Comparison: Michigan Imputation Server / Shapeit+Minimac3

Chromosome 8

Ignacio Tolosana

Imputation

Data exploration

```
michigan
```

```
## class: CollapsedVCF
## dim: 14014 2280
## rowRanges(vcf):
##   GRanges with 5 metadata columns: paramRangeID, REF, ALT, QUAL, FILTER
## info(vcf):
##   DataFrame with 4 columns: AF, MAF, R2, ER2
## info(header(vcf)):
##        Number Type  Description
##    AF  1      Float Estimated Alternate Allele Frequency
##    MAF 1      Float Estimated Minor Allele Frequency
##    R2  1      Float Estimated Imputation Accuracy
##    ER2 1      Float Empirical (Leave-One-Out) R-square (available only ...
## geno(vcf):
##   SimpleList of length 3: GT, DS, GP
## geno(header(vcf)):
##       Number Type   Description
##    GT 1      String Genotype
##    DS 1      Float  Estimated Alternate Allele Dosage : [P(0/1)+2*P(1/1)]
##    GP 3      Float  Estimated Posterior Probabilities for Genotypes 0/0...
```

```
minimac
```

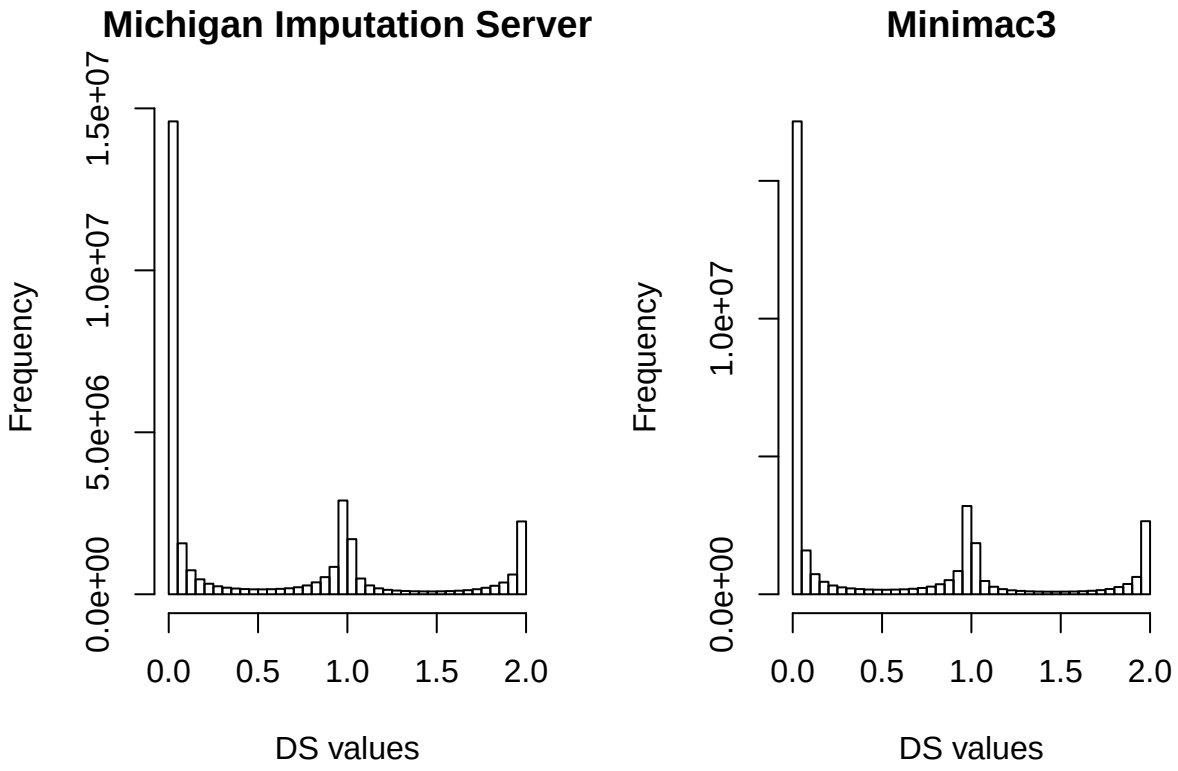
```
## class: CollapsedVCF
## dim: 15576 2280
## rowRanges(vcf):
##   GRanges with 5 metadata columns: paramRangeID, REF, ALT, QUAL, FILTER
## info(vcf):
##   DataFrame with 4 columns: AF, MAF, R2, ER2
## info(header(vcf)):
##        Number Type  Description
##    AF  1      Float Estimated Alternate Allele Frequency
##    MAF 1      Float Estimated Minor Allele Frequency
##    R2  1      Float Estimated Imputation Accuracy
##    ER2 1      Float Empirical (Leave-One-Out) R-square (available only ...
## geno(vcf):
```

```
##   SimpleList of length 3: GT, DS, GP
## geno(header(vcf)):
##       Number Type   Description
##    GT 1      String Genotype
##    DS 1      Float  Estimated Alternate Allele Dosage : [P(0/1)+2*P(1/1)]
##    GP 3      Float  Estimated Posterior Probabilities for Genotypes 0/0...
```

