Bavarian State Research Center
for Agriculture

# Effects of Selective Genotyping and Selective Imputation in Single-Step GBLUP

C. Edel, E.C.G. Pimentel, L. Plieschke, R. Emmerling and K.-U. Götz

Institute for Animal Breeding

# Motivation

- Single-Step Genomic BLUP frequently results in inflated genomic predictions
- ad hoc remedies were proposed
  - ✓ ω -, τ - scaling
  - ✓ data-pruning, etc.
- however, underlying mechanisms remain unclear

# Background

- computing the standard H matrix includes an implicit step of imputation (Fernando et al., 2014)

$$
\begin{aligned}
\mathbf{H} &= \mathbf{G}_{add} + \mathbf{G}_{est} \\
&= \begin{bmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}
\end{aligned}
$$

**`imputation residual`** **covariance of observed and imputed gt**

LfL Animal Breeding

# Background

- computing the standard H matrix includes an implicit step of imputation (Fernando et al., 2014)

$$\begin{aligned}\mathbf{H} &= \mathbf{G}_{add} + \mathbf{G}_{est} \\ &= \begin{bmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}\end{aligned}$$

$\mathbf{A}_{12}\mathbf{A}_{22}^{-1}(gc\text{-}\mu_{gc})$: imputation of gene contents (GC) for **11*** animals (ungenotyped)

# Background

- computing the standard H matrix includes an implicit step of imputation (Fernando et al., 2014)

- Single-step genomic BLUP conceptually comprises two estimation steps
  - estimation of gene contents using all observed genotypes
  - estimation of gEBV using all phenotypic data

# Background

J. Dairy Sci. 100:1–5
https://doi.org/10.3168/jds.2017-12734


***Short communication:*** **The role of genotypes from animals without phenotypes in single-step genomic evaluations**

**T. Shabalina,[1] E. C. G. Pimentel,[1,2] C. Edel, L. Plieschke, R. Emmerling, and K.-U. Götz**
Institute of Animal Breeding, Bavarian State Research Center for Agriculture, 85586 Grub, Germany

see also contribution to poster-session, EAAP 2017

- Simulation study:
  - exploring effects of implicit imputation in single-step genomic BLUP
  - genotypes without phenotypes improve the predictive ability of the system by imputing phenotyped animals without genotypes
  - however, the quality of imputed genotypes varies

# Background

J. Dairy Sci. 100:1–5
https://doi.org/10.3168/jds.2017-12734


***Short communication:*** **The role of genotypes from animals without phenotypes in single-step genomic evaluations**

**T. Shabalina,[1] E. C. G. Pimentel,[1,2] C. Edel, L. Plieschke, R. Emmerling, and K.-U. Götz**
Institute of Animal Breeding, Bavarian State Research Center for Agriculture, 85586 Grub, Germany

see also contribution to poster-session, EAAP 2017

**Table 2.** Number of genotyped sons per sire from generation 3 to 4, calculated and expected correlation between imputed and true genotypes for different scenarios

| Item | Scenario 1 | Scenario 2 | Scenario 3 |
|---|---|---|---|
| Number of sons per sire | 2.41 | 3.10 | 14.51 |
| Calculated correlation (SD) | 0.77 (0.06) | 0.82 (0.06) | 0.93 (0.06) |
| Expected correlation | 0.66 | 0.71 | 0.91 |

with increasing number of genotyped offspring the true genotype of an ancestor is imputed with higher accuracy

LfL Animal Breeding

# Definitions

- selective genotyping
  - animals are selected for genotyping based on a breeding value containing Mendelian Sampling (MS) information

- selective imputation
  - selective genotyping as introduced by imputation
  - genotypes of frequently used sires and dams are imputed with high accuracy (only )

- reference set
  - a group of animals contributing informative ties between phenotype and (observed or imputed) genotype

# Hypothesis I

- selective genotyping in the reference set can have a negative impact on quality and unbiasedness of genomic predictions
- the effect should already be observable in standard two step genomic applications

