Error rate for imputation from lower density marker panels to BovineHD in a multi-breed dataset

Chris Schrooten[1)], Rianne van Binsbergen[2,3)], Phil Beatson[3)], Henk Bovenhuis[2)]

[1)]CRV BV, the Netherlands
[2)] Wageningen University, Animal Breeding and Genomics Centre, the Netherlands
[3)]CRV AmBreed, New Zealand

Contents

- Imputation
  - Definition
  - Principles
  - Tools
- Study on imputation in multi-breed dataset
  - Breeds
    - Jersey, Friesian, Crossbred (New Zealand)
  - Chips
    - BovineLD (6.9k) ➔ BovineHD (777k)
    - Various 50k-chips ➔ BovineHD

Context

- Genomic Selection, from 2007 onwards
  - Animals genotyped with 50k-chips
    - More accurate selection
    - Selection earlier in life
    - Higher selection intensity
      - ➔ Higher genetic progress
- From 2010 onwards
  - Other chips available
    - Lower density (3k, 6.9k), cheaper
    - Higher density (777k), higher reliability (?)
    - Full sequence
    - ➔ Need to convert information between chips
    - ➔ Imputation

Imputation

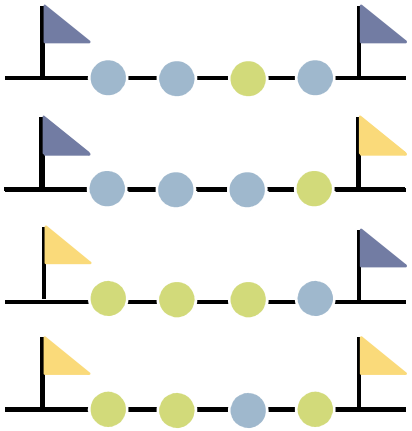
- Definition
  - Derive marker genotype at non-genotyped loci
- Information sources
  - Reference animals
    - Animals with genotype at highest density (e.g. key ancestors)
    - Marker map
    - Pedigree structure
  - Target animals
    - Animals with genotype at lower density
- Result
  - All animals have genotype at highest density

Imputation

- Steps involved
  - Genotype reference set at highest density
  - Determine phased haplotypes in reference set based on
    - Population information (linkage disequilibrium)
    - Pedigree information (linkage)
    - Combination of population and pedigree info
  - Determine phased haplotypes in complete dataset
  - Derive the genotypes at ungenotyped loci

Imputation

4 possible combinations based on information of HD-genotyped animals



Low density (LD) markers (,)
High density (HD) markers (,)

Low density genotype of animal X

Imputed HD-genotype of animal X

Factors affecting imputation accuracy

- Reference set
  - # of animals
  - Relationship with animals to impute
    - single- or multi-breed
- Effective population size
- Imputation tool
  - Parameters
- Chips
  - # SNP
    - On low density and higher density chip
  - Distribution of SNP
    - On low density and higher density chip
- Quality of marker map

Imputation accuracy

- Allelic imputation error rate
- Genotype imputation error rate
- Correlation between imputed and true genotype
  - Mean square error of prediction
  - Regression of imputed on true genotype
- % of alleles not imputed

Imputation accuracy

- Imputation is not 100% accurate
- Allelic imputation errors (range):
  - 3k ➔ 50k: 2-3%
    - Lower when both parents genotyped on 50k
  - BovineLD ➔ 50k: < 1%
  - 50k ➔ BovineHD: 0.5-0.8%
- Impact of imputation errors
  - Lower reliability of genomic EBV
    - 3k ➔ 50k: substantially lower
      - Obtained 85% of reliability gain with imputed info
    - BovineLD ➔ 50k: limited impact

Imputation methods

- Beagle
- Impute
- FastPhase
- HaploRec
- ChromIBD
- Fimpute
- Findhap
- DAGPHASE / PHASEBOOK
- AlphaImpute

Which imputation method?

