

Comparison between SNP array and imputed data to estimate population structure in horse breeds

Chessari G., Reich P., Criscione A., Falker-Gieske C., Mastrangelo S.,
Tumino S., Bordonaro S., Marletta D. and Tetens J.



Overview ● ● ● ●

Origin of *Equus caballus*

The domestication began ~5000 years ago in the *Eurasian steppe*, followed by multiple events that led to the definition of different genetic roots. Horse species has played a pivotal role in human evolution, involving both *production* aspects and *recreational* and *sporting* aspects.

Importance of genetic characterization

- Assessing breeding values
- Establishing management plans
- Safeguarding native breeds
- Improving of existing breeds

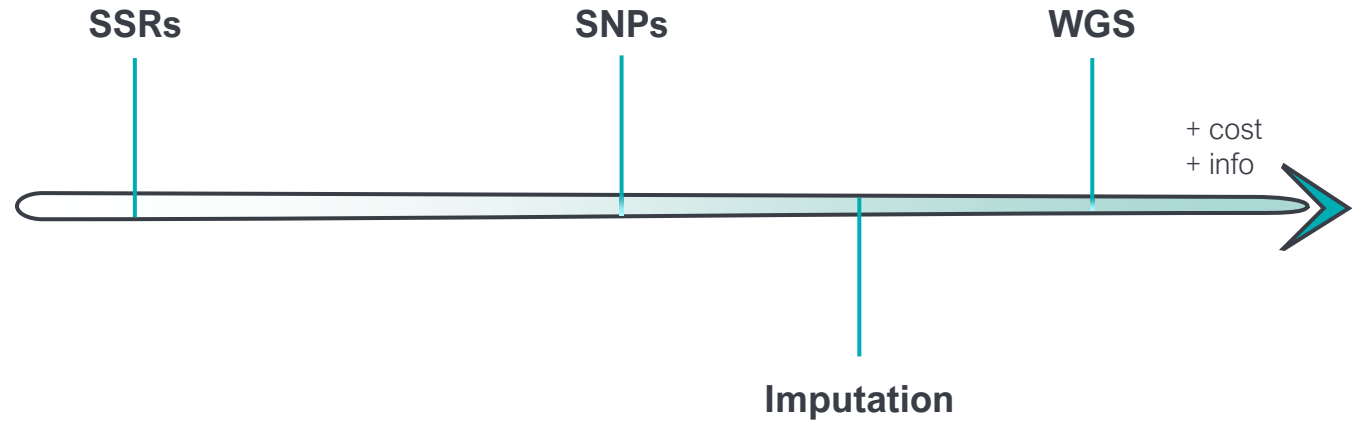
Overview ● ● ● ●

Origin of *Equus caballus*

The domestication began ~5000 years ago in the *Eurasian steppe*, followed by multiple events that led to the definition of different genetic roots. Horse species has played a pivotal role in human evolution, involving both *production* aspects and *recreational* and *sporting* aspects.

Importance of genetic characterization

- Assessing breeding values
- Establishing management plans
- Safeguarding native breeds
- Improving of existing breeds



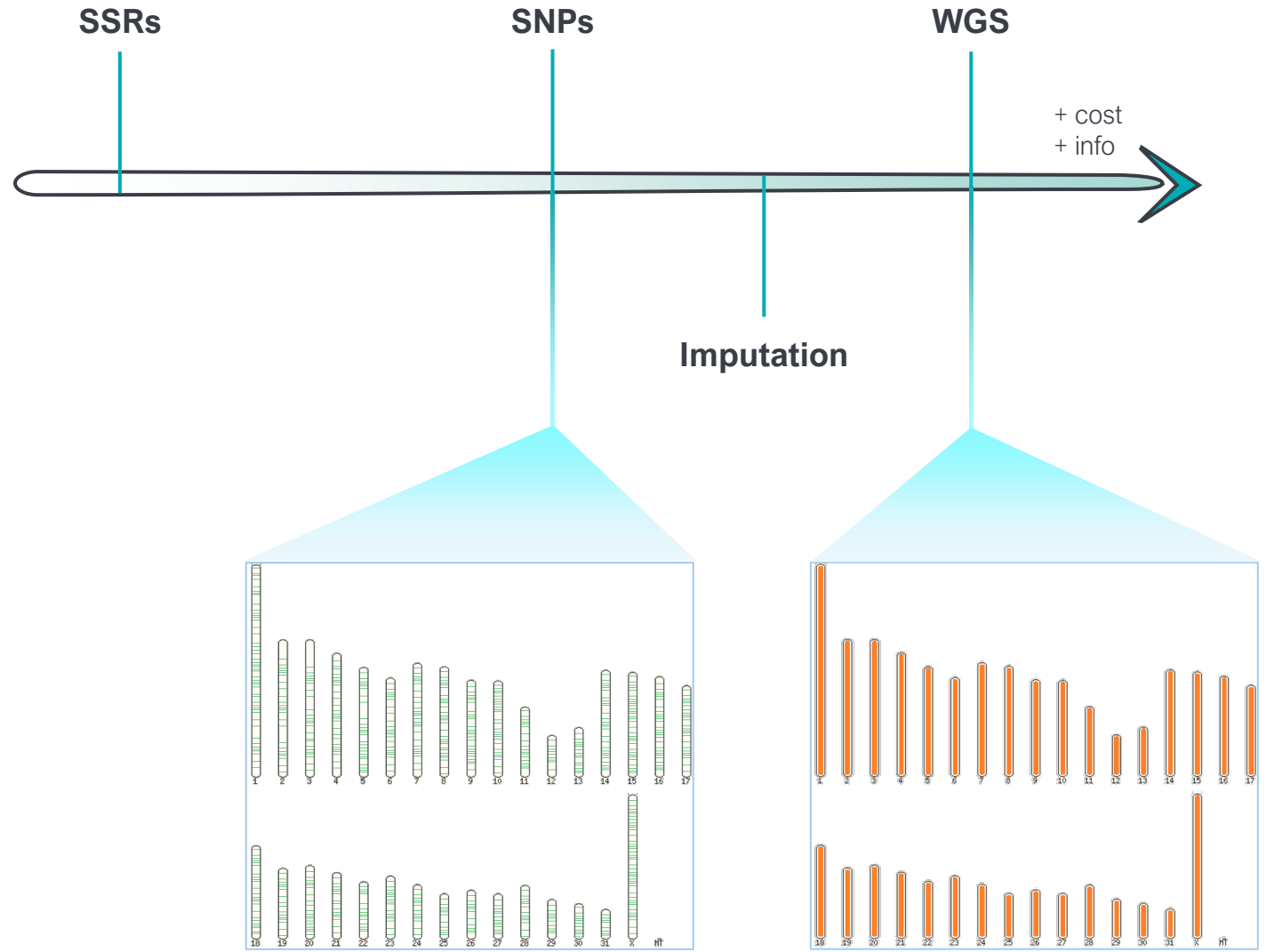
Overview ● ● ● ●

Origin of *Equus caballus*

The domestication began ~5000 years ago in the *Eurasian steppe*, followed by multiple events that led to the definition of different genetic roots. Horse species has played a pivotal role in human evolution, involving both *production* aspects and *recreational* and *sporting* aspects.

Importance of genetic characterization

- Assessing breeding values
- Establishing management plans
- Safeguarding native breeds
- Improving of existing breeds



