Comparison: Michigan Imputation Server / Shapeit+Minimac3

Chromosome 7

Ignacio Tolosana

Imputation

Data exploration

```
michigan
```

```
## class: CollapsedVCF
## dim: 192 2275
## 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
```

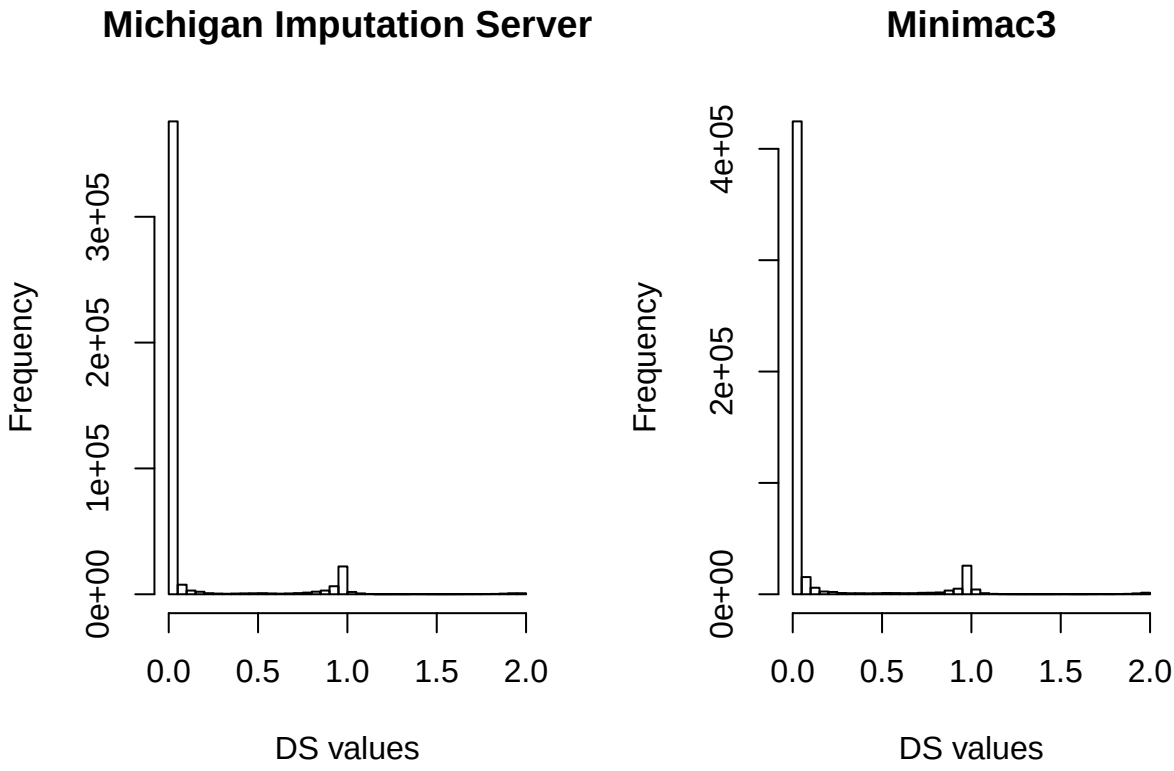
```
## class: CollapsedVCF
## dim: 223 2275
## rowRanges(vcf):
##   GRanges with 5 metadata columns: paramRangeID, REF, ALT, QUAL, FILTER
## info(vcf):
##   DataFrame with 4 columns: AF, MAF, R2, ER2
## info(header(vcf)):
##        Number Type  Description
##    AF  1      Float Estimated Alternate Allele 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] 0.4930321
```

```
max(cor_by_ind)
```

```
## [1] 0.9999955
```

```r
mean(cor_by_ind)
```

```
## [1] 0.9857534
```