DS values

Distribution of the DS values in each imputation

```
par(mfrow=c(1,2))
HIST_MICHIGAN <- hist(DS_michigan, breaks=seq(0, 2, by=0.05), main="Michigan Imputation Server", xlab="
HIST_MINIMAC <- hist(DS_minimac, breaks=seq(0, 2, by=0.05), main="Minimac3", xlab="DS values")
```



DS correlation by individuals

```
min(cor_by_ind)
```

```
## [1] 0.9547537
```

```
max(cor_by_ind)
```

```
## [1] 0.9997556
```

```r
mean(cor_by_ind)
```

```
## [1] 0.9943097
```

```r
sum(cor_by_ind > 0.95)/length(cor_by_ind)
```

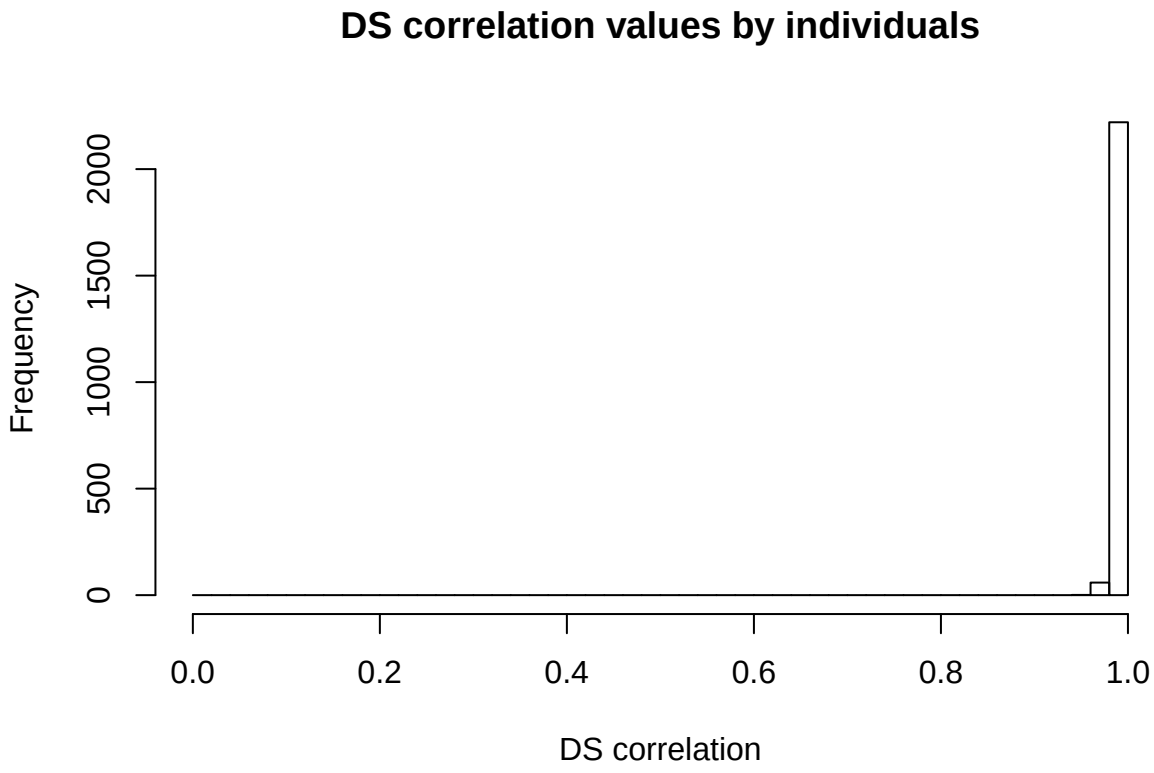
```
## [1] 1
```

```r
sum(cor_by_ind > 0.99)/length(cor_by_ind)
```

```
## [1] 0.8232456
```