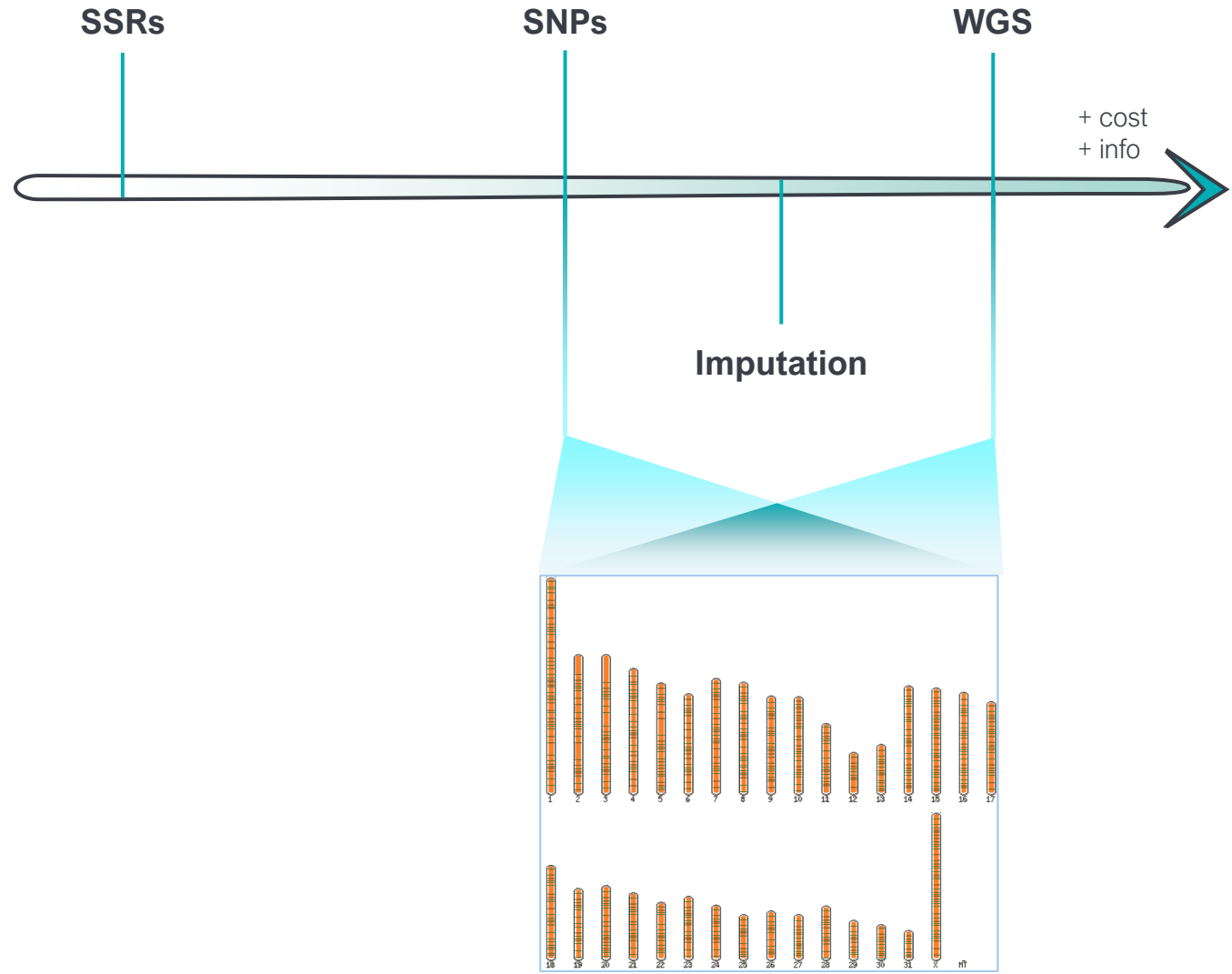
Overview ● ● ● ●

Origin of *Equus caballus*

The domestication began ~5000 years ago in the *Eurasian steppe*, followed by multiple events that led to the definition of different genetic roots. Horse species has played a pivotal role in human evolution, involving both *production* aspects and *recreational* and *sporting* aspects.

Importance of genetic characterization

- Assessing breeding values
- Establishing management plans
- Safeguarding native breeds
- Improving of existing breeds



WGS Reference panel



SEQUENCES

Illumina paired-end x20



③ Purosangue
Orientale
Siciliano



③ Siciliano



④ Sanfratellano

MAPPING and VARIANT CALLING

EquCab3.0 genome assembly

Best practices recommendations by GATK v4.1.7.0

Read group

Mark Illumina adapters

Align and Merge reads

Mark duplicates

BQSR

Haplotype calling

Combine gVCFs

Genotype gVCFs



Merging to a reference
panel including
317 horses belonging to
46 worldwide breeds
(Reich *et al.*, 2022)

Raw SNP panel



SNP genotyping

Medium density arrays



Sicilian horses panel

(Criscione *et al.*, 2022)

- 65'157 SNPs
- 36 individuals
- 3 Sicilian breeds

Worldwide breeds panel

(Petersen *et al.*, 2013)

- 40'816 SNPs
- 814 individuals
- 38 breeds

	Breed	Code	Size
Endurance / Riding horse	Purosangue Orientale Sic	ORI	9
	Sanfratellano	SAN	13
	Siciliano	SIC	14
Dressage / Show Jumping	German Hanoverian	GER	15
Endurance	Akhal-Teke	AKT	19
	Arabian	ARA	24
Draft	Franches-Montagnes	FRA	19
Light-Draft	Icelandic	ICE	25
	Shetland	SHE	27
Riding horse	Quarter Horse	QUA	40
	Standardbred	STA	40
	Thoroughbred	THO	36

= 281 individuals and 40'861 SNPs



Imputation



SNP PANEL

Medium density arrays



IMPUTATION PROCESS

EquCab3.0 genome assembly

Based on the horse ReferencePanel

Software: Beagle v5.1

- 40'861 SNPs
- 281 individuals
- 12 breeds

Final dataset

Update to Reference assembly

Imputation step per chr

Combine VCFs

Calculation of DR2



Imputation



SNP PANEL

Medium density arrays

- 40'861 SNPs
- 281 individuals
- 12 breeds



IMPUTATION PROCESS

EquCab3.0 genome assembly

Based on the horse ReferencePanel

Software: Beagle v5.1

Final dataset

Update to Reference assembly

Imputation step per chr

Combine VCFs

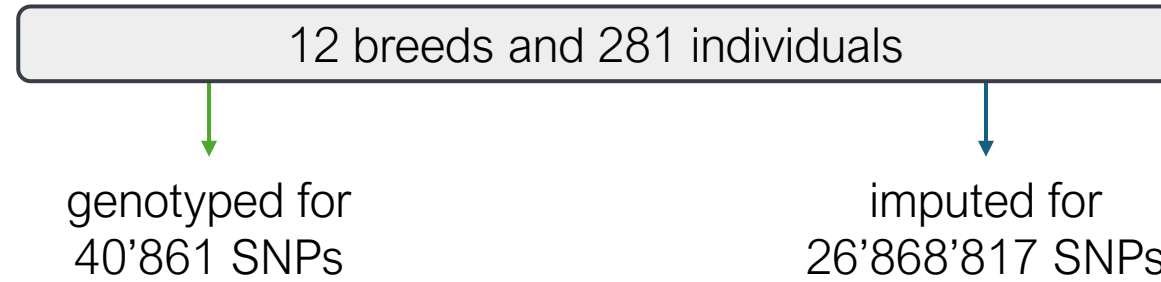
Calculation of DR2



26'868'817 SNPs



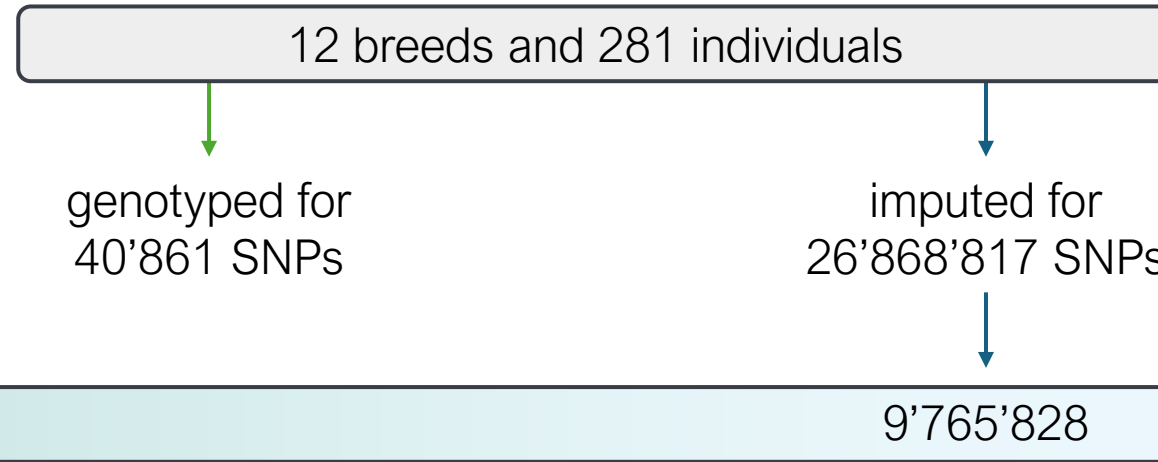
Filtering, Quality control and Pruning



Are imputed data always more informative?



