Quality of reconstructed haplotypes in cattle data

Malena Erbe[1], Magdalena Frąszczak[2], Henner Simianer[1] und Joanna Szyda[2]

1 Department of Animal Sciences, Animal Breeding and Genetics Group, Georg-August-University Göttingen, Germany

2 Wrocław University of Environmental and Life Science, Institute of Animal Genetics, Biostatistics Group

Introduction

- Data from genotyping platforms available for different livestock species ➔ no haplotype information
- Not a disadvantage for genomic breeding value prediction just based on genotypic information, but: disadvantage for applications like LD calculation or haplotype based association studies or prediction approaches
- Possible solution: „in silico“ reconstruction of the haplotypes (“phasing”)

This study

- Several free programs are available
  - most of them also impute missing data
  - in livestock data: accuracy of imputation has often been studied, quality of reconstructed haplotypes only rarely

- Aim of this study:

  Assessment of the quality of haplotyping

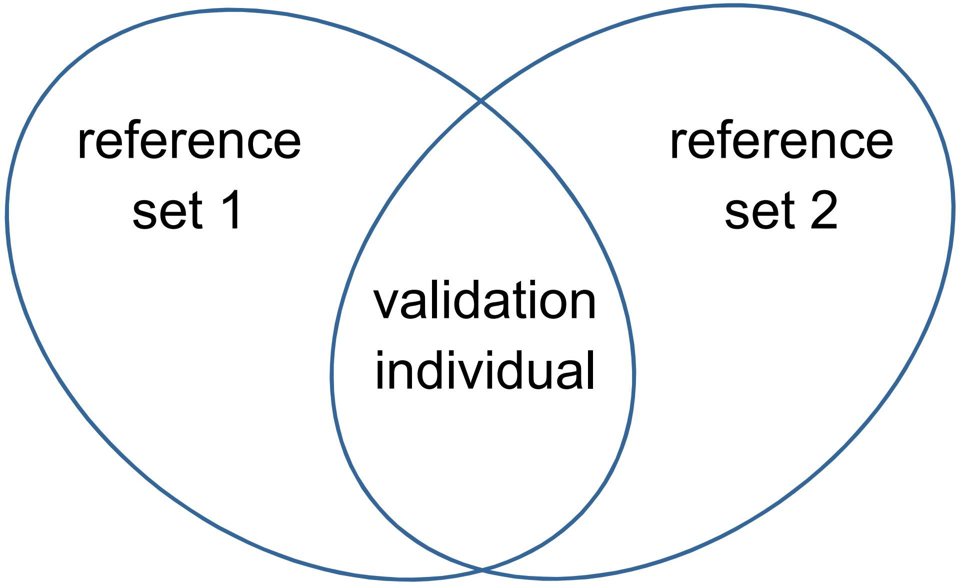
  with different software tools (freely available, reasonably fast)

  in a real cattle data set (true haplotypes not available)

This study

- Idea of implementation:

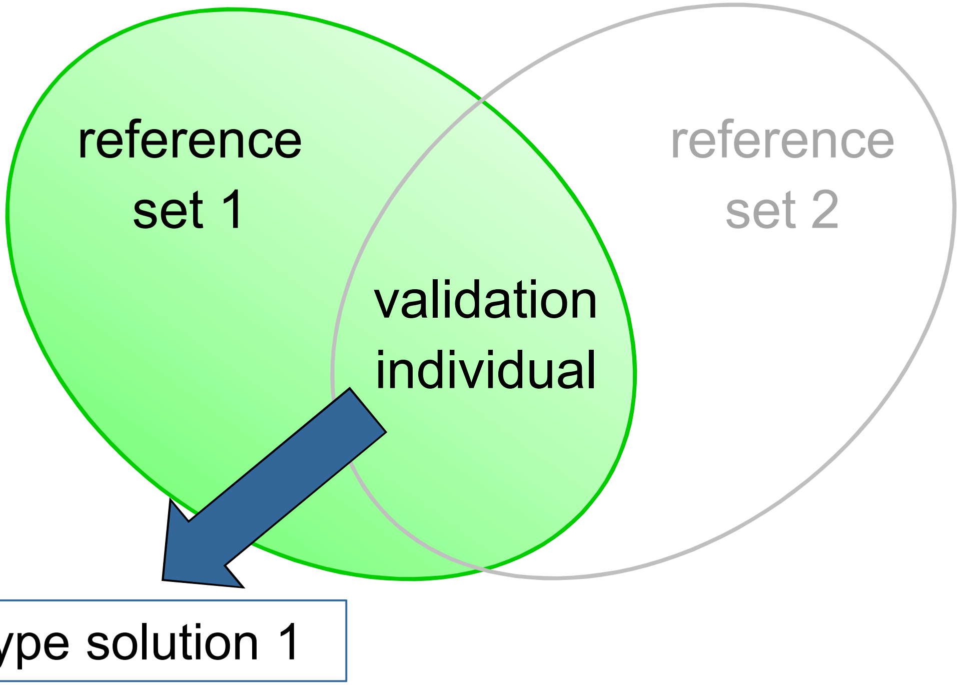
    1. Reconstruct haplotypes for various validation individuals based on two different sets of other individuals ("reference", phased simultaneously with validation set)



This study

- Idea of implementation:

