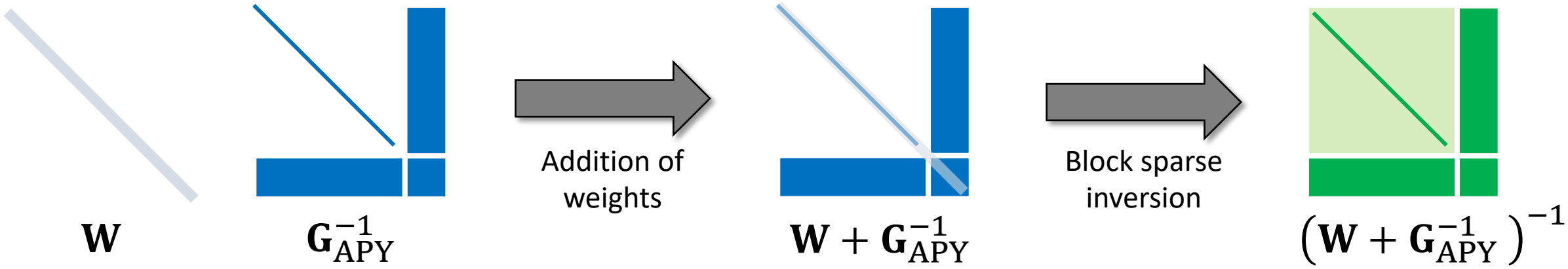
- Reliability based on PEV
  - Approximated for large populations
  - Weights based on approximations
  - Block sparse inversion with APY

JOURNAL ARTICLE

Efficient approximation of reliabilities for single-step genomic best linear unbiased predictor models with the Algorithm for Proven and Young

Matias Bermann, Daniela Lourenco, Ignacy Misztal

Journal of Animal Science, Volume 100, Issue 1, January 2022, skab353, https://doi.org/10.1093/jas/skab353



$$diag\left(\mathbf{W} + \mathbf{G}^{-1}_{\text{APY}}\right)^{-1} = diag\left((\mathbf{W}_{nn} + \mathbf{M}^{-1}_{nn})^{-1} + (\mathbf{W}_{nn} + \mathbf{M}^{-1}_{nn})^{-1}\mathbf{G}^{nc}(\mathbf{W}_{cc} + \mathbf{G}^{cc} - \mathbf{G}^{cn}(\mathbf{W}_{nn} + \mathbf{M}^{-1}_{nn})^{-1}\mathbf{G}^{nc})^{-1}\mathbf{G}^{cn}(\mathbf{W}_{nn} + \mathbf{M}^{-1}_{nn})^{-1}\right)$$

$$diag\left((\mathbf{W}_{cc} + \mathbf{G}^{cc} - \mathbf{G}^{cn}(\mathbf{W}_{nn} + \mathbf{M}^{-1}_{nn})^{-1}\mathbf{G}^{nc})^{-1}\right)$$

Single-step GWAS – Many genotypes

- Genomic evaluation process
  - GEBV using APY ssGBLUP + reliability using block sparse inversion
    - $\widehat{\boldsymbol{a}}|\widehat{\mathbf{u}} = k\mathbf{Z}_c'\mathbf{G}_{cc}^{-1}\widehat{\mathbf{u}}_c$
    - $\mathbf{C}^{\mathbf{u}_{2c}\mathbf{u}_{2c}} \approx \left(\mathbf{W} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{G}_{\mathbf{APY}}^{-\mathbf{1}}\right)^{-\mathbf{1}}$
    - $\mathrm{Var}(\widehat{\boldsymbol{a}}|\widehat{\boldsymbol{u}}) = k\mathbf{Z}_c'\mathbf{G}_{cc}^{-1}\,(\mathbf{G}_{cc} - \mathbf{C}^{\mathbf{u}_{2c}\mathbf{u}_{2c}})\mathbf{G}_{cc}^{-1}\mathbf{Z}_c k$

