AARHUS UNIVERSITY
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

62nd Annual Meeting EAAP 2011
August 29th – September 2nd
Stavanger NORWAY

Influence of Marker Editing Criteria on Accuracy of Genomic Prediction

Vahid Edriss PhD student

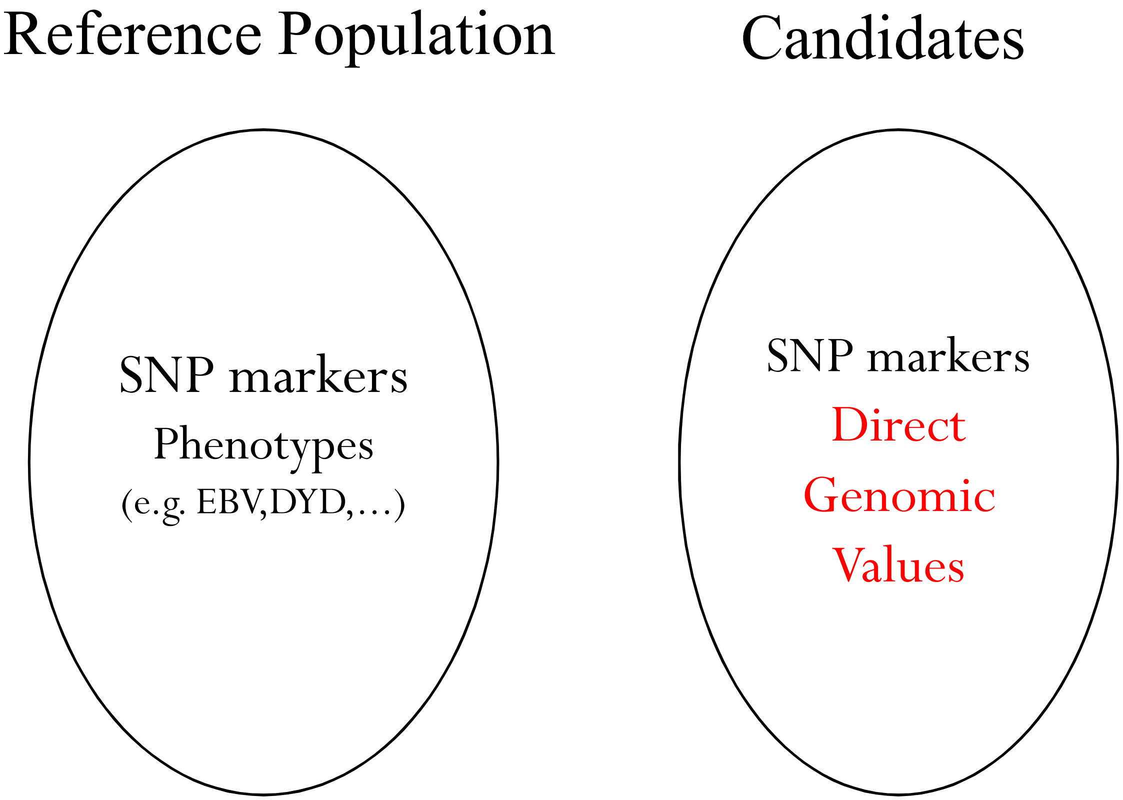
Guosheng Su
Mogens Sandø Lund
Bernt Guldbrandtsen

29 August 2011

Overview

- Genomic Selection
- Data
- Editing markers
- Imputation

Genomic Selection



Reference Population

SNP markers
Phenotypes
(e.g. EBV,DYD,…)