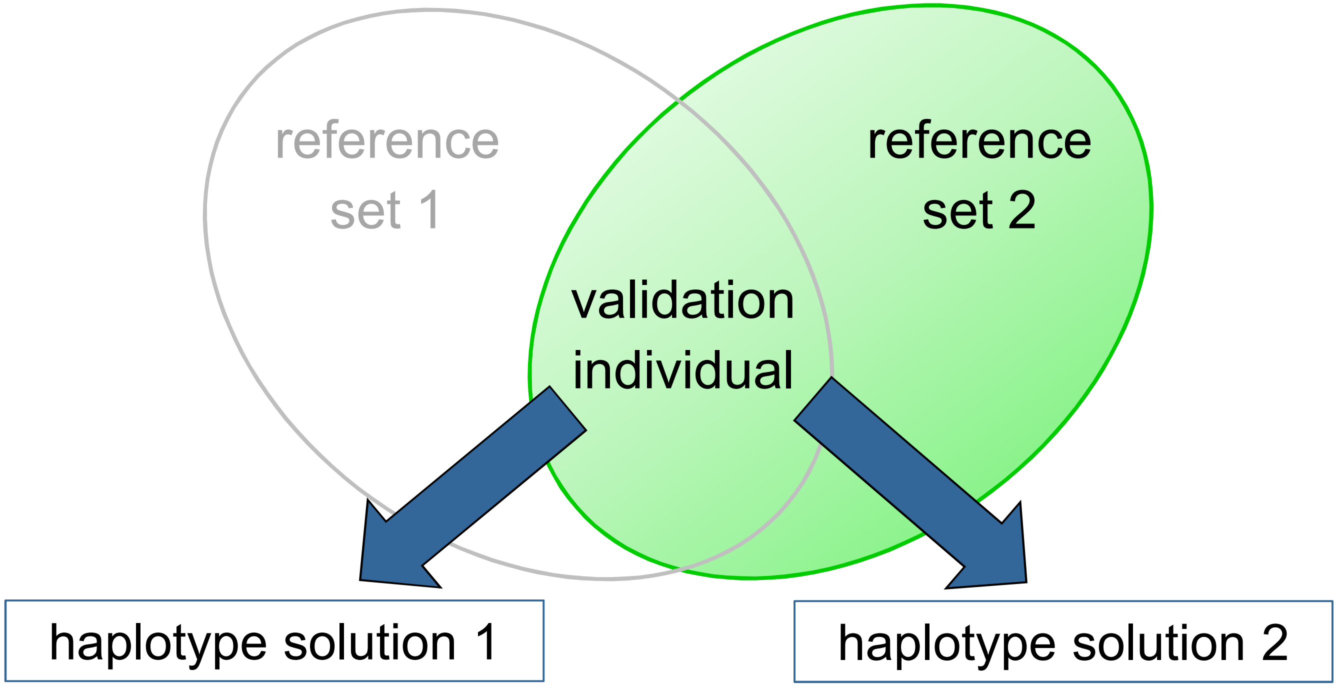
  1. Reconstruct haplotypes for various validation individuals based on two different sets of other individuals ("reference", phased simultaneously with validation set)



This study

- Idea of implementation:

  1. Reconstruct haplotypes for various validation individuals based on two different sets of other individuals ("reference", phased simultaneously with validation set)

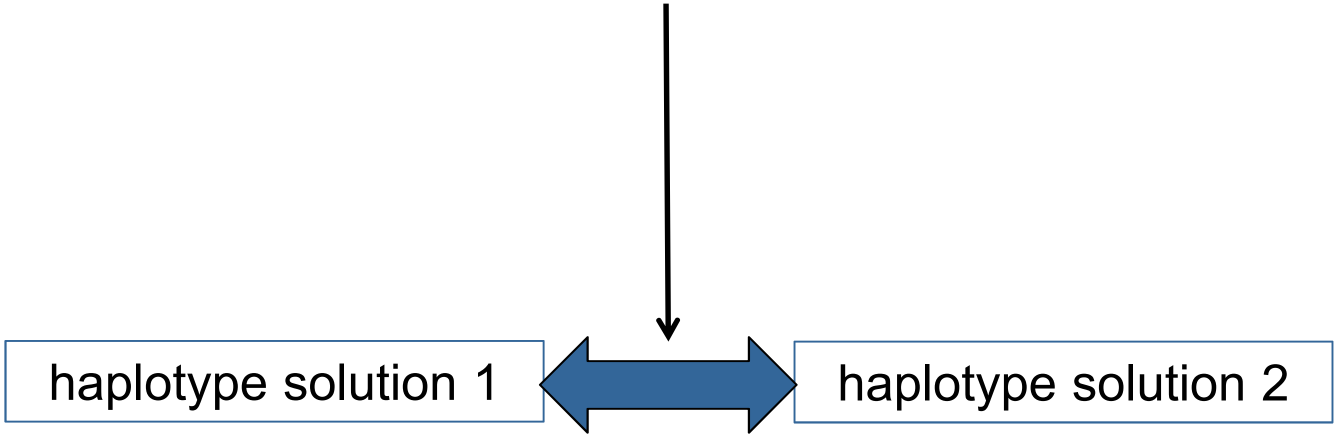


This study

- Idea of implementation:

1. Reconstruct haplotypes for various validation individuals based on two different sets of other individuals ("reference", phased simultaneously with validation set)

2. Compare the two haplotypes obtained for a validation individual when phased with different references



haplotype solution 1

haplotype solution 2

Data set

- 5501 Holstein Friesian bulls genotyped with Illumina 50K Bovine SNP Chip (Matukumalli et al. 2009)
- quality control for SNPs: minor allele frequency > 0.01
  call rate per SNP > 0.95
- only SNPs on Chr 1 (2767 SNPs) were used for further analyses