Histogram of the DS correlation values by individuals

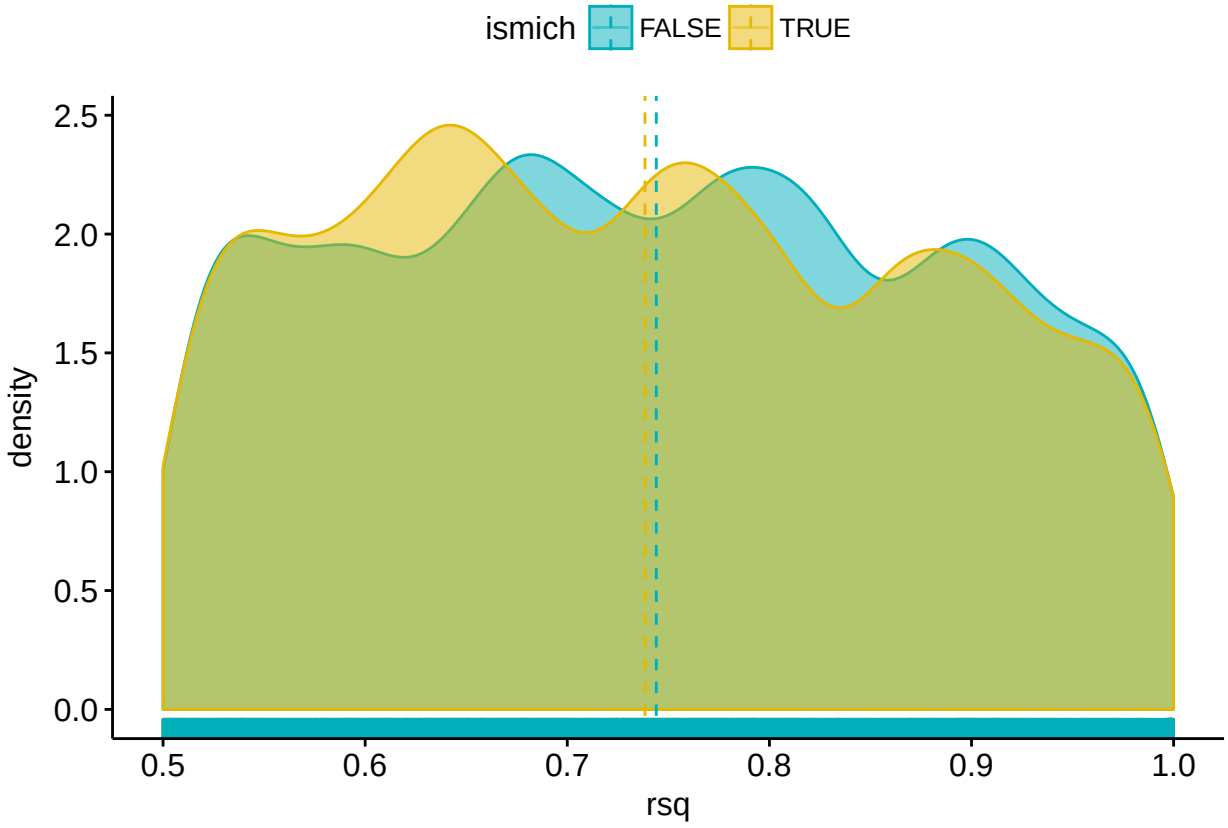
```r
par(mfrow=c(1,1))
CORR_HIST <- hist(cor_by_ind, breaks=seq(0, 1, by=0.02), main="DS correlation values by individuals", x
```