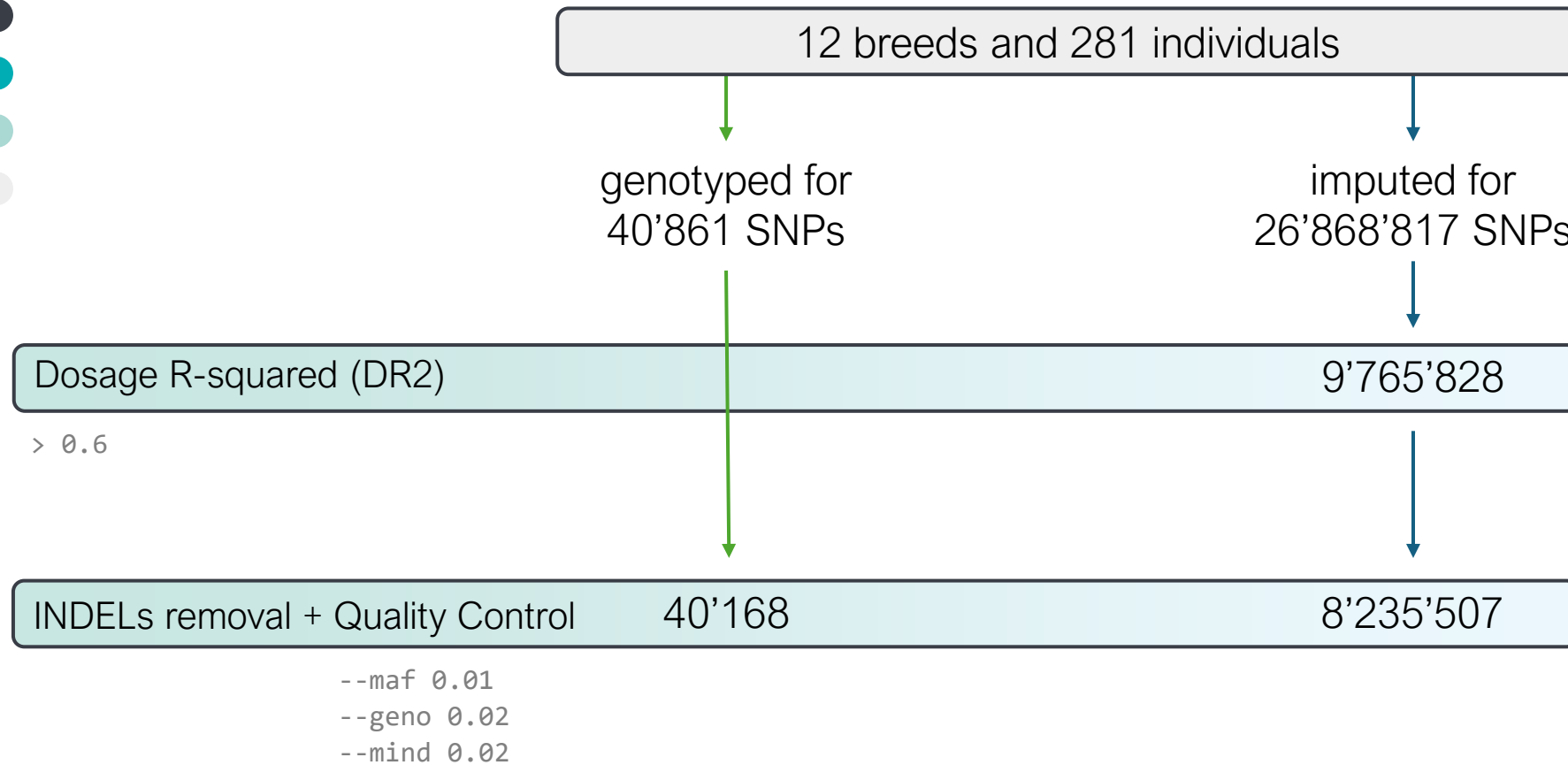
Filtering, Quality control and Pruning



> 0.6



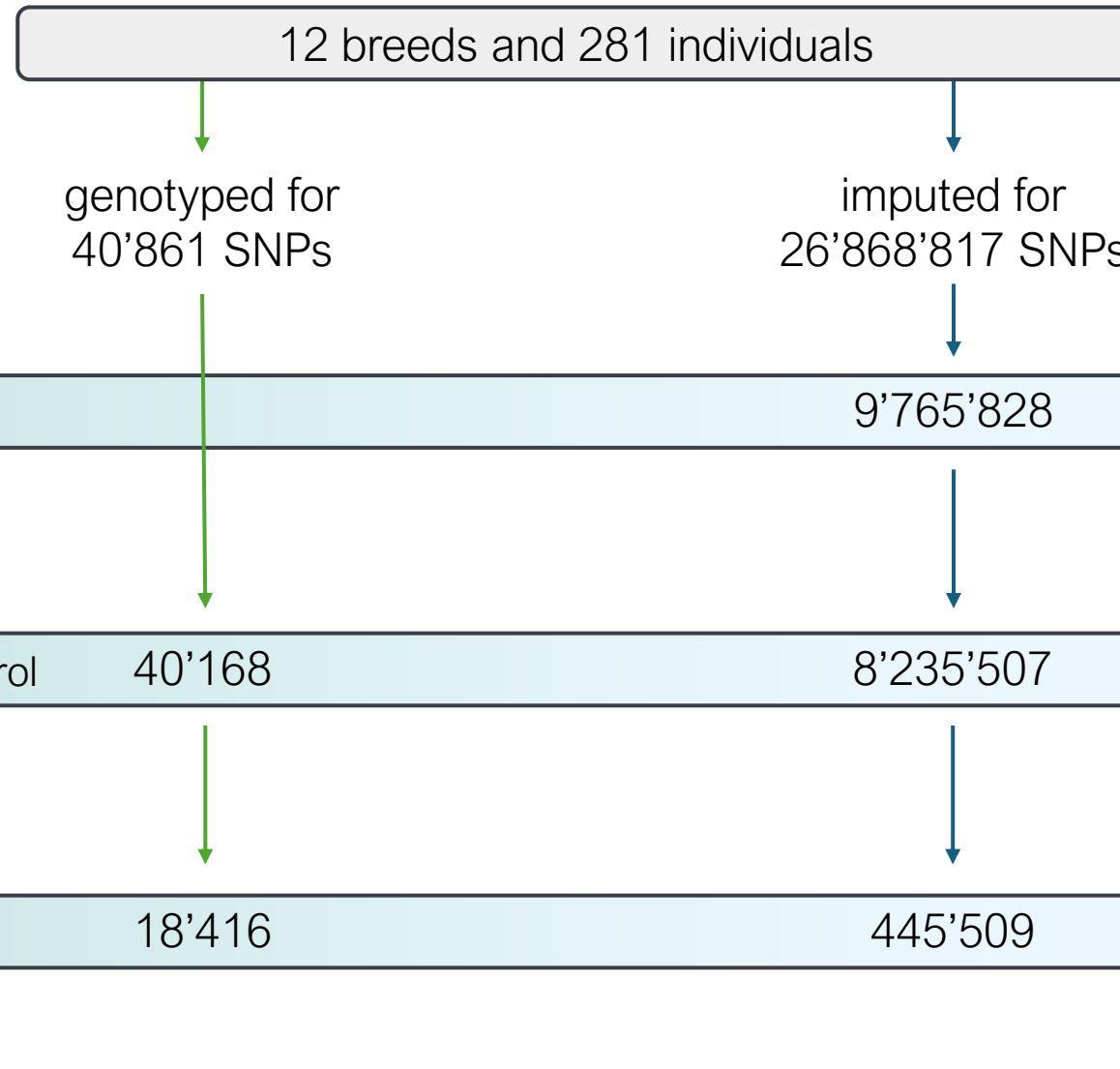
Filtering, Quality control and Pruning



- Genetic diversity indices
- Runs of homozygosity



Filtering, Quality control and Pruning



- Genetic diversity indices
- Runs of homozygosity
- Population structure



Results and Discussion

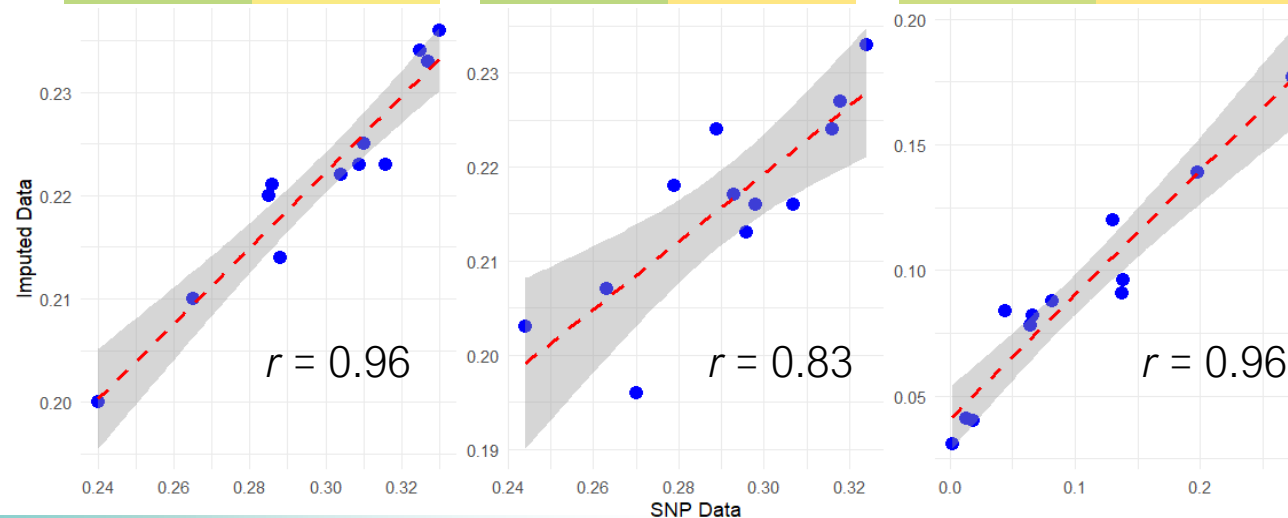
H_O = Observed heterozygosity

H_E = Expected heterozygosity

F_{IS} = Inbreeding coefficient

	H_O		H_E		F_{IS}	
	SNP	IMP	SNP	IMP	SNP	IMP
Purosangue Orientale Sic	0.309	0.223	0.270	0.196	0.066	0.082
Sanfratellano	0.310	0.225	0.296	0.213	0.064	0.078
Siciliano	0.327	0.233	0.316	0.224	0.013	0.041
German Hanoverian	0.330	0.236	0.318	0.227	0.002	0.031
Akhal-Teke	0.304	0.222	0.298	0.216	0.082	0.088
Arabian	0.288	0.214	0.293	0.217	0.130	0.120
Franches-Montagnes	0.286	0.221	0.279	0.218	0.137	0.091
Icelandic	0.265	0.210	0.263	0.207	0.198	0.139
Shetland	0.240	0.200	0.244	0.203	0.274	0.177
Quarter Horse	0.325	0.234	0.324	0.233	0.019	0.040
Standardbred	0.285	0.220	0.289	0.224	0.138	0.096
Thoroughbred	0.316	0.223	0.307	0.216	0.044	0.084