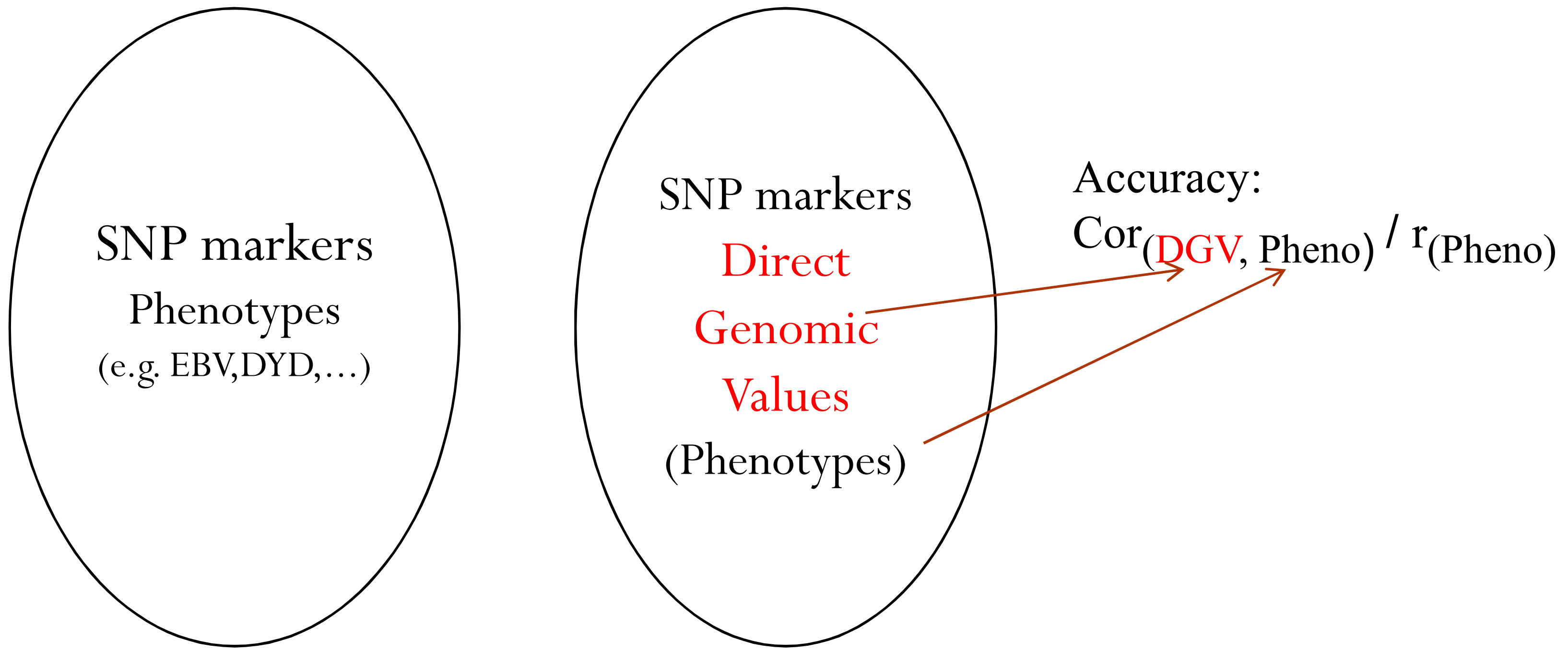
Candidates

SNP markers
Direct
Genomic
Values

Genomic Selection validation

Reference Population

Test Population



SNP markers
Phenotypes
(e.g. EBV,DYD,…)

SNP markers
Direct
Genomic
Values
(Phenotypes)

Accuracy:
$\mathrm{Cor}_{(\mathrm{DGV},\ \mathrm{Pheno})} / r_{(\mathrm{Pheno})}$

Data

- Data from Jersey and Holstein breeds.
- Traits :

    1) Fertility

    2) Milk protein

    3) Mastitis

Data (Jersey)

- Data contains :
  - 1,071 animals with phenotype and genotype.
  - Animals born from 1981 to 2005.
  - 44,305 SNPs from 30 chromosomes.
  - Deregress proof breeding value.

Data (Holstein)

- Data contains :
  - 4,429 animals with phenotype and genotype.
  - Animals born from 1974 to 2006.
  - 48,222 SNPs from 30 chromosomes.
  - Deregress proof breeding value.