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

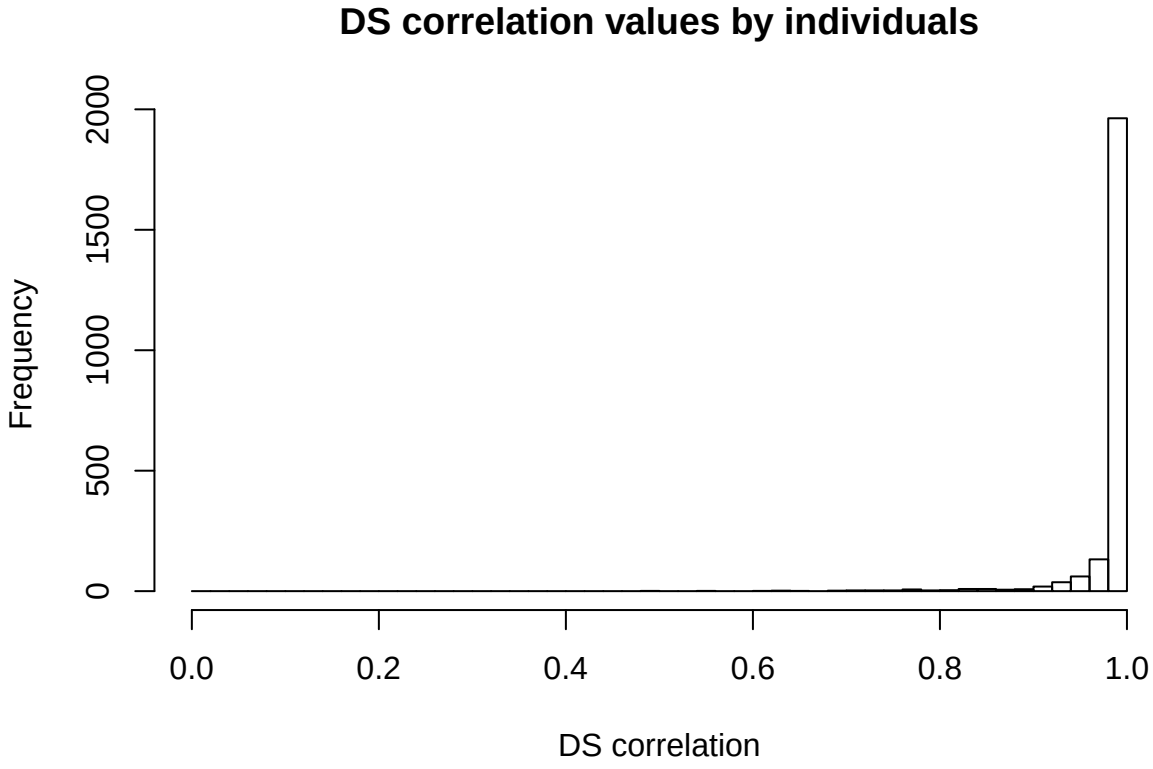
```
## [1] 0.9358242
```

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

```
## [1] 0.7578022
```

