Comparison: Michigan Imputation Server / Shapeit+Minimac3

Chromosome 17

Ignacio Tolosana

Imputation

Data exploration

```
michigan
```

```
## class: CollapsedVCF
## dim: 3041 2286
## 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: 3176 2286
## 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")
```
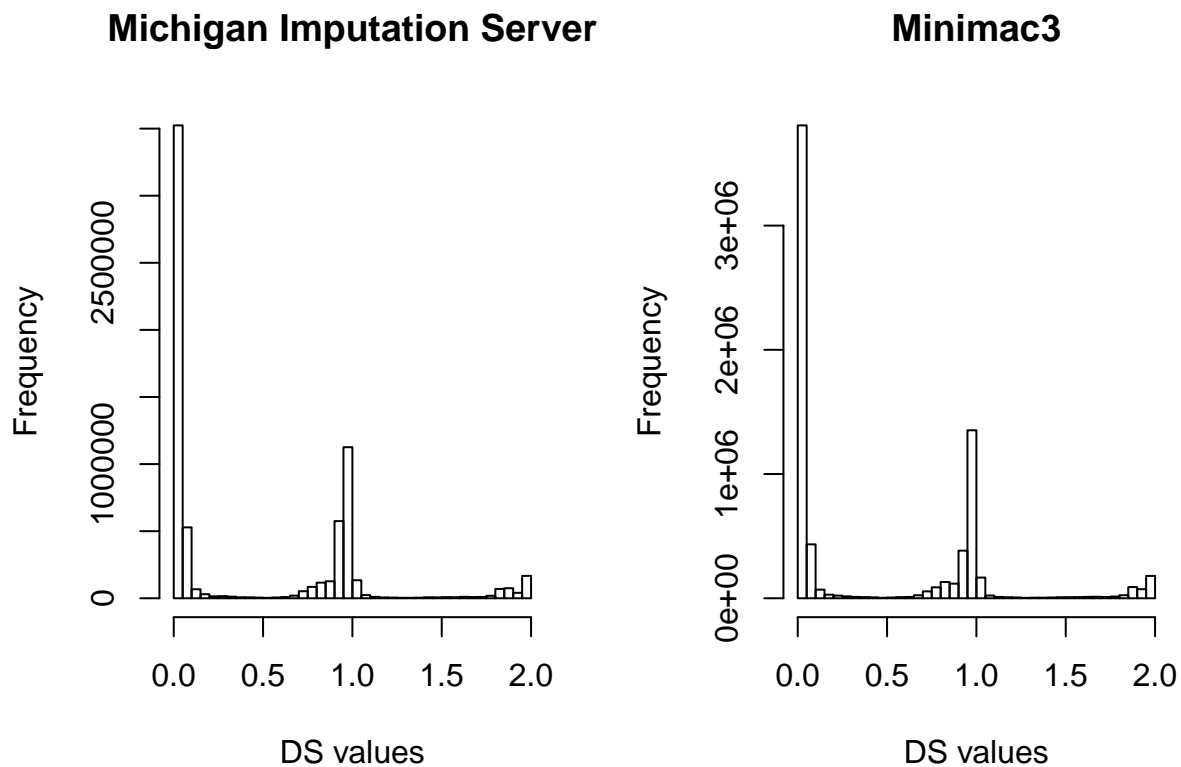