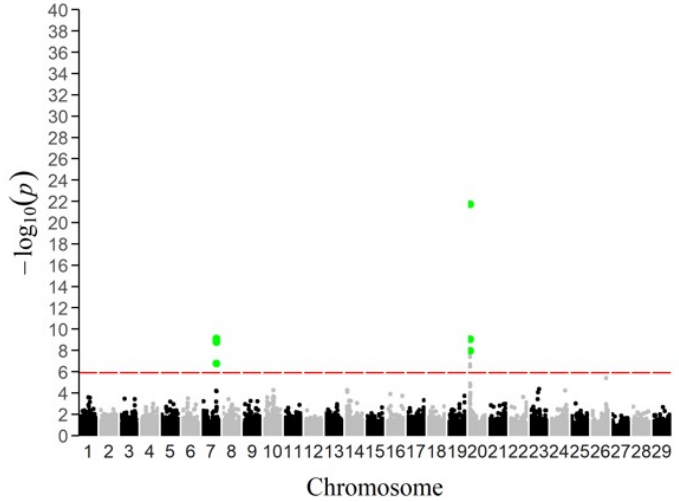
Leite, Bermann et al. (in preparation)

$$p\text{-}value_i = 2\left(1 - \Phi\left(\left|\frac{\hat{a}_i}{sd(\hat{a}_i)}\right|\right)\right)$$

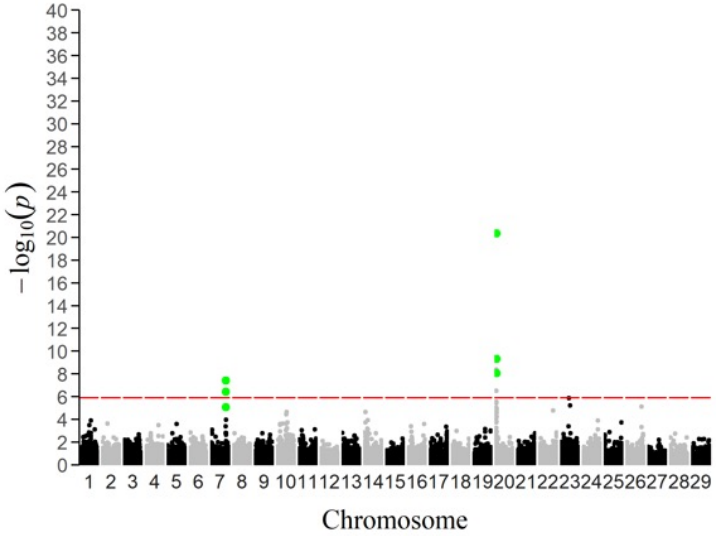
- Initial tests – AGI data
  - 845k phenotypes for post-weaning gain
  - 50k genotyped + 1.58M pedigree
  - 450k genotyped (13k core + 437k noncore) + 1.8M pedigree

