Selecting animals for whole genome sequencing: methods optimization and comparison

Adrien Butty[1], M. Sargolzaei[1,2], F. Miglior[1,3], B. Gredler[4], C. Baes[1]

[1]CGIL - University of Guelph; [2]Semex Alliance, Guelph; [3]Canadian Dairy Network, Guelph, Ontario, Canada; [4]Qualitas AG, Zug, Switzerland

The Efficient Dairy Genome Canada Project

- International project aiming to develop genomic evaluation for Feed Efficiency and Methane Emissions of dairy cattle
- Genome Canada Project will sequence 48 animals to improve imputation of the Holstein population at whole-genome sequence density.
- Currently 451 Holstein bulls from the 1000 Bulls Genome Project are sequenced and available to us.

➔ The task: create a list of the animals to sequence.

Introduction

- Genotyping is getting cheaper and cheaper…
  - Routine genotyping is practised for AI bulls everywhere
  - Mostly use of medium or high density chips
    - for genotyping and for genomic evaluation
- Sequencing is also getting cheaper, but still expensive for large scale sequencing
  - Cost per genome: > $1,000
  - Imputation of medium or high density genotypes to whole-genome sequence (WGS) density feasible, **but**

➔ Good reference population needed to more accurately impute **rare variants**!

Selection methods

- Without genotypic information
  - pedigree based; aim for high genetic contribution

- Within genotyped population
  - based on genomic relationship matrix
  - relying on haplotype frequencies
    - target common haplotypes
    - target rare haplotypes

➔Currently, sequenced animals are mostly key ancestors; common haplotypes are thus expected to be sequenced.

Pedigree-based method

- Also called ”Key Ancestors” method

- Aim to find the *n* animals explaining the most of the genetic variance of a population with: $\mathbf{p}_n = \mathbf{A}^{-1}_n \mathbf{c}_n$

$\mathbf{p}_n$	=	vector of the proportion of gene pool captured by the n animals
$\mathbf{A}^{-1}_n$	=	subset of the numerator relationship matrix
$\mathbf{c}_n$	=	average relationships of n selected animals with the entire population
n	=	number of selected animals

➔ True proportion of genetic variance mostly overestimated

➔ The **A** matrix can be replaced by the **G** matrix.

Haplotype-based methods

- Druet et al. (2014)→ maximize haplotype coverage / sum of the haplotype frequency at every SNP

$$Sample\ Weight = \sum_{i=1}^{NHAP} f_i$$

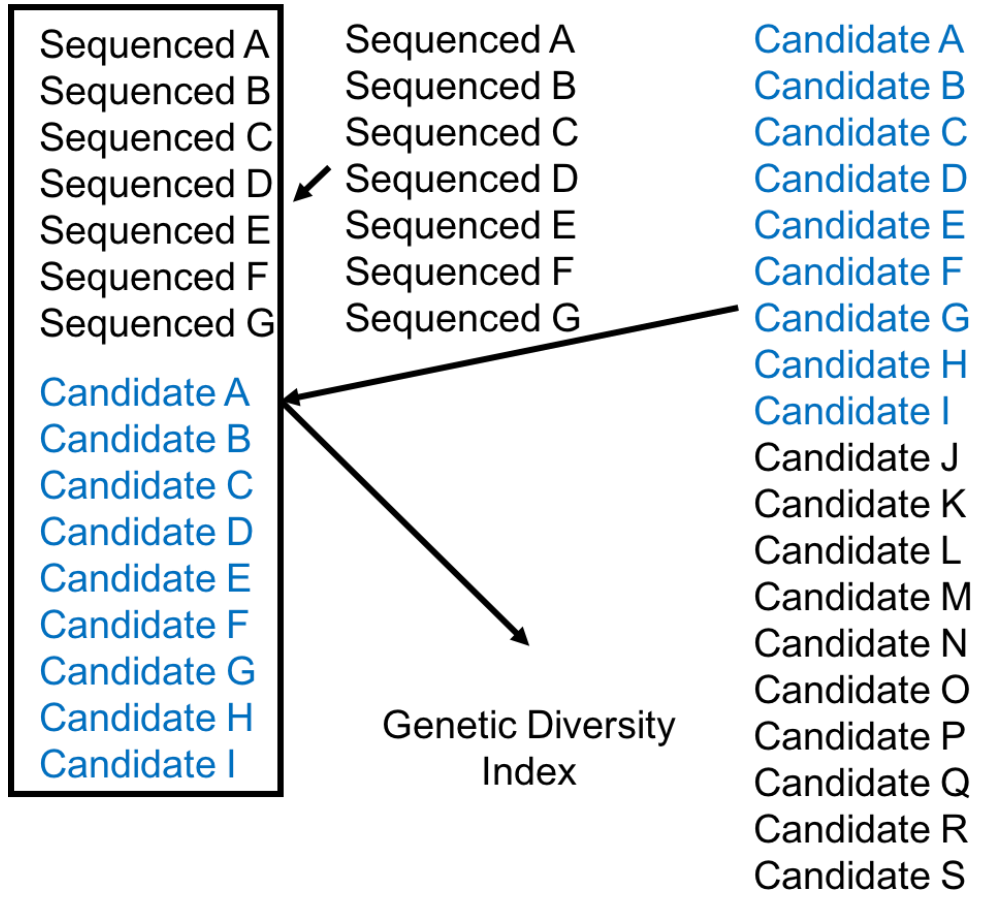
- Bickhart et al. (2016) → prioritize sequencing of rare haplotypes