Programs

- BEAGLE (Browning & Browning, 2007)
  - only based on LD structure
  - whole chromosome phased at once
  - localized haplotype clusters are built, sampling of haplotypes in a hidden Markov model
- findhap (VanRaden, 2011)
  - uses pedigree information and LD structure
  - divides chromosome in smaller parts and reconstructs haplotypes within these parts
  - builds haplotype library against which genotypes are checked

Scenarios

scenario	validation individuals	reference sets
“sire”	70 youngest bulls with genotyped sire	with the respective sire 2x (50, 100, 250, …, 2500)
“no sons”	70 bulls with at least 5 sons in the whole data set	none of the sons of validation bulls 2x (50, 100, 250, …, 1500)
“sons“	70 bulls with at least 5 sons in the whole data set	same number of sons of the validation bulls in the two references 2x (50, 100, 250, …, 2500)

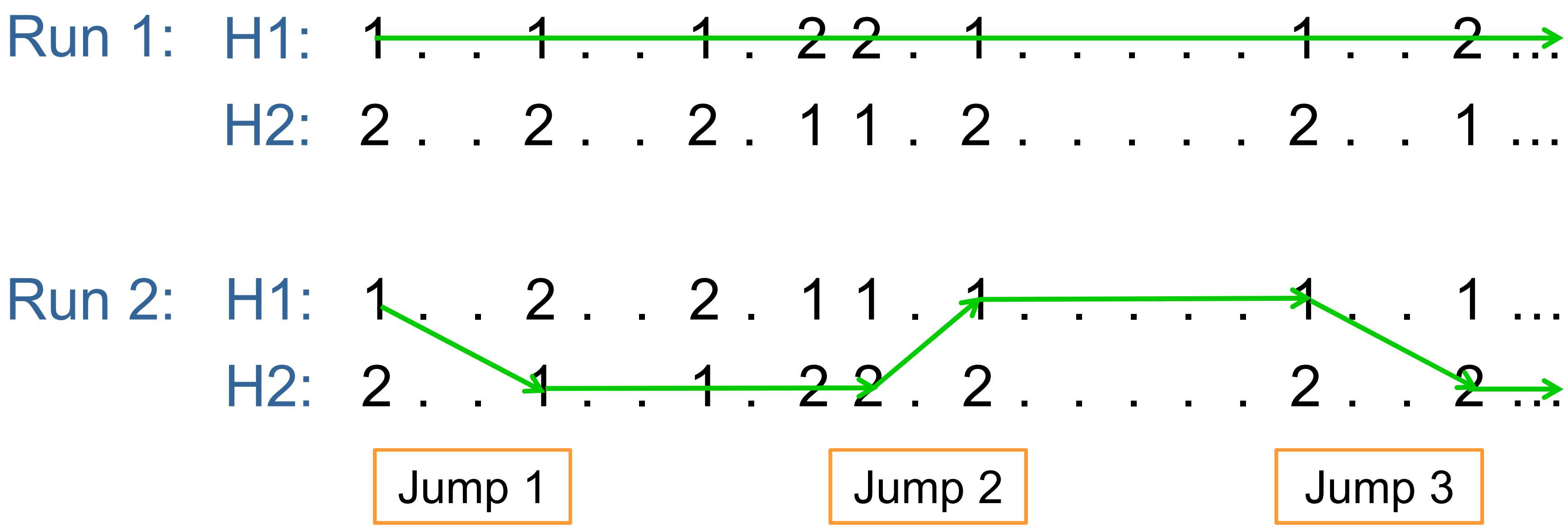
Criteria of comparison

Comparison between the two runs for each validation individual:

➔number of „jumps“ (positions where phase changes)

Criteria of comparison

Example:



Run 1: H1: 1 . . 1 . . 1 . 2 2 . 1 . . . . . 1 . . 2 ...

H2: 2 . . 2 . . 2 . 1 1 . 2 . . . . . 2 . . 1 ...

Run 2: H1: 1 . . 2 . . 2 . 1 1 . 1 . . . . . 1 . . 1 ...

H2: 2 . . 1 . . 1 . 2 2 . 2 . . . . . 2 . . 2 ...

Jump 1

Jump 2

Jump 3

Criteria of comparison

Comparison between the two runs for each validation individual:

➔ number of „jumps“ (positions where phase changes)