- Important factors for choosing a certain imputation method
  - Imputation errors / accuracy
  - % of alleles not imputed
  - Speed of imputation
  - Ease of use
  - Possibility to combine with other method(s)
  - Possibility to integrate into routine pipeline

Study on imputation

- Objective
  - Investigate the error rate for imputation from lower density marker panels to the BovineHD marker panel in a multi-breed dataset

Study on imputation

- Context
  - Genomic Selection in New Zealand, CRV Ambreed
    - Breeding programs for Friesian and Jersey
    - GS started in 2007
    - Relatively small reference populations for GS
      - 2200 Friesians, 1200 Jerseys
      - Limited increase in reliability due to genomics
      - Higher reliability with combined ref. populations?
        - Need high density SNP-chip
        - Impute to BovineHD
          - 50k: reference animals, selection candidates
          - BovineLD: genotypes of cows or candidates

Material & Methods – animals and subsets

- BovineHD genotypes available for
  - 463 Friesians
  - 229 Jersey animals
  - 57 crossbred animals
- Five alternative animal subsets, varying
  - Breeds in reference set
  - # animals in reference set
- Each animal subset
  - Imputation studied for 5 lower density chips
- Four replicates per subset
  - Different validation and reference animals
    - validation animals in replicate 1 of subset x the same as validation animals in replicate 1 of subset y

Material & Methods - alternatives

Alternative	Reference		Validation
	Breed	# animals	# animals
Complete	Friesian	438	25
	Jersey	204	25
	Crossbred	47	10
Friesian+Jersey	Friesian	438	25
	Jersey	204	25
	Crossbred	0	10
Friesian	Friesian	438	25
	Jersey	0	25
	Crossbred	0	10

Material & Methods - chips

Chip	# SNP (also) on BovineHD*)
BovineHD	625
BovineSNP50 v2	41.2
BovineSNP50 v1	40.9
CRV v2	33.5
CRV v1	27.5
BovineLD	6.6

*) BTA 1-29, after applying all edits

Material & Methods - chips

Chip	# SNP (also) on BovineHD*)
BovineHD	625
BovineSNP50 v2	41.2
CRV v2	33.5
BovineLD	6.6

