Single-step evaluation in the US Holstein population with 570k genotyped animals

*Yutaka Masuda**[1], I. Misztal[1], S. Tsuruta[1], D. A. L. Lourenco[1], B. Fragomeni[1], A. Legarra[2], I. Aguilar[3], T. J. Lawlor[4]

[1] University of Georgia; [2] INRA; [3] INIA; [4] Holstein Association USA Inc.

Objectives

- To show an efficient implementation of ssGBLUP applied to the US dairy populations
  - Final score in Holsteins
  - Milk yield in Jerseys
- To demonstrate the calculation of genomic predictions with ssGBLUP including more than 500K genotyped animals
  - Using APY G-inverse
- To validate genomic predictions for young bulls
  - With truncated data

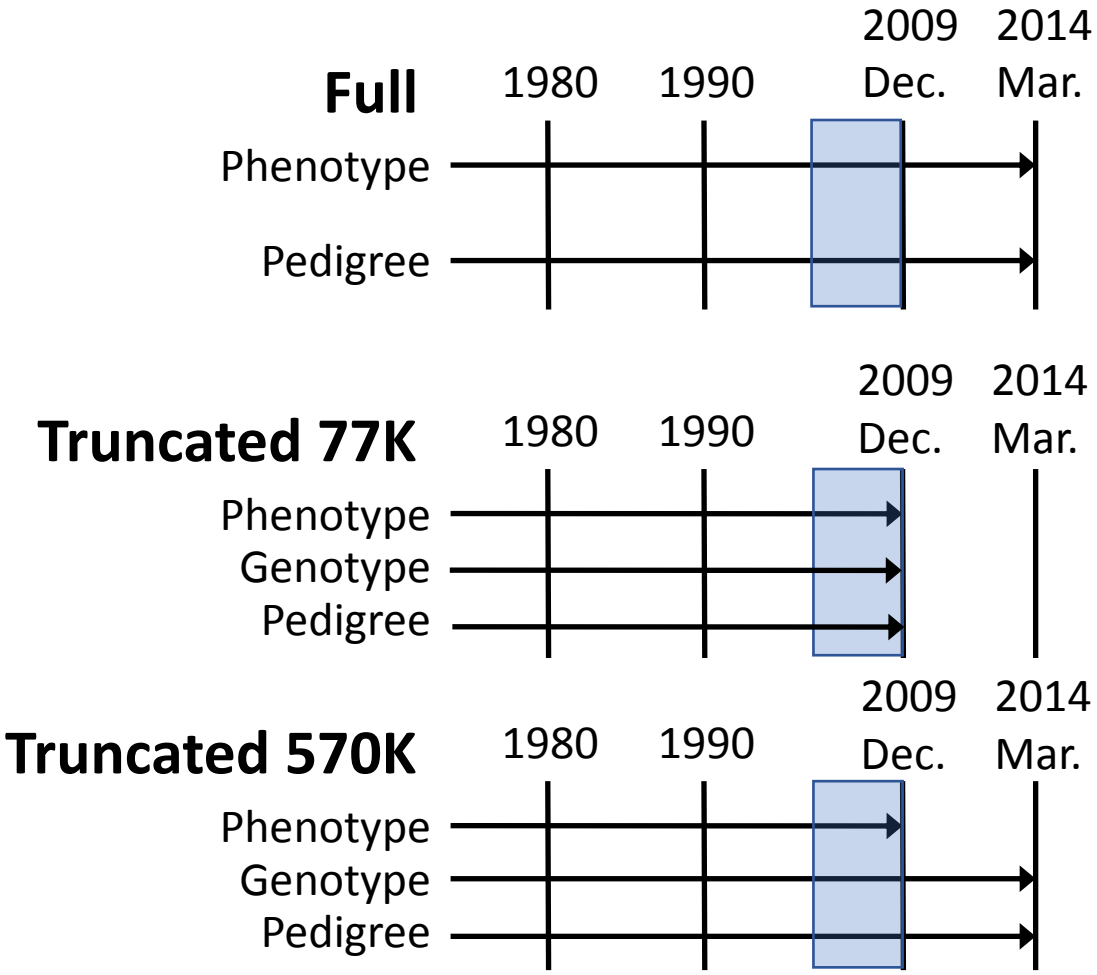
Full Data (HOL)