Pearson correlation

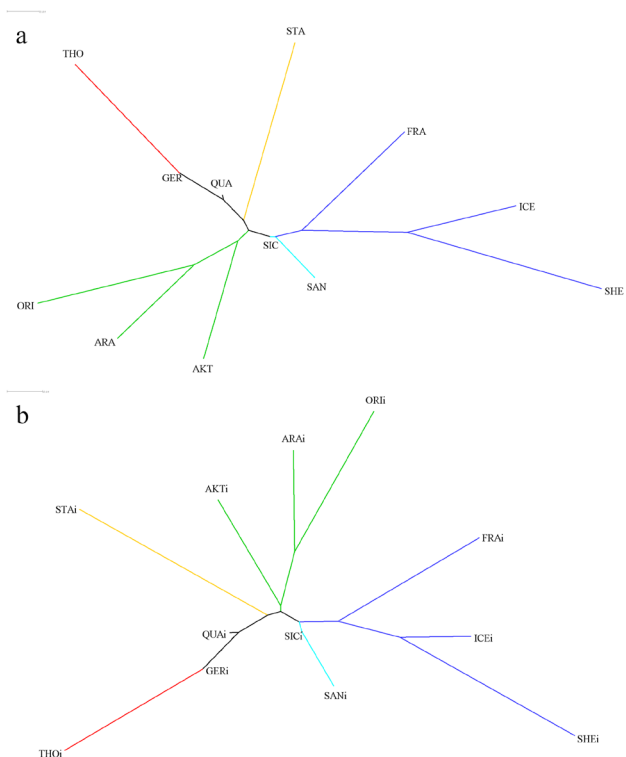


Results and Discussion

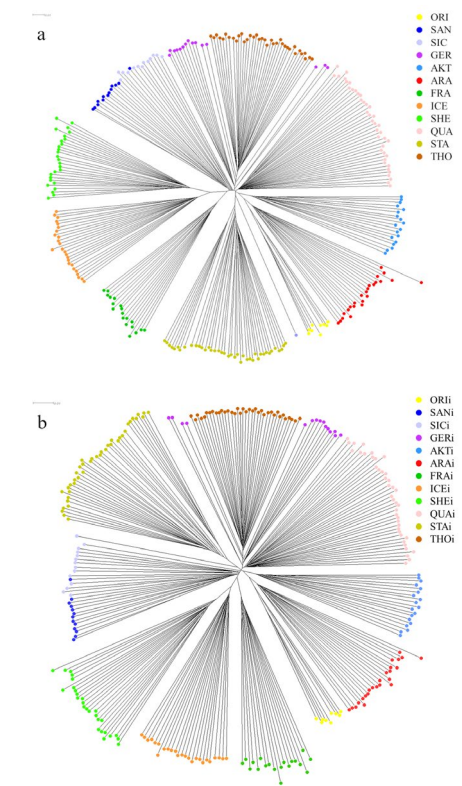
Pairwise F_{ST} distances



Neighbor-Joining tree on Reynolds' genetic distances



Allele Sharing Distances (ASD) pairwise distances



Mantel test

r : 0.962
 p -value: 0.001

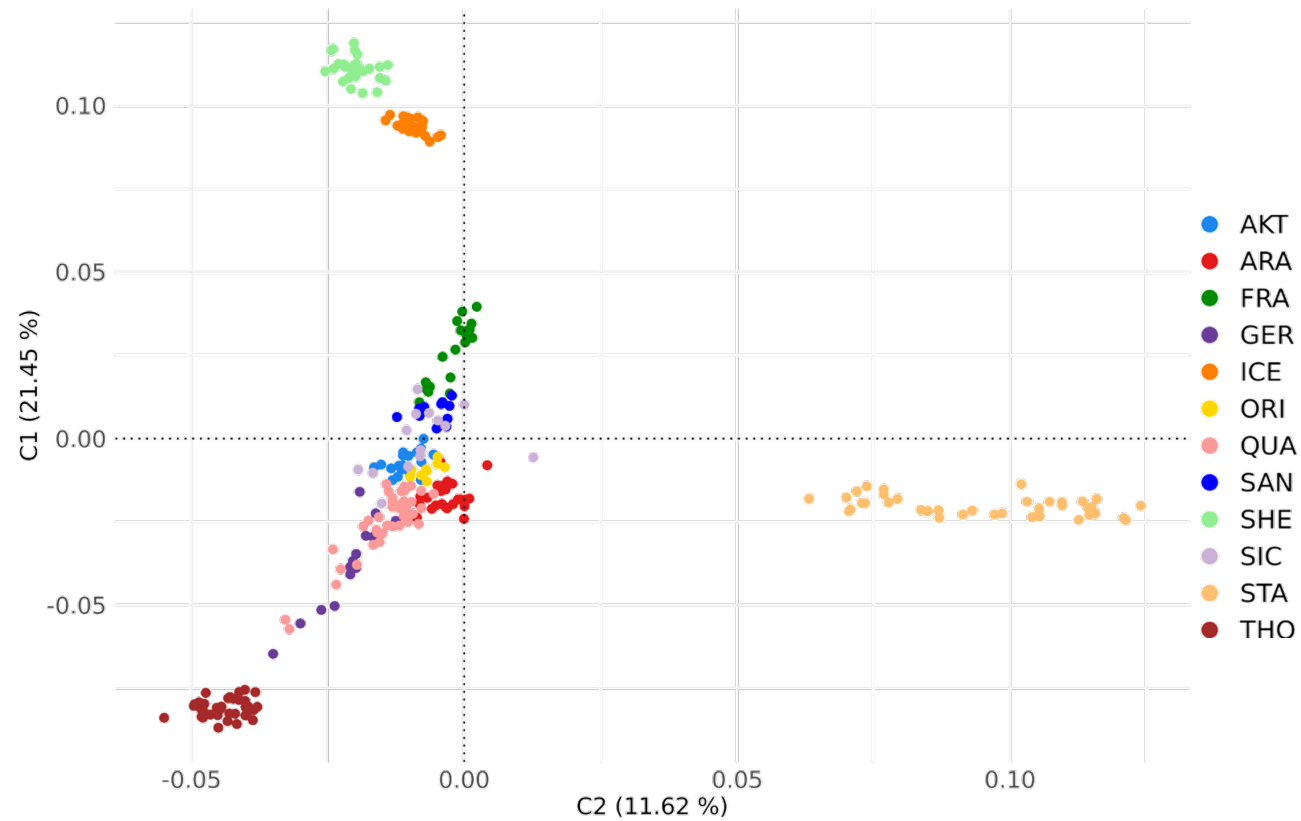
r : 0.961
 p -value: 0.001

r : 0.813
 p -value: 0.001