Histogram of the DS correlation values by individuals

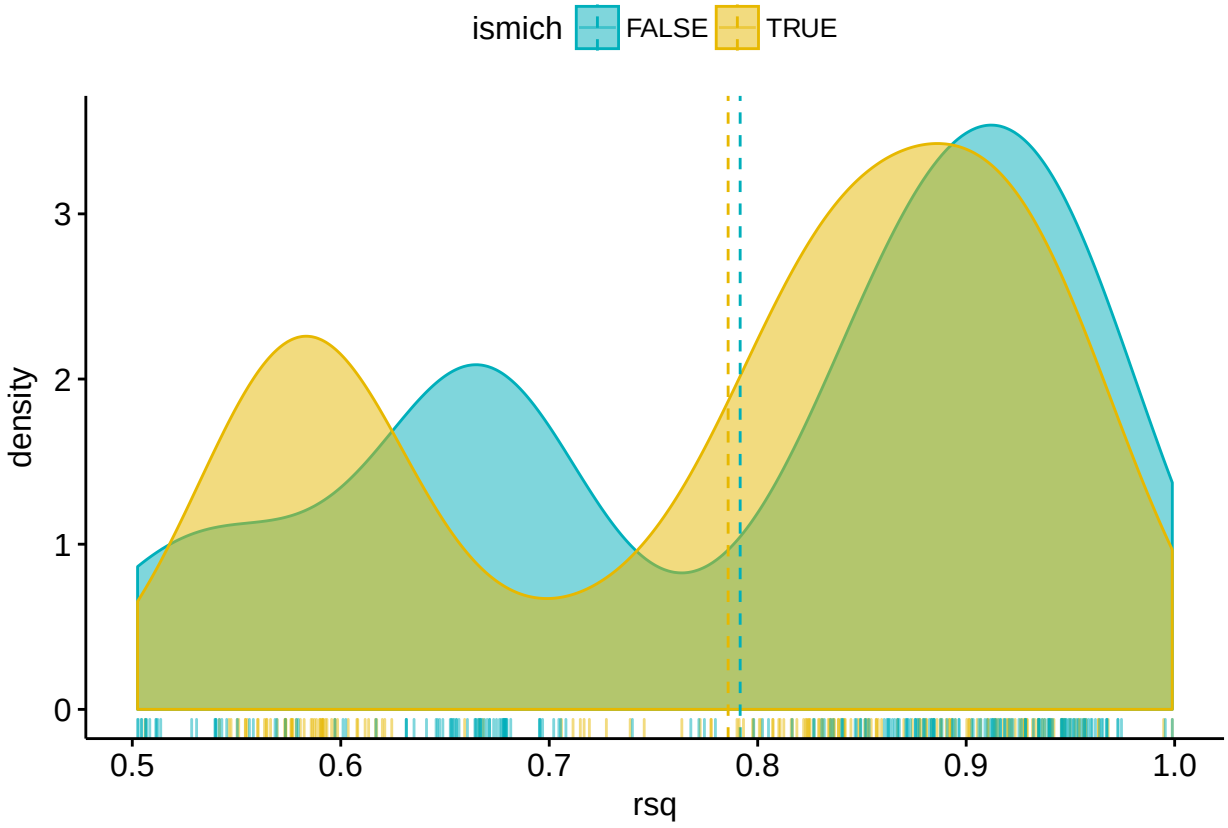
```r
par(mfrow=c(1,1))
CORR_HIST <- hist(cor_by_ind, breaks=seq(0, 1, by=0.02), main="DS correlation values by individuals", x
```