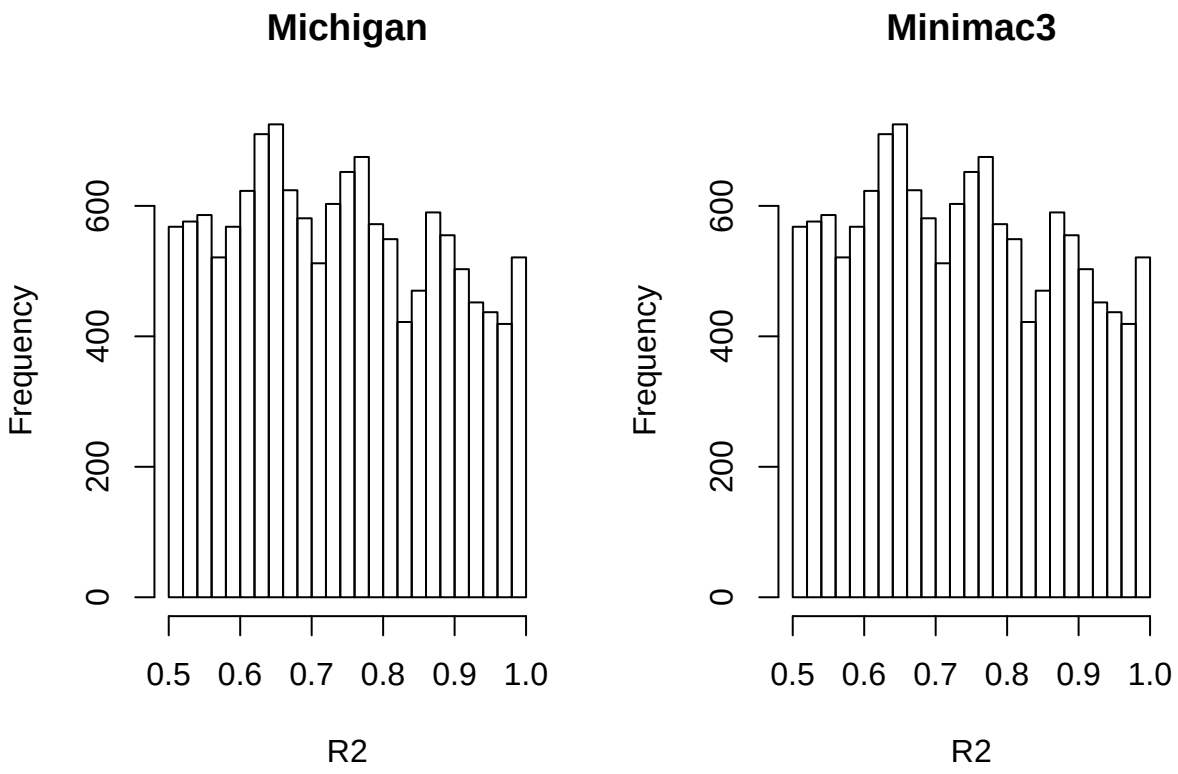
R^2

Density and histogram plots comparing the RS2 values in both methods ("ismich = TRUE" indicates the values for the Michigan imputation, whereas "ismich = FALSE" shows the values for the Minimac imputation)

```
ggdensity(comparison, x = "rsq",
          add = "mean", rug = TRUE,
          color = "ismich", fill = "ismich",
          palette = c("#00AFBB", "#E7B800"))
```




```
par(mfrow=c(1,2))
HIST_R2_MICHIGAN <- hist(info(michigan)$R2, breaks=seq(0.5, 1, by=0.02), main="Michigan", xlab="R2")
HIST_R2_MINIMAC <- hist(info(michigan)$R2, breaks=seq(0.5, 1, by=0.02), main="Minimac3", xlab="R2")
```



Genotype predictions

Compare the genotype predictions with each method by individuals. "perc_by_ind" is the % of SNPs by individual predicted equally in both methods

```
min(perc_by_ind)
```

```
## [1] 0.9047722
```

```
max(perc_by_ind)
```

```
## [1] 0.9995688
```

```
mean(perc_by_ind)
```

```
## [1] 0.9845408
```

```
sum(perc_by_ind > 0.95)/length(perc_by_ind)
```

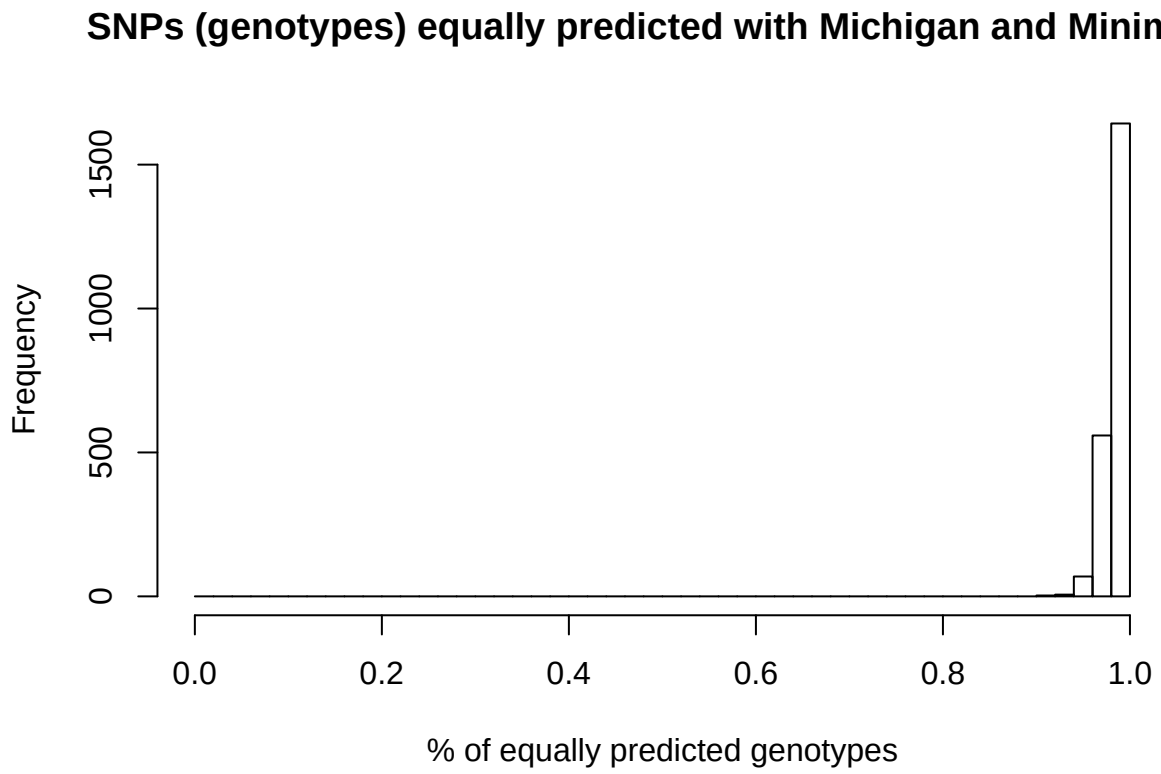
```
## [1] 0.9877193
```

```
sum(perc_by_ind > 0.99)/length(perc_by_ind)
```

```
## [1] 0.372807
```