DS correlation by individuals

```
min(cor_by_ind)
```

```
## [1] 0.8202466
```

```
max(cor_by_ind)
```

```
## [1] 0.9999577
```

```
mean(cor_by_ind)
```

```
## [1] 0.9979358
```

```
sum(cor_by_ind > 0.95)/length(cor_by_ind)
```
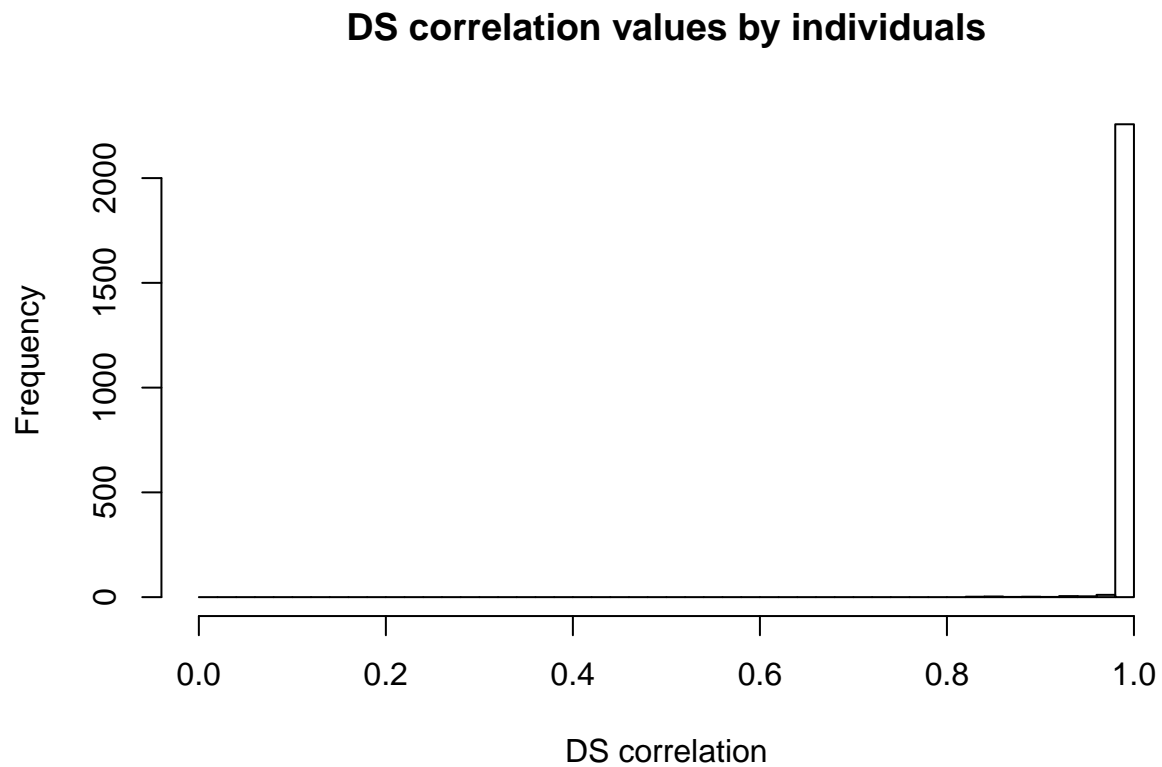
```
## [1] 0.9934383
```

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

```
## [1] 0.9741907
```

Histogram of the DS correlation values by individuals

```
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
```