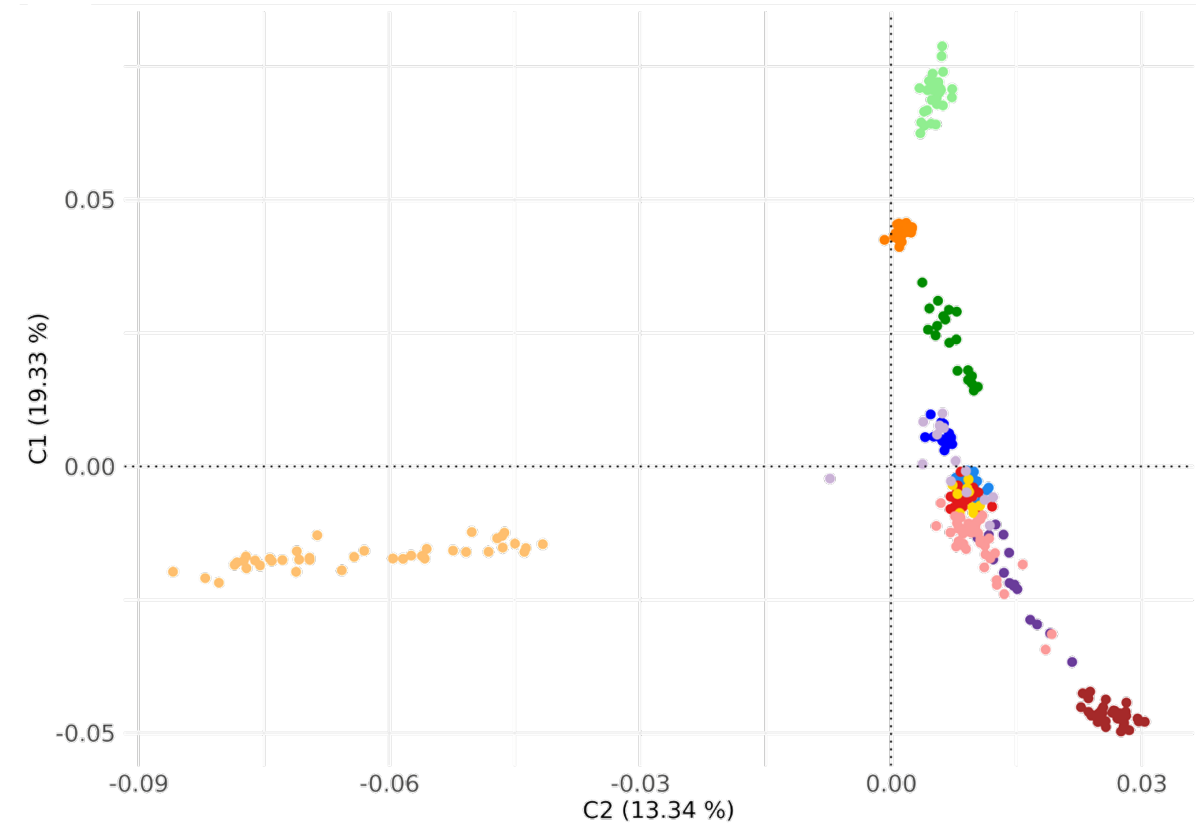


Results and Discussion

SNP

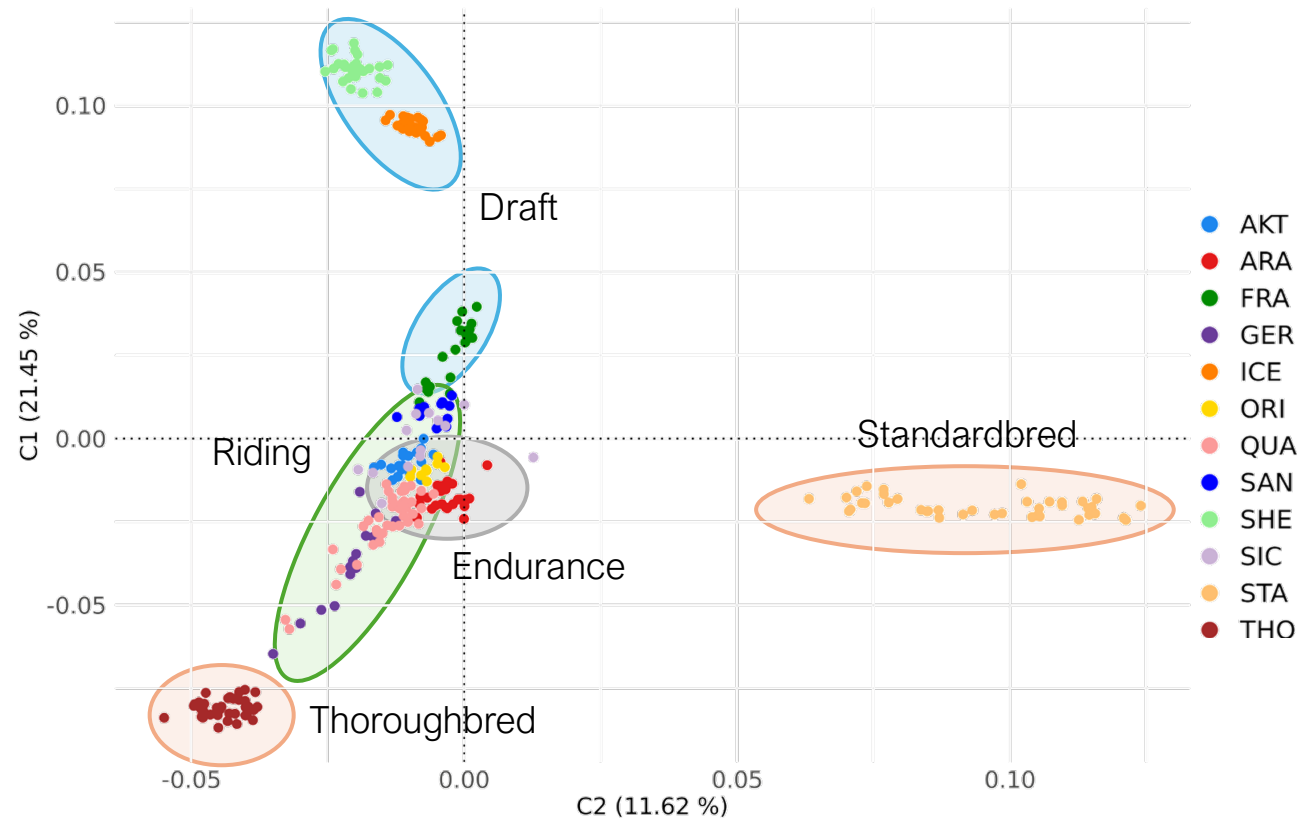


IMP

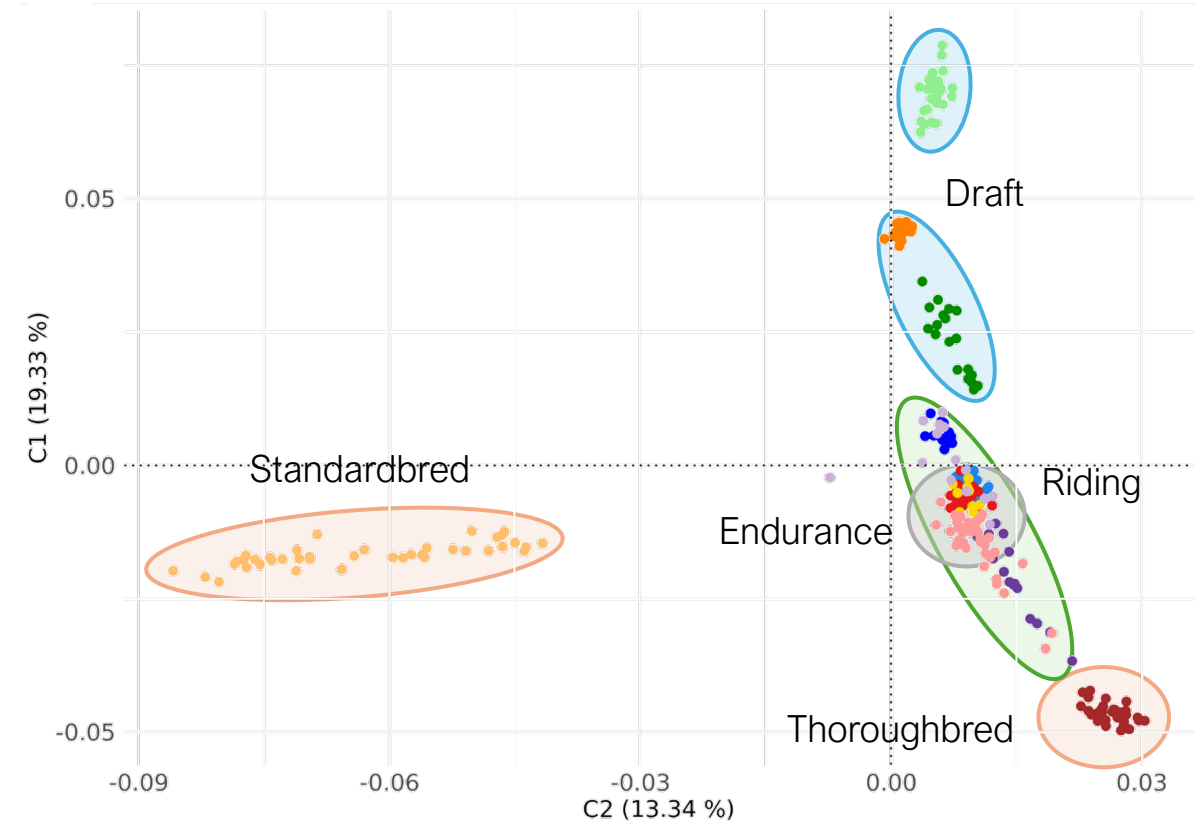


Results and Discussion

SNP

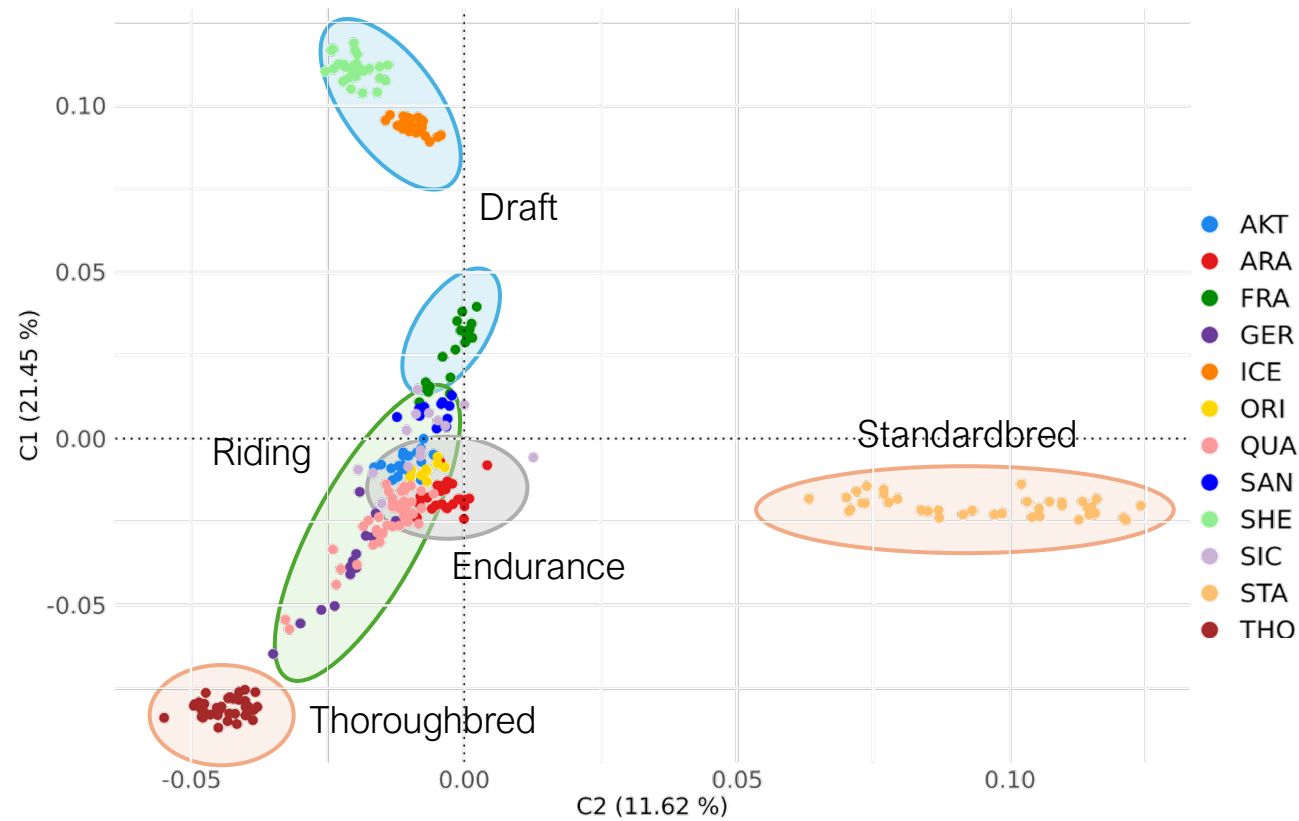


IMP

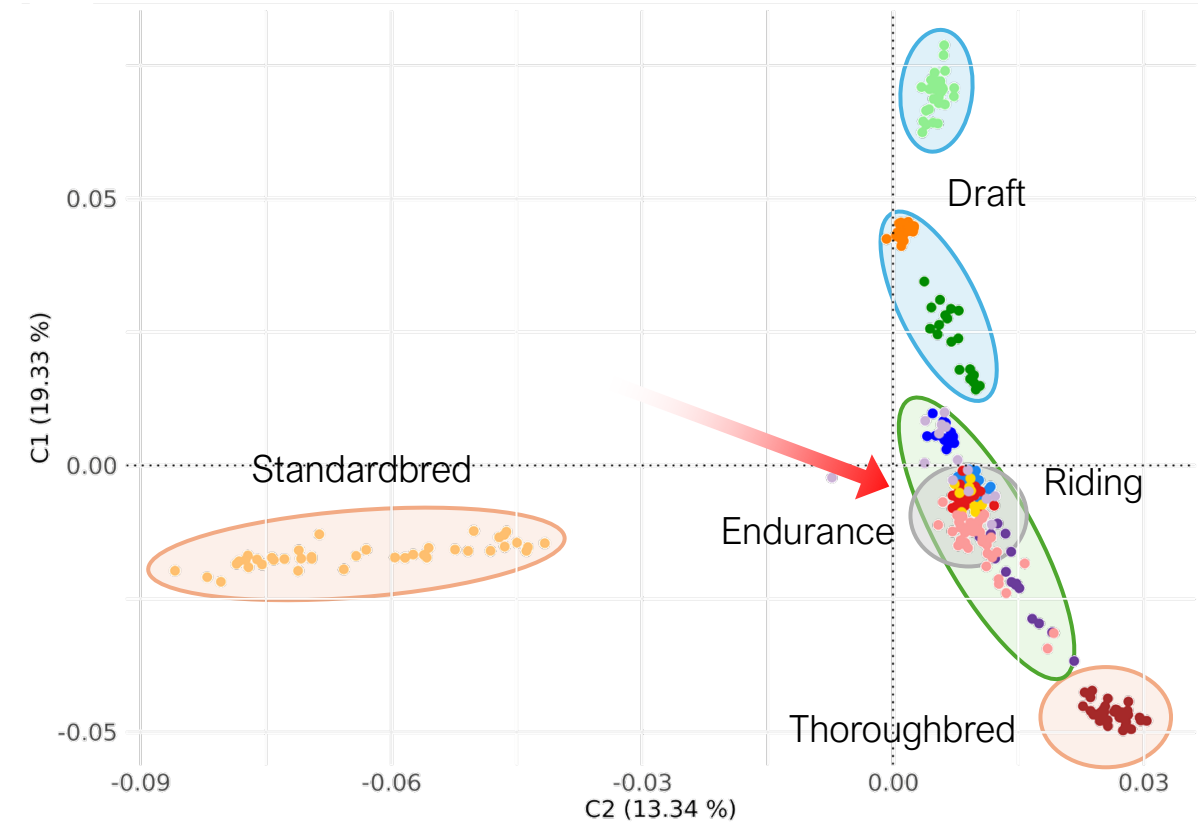


Results and Discussion

SNP



IMP



Runs of homozygosity



SNP dataset

after quality control

Sliding Windows

Threshold	0.05
windowSize	25
minSNP	20
ROHet	FALSE
maxOppRun	0
maxMissRun	0
maxOppWindow	0
maxMissWindow	0
maxGap	1'000'000
minLengthBps	100'000
minDensity	1/100

For a total of 23'026 ROHs

IMP dataset

after quality control

Sliding Windows

Threshold	0.05
windowSize	500
minSNP	50
ROHet	FALSE
maxOppRun	0
maxMissRun	0
maxOppWindow	0
maxMissWindow	0
maxGap	1'000'000
minLengthBps	100'000
minDensity	1/50