# Evidence from Empirical Data: Two step

- Fleckvieh, routine application, forward prediction (4 y)

| | $b_1$ | | $Rel_{real}$ | |
|---|---|---|---|---|
| | **MY** | **PY** | **MY** | **PY** |
| **raw** | .87 | .89 | 63 | 63 |
| | | | | |
| | | | | |

**$b_1$:** regression slope ITB GEBV-Test (Mäntysaari et. al, 2010)
**$Rel_{real}$:** realized reliability (VanRaden et al., 2009)

LfL Animal Breeding

# Evidence from Empirical Data: Two step

- Fleckvieh, routine application, forward prediction (4 y)

| | $b_1$ | | $Rel_{real}$ | |
|---|---|---|---|---|
| | **MY** | **PY** | **MY** | **PY** |
| **raw** | .87 | .89 | 63 | 63 |
| **scaled** | .93 | .96 | 63 | 63 |
| | | | | |

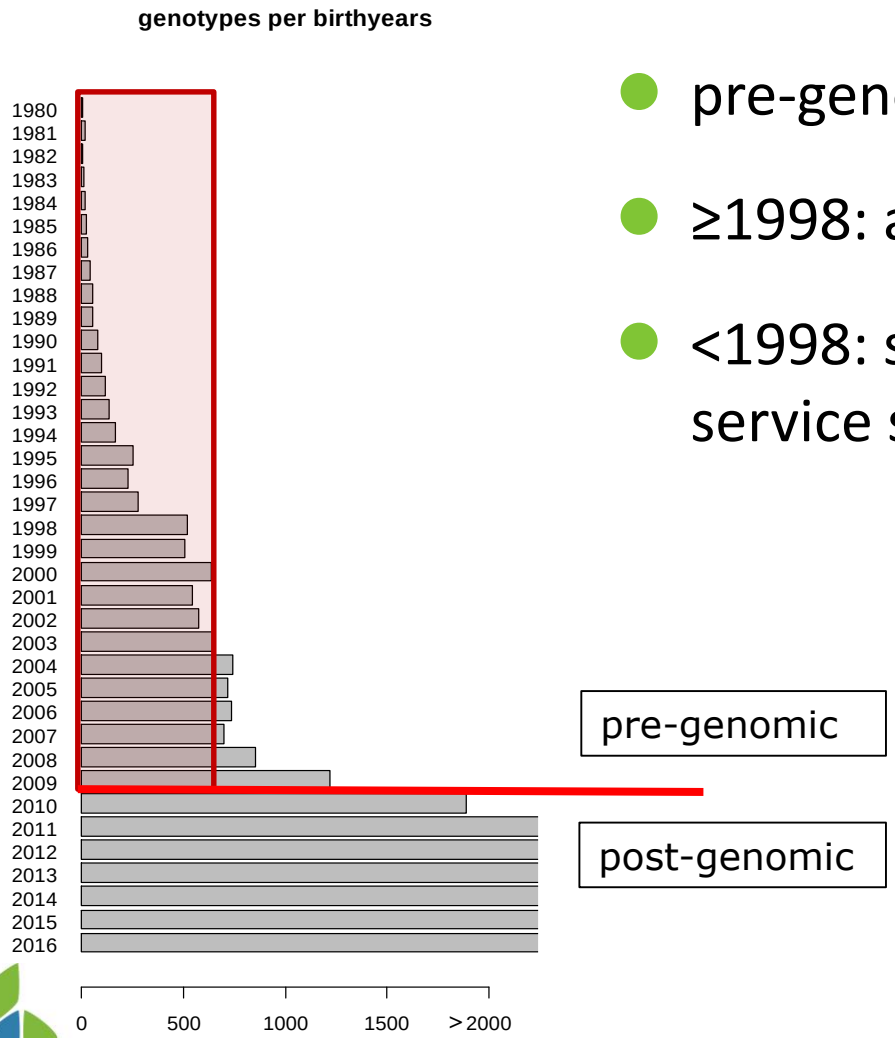
Scaling of predicted values

- to compensate for negative effects of selective genotyping in reference population
- uses systematic difference between PA and EBV (reference animals) as indicator/measure of (pre-)selection on MS