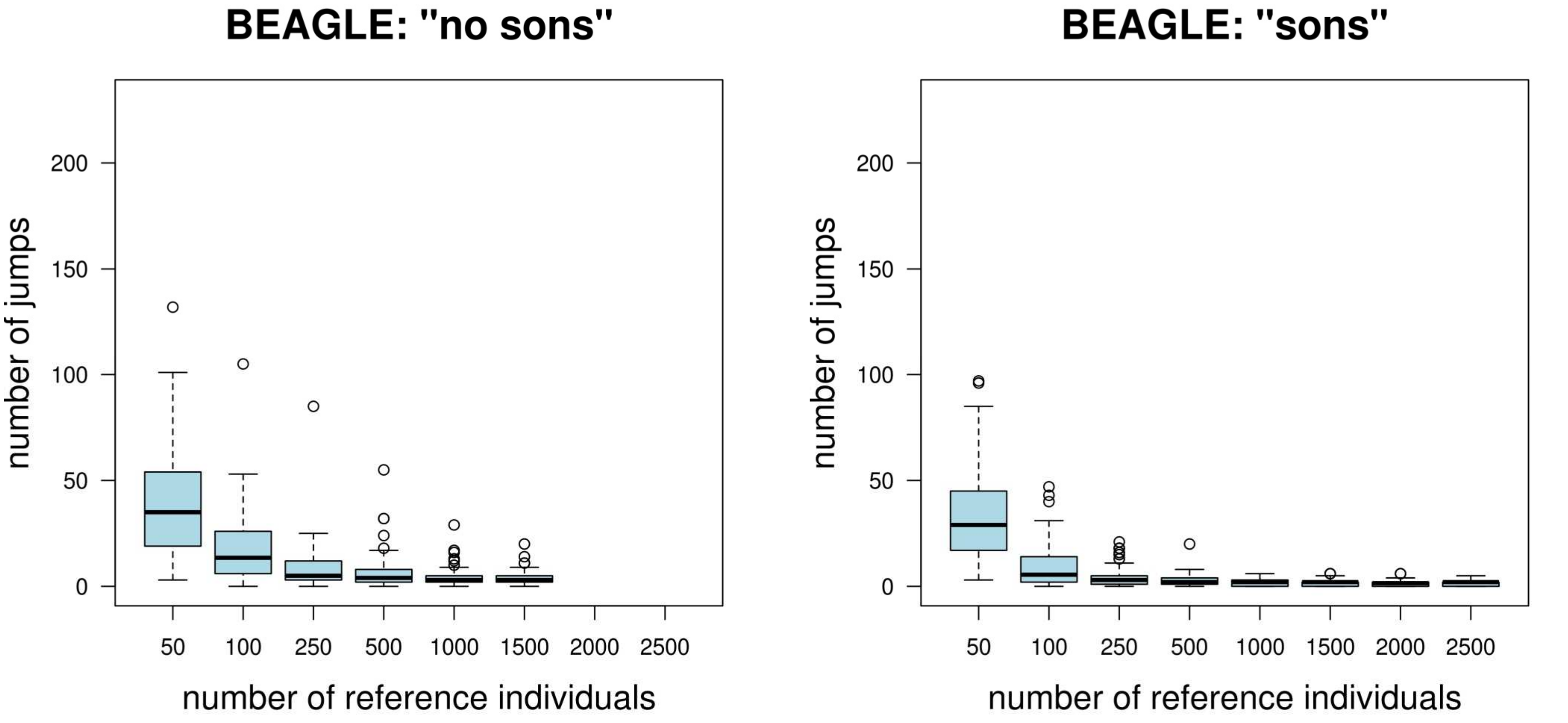
➔ percentage of positions equally phased

Criteria of comparison

		Example A, 2 jumps:		Example B, 2 jumps:
Run 1:	H1:	1 1 1 2 2 1 1 2 …	H1:	1 1 1 2 2 1 1 2 …
	H2:	2 2 2 1 1 2 2 1 …	H2:	2 2 2 1 1 2 2 1 …
Run 2:	H1:	1 2 1 2 2 1 1 2 …	H1:	1 1 1 1 1 2 2 2 …
	H2:	2 1 2 1 1 2 2 1 …	H2:	2 2 2 2 2 1 1 1 …
		87,5% equally phased		50% equally phased