For a total of 117'040 ROHs

Runs of homozygosity



SNP dataset
after quality control

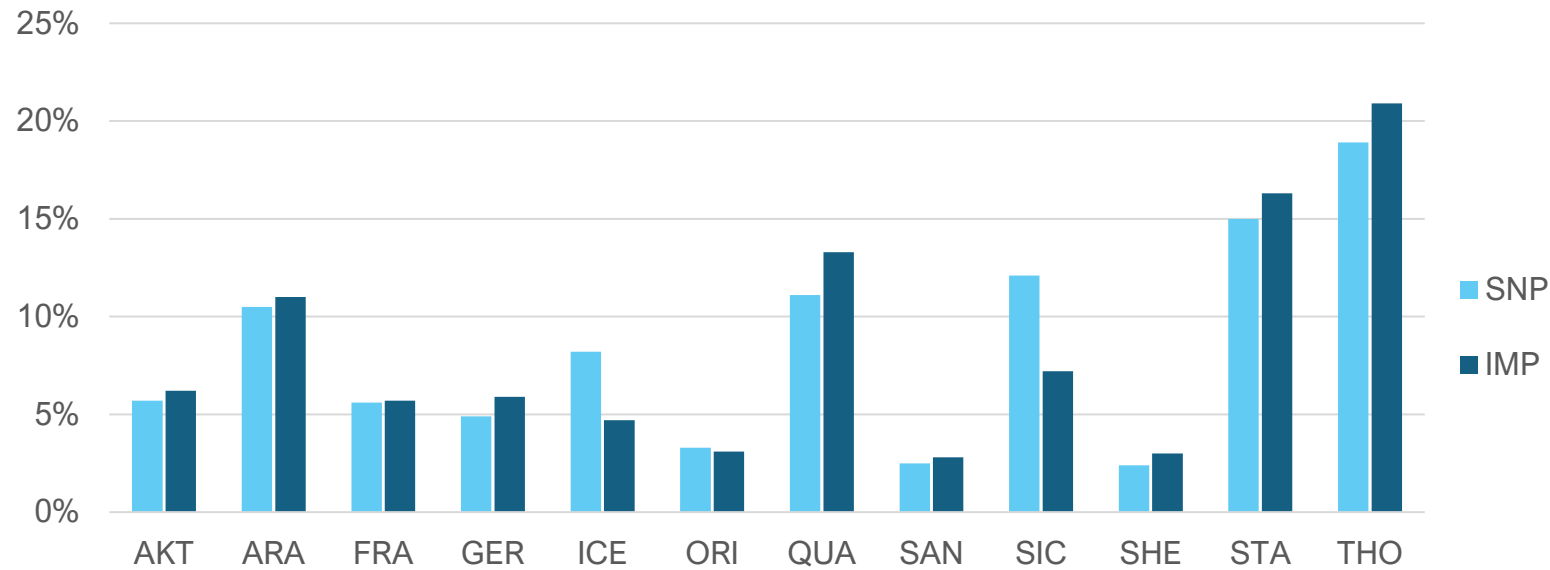
Sliding Windows

For a total of 23'026 ROHs

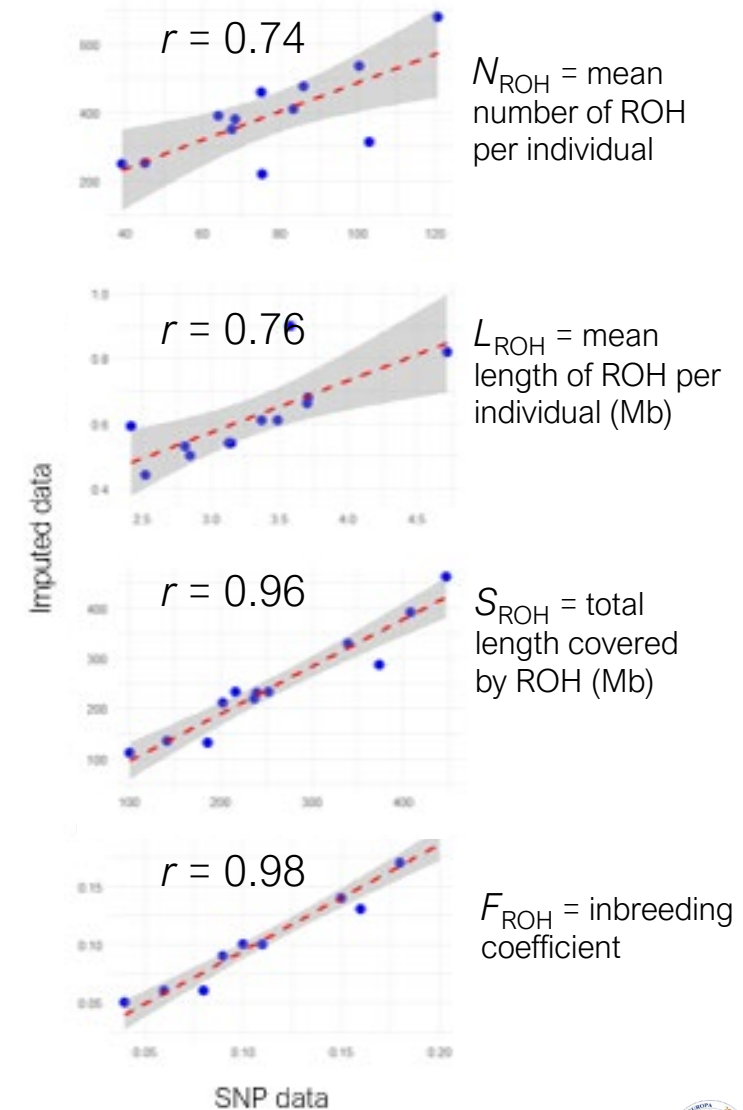
IMP dataset
after quality control

Sliding Windows

For a total of 117'040 ROHs



Pearson
correlation



ROH islands



SNP dataset

after quality control

Sliding Windows

Threshold	0.05
windowSize	25
minSNP	20
ROHet	FALSE
maxOppRun	0
maxMissRun	0
maxOppWindow	0
maxMissWindow	0
maxGap	1'000'000
minLengthBps	100'000
minDensity	1/100

For a total of 23'026 ROHs

IMP dataset

after quality control

Sliding Windows

Threshold	0.05
windowSize	500
minSNP	50
ROHet	FALSE
maxOppRun	0
maxMissRun	0
maxOppWindow	0
maxMissWindow	0
maxGap	1'000'000
minLengthBps	100'000
minDensity	1/50

For a total of 117'040 ROHs

WGS dataset

after quality control

Sliding Windows

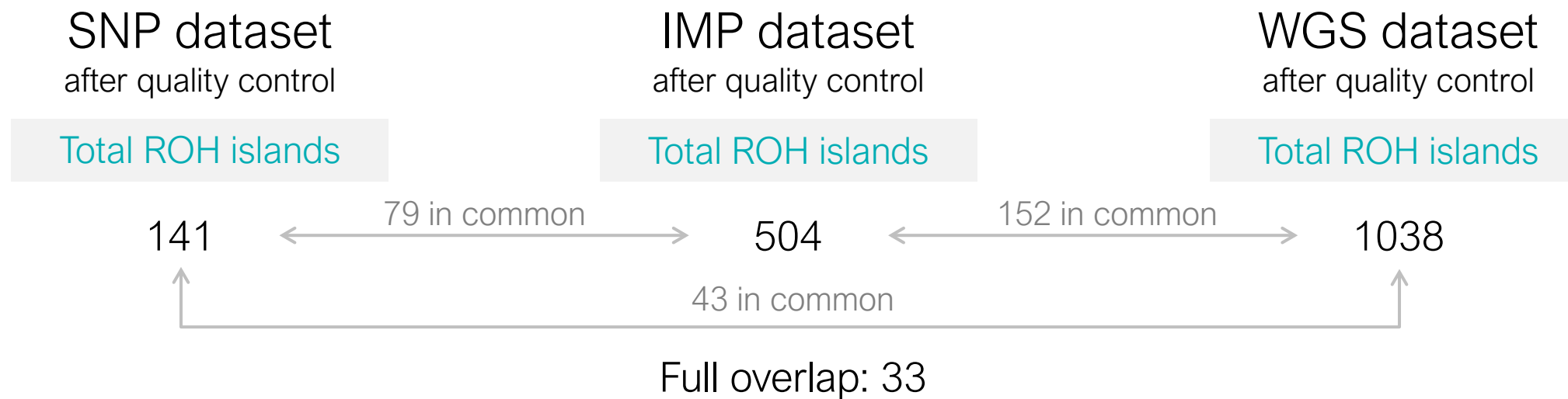
Threshold	0.05
windowSize	500
minSNP	50
ROHet	FALSE
maxOppRun	0
maxMissRun	0
maxOppWindow	0
maxMissWindow	0
maxGap	1'000'000
minLengthBps	100'000
minDensity	1/50

For a total of 137'959 ROHs

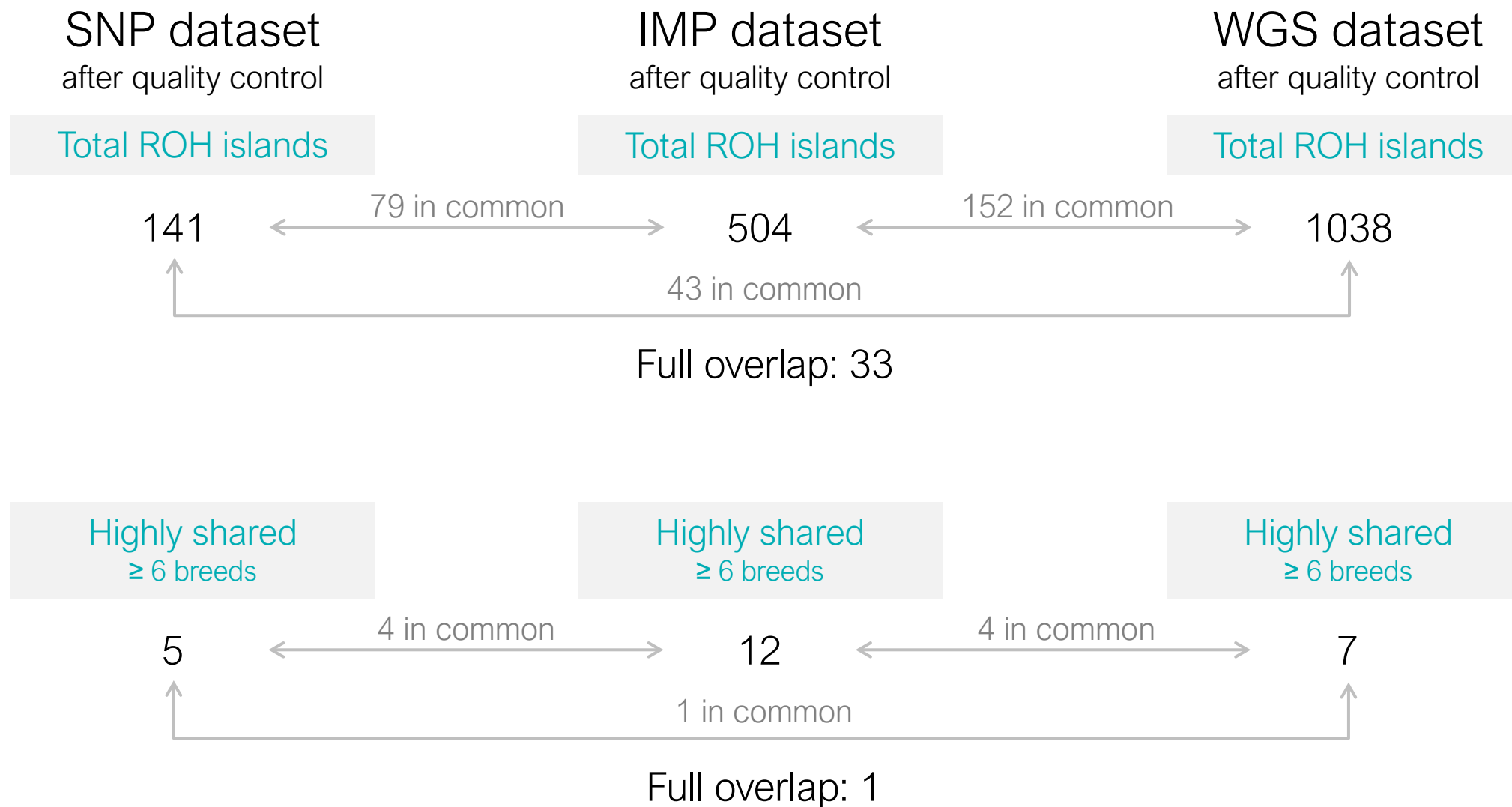
Highly homozygous genomic regions (**ROH islands**) were identified by transforming the SNP-within-ROH incidences per population into standard normal z-scores and thus calculating the *p*-value; only the top 0.5% of SNPs were considered to constitute an island