R²

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"))
```
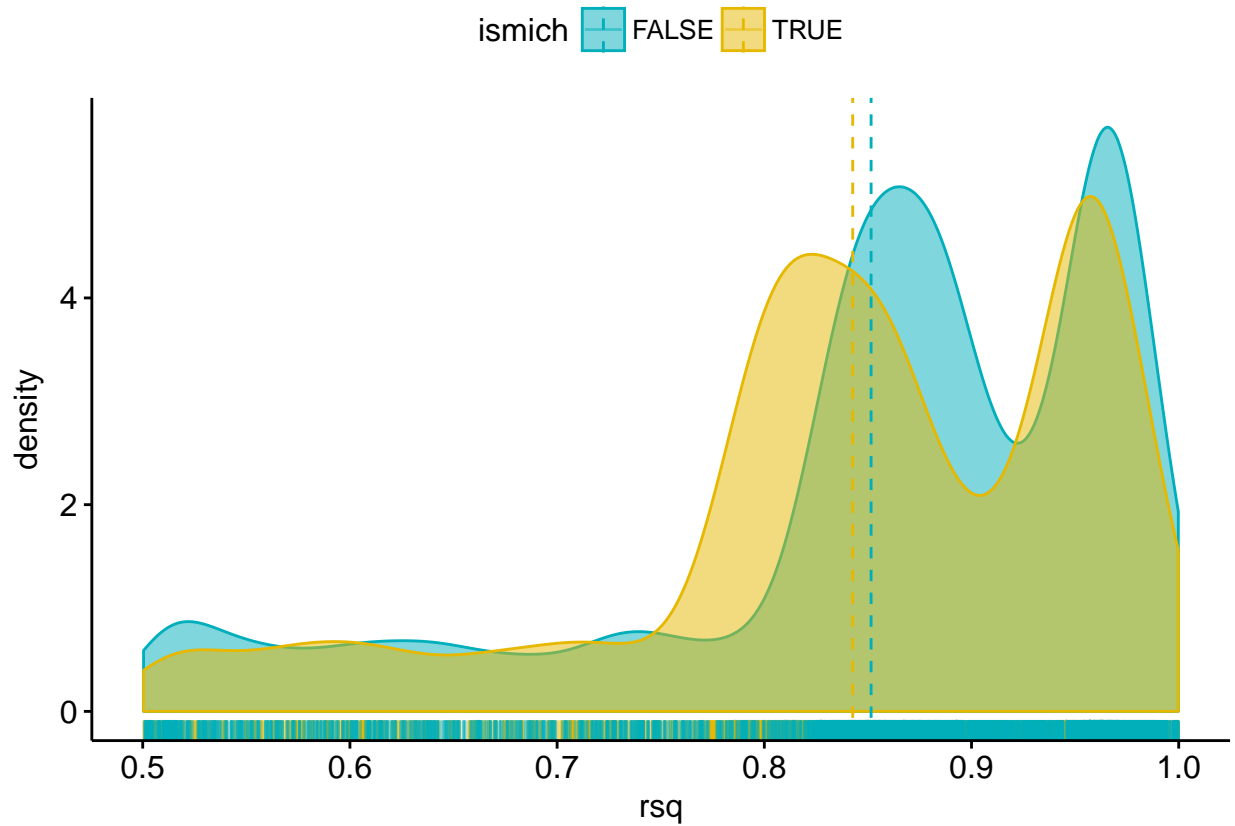