	Description	Number of records/animals
Phenotype	Final score for US Holstein cows classified in Mar. 2014 or earlier	11,626,576
	Cows classified	6,946,841
Pedigree	Animals born in Mar. 2014 or earlier	10,710,381
Genotype	Animals born in Mar. 2014 or earlier	569,404
	SNP loci	60,671

Definition of “base” animals (HOL)

Group	Definition	The number of animals
Base 9K	Genotyped bulls with at least 1 classified daughter up to 2009	9,406
Base 10K	Base 1 + genotyped and classified dams of above bulls up to 2009	10,458
Base 17K	Base 2 + genotyped & classified cows up to 2009	16,828
Rand 5K 10K, 15K, 20K, 30K	Randomly sampled animals born in 2009 or earlier	5,000 to 30,000

Validation (HOL)



For Daughter Deviation (DD2014)

For GEBV & PA (GEBV2009)

Full & APY inverses

For GEBV & PA (GEBV2009)

APY inverses

Validation Bulls: Genotyped young bulls with no classified daughters in 2009 but with at least 30 classified daughters in 2014 (N=2,948)

Validation analysis (HOL)

- A linear regression analysis:

$$DD2014 = b_1 \times GEBV2009 + b_0$$

  - R^2 value: validation reliability
  - Slope (b_1): Bias of prediction

Validation Bulls: Genotyped young bulls
with no classified daughters in 2009 but with at least 30 classified daughters in 2014 (N=2,948)

Indirect computation of $\mathbf{A}_{22}^{-1}$

$$\mathbf{A}_{22}^{-1}\mathbf{q} = [\mathbf{A}^{22} - \mathbf{A}^{21}(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}]\mathbf{q}$$

where $\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix}$

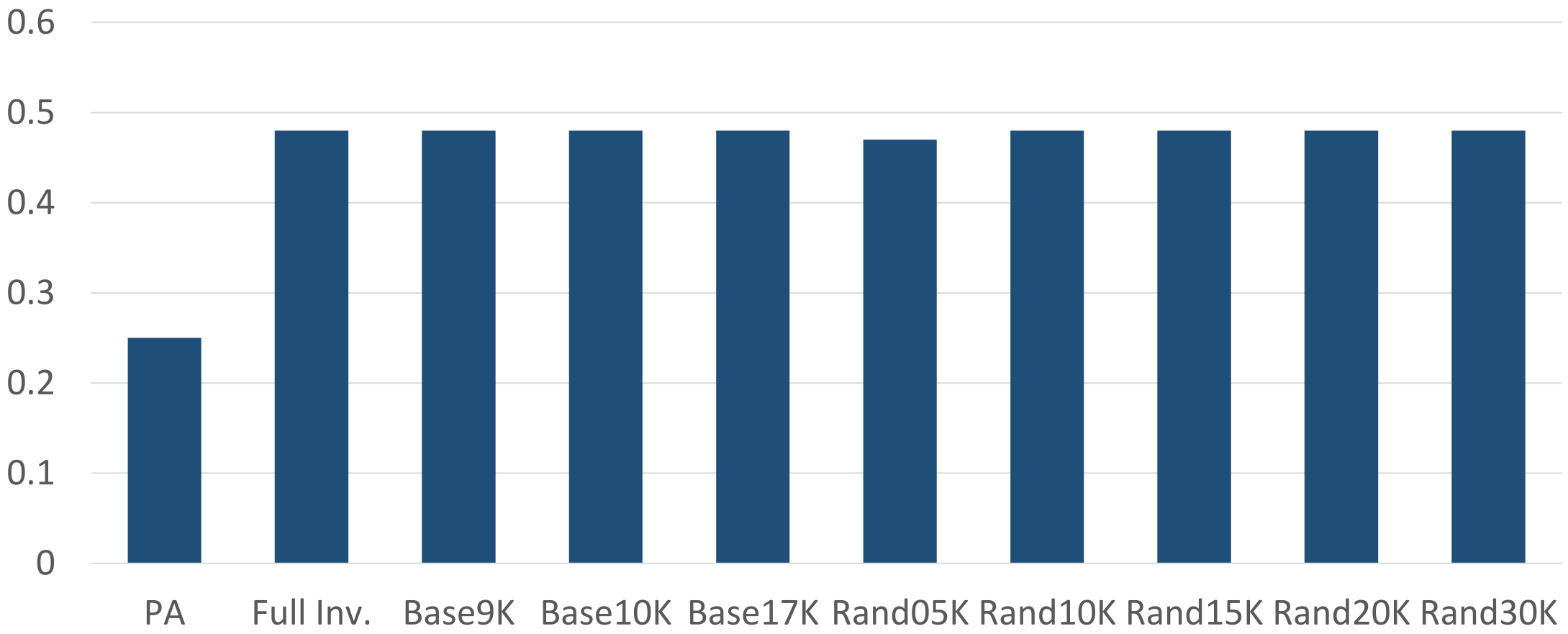
- “2”: genotyped animals & “1”: non-genotyped animals
- Create $\boldsymbol{A}^{22}$, $\boldsymbol{A}^{21}$, and $\boldsymbol{A}^{11}$ with the Henderson’s algorithm
- See Strandén & Mäntysaari (2014)