$$Sample\ Weight = \sum_{i=1}^{NHAP} f_i^2 - 2f_i + 1$$

Our approach

- Optimize the genetic diversity of the sequenced animals
  - use of probabilistic algorithm; simulated annealing
  - optimization based on a group of animals and not on individuals
    - already sequenced and newly selected animals accounted for
  - genetic diversity = sum of count of unique haplotypes

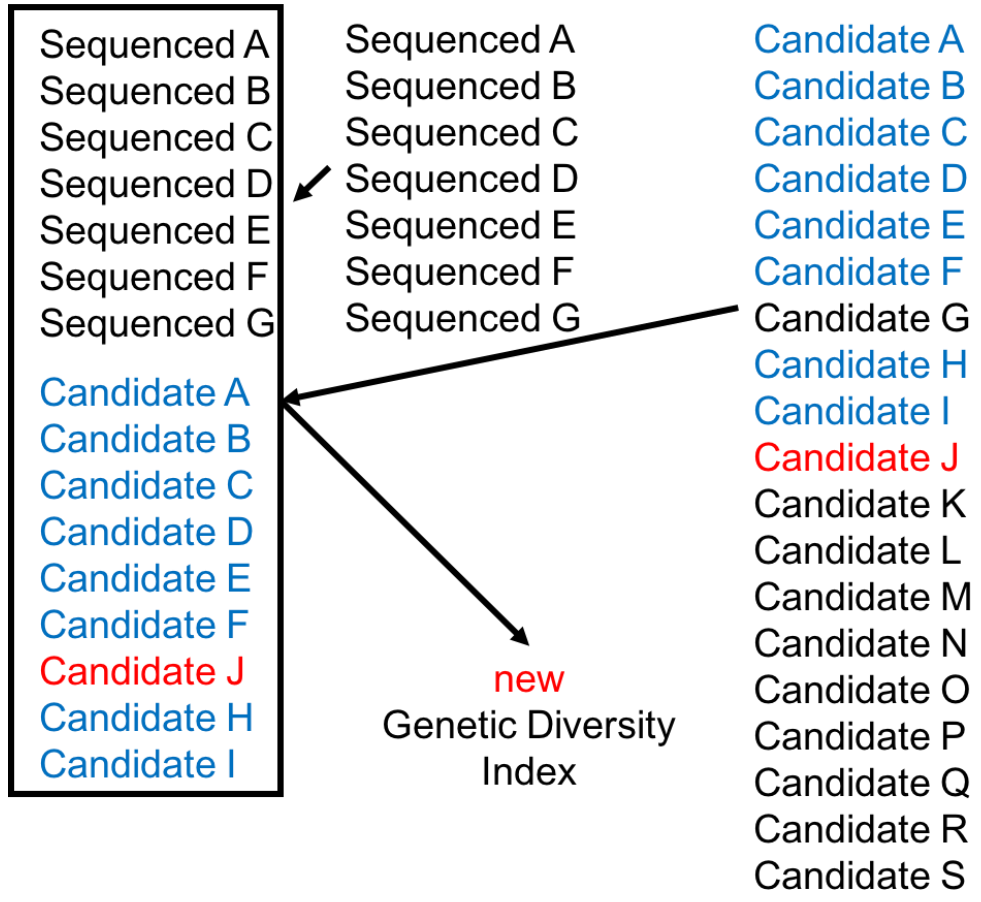
➔ Select animals that will enable **higher accuracy of imputation of rare variants** within the Holstein North-American population