Editing Markers

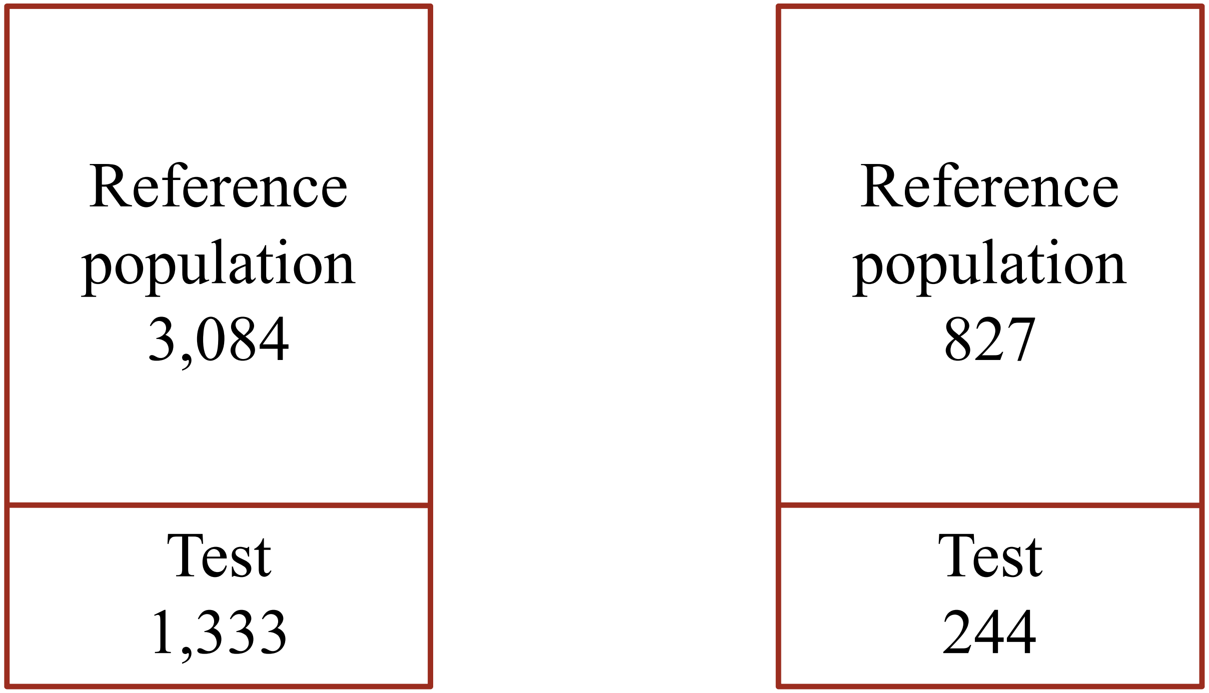
- Minor Allele Frequency (MAF)

- GenCall Score

Minor Allele Frequency

- Allele frequency for SNPs (p and q). Smaller one is MAF.
- To avoid spurious assoc. between SNP and family effects.
- Different countries use different thresholds for MAF (e.g.: USA 0.01, Australia 0.025, Norway 0.025, Nordic 0.05).
- Thresholds of no limitation, 0.001, 0.01, 0.02, 0.05 and 0.1

Calculating DGV's



Holstein

Reference population 3,084

Test 1,333

Total = 4,429

Jersey

Reference population 827

Test 244

Total = 1,071

Calculating DGV

- Use iBay to calculate the DGV for all the thresholds.

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{m} \mathbf{X}_i \mathbf{q}_i v_i + \mathbf{e}$$

$$\mathbf{q}_i \sim N(\mathbf{0}, \mathbf{I}) \qquad v_i \sim TN(0, \sigma_v^2) \qquad \mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$$

y = vector of phenotypic values
μ = overall mean
m = number of SNP markers
X_i = design matrix of the number of alleles.
q_i = vector of scaled SNP effects of marker i
v_i = scaling factor for the SNP effect i
e = residuals

Accuracy for different MAF thresholds in Jersey

	No limit	0.001	0.01	0.02	0.05	0.1
Number of SNP	44,305	42,100	39,097	37,951	35,267	31,105
Fertility	0.470	0.471	0.471	0.471	0.464	0.455
Milk protein	0.575	0.578	0.575	0.575	0.579	0.569
Mastitis	0.487	0.485	0.480	0.474	0.474	0.450