Single-step GWAS – Scenarios

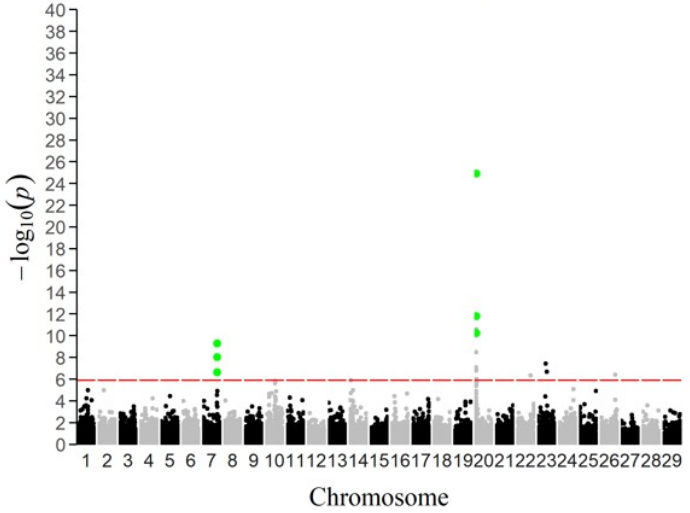
Exact - $\mathbf{G}^{-1}$ 50k



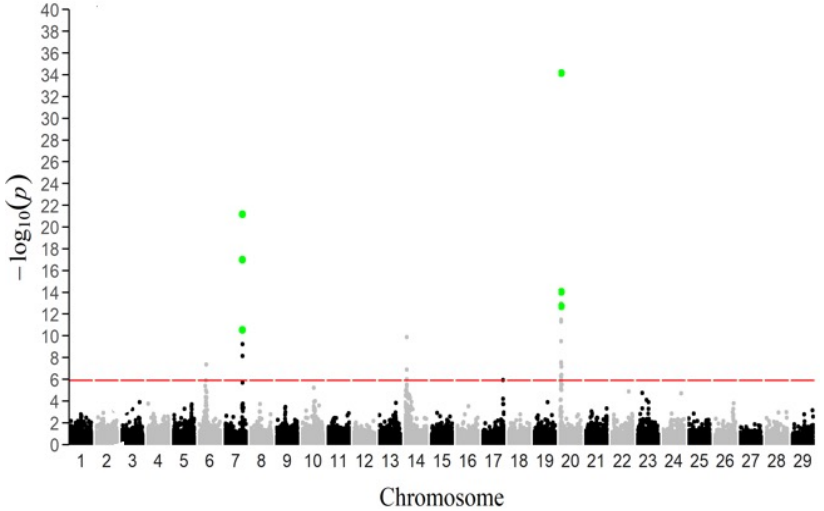
Exact - $\mathbf{G}_{\mathrm{APY}}^{-1}$ 50k



Approximation - $\mathbf{G}_{\mathrm{APY}}^{-1}$ 50k



Approximation - $\mathbf{G}_{\mathrm{APY}}^{-1}$ 450k



Single-step GWAS – Computations

Method	Elapsed time, h:min*	Peak memory, GB*
Exact - $\mathbf{G}^{-1}$ 50k	106:46	159.66
Exact - $\mathbf{G}^{-1}_{\mathrm{APY}}$ 50k	110:59	178.30
Approx - $\mathbf{G}^{-1}_{\mathrm{APY}}$ 50k	2:50	16.62

- Computing cost still high?
- Eliminates the limitation on the amount of data for ssGWAS

Take home messages

- ssGWAS allows for using all data and any model
  - Phenotypes
  - Pedigree
  - Genotypes

- Virtually any number of genotyped animals
  - Improvement in computing time and memory requirements

- Possible because of the limited dimensionality of genomic information
  - Depends on the quality of approximated reliabilities
  - Already implemented in BLUPF90

UNIVERSITY OF GEORGIA

UGA AB&G team

USDA

NIFA

ANGUS
THE BUSINESS BREED
AMERICAN ANGUS ASSOCIATION
3201 Frederick Avenue • St. Joseph, MO 64506 • (816) 383-5100 • Fax (816) 233-9703 • E-mail: angus@angus.org

AGI

PIC

USDA United States Department of Agriculture
Agricultural Research Service
Warmwater Aquaculture Research Unit

USDA Agricultural Research Service
U.S. DEPARTMENT OF AGRICULTURE
Cool and Cold Water Aquaculture Research

zoetis

Smithfield®

Cobb

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

DNA GENETICS
Greater gains. Improved performance.

ABCZ

Rural Development Administration
National Institute of Animal Science

WAGYU
AMERICAN WAGYU ASSOCIATION

BAYER Crop Science

U.S. Registered Holsteins
FOR MAXIMUM PROFIT
HOLSTEIN ASSOCIATION USA, INC.

THE MASCHHOFFS
Progressive Farming. Family Style.

ABS

SEMEX®
Genetics for Life®

PEAK

PFHBiPM

PLEMDAT

Topigs Norsvin

methodgenetics

ANCP