Our approach

- Optimize the genetic diversity of the sequenced animals.
  - use of probabilistic algorithm; simulated annealing
  - optimization based on a group of animals and not on individuals
    - already sequenced and newly selected animals accounted for
  - genetic diversity = sum of count of unique haplotypes

➔ Select animals that will enable **higher accuracy of imputation of rare variants** within the Holstein North-American population



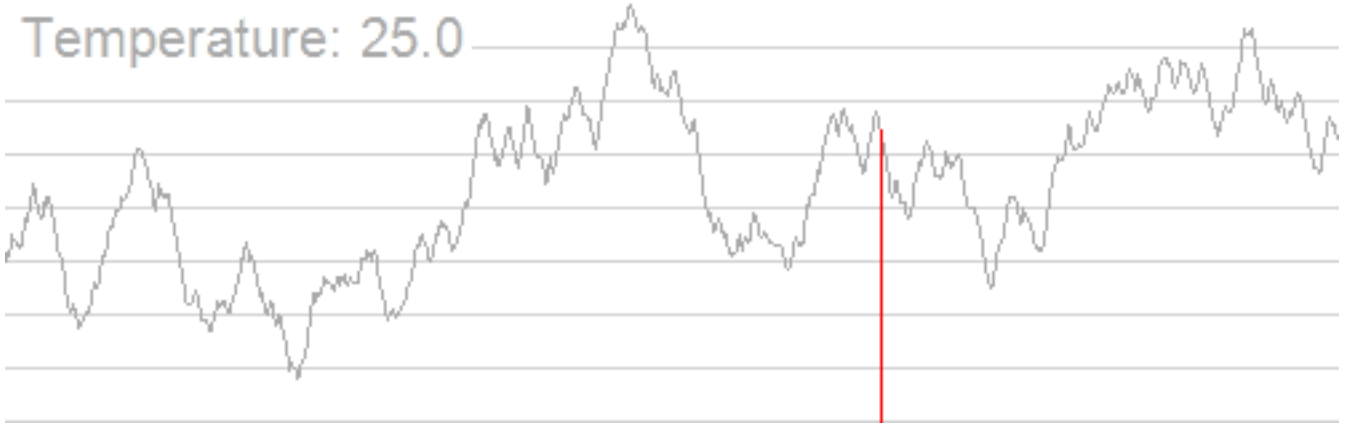
Index of genetic diversity

		Haplotype A	Haplotype B	Haplotype C	Haplotype D
Animal1	H1	10001110010010010001	00101001010110011100	10110000110100111010	01010001010101111110
	H2	10001110010010010001	00101001010110011100	10110000110100111010	01010001010101111110
Animal2	H1	10001110010010010001	00101001010110011100	10110000110100111010	01010001010101111110
	H2	10110010110001111001	00001101010110011100	01000100000100001100	10000000110001010110
Animal3	H1	10000010110010011001	00001101010110011100	10000000000100111111	11111110001010001111
	H2	10110010110001111001	00001101010110011100	01000100000100001100	10000000110001010110
Animal4	H1	10010010010000011001	11001000000110011001	00011000101100001100	01010001010101111110
	H2	00000011110010011001	11101101010110011100	11100100101101001101	01010001010101111110
Animal5	H1	00000011110010011001	11101101010110011100	10000000100100001101	11111110001010001111
	H2	10001110010010010001	00101001010110011100	10110000110100111010	01010001010101111110