Computing environment

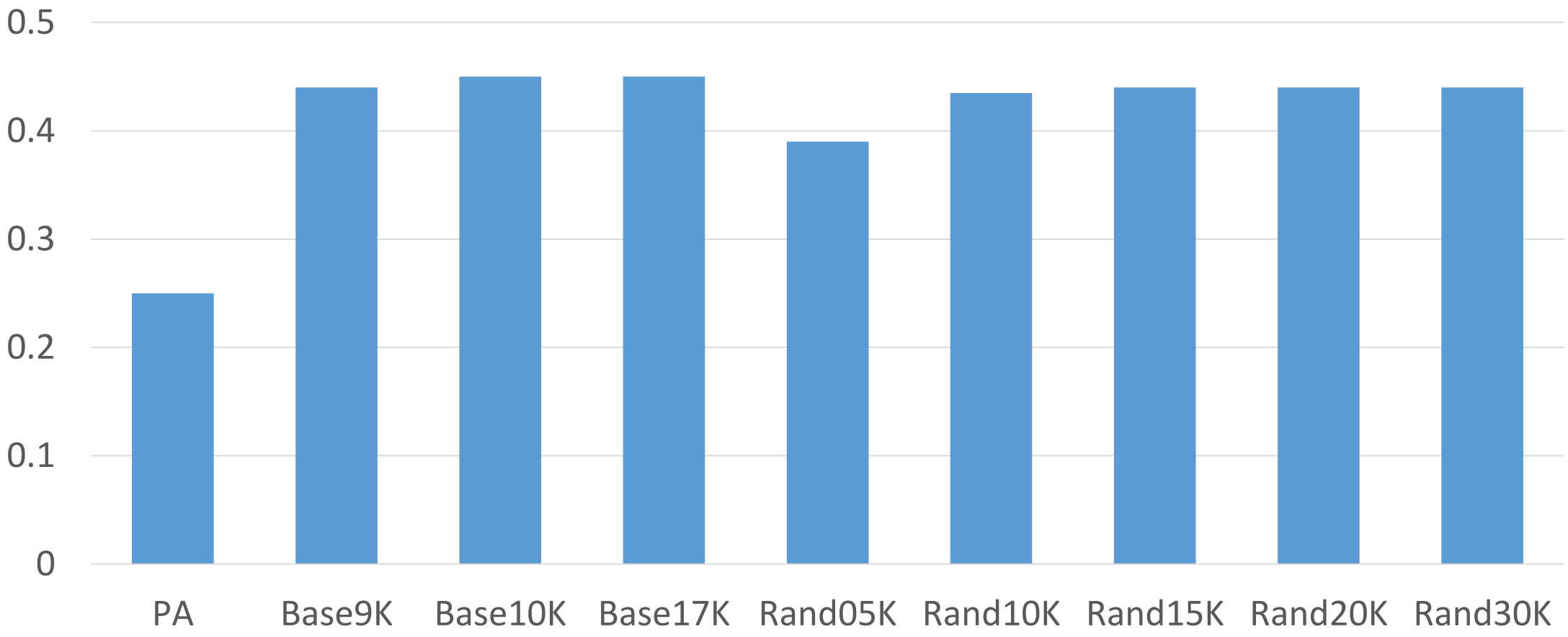
- A server:
  - Intel Xeon @ 3.0GHz (24 cores)
  - Main memory: 1024 GB
  - Disk-array (RAID 5)
- BLUPF90 family programs
  - genomic.f90
  - Blup90iod2 (PCG)
  - Compiled with Intel Fortran Compiler 15.0
  - BLAS/LAPACK in MKL

http://www.dell.com

Results: R^2 from Truncated **77K**



Results: R^2 from Truncated **570K**



Results: b_1 (HOL)