*) BTA 1-29, after applying all edits

Material & Methods - validation animals

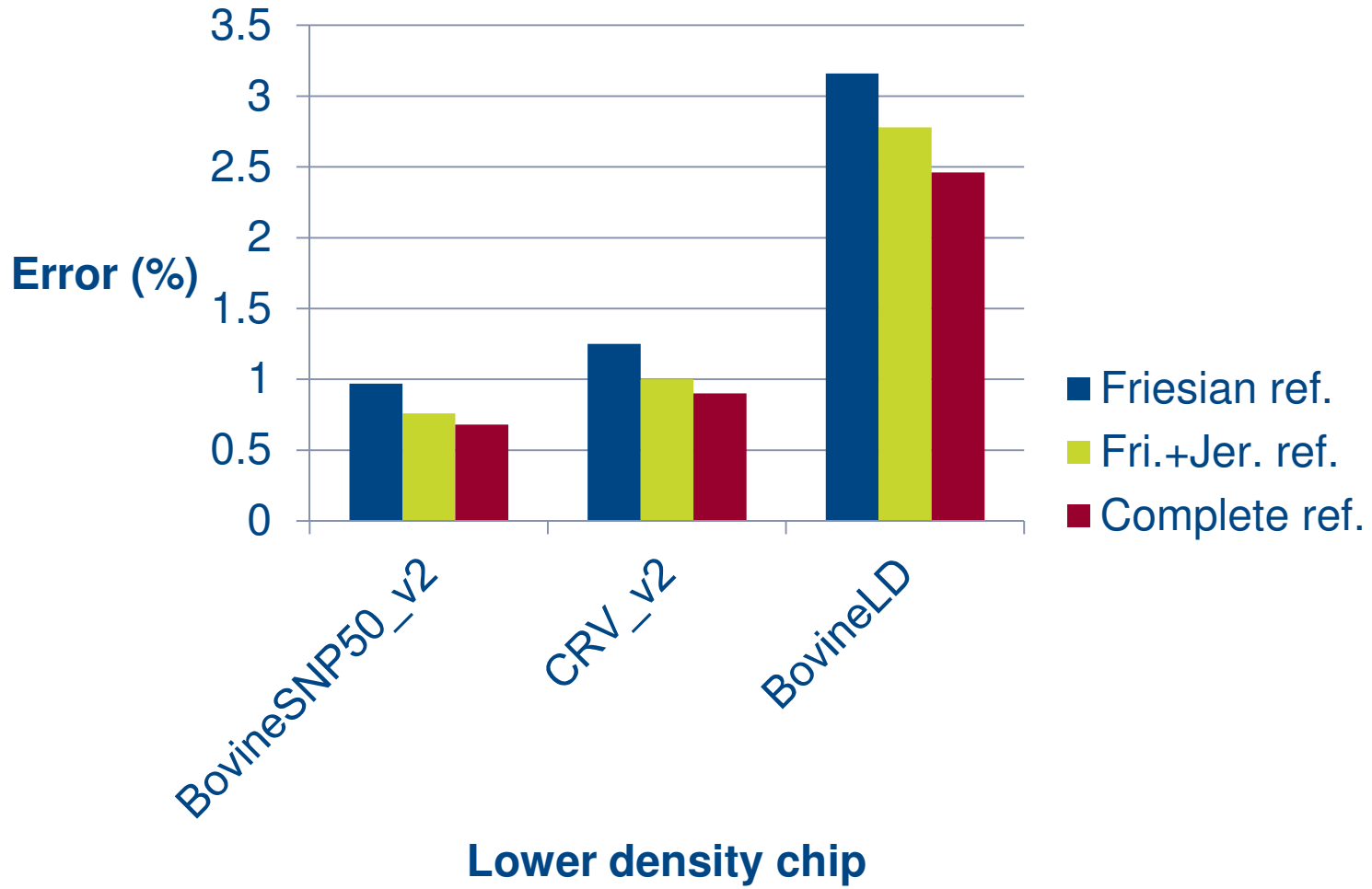
- Randomly chosen, but
  - No descendants genotyped on HD
  - Validation animal only in one replicate
- To mimic lower-density genotypes
  - Mask all the genotypes of SNP that are only on HD
  - Impute HD genotype based on
    - HD-genotype of reference animals
    - Low-density genotype of validation animals
- Categorize validation animals
  - Traceability = expected proportion of genome inherited from HD-genotyped ancestors
    - Only sire on HD: 0.5
    - Sire + Maternal Grandsire: 0.50+0.25=0.75