10001110010010010001		00101001010110011100		10110000110100111010		01010001010101111110		
10110010110001111001		00001101010110011100		01000100000100001100		10000000110001010110		
10000010110010011001		11001000000110011001		10000000000100111111		11111110001010001111		
10010010010000011001		11101101010110011100		00011000101100001100				
00000011110010011001				11100100101101001101				
				10000000100100001101				
5	+	4	+	6	+	3	=	18

Simulated annealing

- Find the global optimum when several local optima exist
- Accept “bad” steps when these are probabilistically near to another optimum (the good one, who knows…?)
  - high temperature → big “bad” steps accepted
  - low temperature → only small “bad” steps accepted

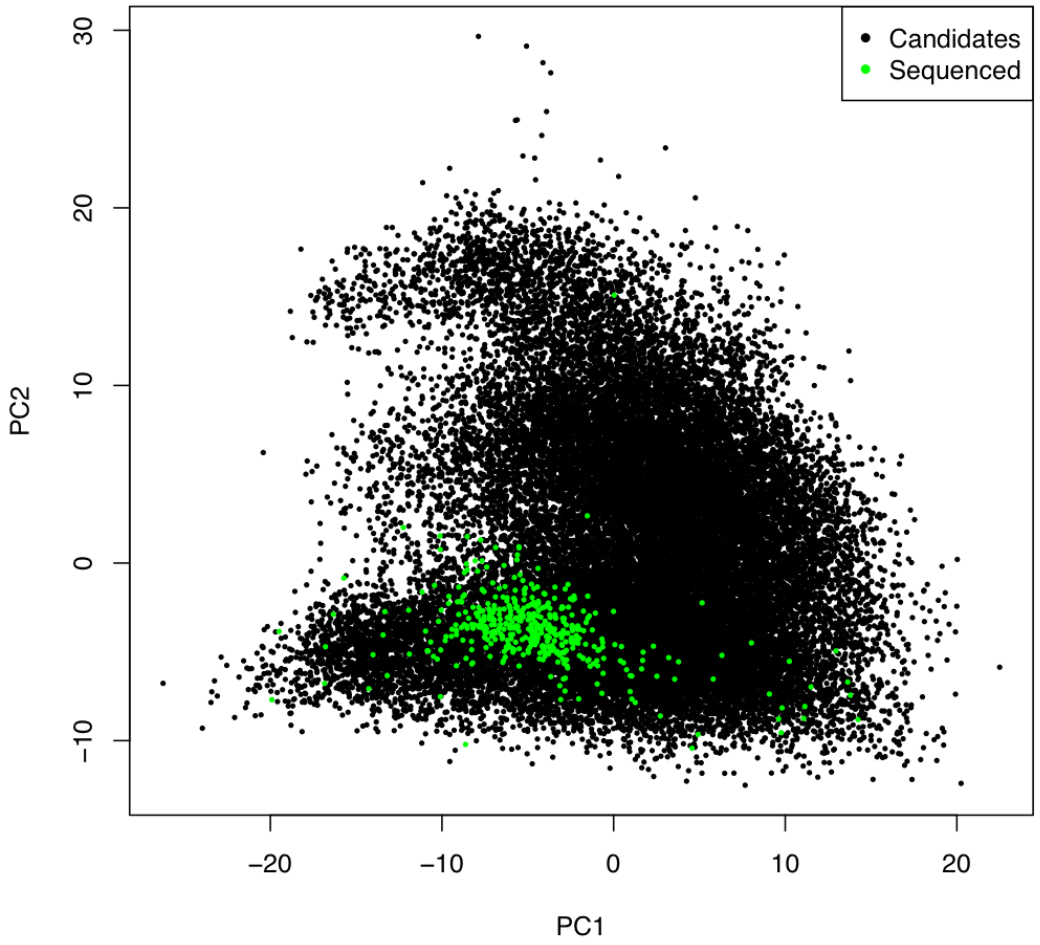


Source: Wiki User Kingpin13

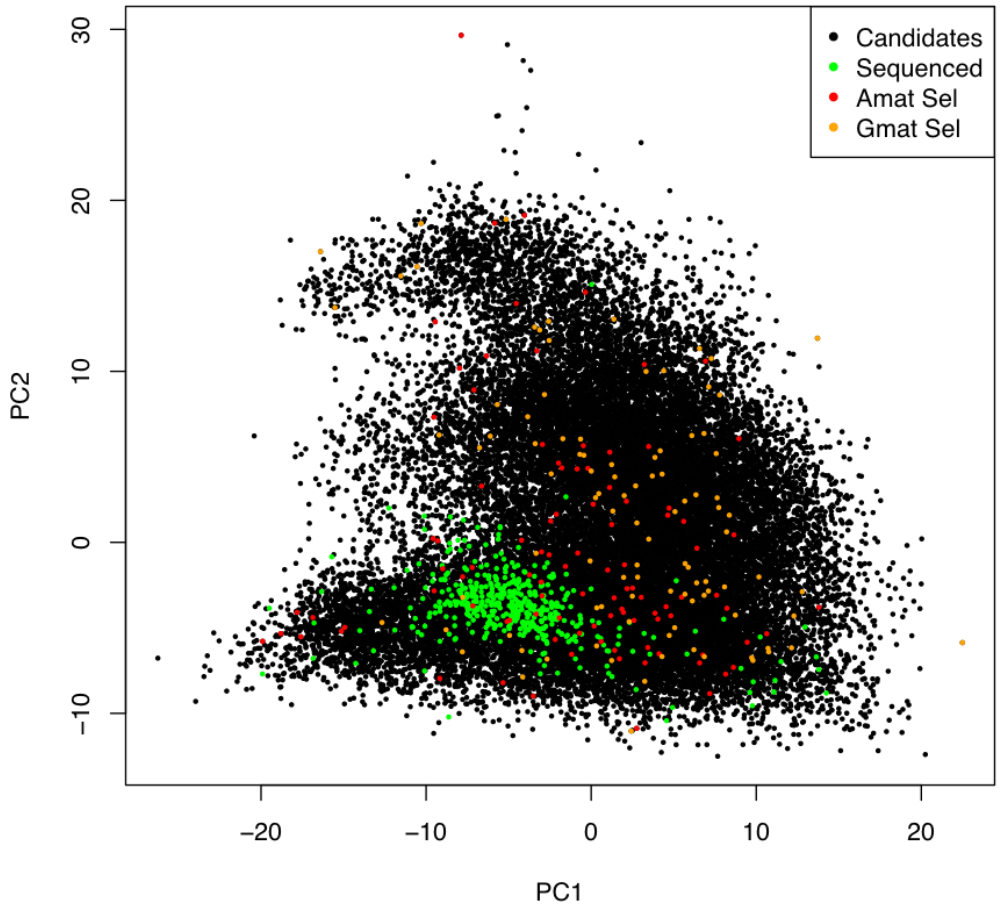
Animals and genotypes

- Candidates
  - Holstein bulls born after 01/01/2011 in Canada or USA and genotyped with 50K or higher density — 35,706 animals
  - After filtering possible crossbreds — 32,000 animals
- Sequenced animals
  - HOL or RED animals from Run5 of the 1'000 Bulls Genome Project — 451 animals
- All genotypes (back-) imputed to 50K panel with FImpute
  - Pedigree contained 151,436 animals, up to 48 generations
  - 44,347 autosomal SNP