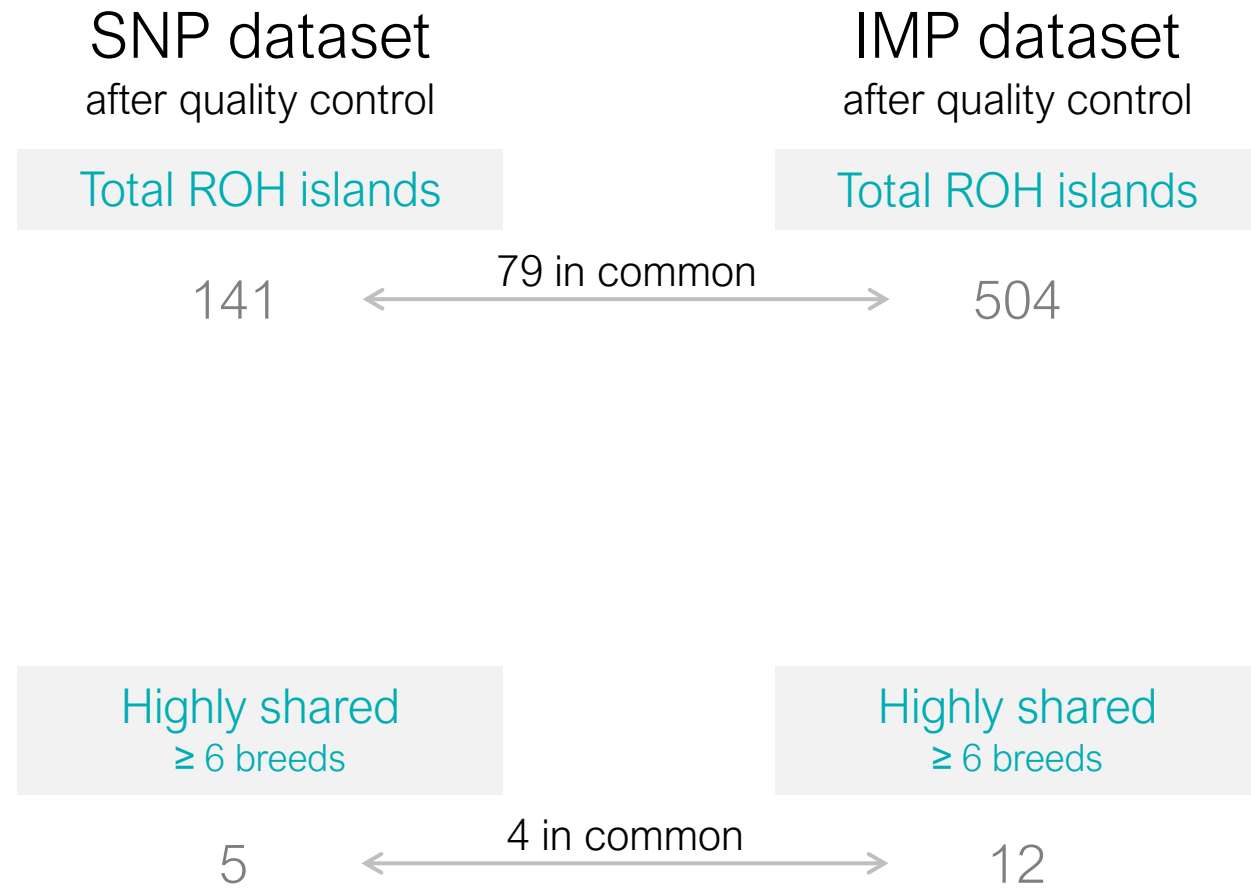
ROH islands



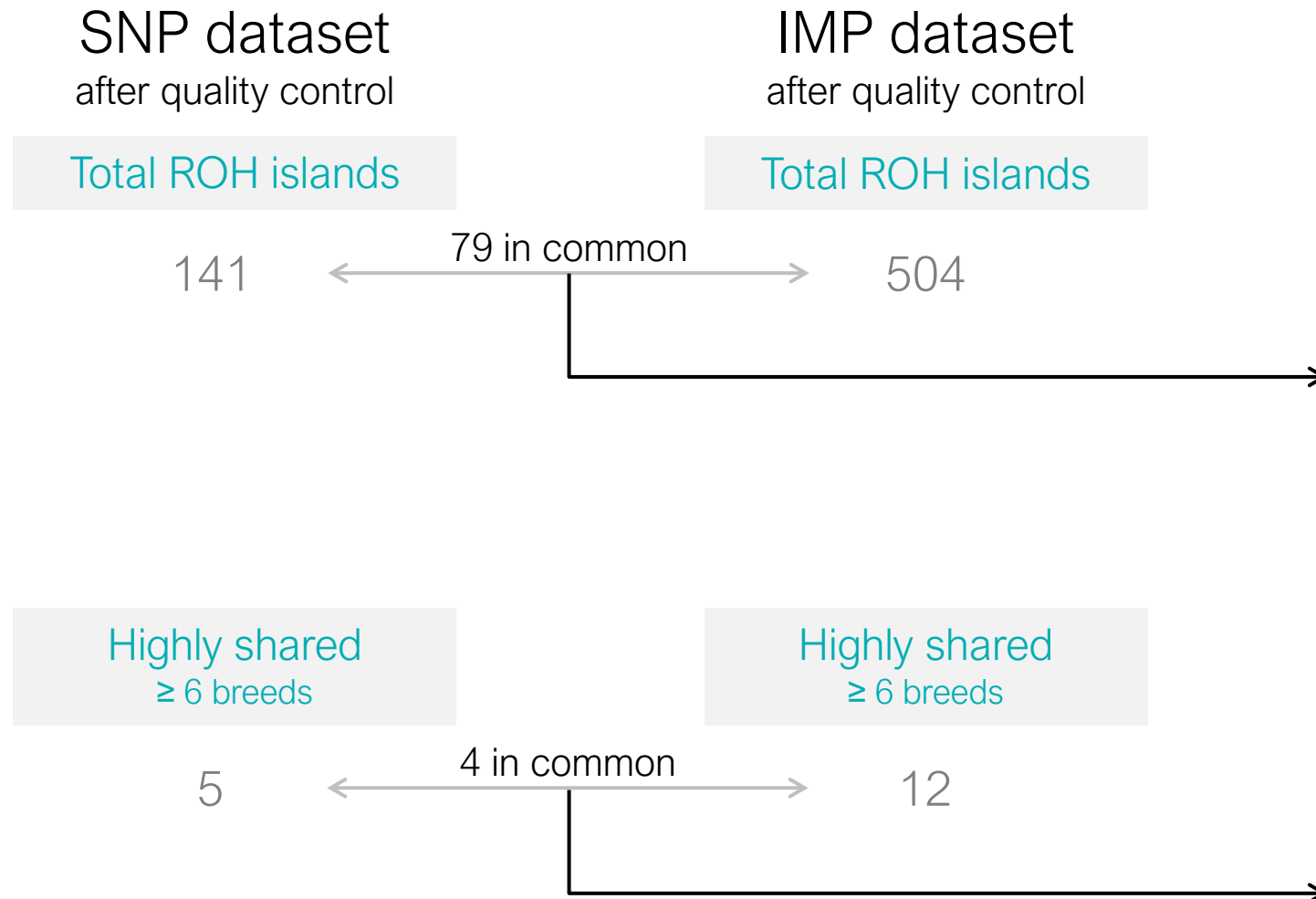
ROH islands



ROH islands



ROH islands



- From 3'039 to 316'045 markers
- From 11 to 973 markers with high or moderate impact on genes
- From 9 to 408 annotated genes

- From 503 to 74'892 markers
- From 3 to 468 markers with high or moderate impact on genes
- From 3 to 218 annotated genes



Conclusion

Are imputed data always more informative?

1. Reliability of imputed dataset
2. Application for predicting selection signatures
3. Limitation of the method: accuracy

Thanks for the attention



Any questions?

Contacts



giorgio.chessari@phd.unict.it



giorgio.chessari@agr.uni-goettingen.de