Plot histogram of the genotypes equally predicted by individuals in both methods

```
GENO_HIST <- hist(perc_by_ind, breaks=seq(0, 1, by=0.02), main="SNPs (genotypes) equally predicted with
```



Inversion prediction

Predicted inversions with scoreInvHap

```
michigan_inv
```

```
## scoreInvHapRes
## Samples:  2280
## Genotypes' table:
##  NI/NI    NI/I    I/I
##  729  1064    487
## - Inversion genotypes' table:
##  NN   NI  II
##  729  1064    487
## - Inversion frequency: 44.69%
```

```
minimac_inv
```

```
## scoreInvHapRes
## Samples:  2280
## Genotypes' table:
##  NI/NI    NI/I    I/I
##  731  1063    486
## - Inversion genotypes' table:
##  NN   NI  II
##  731  1063    486
## - Inversion frequency: 44.63%
```

Comparison table

```
scoreinvhap_table
```

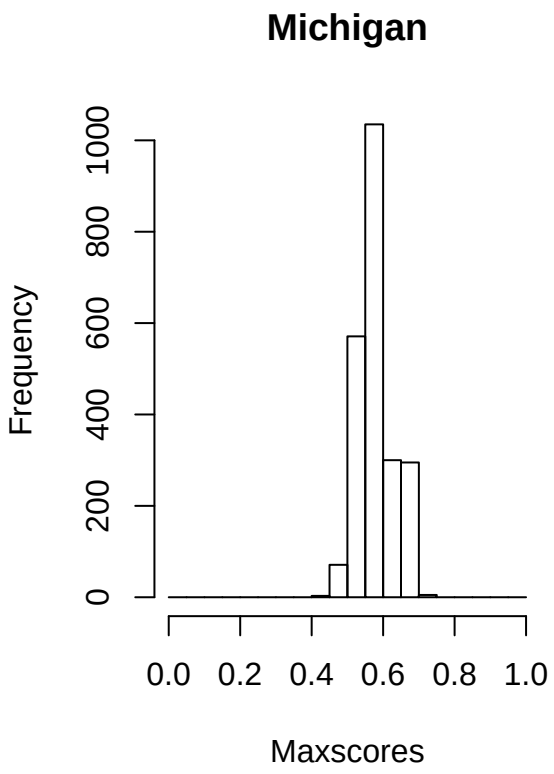
```
##         Minimac
## Michigan NI/NI NI/I  I/I
##    NI/NI   727    2    0
##    NI/I      4 1060    0
##    I/I       0    1  486
```

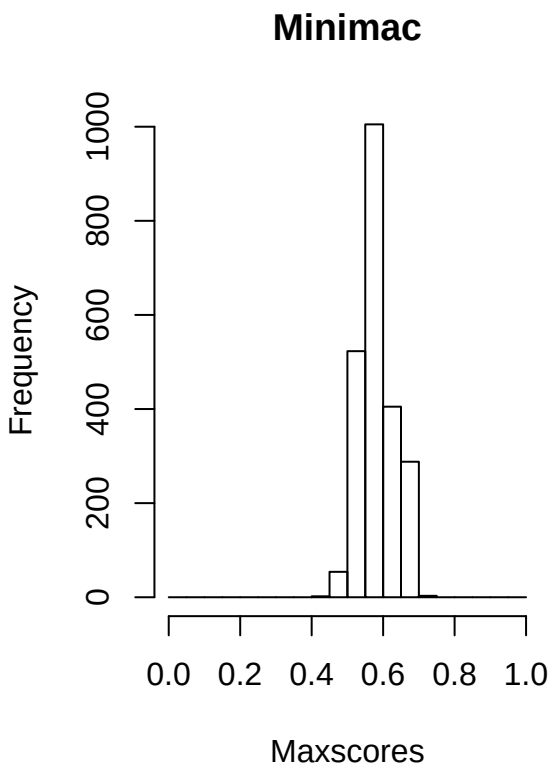
```
sum(diag(scoreinvhap_table))/sum(scoreinvhap_table)
```

```
## [1] 0.9969298
```

