Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Department of Economic Affairs,
Education and Research EAER

Agroscope

Agroscope Swiss national Stud Farm SNSTF

Increase phasing accuracy of selected reference populations

Markus Neuditschko

31. August 2016, EAAP Annual Meeting 2016, Belfast

Current practice

- Existing methods for the selection of informative individuals for re-sequencing and genotype imputation solely focus on the identification of key ancestors.

- Selecting only key ancestors can lead to a loss of phasing accuracy of the reference population.

- To increase phasing accuracy of the selected reference populations, we developed a novel approach to select key contributors based on the eigenvalue decomposition of a genomic relationship matrix.

Agroscope Swiss national Stud Farm SNSTF

Eigenvalue Decomposition (EVD)

- Eigenvalue Decomposition (**EVD**) like **PCA** is a multivariate technique that provides an optimal subspace to investigate population structures.

- Based upon this mathematical principle, we identified individuals that maximize the variation of the genetic relationship structure.

- As such individuals capture most of the relevant genetic relationship structure we called them “**key contributors**”.

Identification of key contributors

- The EVD of genomic relationship matrix (G) returns n non-negative eigenvalues λ_i and n singular eigenvectors u_i, such that:

$$G = \boldsymbol{U} \lambda \boldsymbol{U}^T \qquad \textbf{1.1}$$

- Based upon this principle we derived standardized eigen-vectors (s_i) and calculated the correlation coefficients (r_j) between s_i and each individual (g_j) limiting the number of s_i to k significant components

$$r_j = \sum_{i=1}^{k} s_i g_j \qquad \textbf{1.2}$$

Swiss national Stud Farm SNSTF

Agroscope

Identification of key contributors

- Finally, we rank all individuals according to the genetic contribution score (gc_j) and consider individuals correlated with top k significant components as key contributors

$$gc_j = \sum_{i=1}^{k} (r_i)^2 \qquad \textbf{1.3}$$

- The method to identify key contributors within populations is available online at https://github.com/esteinig/netview.

Phasing accuracy

- To demonstrate the utility of our strategy to increase phasing accuracy, we compared the phasing accuracy of selected individuals with two common applied methods.

  (1) Pedigree-based marginal gene contributions (PED)

  (2) Maximization of the expected genetic relationship to the reference population (REL)

- After selecting sets of informative individuals (20 – 80) the inferred haplotype phase was compared with the true or most likely haplotype phase.
- Phasing accuracy was examined using switch-error metric.

Datasets

(1) Simulated population

Base population (F0) of 1,020 individuals (20 males and 1,000 females), by mating each male with 50 females. Each of the next four generations (F1-F4) also consisted of 20 males and 1,000 females and was generated following the same principle. Resulting in a total of 4,100 individuals and 10,000 SNPs.

(2) Sheep population

The sheep population represents and experimental backcorss/intercross sheep resource flock, where 4 F1 sires and 3 F2 sires were selected for mating. Here, we studied 1,421 individuals genotyped for 44,693 SNPs.