Model	Prediction	# of "Base" animals	77K	570K
No genomics	PA2009		0.63	0.63
Single-step $\mathbf{G}^{-1}$	Full		0.69	N/A
Single-step $\mathbf{G}_{APY}^{-1}$	Base 1	9,406	0.69	0.82
	Base 2	10,458	0.69	0.82
	Base 3	16,828	0.69	0.83
	Rand05K	5,000	0.71	0.75
	Rand10K	10,000	0.71	0.84
	Rand15K	15,000	0.71	0.84
	Rand20K	20,000	0.70	0.83
	Rand30K	30,000	0.70	0.83

Wall-clock time (min) in 570K data

G-part	10K "base"	20K "base"
Read & store SNPs	18	18
Setting-up $\mathbf{G}_{APY}$	26	48
Blend with $\mathbf{A}_{22}$	26	25
Setting-up $\mathbf{G}_{APY}^{-1}$	7	24
Total time	**1 h 17 min**	**1 h 55 min**

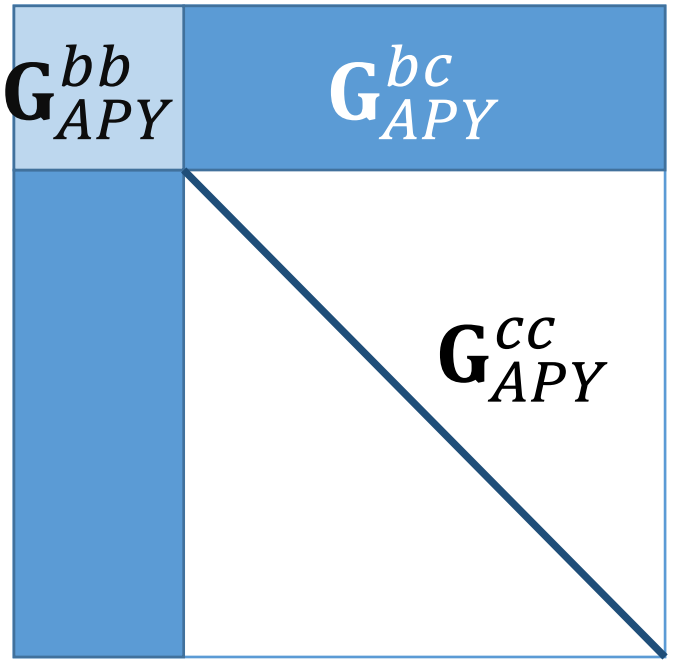
A22-part	Time
Set-up the components	9
Other operations	2
Total computing time	**11 min**

PCG iteration	10K "base"	20K "base"
Time per round	11.7 sec	13.7 sec
Total time in 1,000 rounds	**3 h 15 min**	**3 h 48 min**

Peak memory requirement in 570K data

Step	10K “base”	20K “base”
Marker genotypes	16.1 GB	16.1 GB
Storage for $\mathbf{G}_{APY}$ and $\mathbf{G}_{APY}^{-1}$	42.4 GB	84.9 GB
Peak temporary memory	6.4 GB	10.9 GB
A22 related	1.8GB	1.8GB
Peak requirement	**66.7 GB**	**113.7 GB**