Comparison of the results for both imputation methods

```
par(mfrow=c(1,2))
hist(maxscores(michigan_inv), breaks=seq(0, 1, by=0.05), main="Michigan", xlab="Maxscores")
hist(maxscores(minimac_inv), breaks=seq(0, 1, by=0.05), main="Minimac", xlab="Maxscores")
```





Score correlation by individuals between both imputation methods

```
min(score_corr)
```

```
## [1] 0.9988036
```

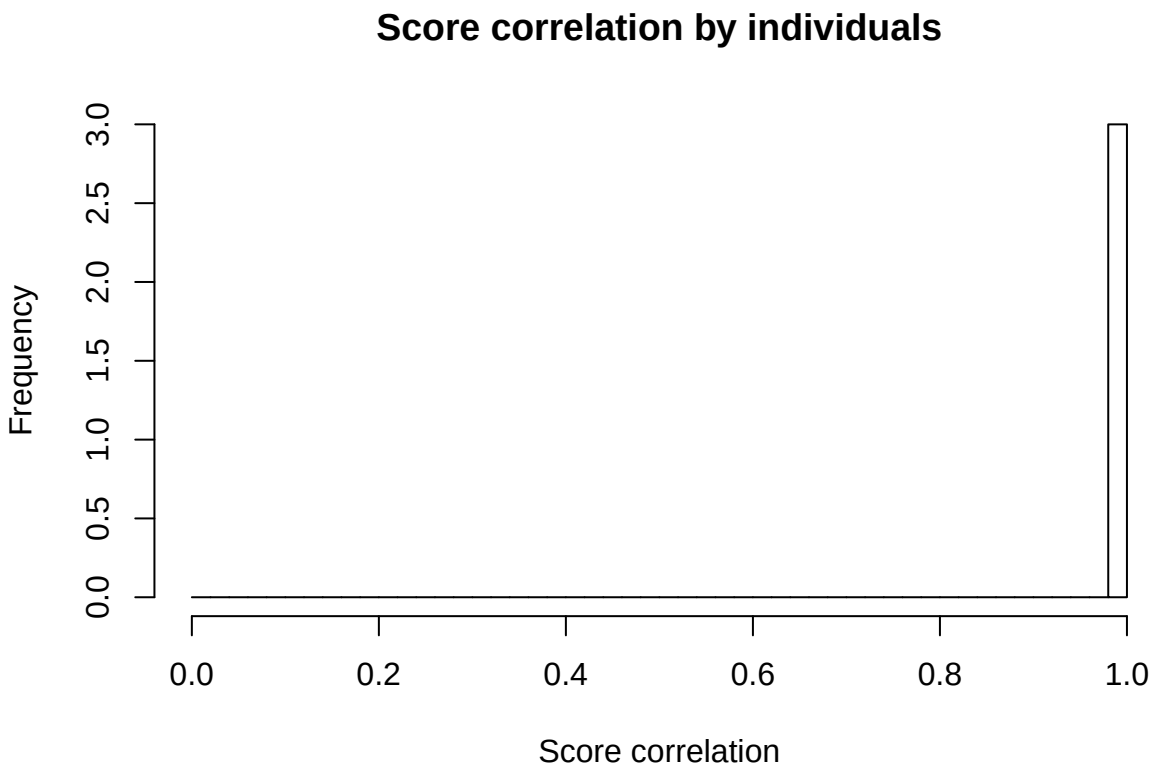
```
max(score_corr)
```

```
## [1] 0.9996256
```

```
mean(score_corr)
```