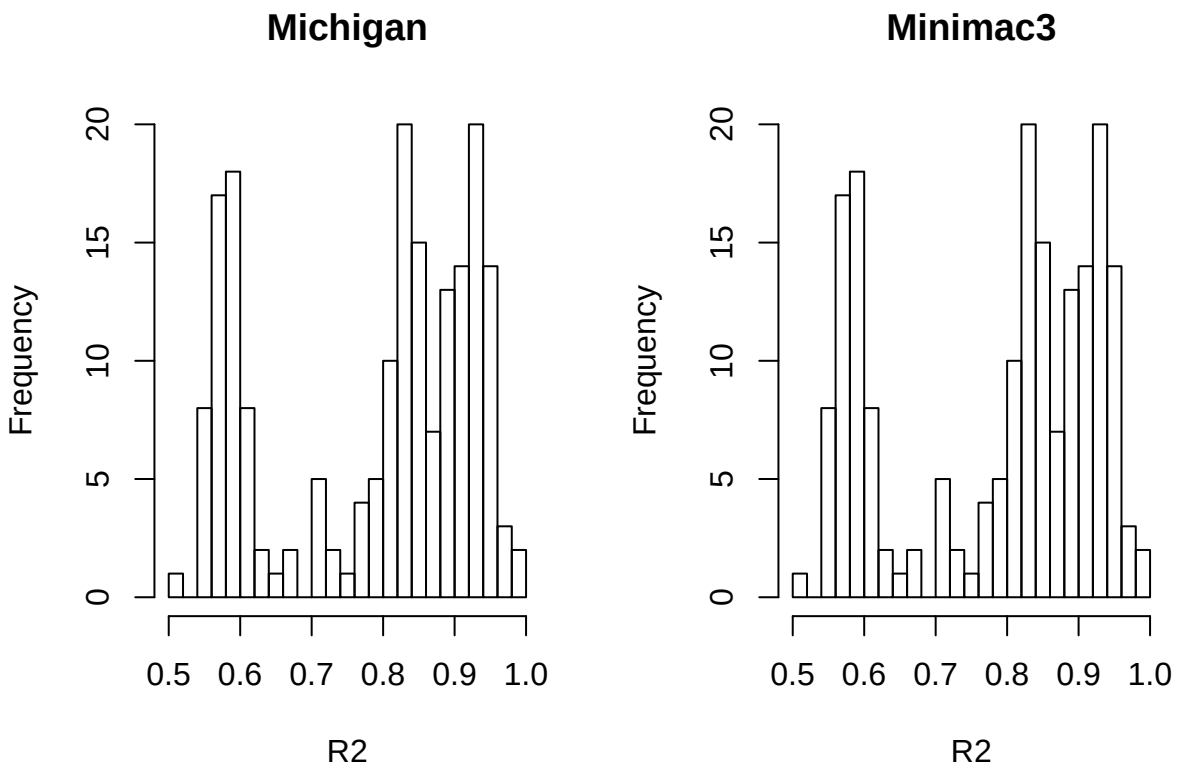
R^2

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] 0.7864583
```

```
max(perc_by_ind)
```

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

```
mean(perc_by_ind)
```

```
## [1] 0.9972436
```

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

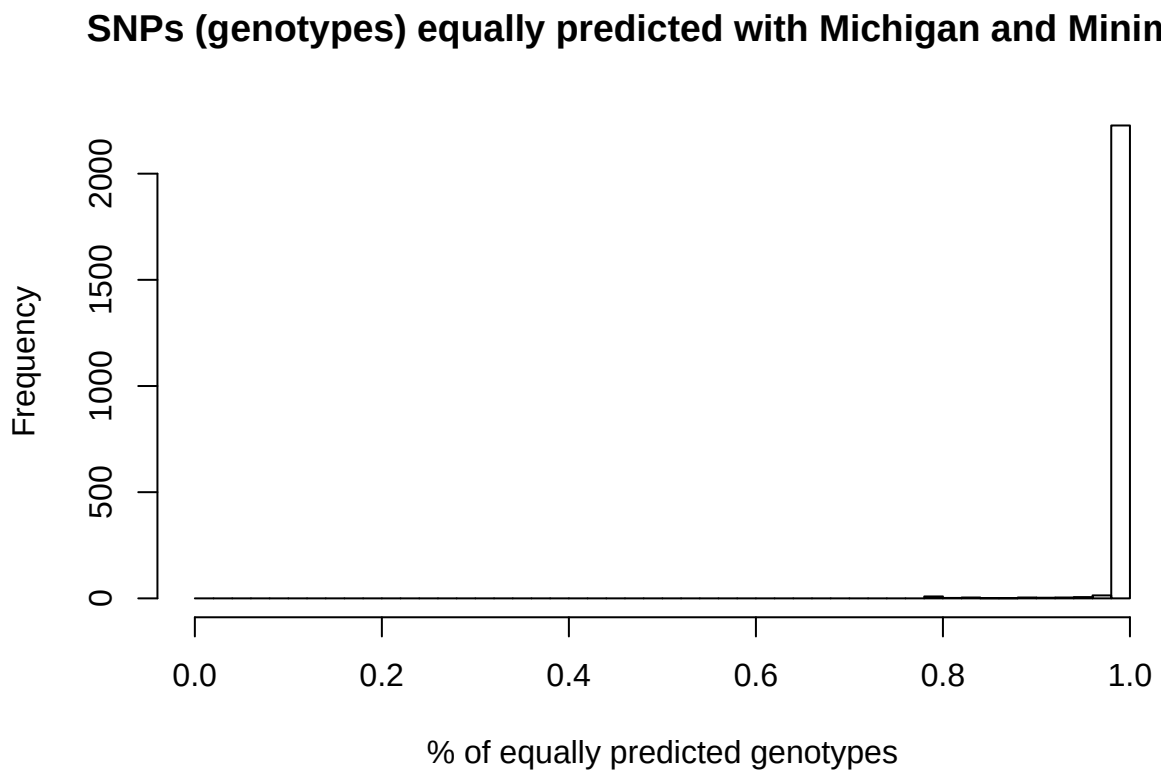
```
## [1] 0.9863736
```

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