APY G-Inverse: more genotypes



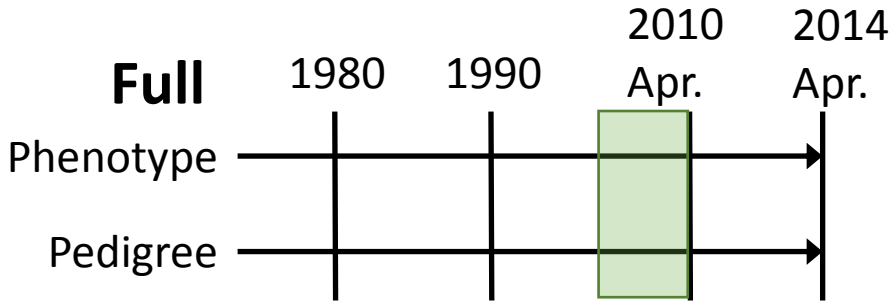
- Assume 2M genotypes with 10K “base”
- Required storage:
  - $\mathbf{G}_{APY}^{-1}$: 149 GB vs $\mathbf{G}^{-1}$: 29 TB
  - Temporary memory: negligible
- Computing cost:
  - $\mathbf{G}_{APY} \sim O(n_b)$ vs $\mathbf{G} \sim O(n^3)$
  - $\mathbf{G}_{APY}^{-1} \sim O(n_b^2)$ vs $\mathbf{G}^{-1} \sim O(n^3)$

  If the total number of genotyped animals is fixed in the comparison.

Full Data (JER)

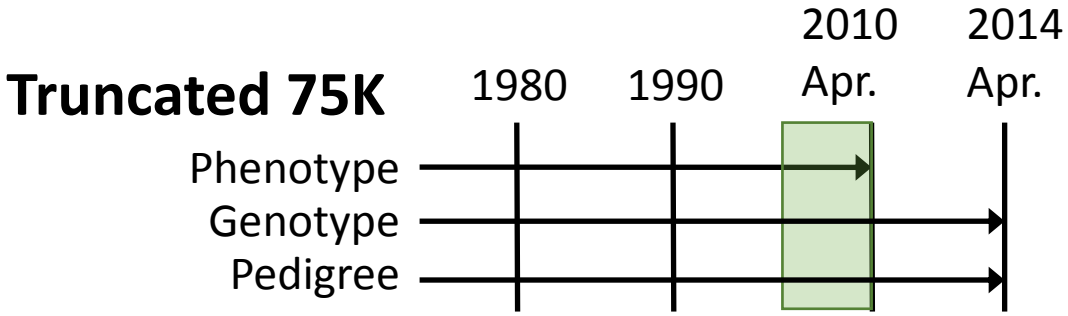
	Description	Number of records/animals
Phenotype	305-d milk yield for US Jersey cows recorded in 2014 or earlier	4,168,048
	Cows tested	2,468,914
Pedigree	Animals born in 2014 or earlier	272,579
genotype	Animals born in 2014 or earlier	75,053
	SNP loci	60,000

Validation



For Deregressed EBV (DEBV2014)

Validation Bulls: Genotyped young bulls with no daughters tested in 2010 but with at least 1 daughter tested in 2014 (N=457)



For GEBV (GEBV2010)

Full & APY inverses

Results: R^2 and b_1 (JER*)

Model	Prediction	# of "base" animals	R^2	b_1
No genomics	PTA2010		0.40	0.78
Multi-step[1]	GPTA2010		0.54	0.89
Single-step $\mathbf{G}^{-1}$	Full inverse	75,053	0.56	0.84
Single-step $\mathbf{G}_{APY}^{-1}$	Selected bulls[2]	10,677	0.55	0.84
	Selected bulls[3]	15,960	0.56	0.84
	Rand10K[4]	10,000	0.55	0.84
	Rand15K[4]	15,000	0.56	0.84

* Predicted young bulls with at least 1 daughter with records in 2014 (N = 457)
1 All tests predicted 482 validation bulls that had no daughters in 2010; 2 Old bulls with at least 1 progeny; 3: All the bulls with at least 1 progeny; 4 Randomly selected from all the genotyped animals;

Conclusion

- **Genomic evaluation using ssGBLUP with 1M genotypes is feasible.**
- APY-inverse provides GEBV with higher accuracy than PA in final score for the US Holsteins and milk yield for the US Jerseys.
- In Jersey, ssGBLUP has similar accuracy to multi-step method + MACE evaluations.
- Choice of “base” animals has almost no effect on the accuracy.
- Only 10,000 “base” animals are enough to achieve the highest R^2 in the validation.
- Quality of genotypes may have an impact on the accuracy.

Acknowledgement

- USDA AGIL to provide official GPTA for Jersey
  - Paul VanRaden
  - Melvin Tooker
  - George Wiggans