Material & Methods

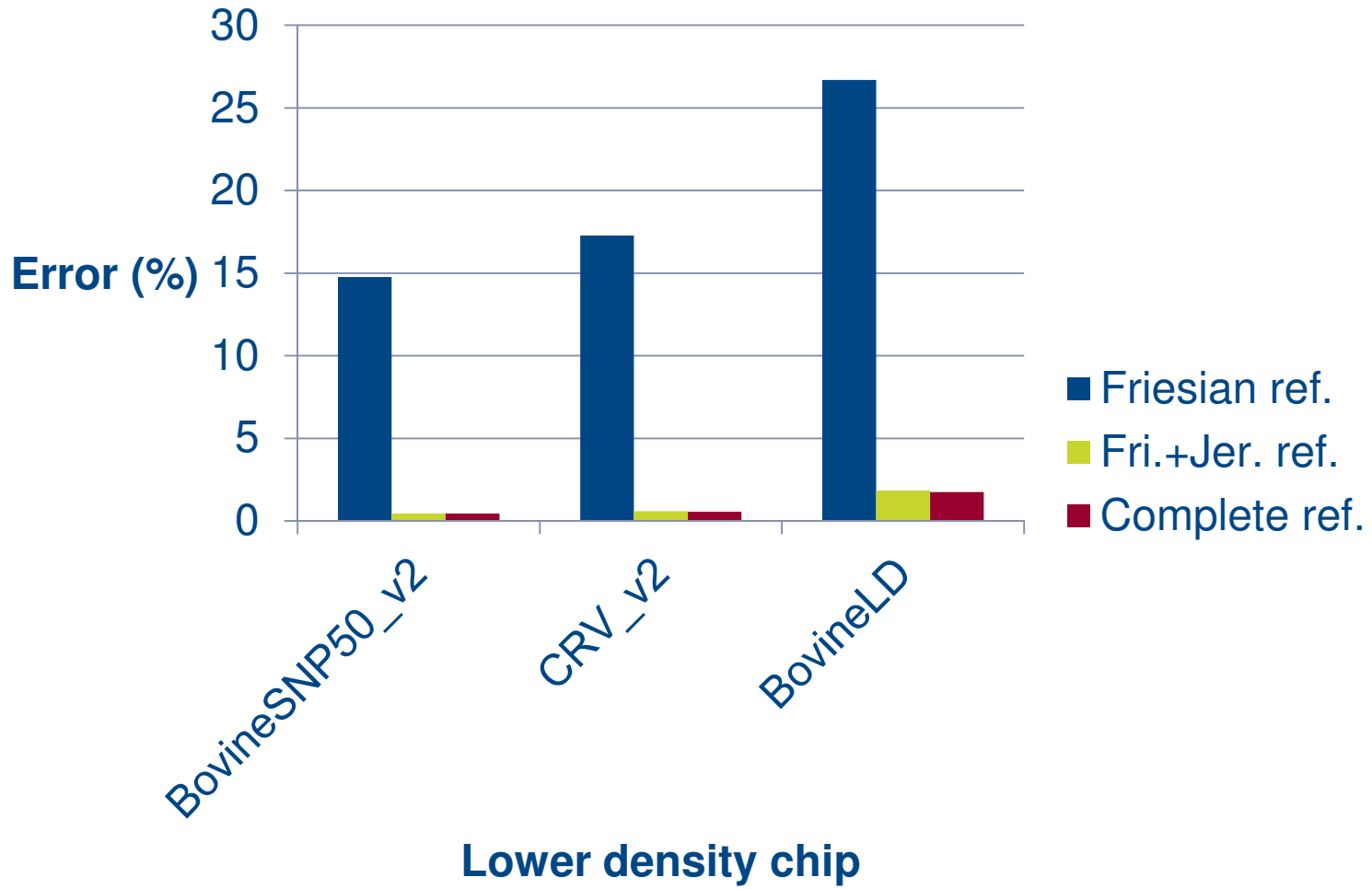
- Subset of chromosomes used
  - To limit computer time and space
  - BTA 1, 6, 11, 14, 20, 29
- Imputation with Beagle 3.3.0
- Evaluate

$$\text{Allelic imputation error (\%)} = \frac{n_{\text{imputed}\neq\text{observed}}}{n_{\text{imputed.and.observed}}}$$