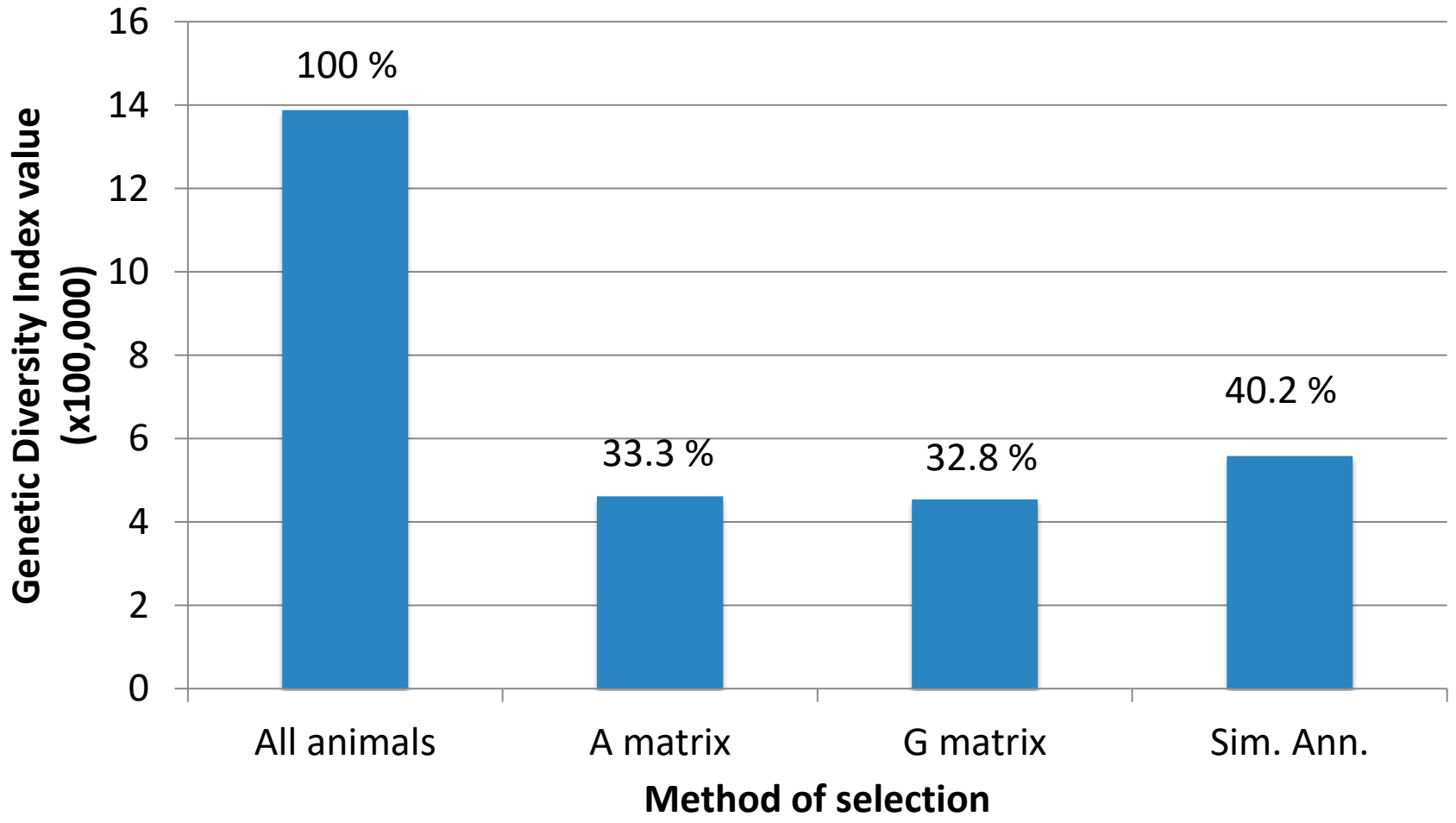
PCA distributions



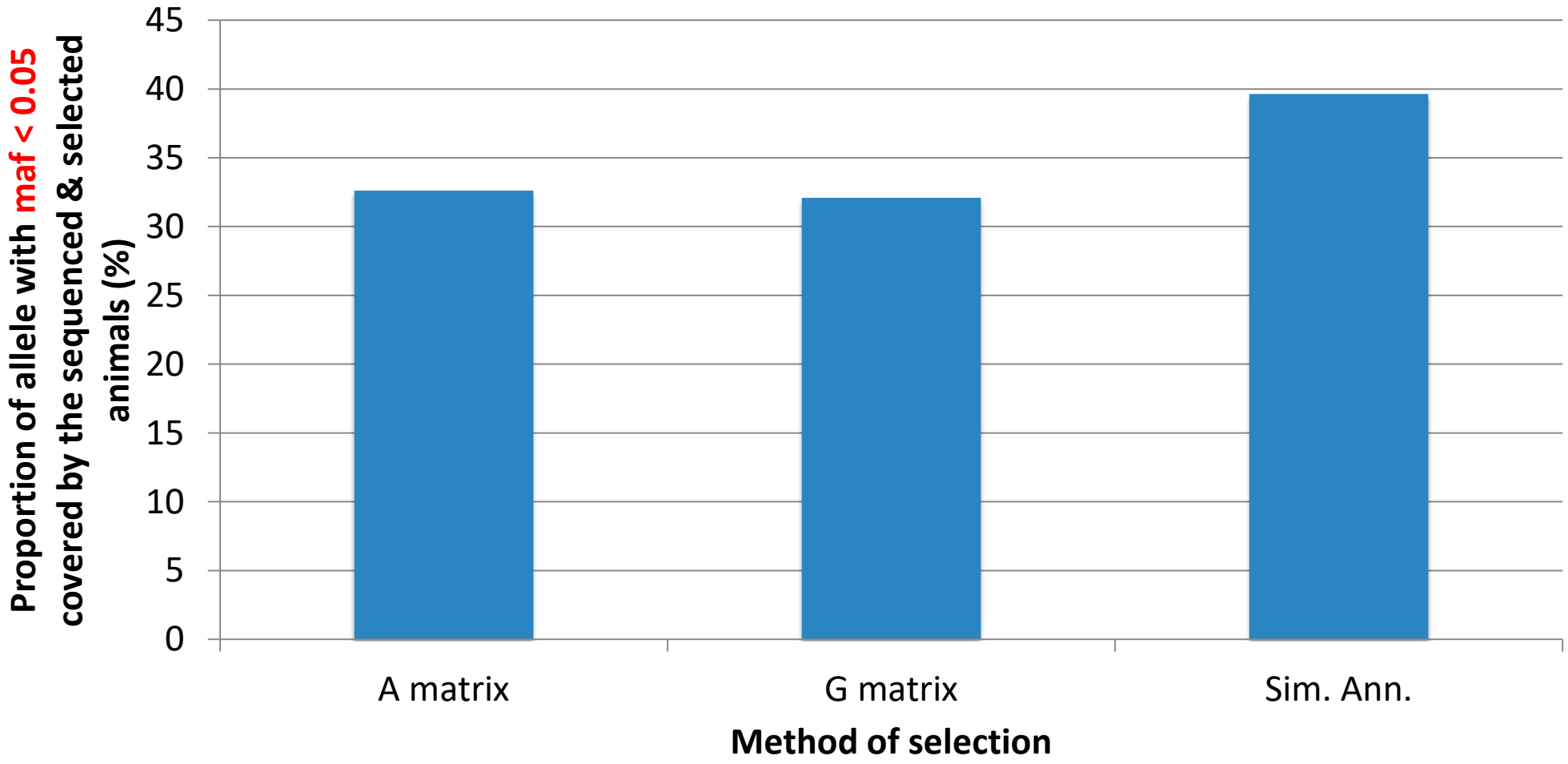
PCA distributions



Genetic Diversity Index



Proportion of rare haplotype alleles



In conclusion …

- Methods of selection for sequencing have shifted **from pedigree or SNP-based to haplotype-based methods**
- Aiming discovery and coverage of rare variants at sequencing should **improve their imputation accuracy**
- Our method identifies animals **representing the whole population** as well as previous methods
- Using our approach, we hope to better cover **rare haplotypes** in future sequenced populations

Acknowledgements

We very gratefully acknowledge support from our funders and collaborators:

GenomeCanada

GenomeAlberta

Ontario Genomics

QUALITAS

ALMA Alberta Livestock and Meat Agency Ltd.

GrowSafe Systems

SRUC

Ontario Ministry of Research & Innovation

USDA

Réseau laitier canadien Canadian Dairy Network

CANWEST DHI

State Government Victoria
Department of Economic Development, Jobs, Transport & Resources

SEMEX ALLIANCE

valacta

University of Guelph

University of Alberta

Wageningen UR For quality of life

Ontario Ministry of Agriculture, Food and Rural Affairs

Aarhus University

Thank you for your attention!

more about the Efficient Dairy Genome Project:

« An international initiative to decrease the environmental footprint of dairy cattle using genomics »

by Filippo Miglior,
Wednesday at **14:45** in room **3B**

University of Guelph
Changing Lives Improving Life
DAIRY at GUELPH
CANADA'S DAIRY UNIVERSITY

THE Efficient Dairy GENOME PROJECT