LfL Animal Breeding

# Evidence from Empirical Data: Two step

- Background: selected reference population

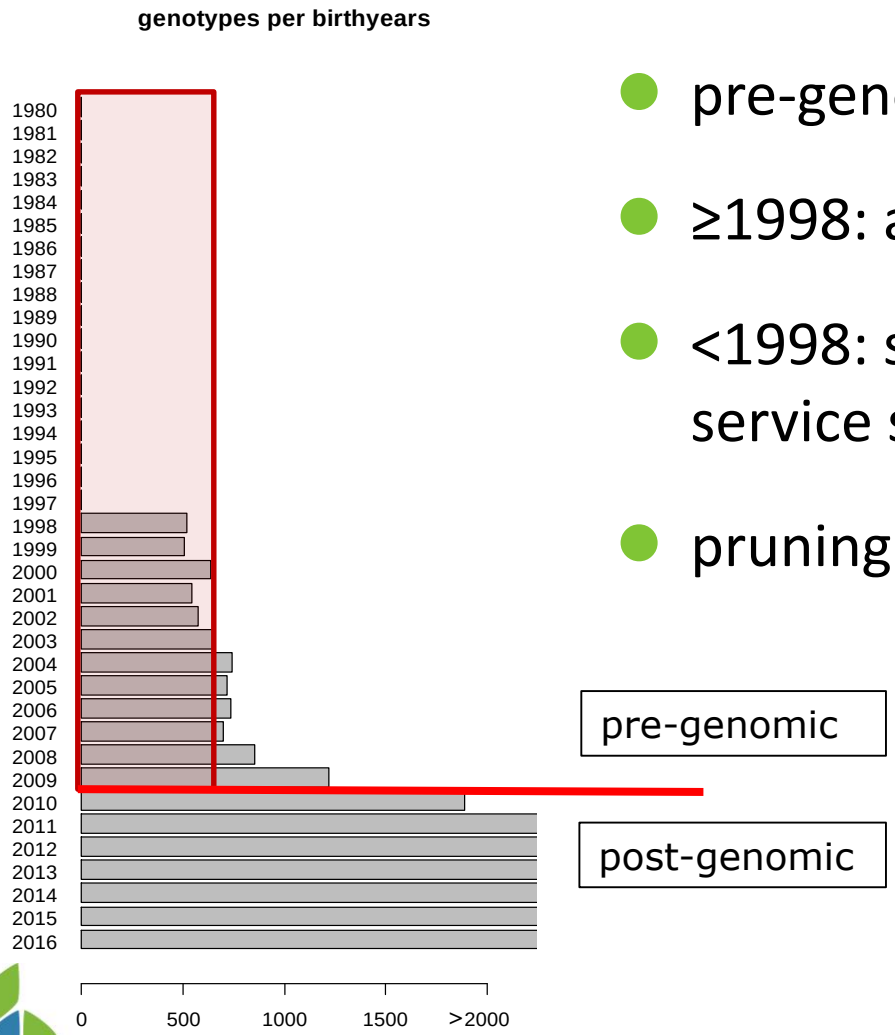



- pre-genomic: ~600 bulls per year tested
- ≥1998: approx. completely genotyped
- <1998: selective genotyping (~second service sires only)

LfL Animal Breeding

# Evidence from Empirical Data: Two step

- Background: selected reference population




- pre-genomic: ~600 bulls per year tested
- ≥1998: approx. completely genotyped
- <1998: selective genotyping (~second service sires only)
- pruning of birthyears < 1998