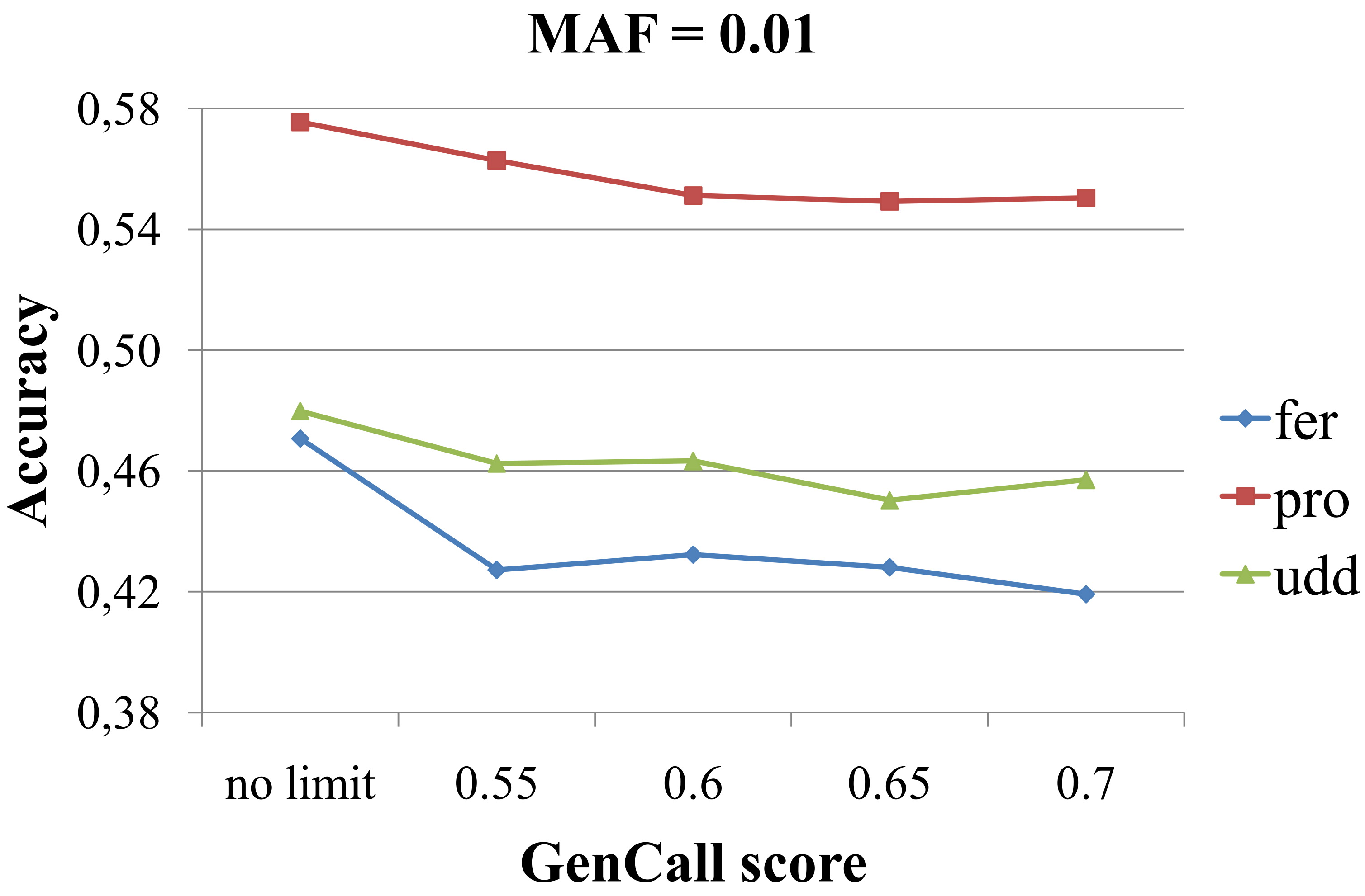
Accuracy for different MAF thresholds in Holstein

	No limit	0.001	0.01	0.02	0.05	0.1
Number of SNP	48,222	46,100	44,321	43,286	40,858	36,818
Fertility	0.614	0.614	0.614	0.613	0.613	0.609
Milk protein	0.648	0.647	0.649	0.650	0.649	0.650
Mastitis	0.620	0.622	0.623	0.622	0.623	0.620