```
## [1] 0.952967
```

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:  2275
## Genotypes' table:
##  IaIa    IaIb    IbIb    NaIa    NaIb    NaNa    NaNb    NbIa    NbIb    NbNb
##  89   5  367     631     0   1036    0   21  123     3
## - Inversion genotypes' table:
##  NN   NI  II
##  1039     775     461
## - Inversion frequency: 37.30%
```

```
minimac_inv
```

```
## scoreInvHapRes
## Samples:  2275
## Genotypes' table:
##  IaIa     IaIb    IbIb    NaIa    NaIb    NaNa    NaNb    NbIa    NbIb    NbNb
##  89   70  433     565     0   970     2   21  118     7
## - Inversion genotypes' table:
##  NN   NI  II
##  979  704     592
## - Inversion frequency: 41.49%
```

Comparison table

```
scoreinvhap_table
```

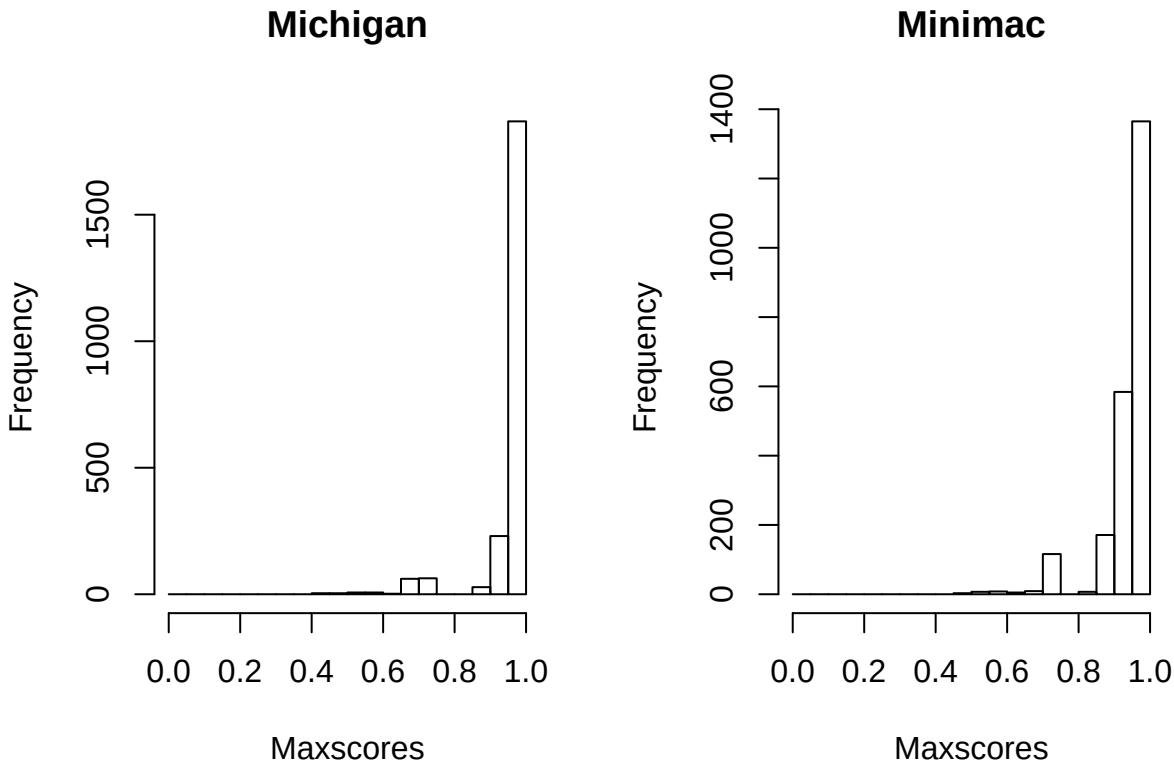
```
##         Minimac
## Michigan IaIa IaIb IbIb NaIa NaIb NaNa NaNb NbIa NbIb NbNb
##     IaIa   89    0    0    0    0    0    0    0    0    0
##     IaIb    0    4    0    1    0    0    0    0    0    0
##     IbIb    0    0  344    0    0   23    0    0    0    0
##     NaIa    0   66    1  564    0    0    0    0    0    0
##     NaIb    0    0    0    0    0    0    0    0    0    0
##     NaNa    0    0   88    0    0  947    1    0    0    0
##     NaNb    0    0    0    0    0    0    0    0    0    0
##     NbIa    0    0    0    0    0    0    0   21    0    0
##     NbIb    0    0    0    0    0    0    1    0  115    7
##     NbNb    0    0    0    0    0    0    0    0    3    0
```