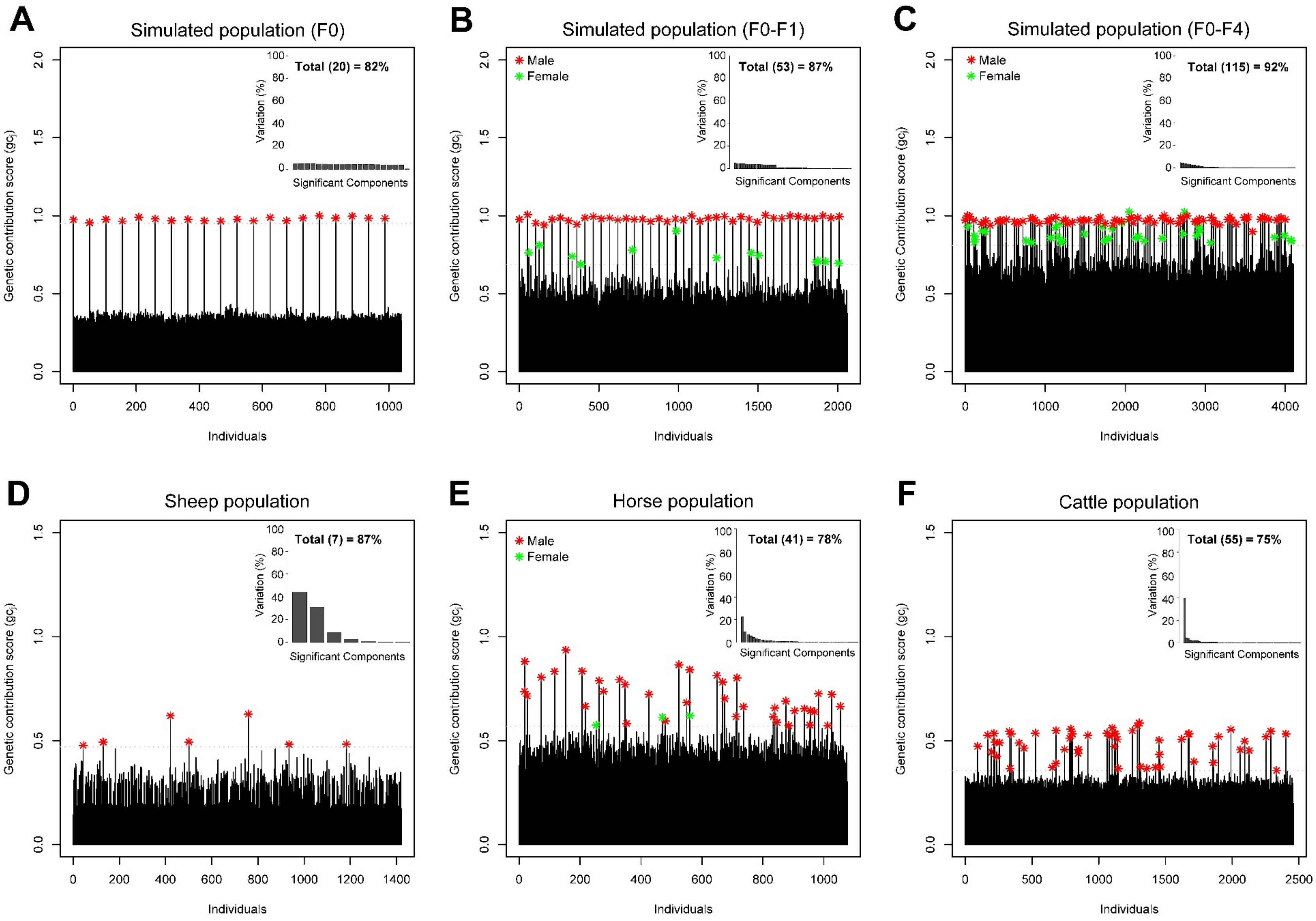
(3) Horse population

The horse population consisted of a sample collection of 1,077 Franches-Montagnes horses genotyped for 38,124 SNPs.