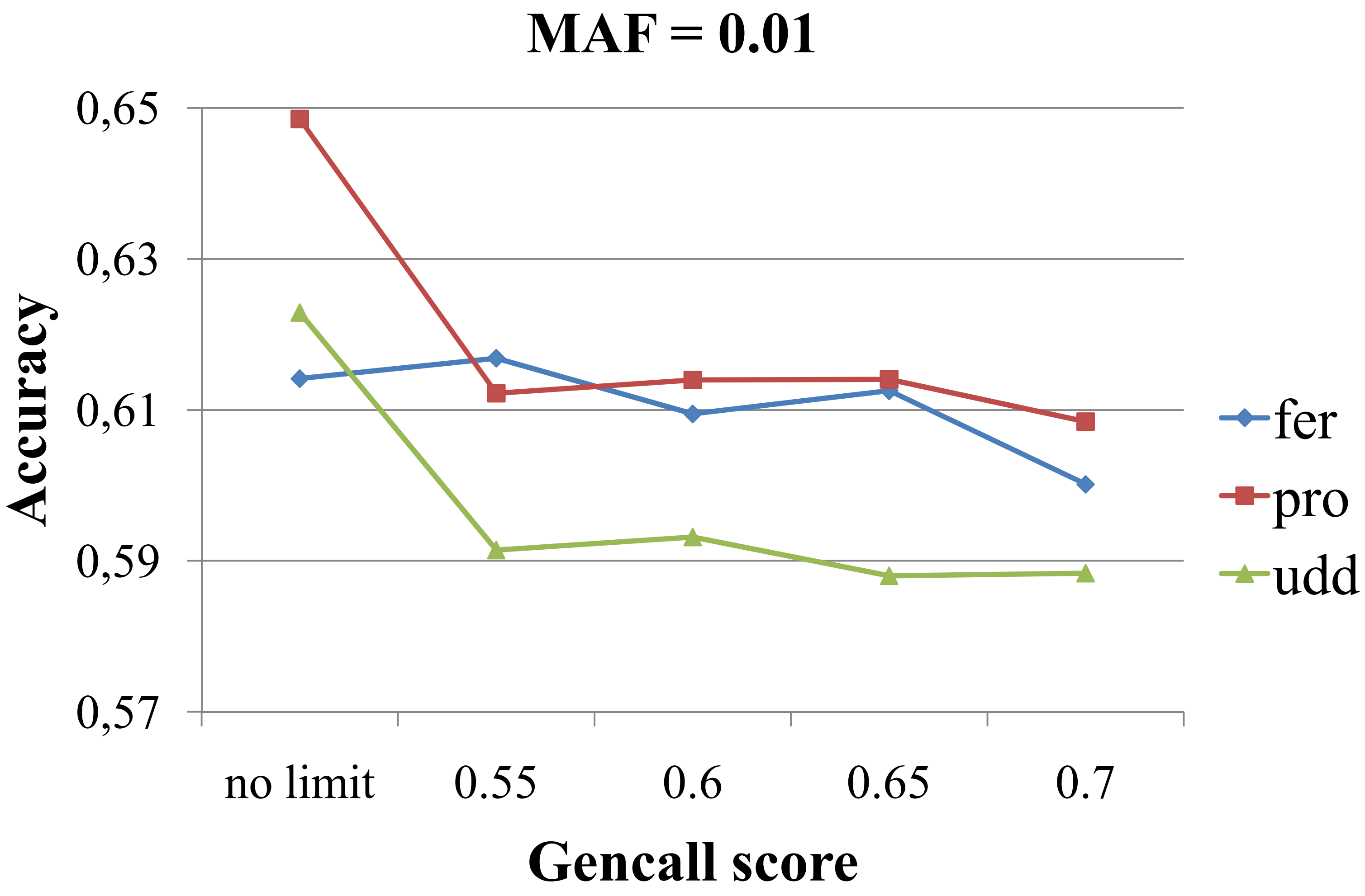
GenCall

- GenCall score is used to rank and fillter out failed genotypes and loci.
- Between 0 and 1.
- Lower GC Score have a lower reliability.
- Calculate accuracy based on 4 threshold for individual typing (GC Score less than 0.55, 0.6, 0.65 and 0.7)

GenCall (Jersey)