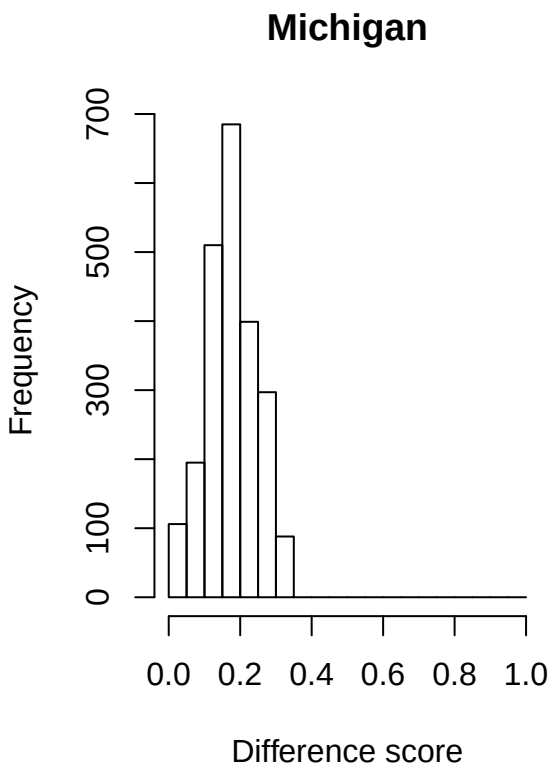
```
## [1] 0.9993435
```

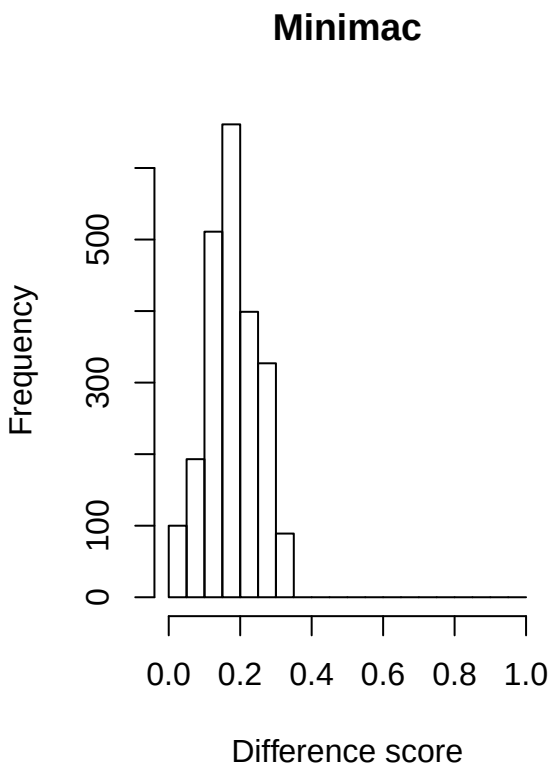
```
par(mfrow=c(1,1))
SCORE_CORR_HIST <- hist(score_corr, breaks=seq(0, 1, by=0.02), main="Score correlation by individuals",
```



Difference score between the highest similarity score and the second highest, in both imputation methods

```
par(mfrow=c(1,2))
hist(diffscores(michigan_inv), breaks=seq(0, 1, by=0.05), main="Michigan", xlab="Difference score")
hist(diffscores(minimac_inv), breaks=seq(0, 1, by=0.05), main="Minimac", xlab="Difference score")
```





Numbers of scores used

```
mean(numSNPs(michigan_inv))
```

```
## [1] 10011
```

```
max(numSNPs(michigan_inv))
```

```
## [1] 10011
```

```
min(numSNPs(michigan_inv))
```

```
## [1] 10011
```

```
mean(numSNPs(minimac_inv))
```

```
## [1] 10660
```

```
max(numSNPs(minimac_inv))
```

```
## [1] 10660
```

```
min(numSNPs(minimac_inv))
```

```
## [1] 10660
```

Number of samples in both imputation methods before and after QC filtering

```
length(classification(michigan_inv))
```

```
## [1] 2280
```

```
length(classification(michigan_inv, minDiff = 0.1, callRate = 0.9))
```

```
## [1] 1979
```

```
length(classification(michigan_inv, minDiff = 0.1, callRate = 0.9))/length(classification(michigan_inv))
```

```
## [1] 0.8679825
```

```
length(classification(minimac_inv))
```

```
## [1] 2280
```

```
length(classification(minimac_inv, minDiff = 0.1, callRate = 0.9))
```

```
## [1] 1987
```