LfL Animal Breeding

# Evidence from Empirical Data: Two step

- Fleckvieh, routine application, forward prediction (4 y)

| | $b_1$ | | $Rel_{real}$ | |
|---|---|---|---|---|
| | **MY** | **PY** | **MY** | **PY** |
| **raw** | .87 | .89 | 63 | 63 |
| **scaled** | .93 | .96 | 63 | 63 |
| **pruned** | .92 | .94 | 62 | 61 |

**$b_1$:** regression slope ITB GEBV-Test (Mäntysaari et. al, 2010)
**$Rel_{real}$:** realized reliability (VanRaden et al., 2009)

LfL Animal Breeding

# Conclusion I

- selectively genotyped sires from older birth years inflate genomic predictions if they are included in the reference
- omitting these sires reduces inflation and has only a small impact on reliability
- similar effects can be observed with scaling, depending on the measure of selectedness of reference

# Hypothesis II

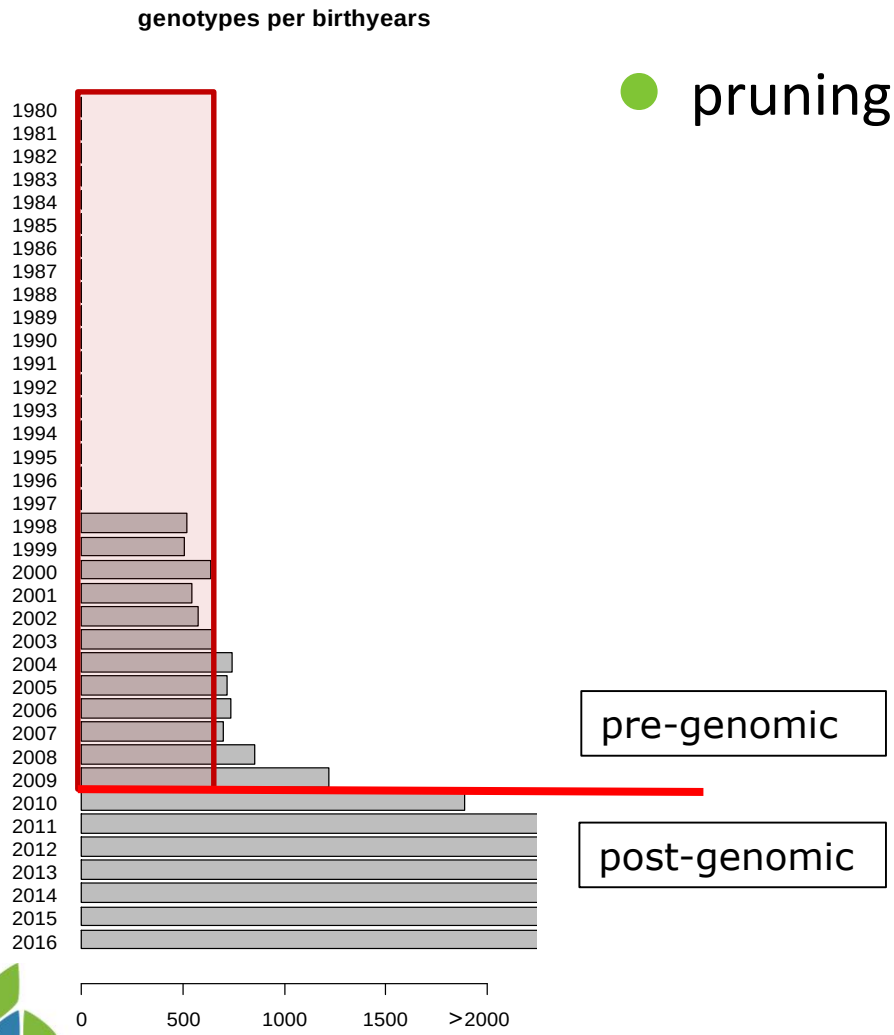
- in Single-Step Genomic BLUP selective genotyping of reference animals should have a similar impact as in Two-Step GBLUP