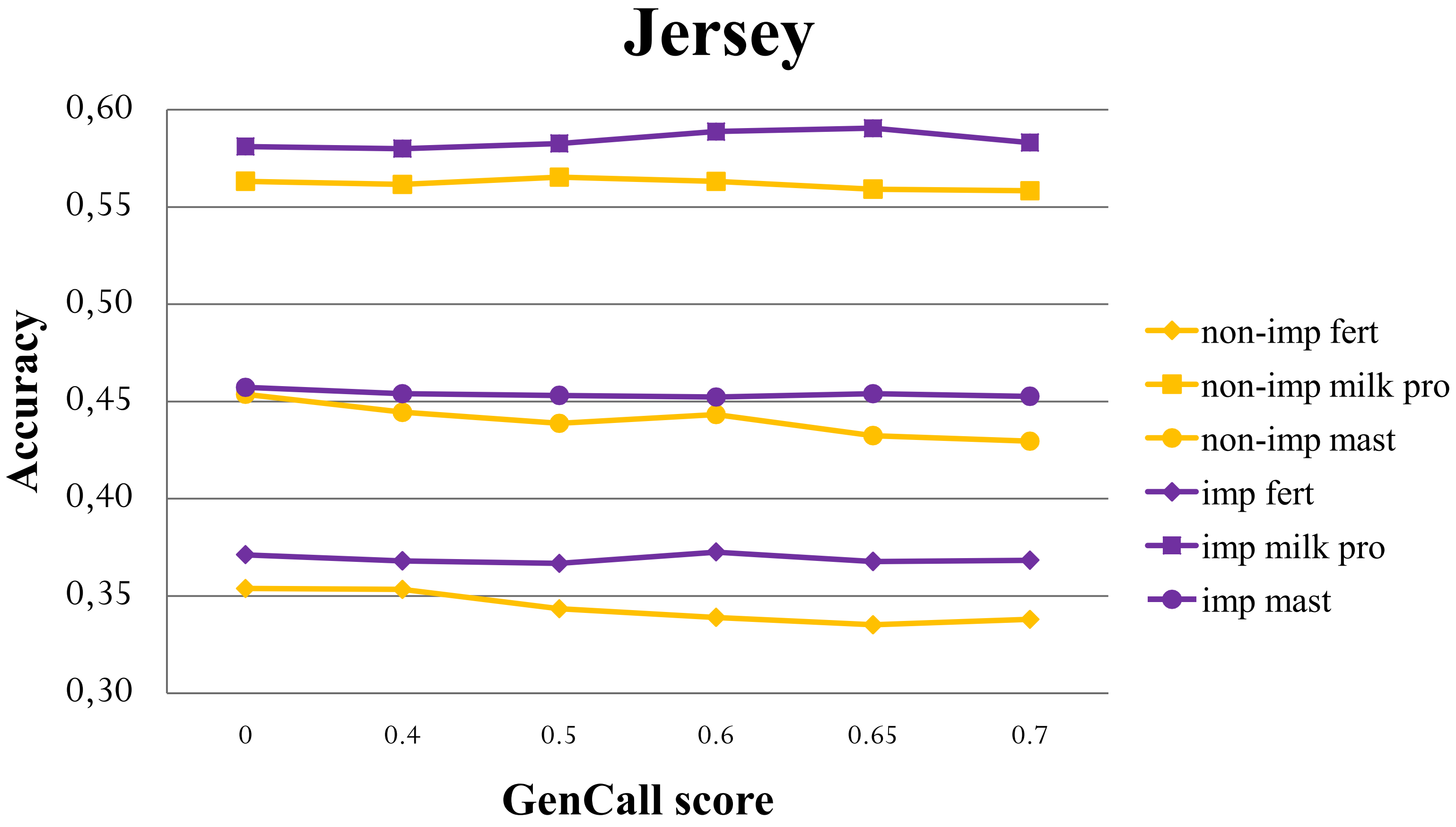
GenCall (Holstein)



Imputation

- Remove indiviual SNP with low GC score.
- Impute by Beagle package.
- Calculate DGV and accuracy.

Imputation


Jersey
Accuracy
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0
0.4
0.5
0.6
0.65
0.7
GenCall score
non-imp fert
non-imp milk pro
non-imp mast
imp fert
imp milk pro
imp mast

Conclutions

- Small difference and no clear trend between MAF’s.

- By taking out individual typing with low GC scores and replacing it with missing, accuracy goes down.

- Imputing missing values improve the accuracy.