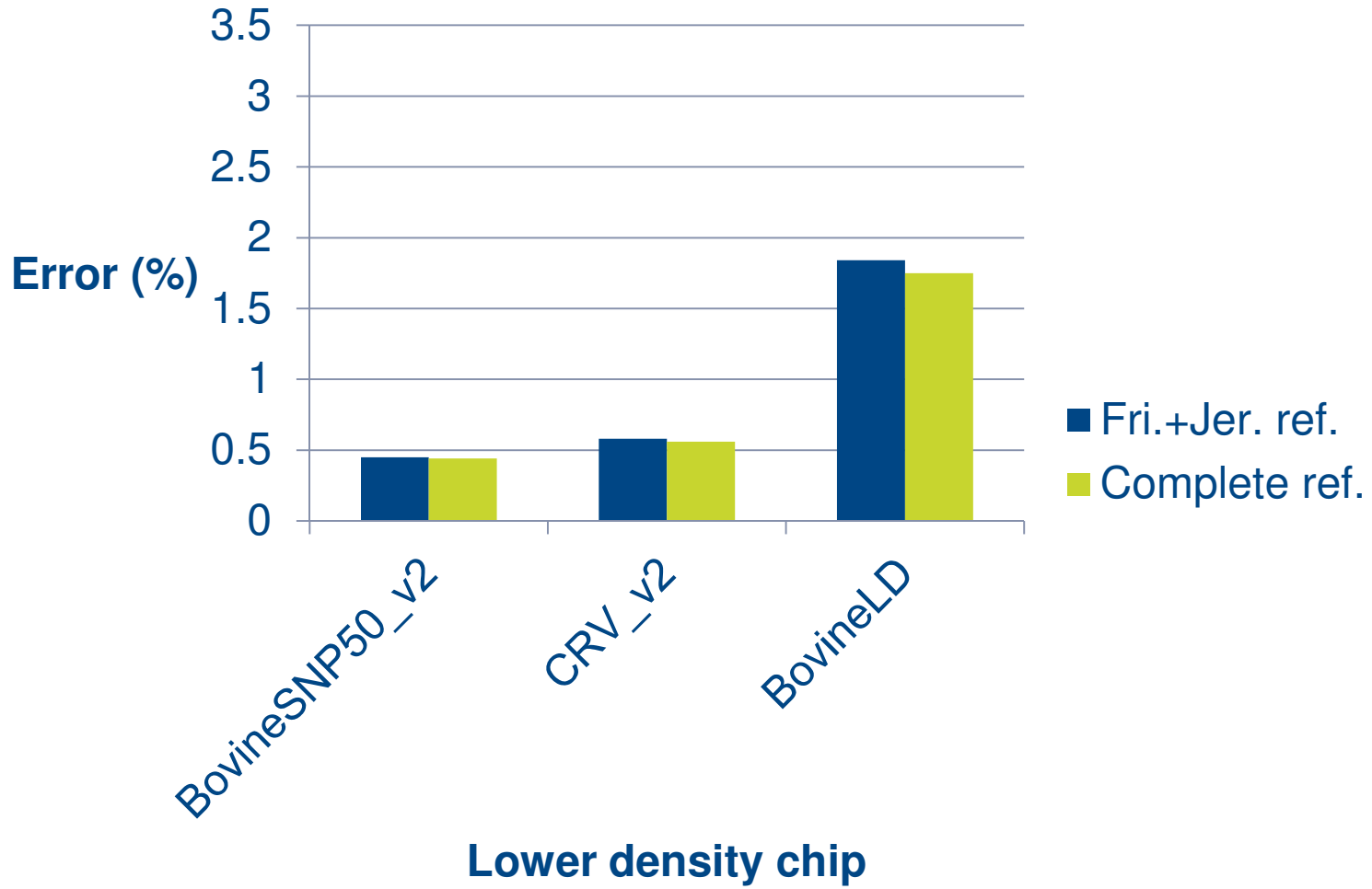
Imputation error (%) for Friesians



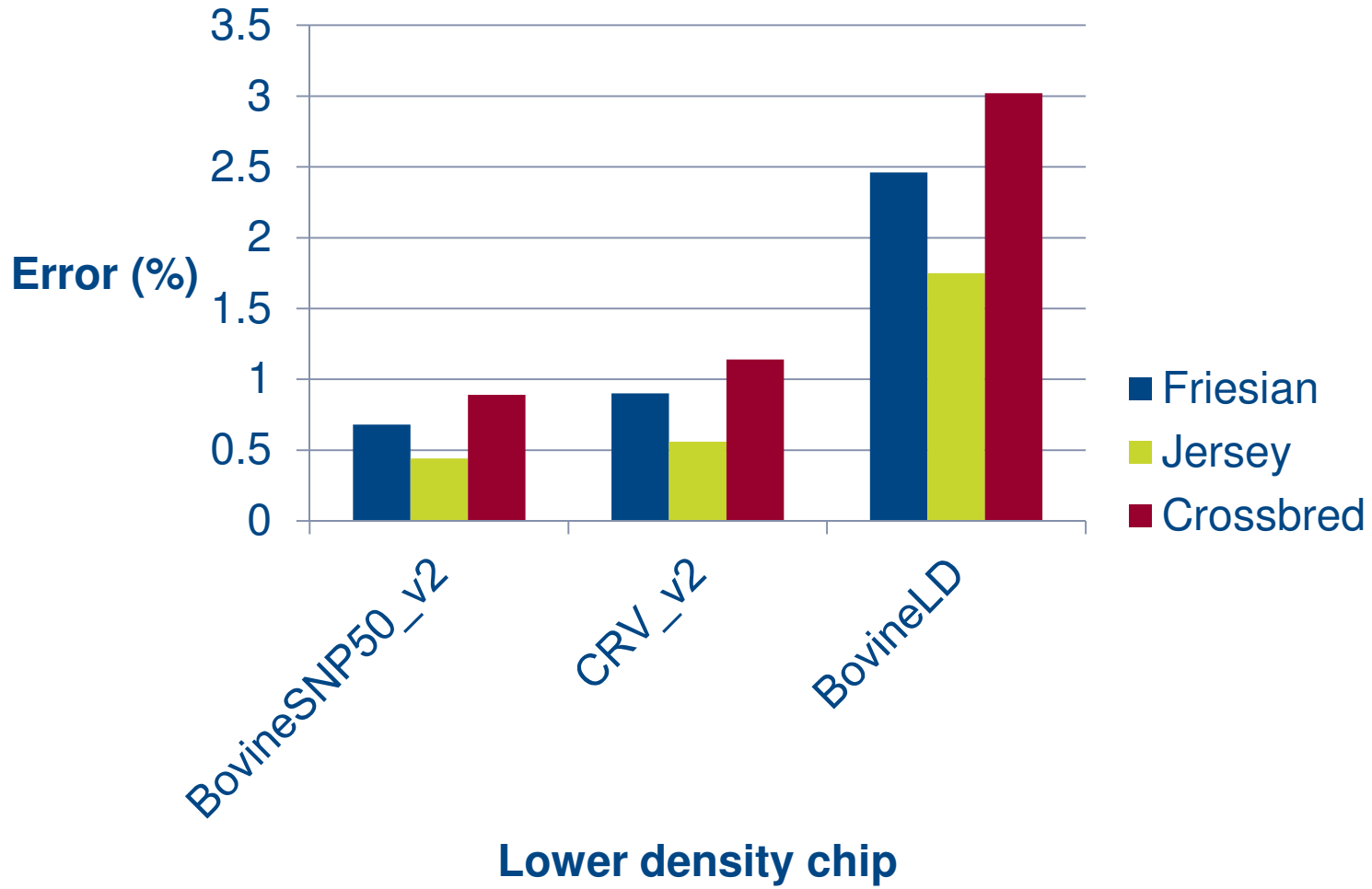
Imputation error (%) for Jerseys



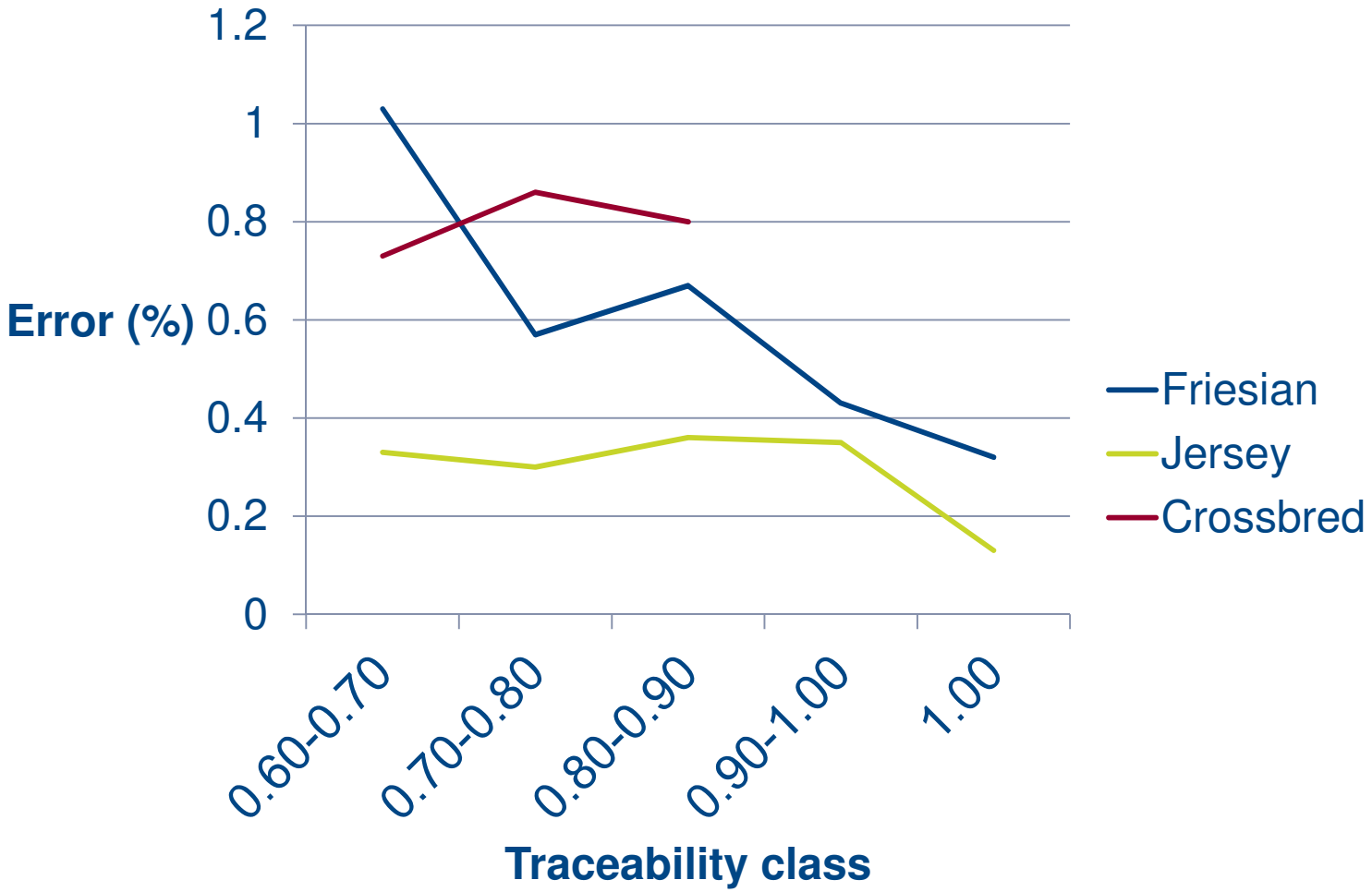
Imputation error (%) for Jerseys



Imputation error (%) using complete reference



Imputation error per traceability class, complete reference, BovineSNP50_v2



Results - summary

- Lower imputation error if
  - More animals in the reference set
    - Even if additional animals are from different breed
    - Breed to impute needs to be present in the reference set
  - Lower density chip has more markers in common with BovineHD
  - More ancestors genotyped on BovineHD
- Lower imputation error in
  - Jersey vs. Friesian
    - Effective population size is smaller in Jersey
    - Relatively many population haplotypes present in the BovineHD-reference set

Implications

- Imputation from 50k to BovineHD is possible at low error rates
  - Jersey 0.4%; Friesian 0.7%
  - Little impact of imputation error expected on reliability of genomic EBV
- Imputation from BovineLD to BovineHD shows higher error rates
  - Jersey 1.7%; Friesian 2.5%
  - Effect on reliability of genomic EBV too large?
- (Imputed) BovineHD-data useful for across-breed genomic evaluation?
  - Currently being evaluated

Thank you for your attention!