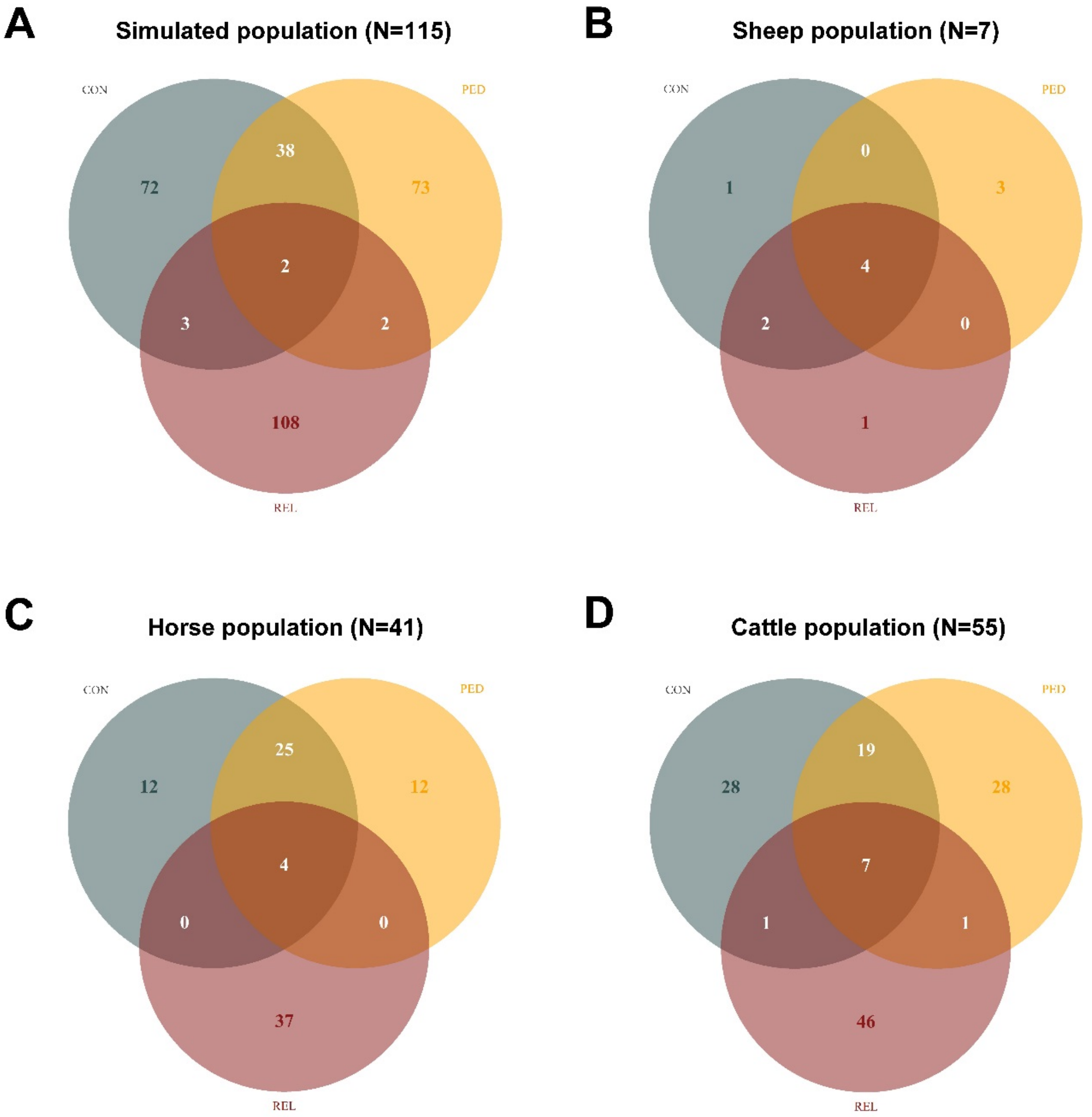
(4) Cattle population

The cattle population represents 2,457 progeny-tested Australian Holstein-Friesian bulls genotyped for 45,765 SNPs.

Results – Identification of key contributors



Results – Identification of key contributors

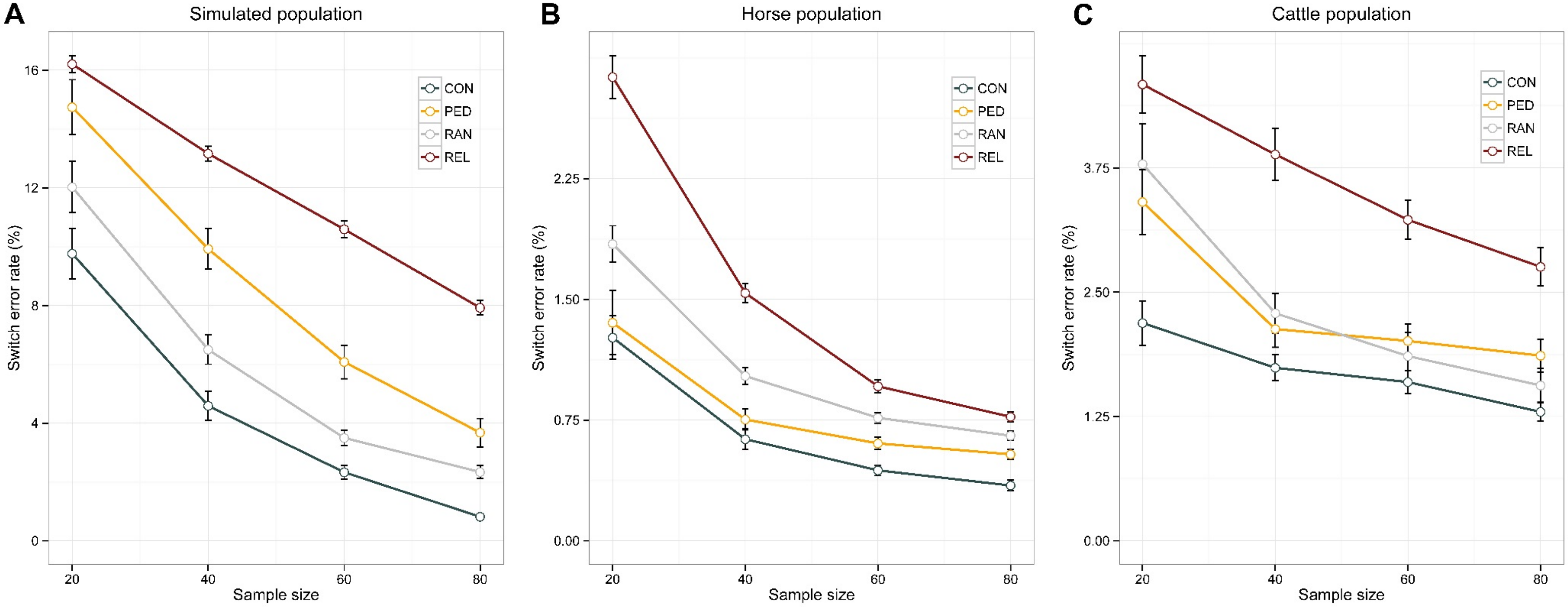


Results – Phasing accuracy

Table. Switch error rates of the selected reference populations within the four datasets.