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

```
## [1] 0.916044
```

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] 0.9771856
```

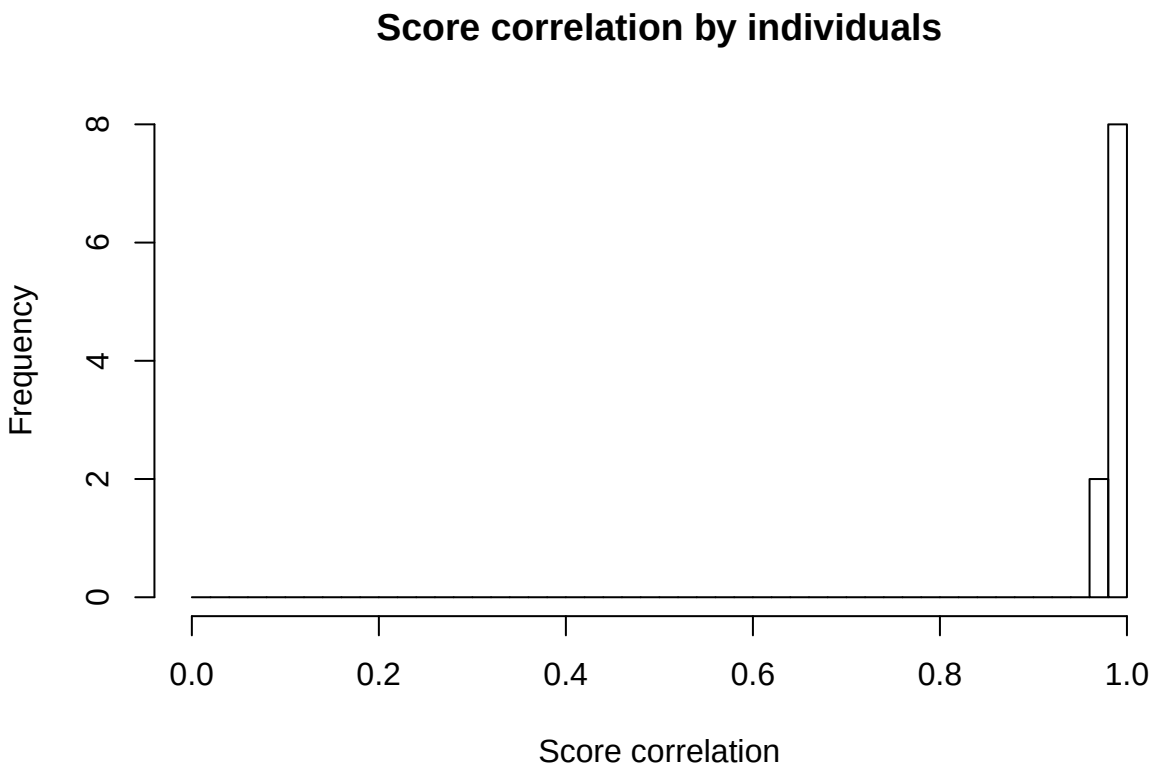
```
max(score_corr)
```