```
par(mfrow=c(1,2))
HIST_R2_MICHIGAN <- hist(info(michigan)$R2, breaks=seq(0.5, 1, by=0.02), main="Michigan", xlab="R2")
HIST_R2_MINIMAC <- hist(info(michigan)$R2, breaks=seq(0.5, 1, by=0.02), main="Minimac3", xlab="R2")
```

Michigan	**Minimac3**
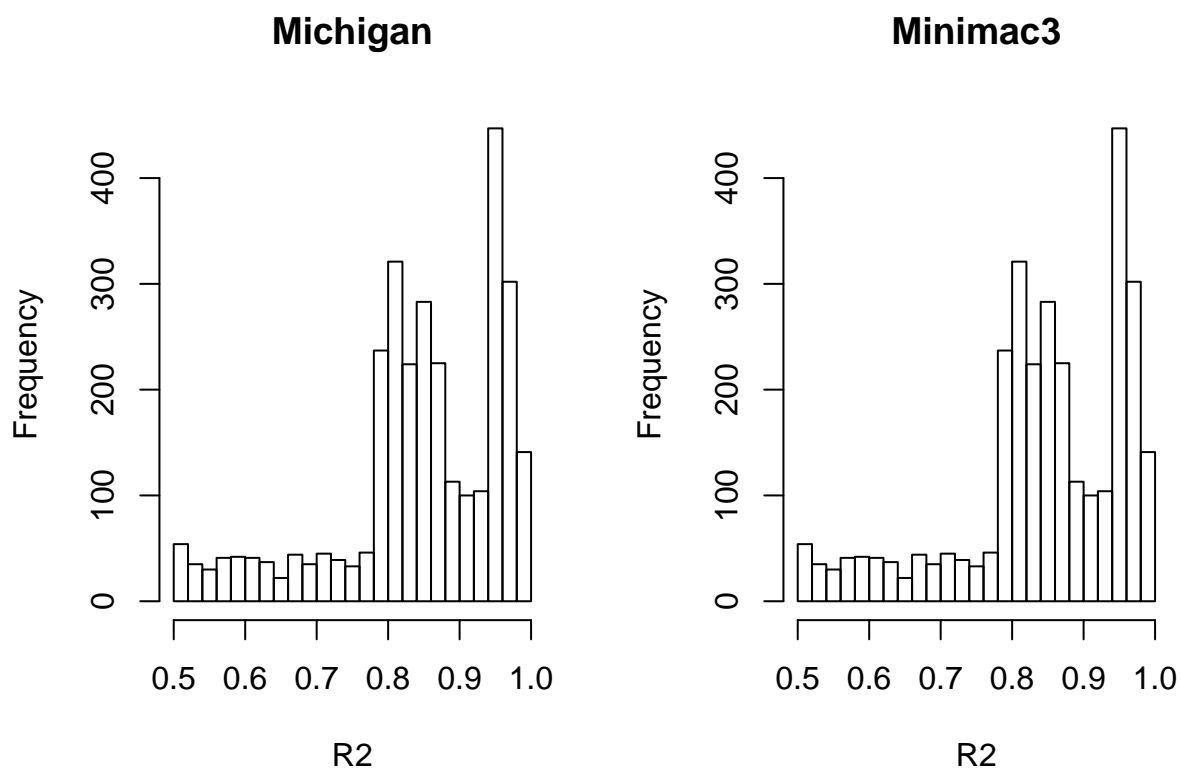
Genotype predictions

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

```r
min(perc_by_ind)
```

```
## [1] 0.8913759
```

```r
max(perc_by_ind)
```

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

```r
mean(perc_by_ind)
```

```
## [1] 0.9995284
```

```r
sum(perc_by_ind > 0.95)/length(perc_by_ind)
```
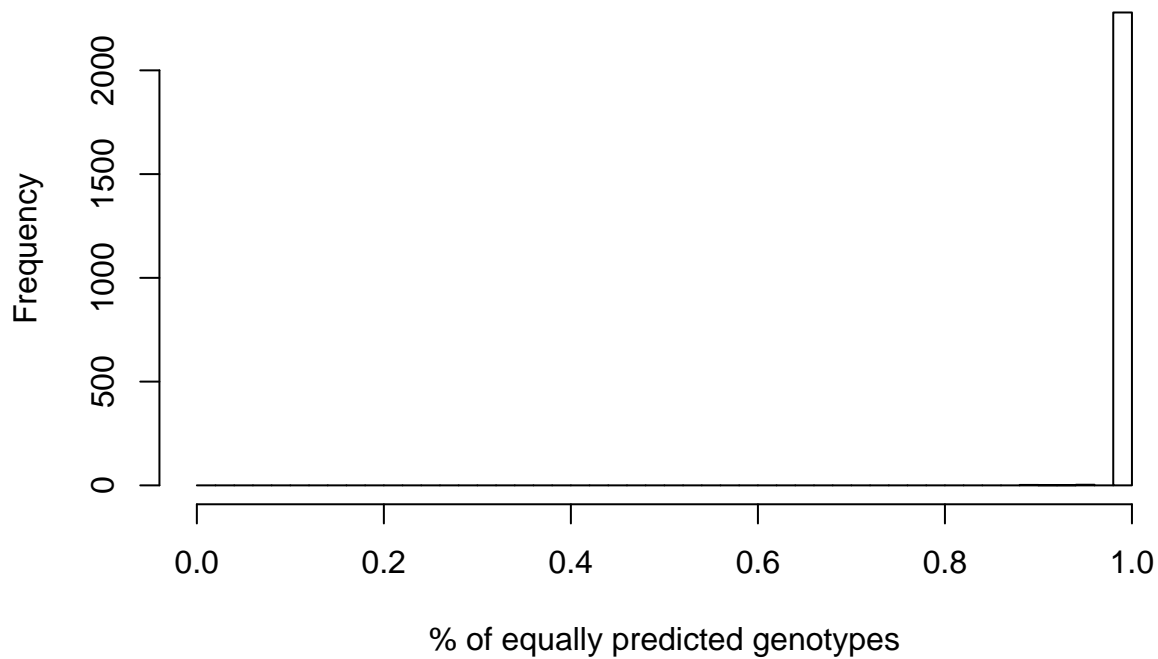
```
## [1] 0.9973753
```

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

[1] 0.9951881

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
```