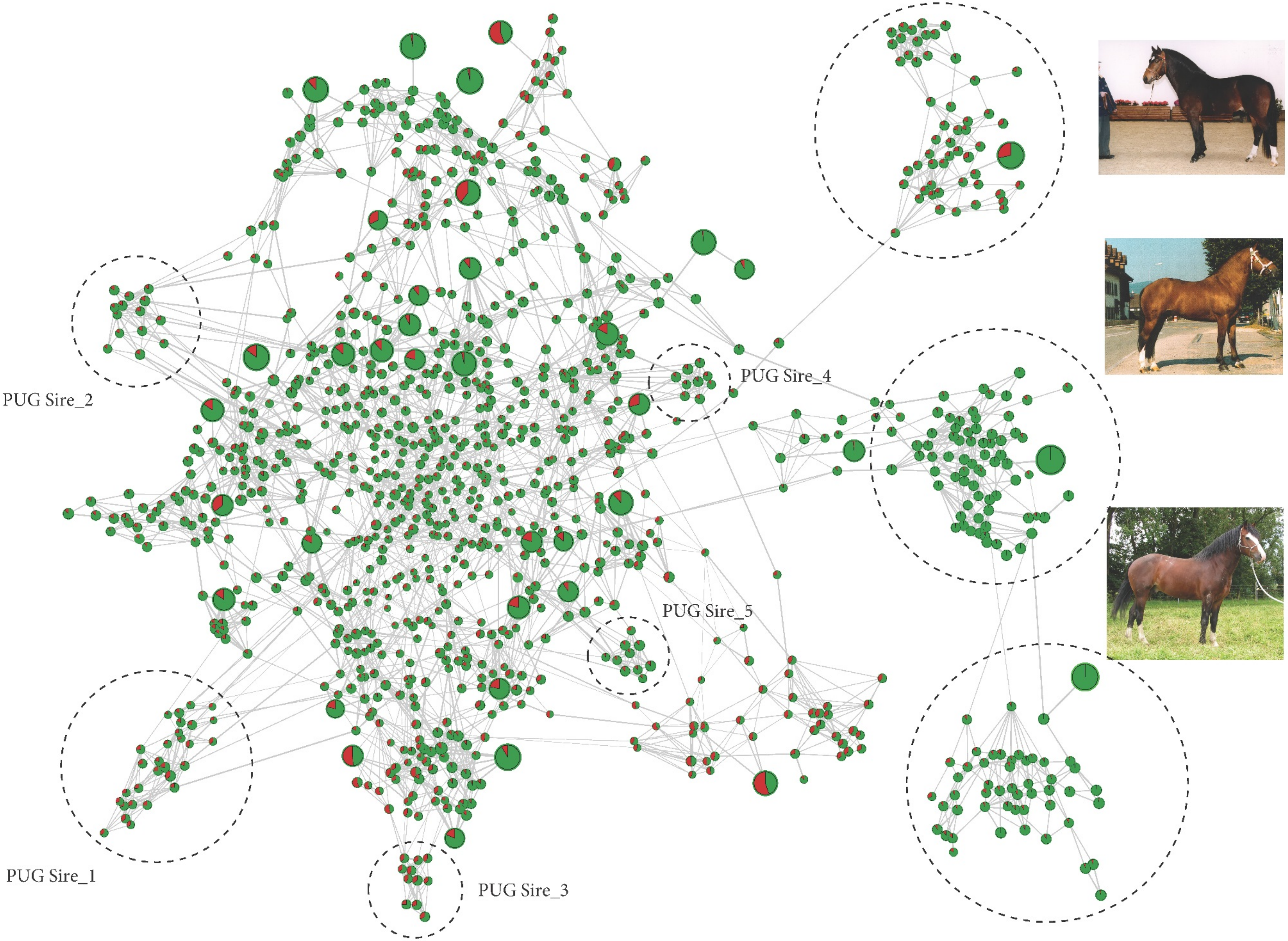
Strategy	Simulated Data (N=115)	Sheep Data (N=7)	Horse Data (N=41)	Cattle Data (N=55)
CON	0.35%	0.26%	0.62%	1.64%
REL	4.27%	0.31%	1.52%	3.48%
PED	0.41%	0.27%	0.74%	2.10%
RAN	1.39%	0.94%	0.97%	1.70%

Results – Phasing accuracy



Conclusion

- Our approach can be successfully applied to identify key contributors (ancestors and influential progeny) within complex population structures.
- With the application of key ancestors it becomes feasible to increase phasing accuracy of selected reference populations.
- REL strategy maximizes genetic diversity, which is not necessarily connected with the identification of key ancestors (simulated data).
- The identification of key ancestors can also support high-resolution population structure analyses (e.g. in combination with model-based clustering and network visualization).


PUG Sire_2
PUG Sire_4
PUG Sire_5
PUG Sire_1
PUG Sire_3

Thank you for your attention

Agroscope Swiss national Stud Farm