- selective genotyping can occur in two ways
  - ✓ directly or
  - ✓ as a consequence of selective imputation

- as a consequence **pruning of genotypes** should have a different effect from **pruning of data** (phenotypes)

# Evidence from Empirical Data: Single Step

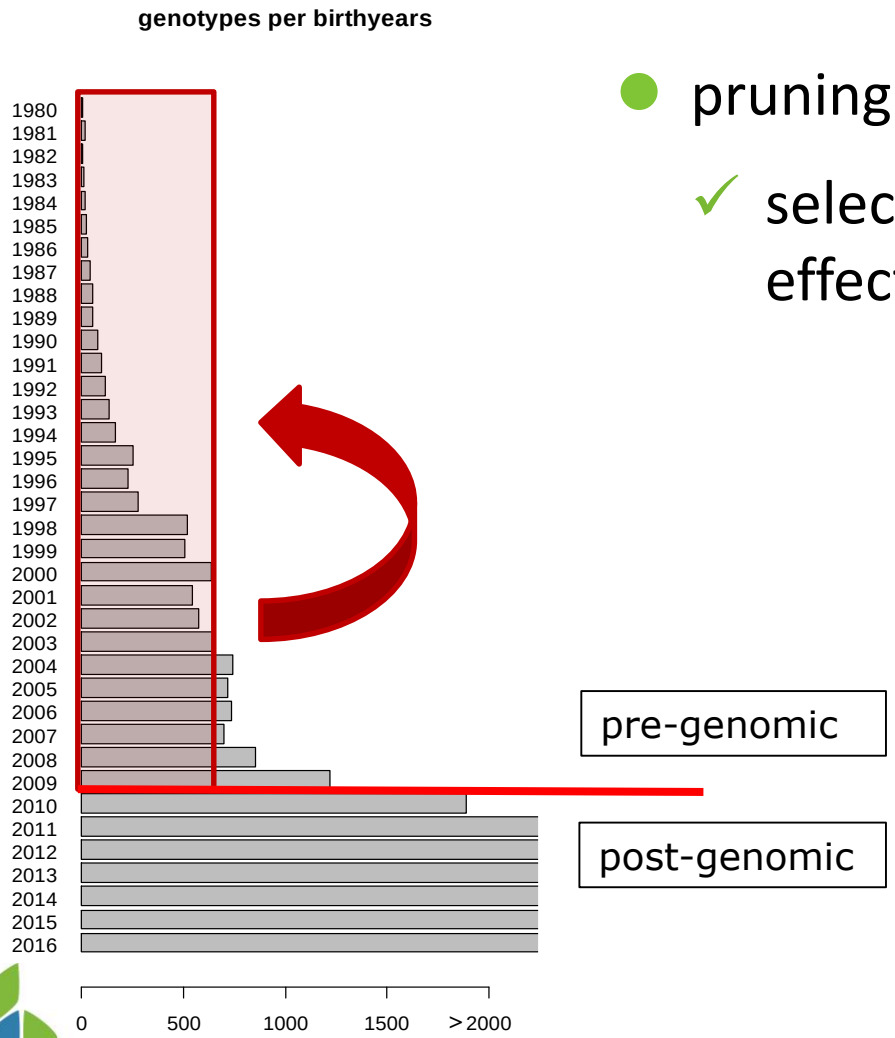
- Background: Effects of pruning




- pruning of genotypes only

# Evidence from Empirical Data: Single Step

- Background: Effects of pruning

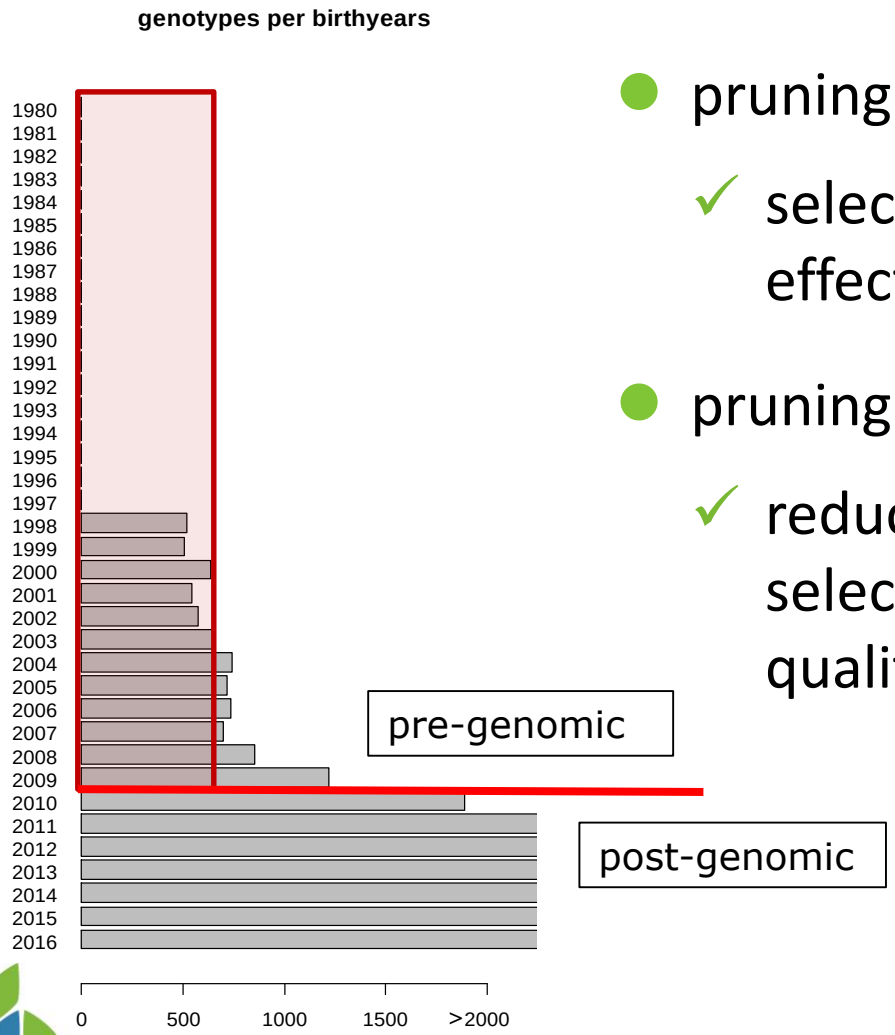



- pruning of genotypes only
  - ✓ selective imputation reestablishes the effects of selective genotyping

LfL Animal Breeding

# Evidence from Empirical Data: Single Step

- Background: Effects of pruning




- pruning of genotypes only
  - ✓ selective imputation reestablishes the effects of selective genotyping
- pruning of data
  - ✓ reduces the negative impact of selective imputation and improves quality of GEBV

LfL Animal Breeding

# Evidence from Empirical Data: Single Step

- Fleckvieh, test application, forward prediction (4 y)

| | $b_1$ | | $Rel_{real}$ | |
|---|---|---|---|---|
| | **MY** | **PY** | **MY** | **PY** |
| **raw** | .81 | .80 | 64 | 61 |
| **pruned: P** | .91 | .89 | 65 | 61 |
| | | | | |

**$b_1$:** regression slope ITB GEBV-Test (Mäntysaari et. al, 2010)
**$Rel_{real}$:** realized reliability (VanRaden et al., 2009)

LfL Animal Breeding

# Evidence from Empirical Data: Single Step

- Fleckvieh, test application, forward prediction (4 y)

| | $b_1$ | | $Rel_{real}$ | |
|---|---|---|---|---|
| | **MY** | **PY** | **MY** | **PY** |
| **raw** | .81 | .80 | 64 | 61 |
| **pruned: P** | .91 | .89 | 65 | 61 |
| **pruned: G** | .82 | .80 | 64 | 61 |