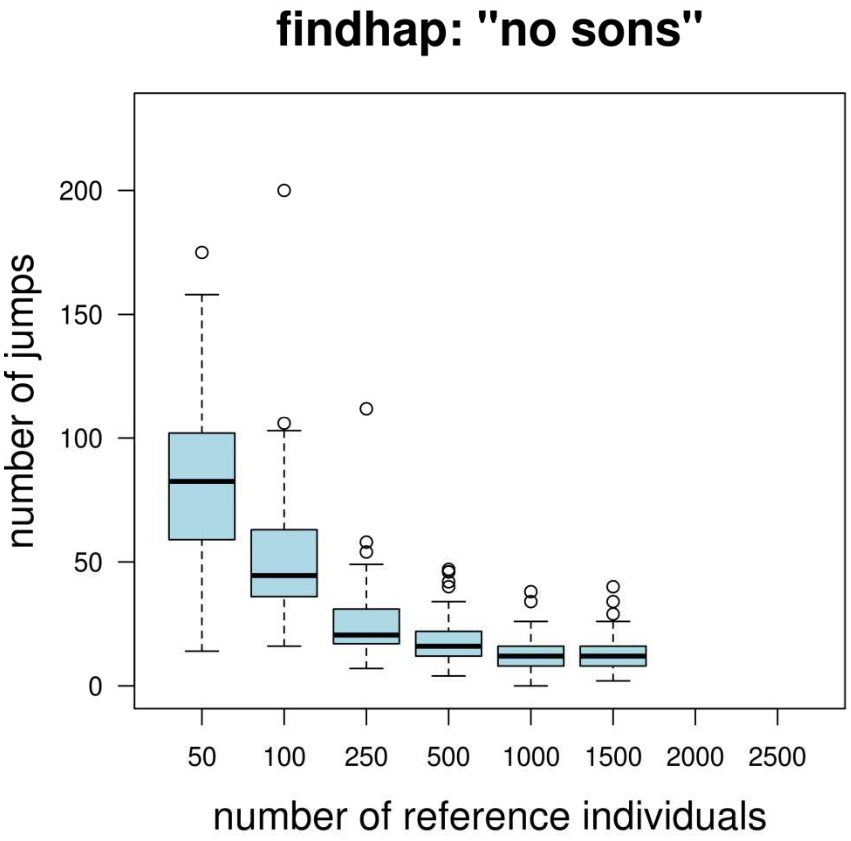
Results: Number of Jumps

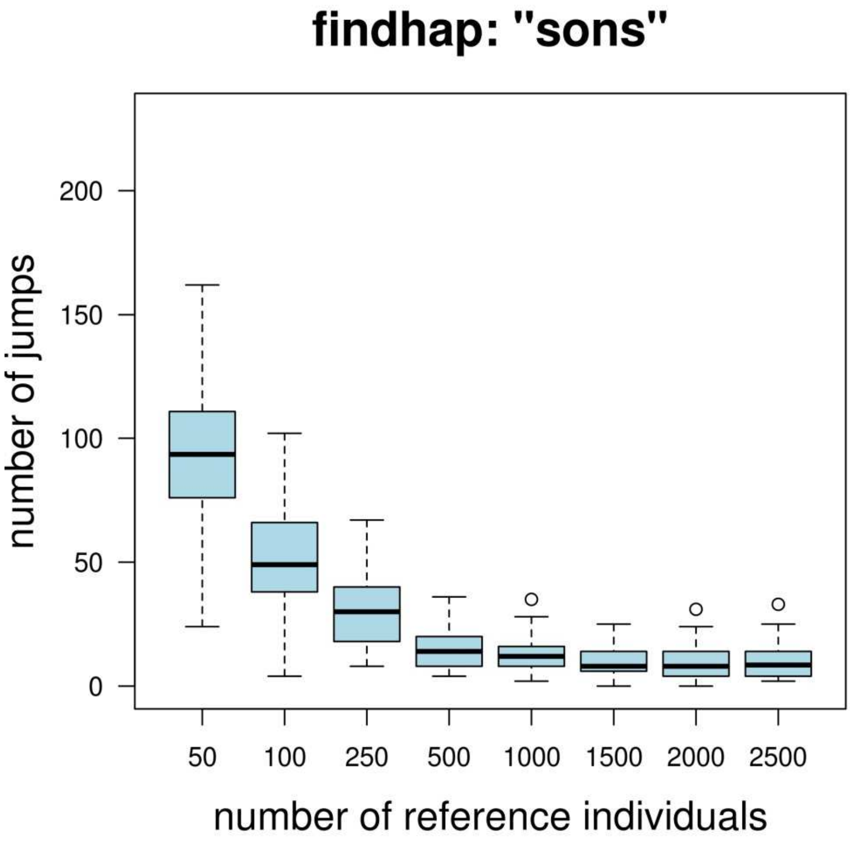
- Number of jumps with BEAGLE



Results: Number of Jumps

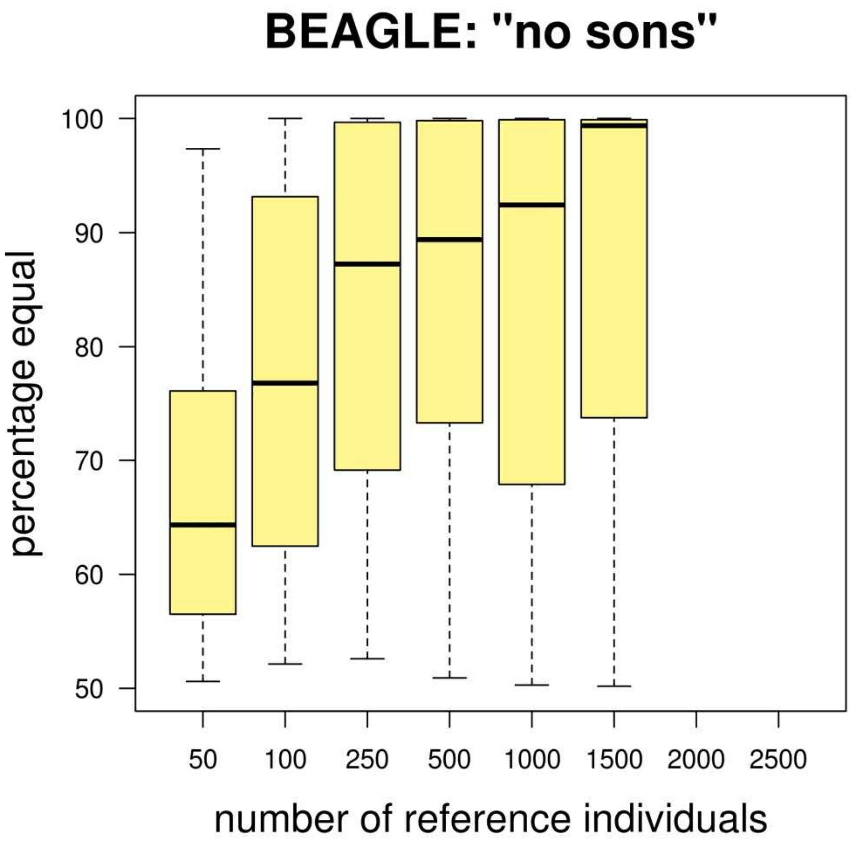
- Number of jumps with findhap

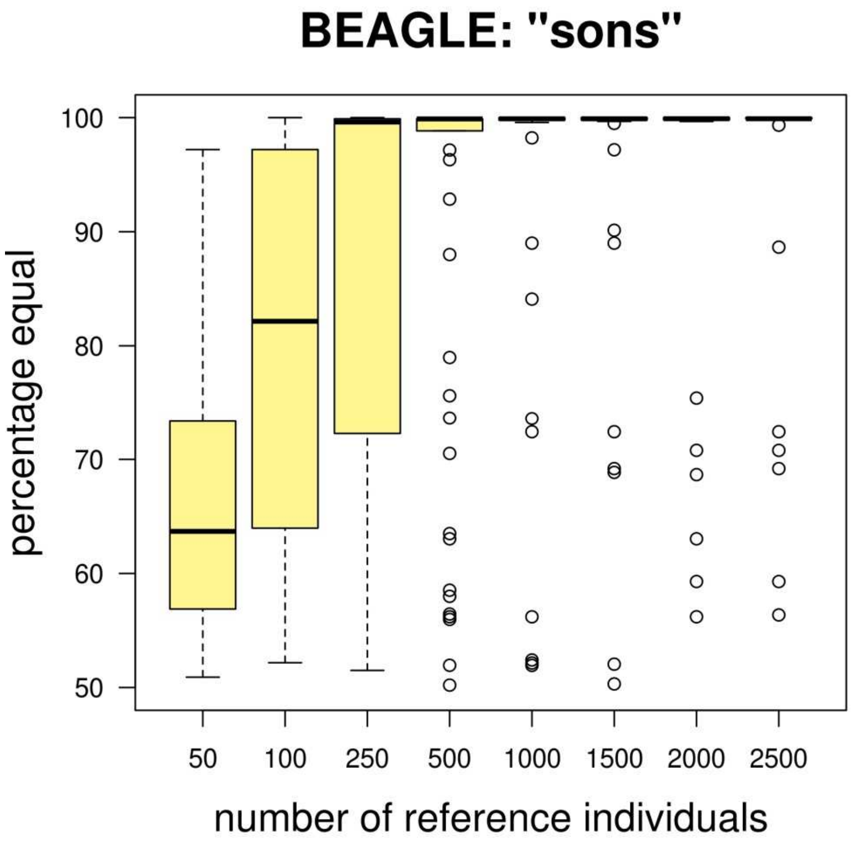




Results: Percentage equally phased

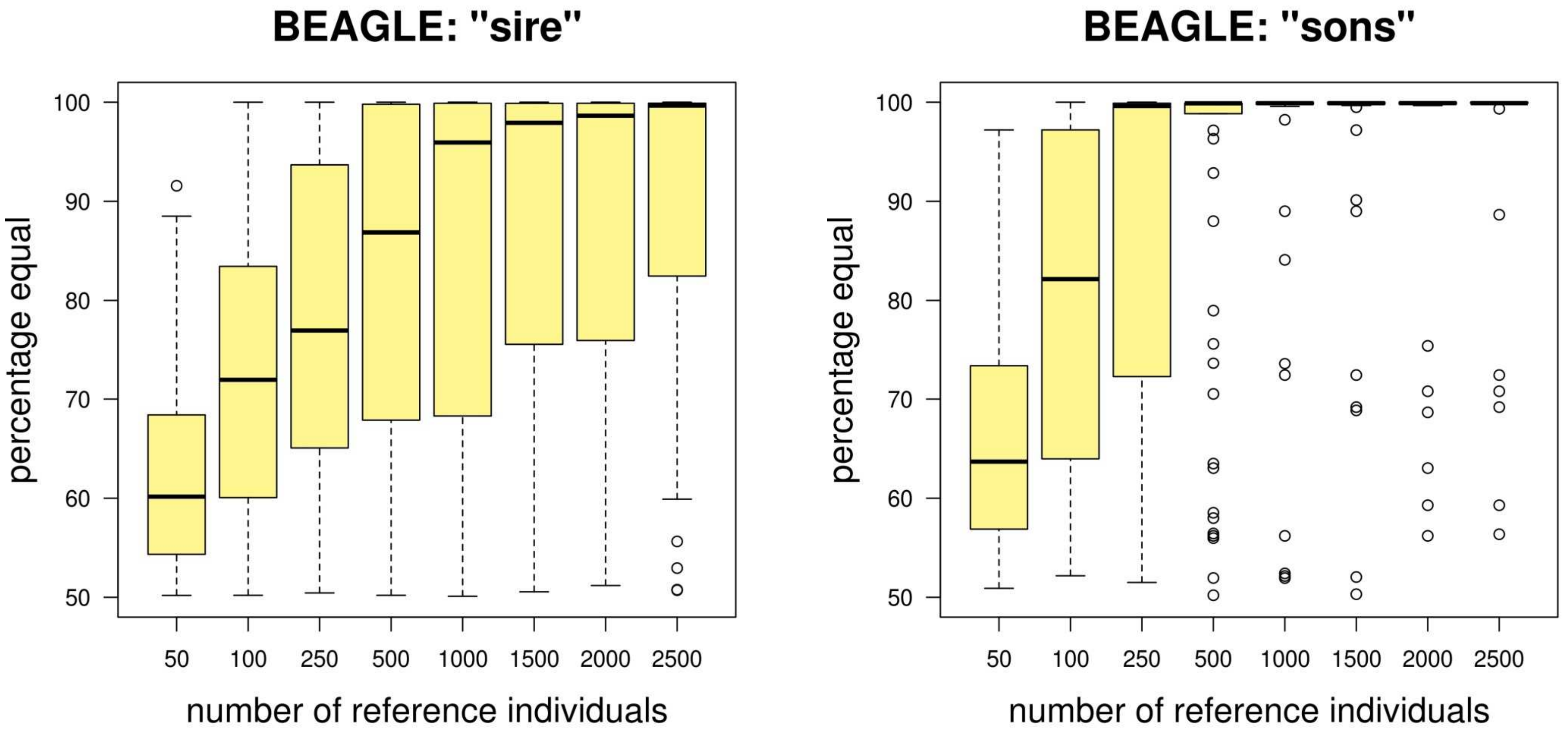