SNPs (genotypes) equally predicted with Michigan and Minimac3

Inversion prediction

Predicted inversions with scoreInvHap

```
michigan_inv
```

```
## scoreInvHapRes
## Samples:  2286
## Genotypes' table:
##  NI/NI    NI/I    I/I
##  1407     760     119
## - Inversion genotypes' table:
##  NN    NI    II
##  1407     760     119
## - Inversion frequency: 21.83%
```

```
minimac_inv
```

```
## scoreInvHapRes
## Samples:  2286
## Genotypes' table:
##  NI/NI    NI/I     I/I
##  1407     760      119
## - Inversion genotypes' table:
##  NN    NI   II
##  1407     760      119
## - Inversion frequency: 21.83%
```

Comparison table

```
scoreinvhap_table
```
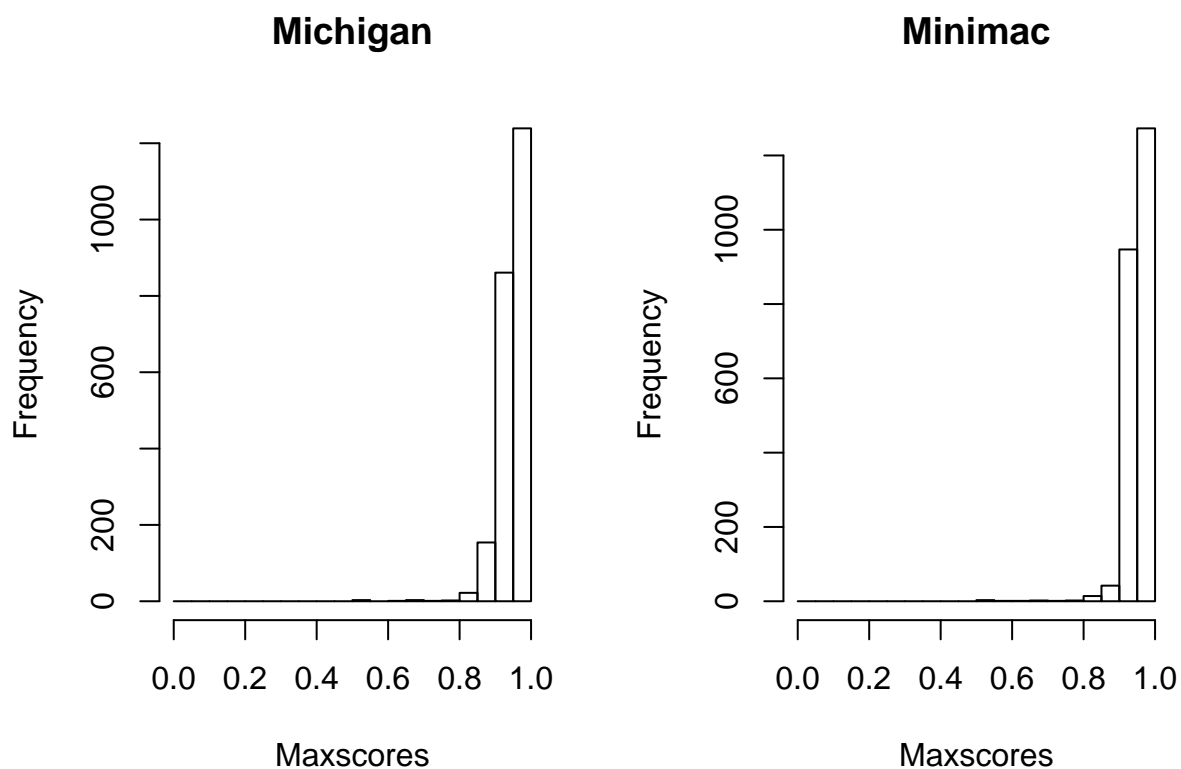
```
##         Minimac
## Michigan NI/NI NI/I  I/I
##    NI/NI  1407    0    0
##    NI/I      0  760    0
##    I/I       0    0  119
```