**$b_1$:** regression slope ITB GEBV-Test (Mäntysaari et. al, 2010)
**$Rel_{real}$:** realized reliability (VanRaden et al., 2009)

LfL Animal Breeding

# Conclusion II

- Single-Step
  - shows similar effects of selective genotyping
  - estimates are generally more inflated than in two step
- removing older animals (P+G) from selectively genotyped birth years reduces inflation considerably
- removal of genotypes only is <u>not</u> sufficient
  - information restored from 'historical' $\mathbf{A_{11}}$ block is selective
  - reestablishes negative effects of selective genotyping on genomic estimates

LfL Animal Breeding

# General Conclusion

- selective genotyping
  - often neglected as a potential source of inflation
  - still an aspect to consider
    - ✓ elite cows (genotyped or selectively imputed)
    - ✓ (unintentional) preselection for cow reference population

- Single Step genomic prediction
  - implicitly restores information of pruned genotypes by imputation
  - therefore, pruning of genotypes is not sufficient
  - data-pruning (P+G) appears to be the only way to control negative effects

# Thank you for your attention

We gratefully acknowledge:

- Arbeitsgemeinschaft Süddeutscher Rinderzucht- und Besamungsorganisationen for financial support within the research cooperation „Zukunftswege“
- Contributors of the genotype pool Germany-Austria

# Evidence from Empirical Data: Single Step

- Fleckvieh, test application, forward prediction (4 y)
  - additional results

| | $b_1$ | | $Rel_{real}$ | |
|---|---|---|---|---|
| | **MY** | **PY** | **MY** | **PY** |
| **raw** | .81 | .80 | 64 | 61 |
| **pruned: P** | .91 | .89 | 65 | 61 |
| **pruned: G** | .82 | .80 | 64 | 61 |
| **expected** | .91 | .93 | -- | -- |
| **pruned: P plus** | .92 | .92 | 66 | 62 |
| **two-step (plus cow gt)** | .93 | .95 | 65 | 63 |

**$b_1$:** regression slope ITB GEBV-Test (Mäntysaari et. al, 2010)
**$Rel_{real}$:** realized reliability (VanRaden et al., 2009)

LfL Animal Breeding

# Evidence from Simulation

- strong effects of phenotypic preselection: cow reference

Plieschke *et al. Genet Sel Evol (2016) 48:73*
DOI 10.1186/s12711-016-0250-9

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

## Systematic genotyping of groups of cows to improve genomic estimated breeding values of selection candidates

Laura Plieschke[1*], Christian Edel[1], Eduardo C. G. Pimentel[1], Reiner Emmerling[1], Jörn Bennewitz[2] and Kay-Uwe Götz[1]

**Table 5 Model-derived reliabilities ($R^2$ were virtually equal across all scenarios), validation reliability ($\rho^2$) and regression slopes of the –/50 scenario and the three additional scenarios**

| Scenarios | | | | –/50 | | $-/50_s$ | | $-/25_r25_s$ [a] | | $-/50_{ub}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Validation set | Sire status | Number of individuals | $R^2$ | $\rho^2$ | b | $\rho^2$ | b | $\rho^2$ | b | $\rho^2$ | b |
| 9 | Reference | 1050 | 81 | 53 | 0.82 | 35 | 0.60 | 40 | 0.98 | 53 | 0.79 |
| 10a | Reference | 4516 | 81 | 65 | 0.95 | 42 | 0.76 | 48 | 1.22 | 65 | 0.95 |
| 10b | Not reference | 10,484 | 76 | 60 | 0.92 | 37 | 0.70 | 44 | 1.14 | 60 | 0.91 |

Validation animals were divided according to whether their sire was in the reference set or not

[a] Higher standard error compared to the other scenarios

LfL Animal Breeding