```
## [1] 0.9973483
```

```
mean(score_corr)
```

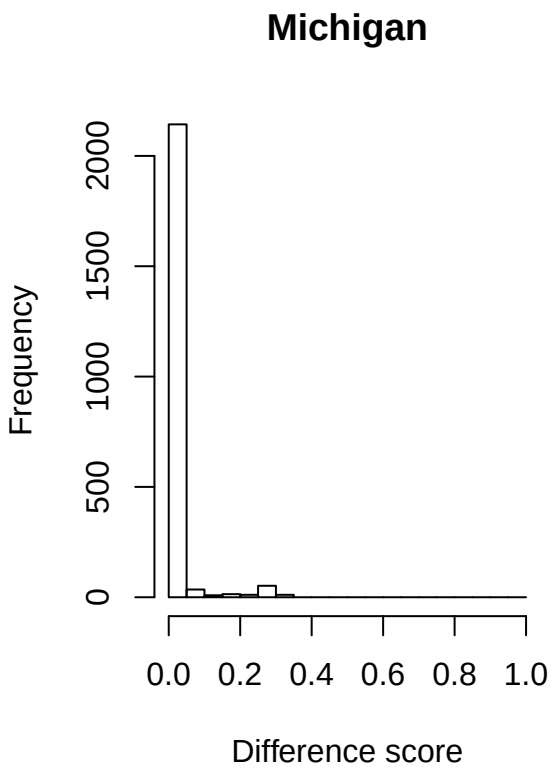
```
## [1] 0.9866348
```

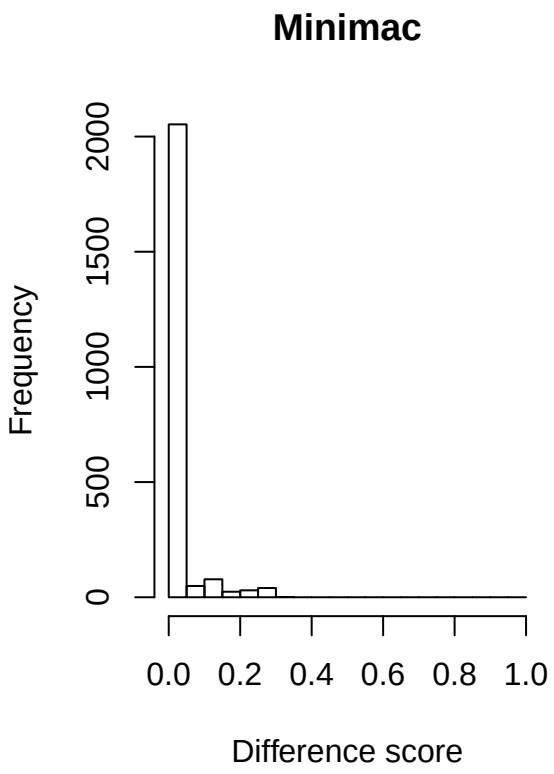
```
par(mfrow=c(1,1))
SCORE_CORR_HIST <- hist(score_corr, breaks=seq(0, 1, by=0.02), main="Score correlation by individuals",
```



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] 164
```

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

```
## [1] 164
```

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

```
## [1] 164
```

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

```
## [1] 184
```

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

```
## [1] 184
```

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

```
## [1] 184
```

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

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

```
## [1] 2275
```

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

```
## [1] 97
```

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

```
## [1] 0.04263736
```

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

```
## [1] 2275
```

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

```
## [1] 173
```

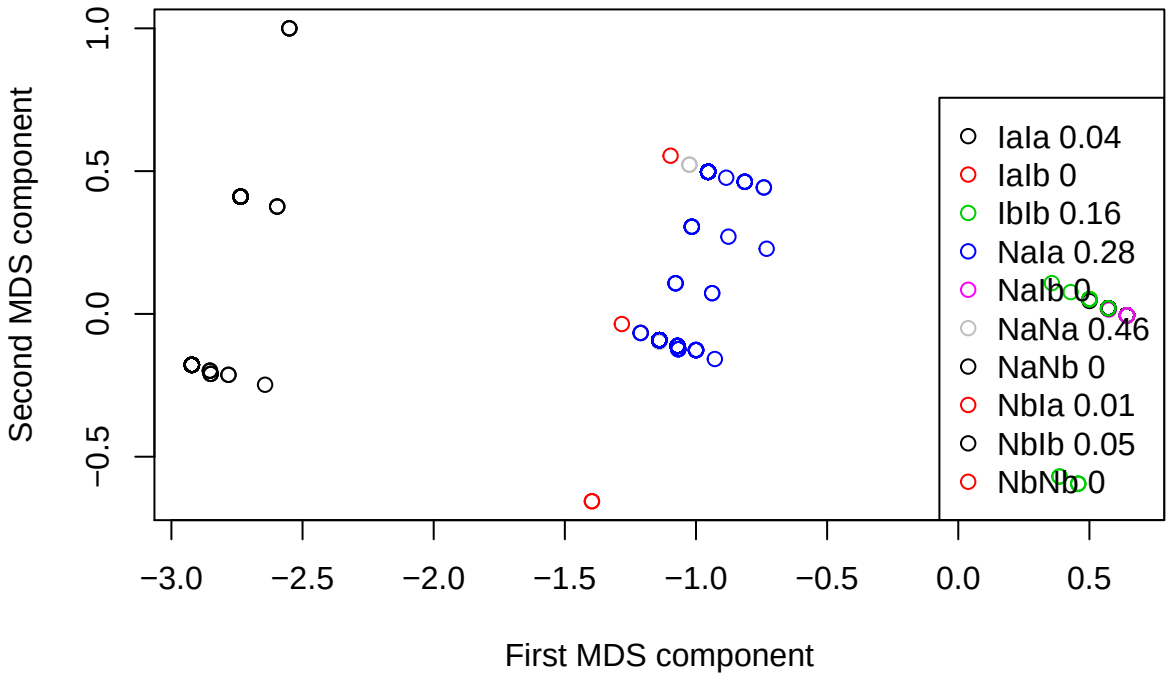
```
length(classification(minimac_inv, minDiff = 0.1, callRate = 0.9))/length(classification(minimac_inv))
```

```
## [1] 0.07604396
```