- Percentage equally phased with BEAGLE





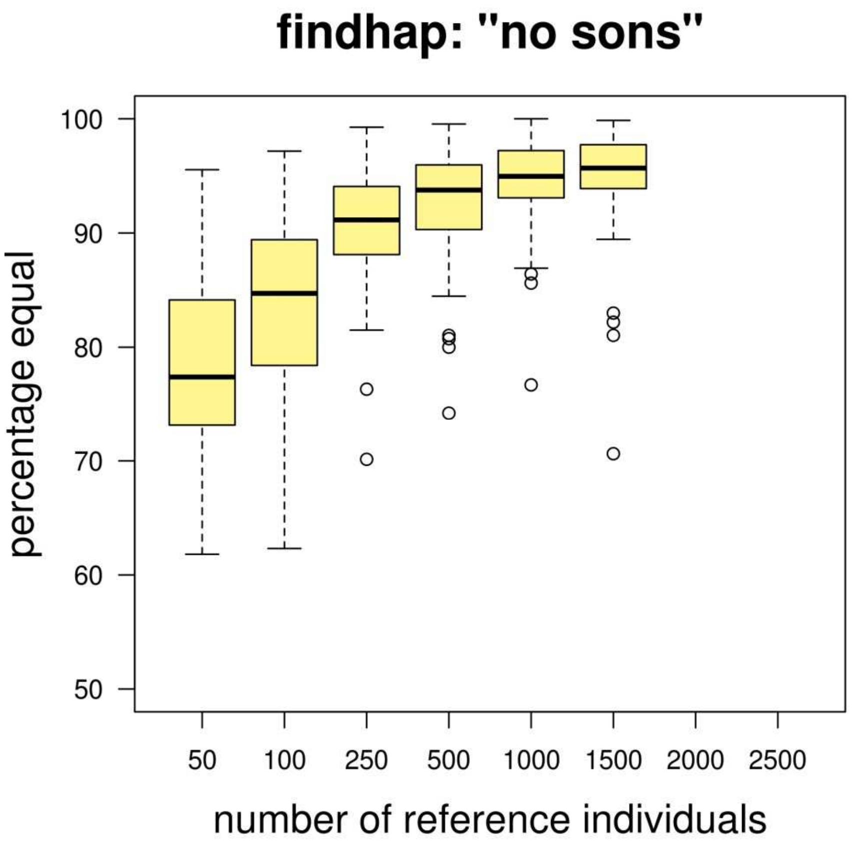
Results: Percentage equally phased

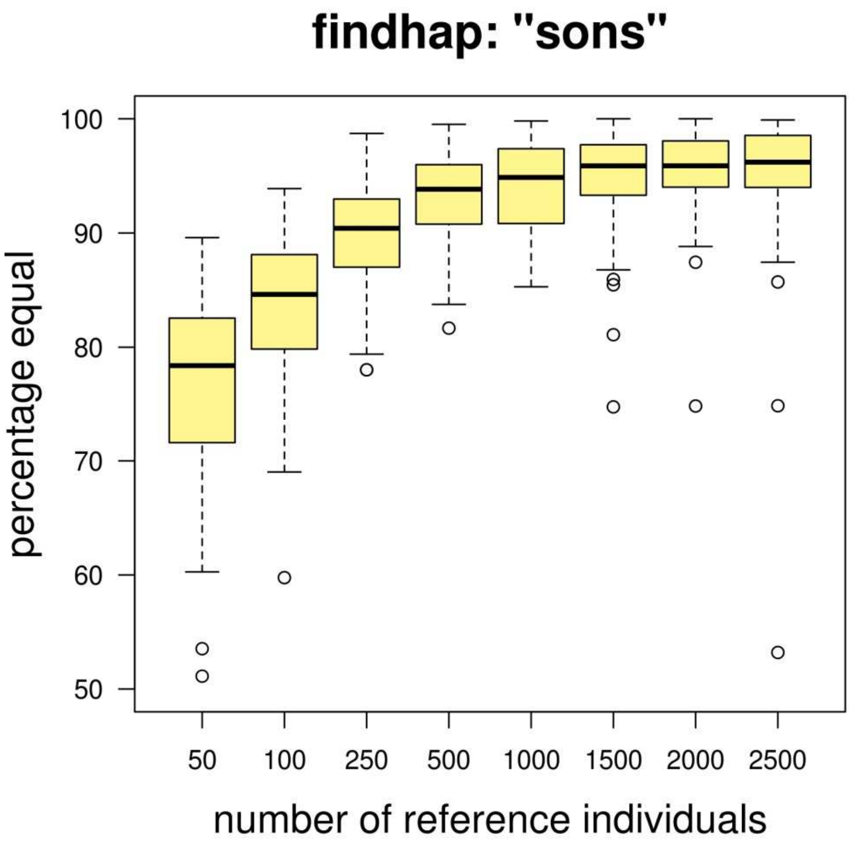
- Percentage equally phased with BEAGLE



Results: Percentage equally phased

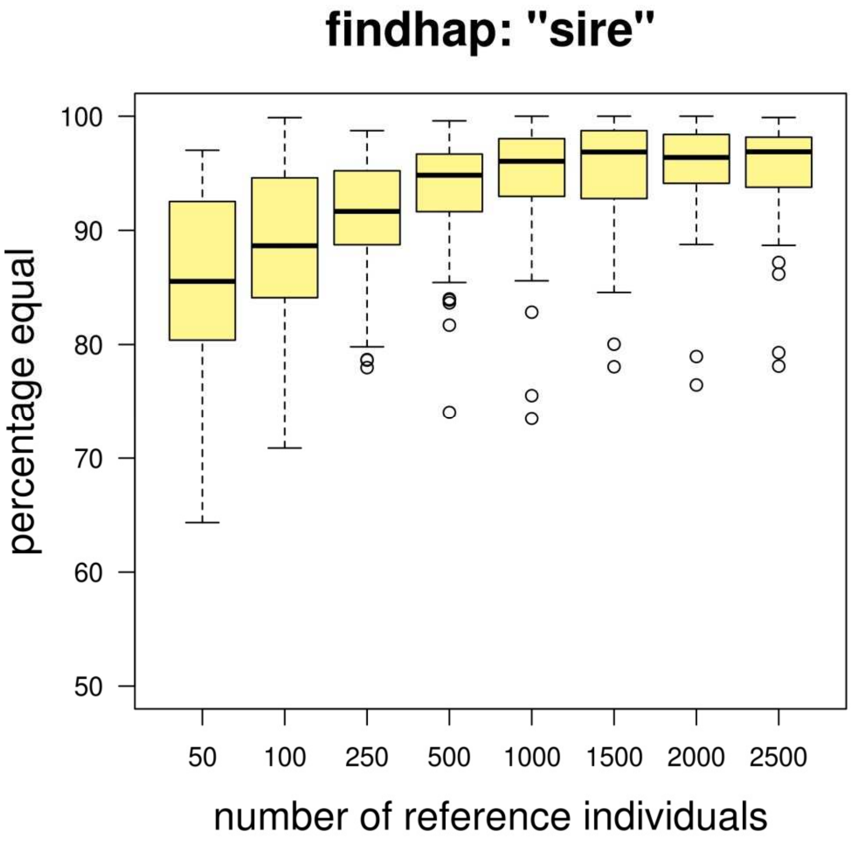
- Percentage equally phased with findhap

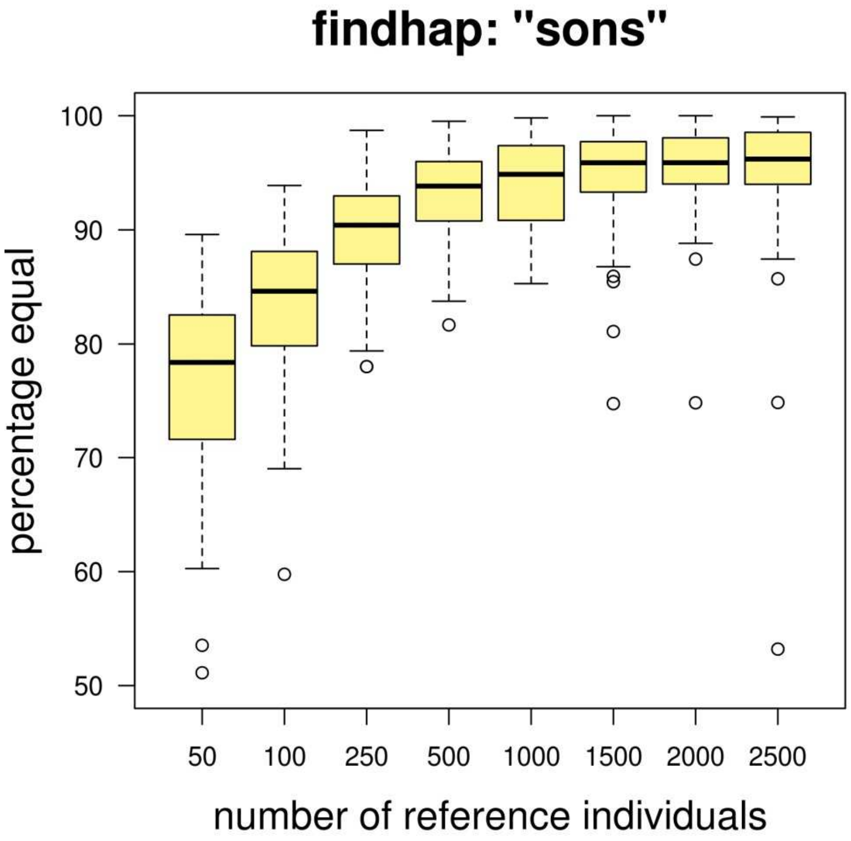




Results: Percentage equally phased

- Percentage equally phased with findhap





Summary

- Number of jumps: strongly influenced by size of reference set
- Percentage equally phased: higher with larger reference sets and higher relationship between reference and validation individuals
- BEAGLE ⇔ findhap: BEAGLE performed better in terms of number of jumps, but in many scenarios worse in terms of percentage equally phased

Summary

- stable version of reconstructed haplotype: high relationship beneficial, but number of genotypes available remains the crucial point
- freely available programs seems to be able to handle large scale genomic data in cattle
- do not overvalue phasing results for long haplotypes

Thank you for your attention!