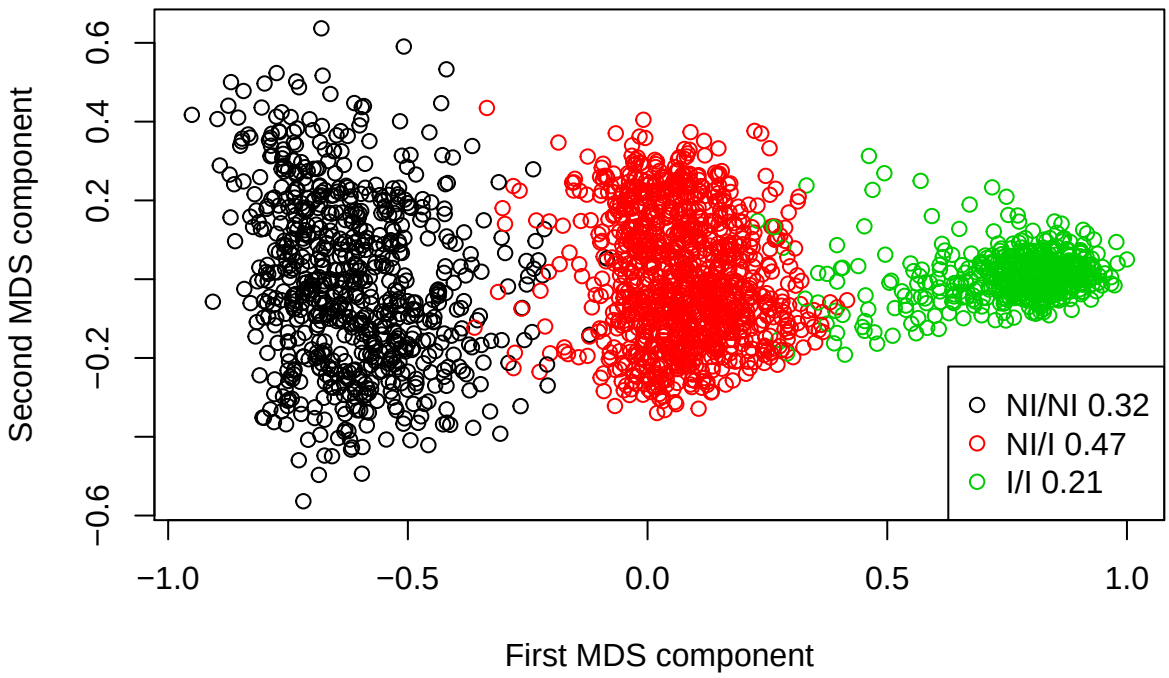
```
length(classification(minimac_inv, minDiff = 0.1, callRate = 0.9))/length(classification(minimac_inv))
```

```
## [1] 0.8714912
```

Plots with invClust

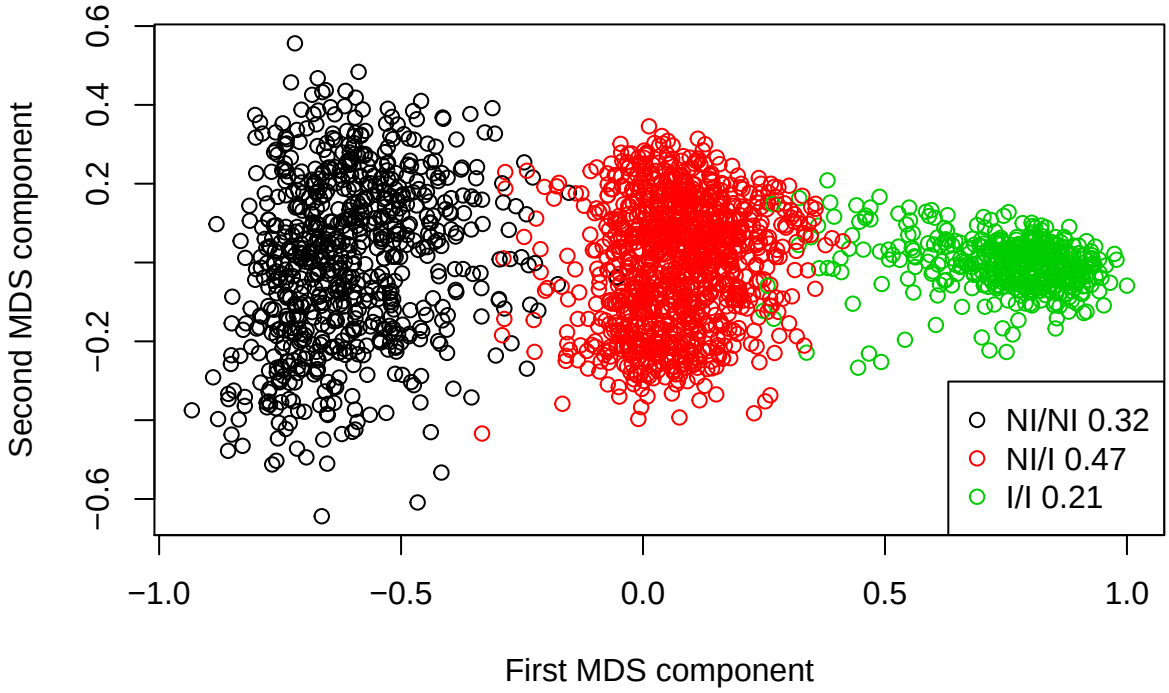
Michigan:

```
plotInv(michigan_invclust, classification = classification(michigan_inv))
```



Minimac:

```
plotInv(minimac_invclust, classification = classification(minimac_inv))
```


Second MDS component
0.6
0.4
0.2
−0.2
−0.6
−1.0
−0.5
0.0
0.5
1.0
First MDS component
NI/NI 0.32
NI/I 0.47
I/I 0.21