Plots with invClust

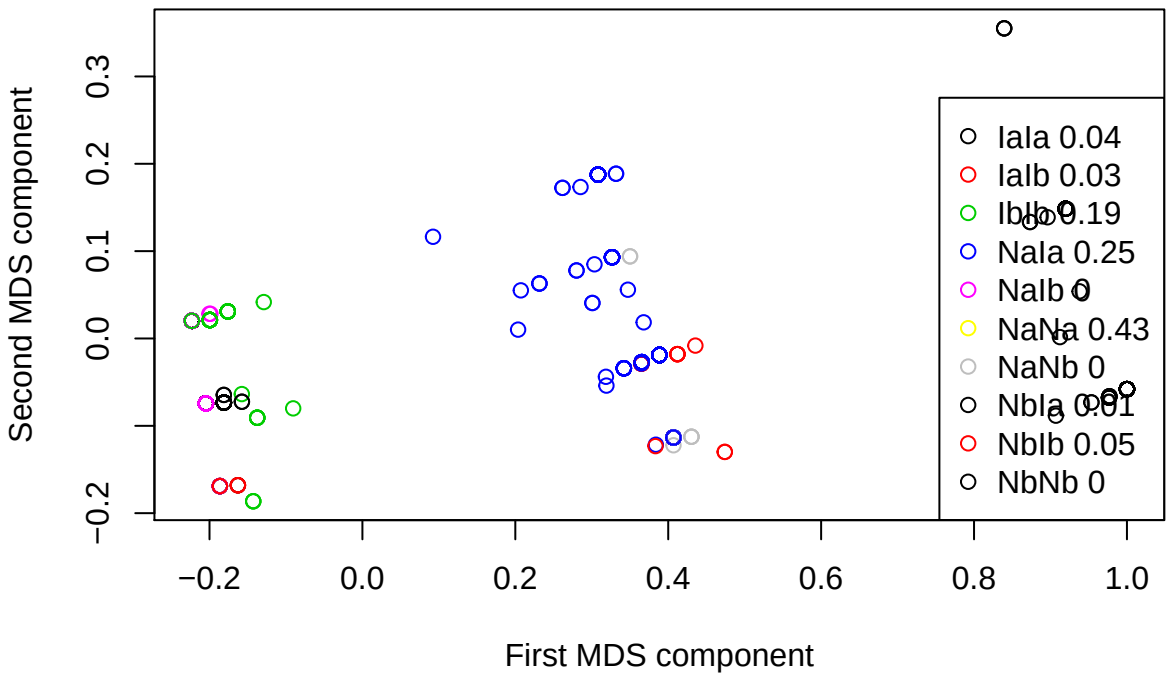
Michigan:

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



Minimac:

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


Second MDS component
First MDS component
−0.2
0.0
0.1
0.2
0.3
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Ia|a 0.04
Ia|b 0.03
Ib|b 0.19
Na|a 0.25
Na|b 0
Na|Na 0.43
Na|Nb 0
Nb|a 0.01
Nb|b 0.05
Nb|Nb 0