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

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

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")
```

Michigan   **Minimac**

Score correlation by individuals between both imputation methods

```r
min(score_corr)
```

```
## [1] 0.999816
```

```r
max(score_corr)
```
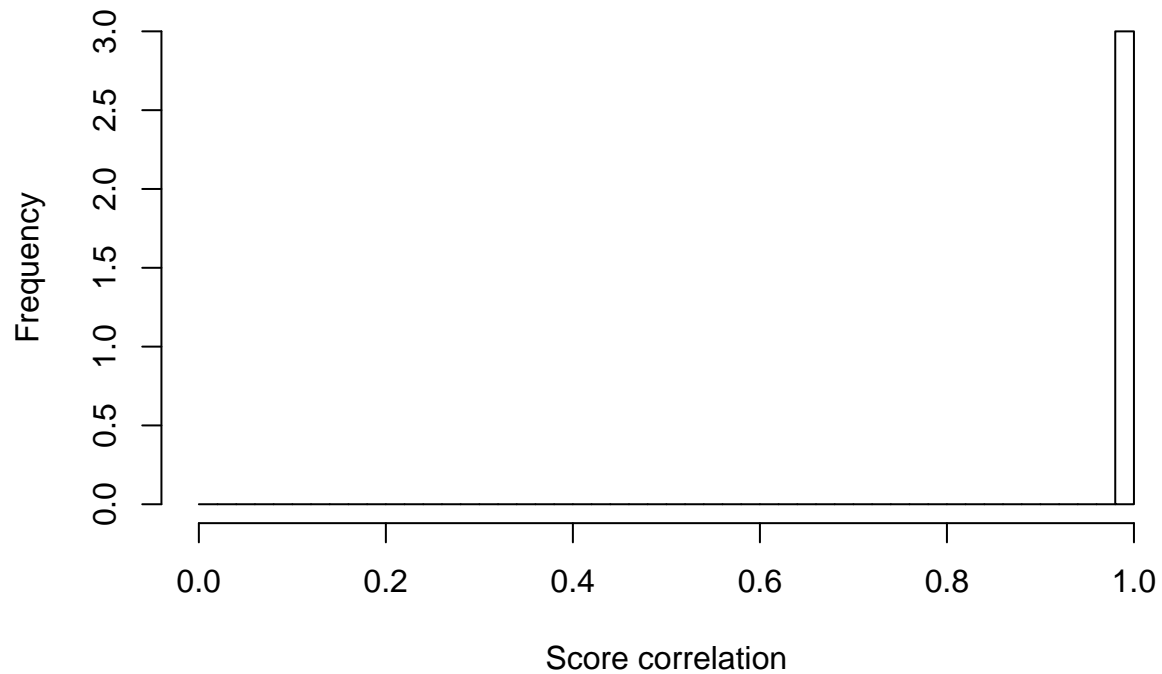
```
## [1] 0.9999092
```

```r
mean(score_corr)
```

```
## [1] 0.9998653
```

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

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")
```
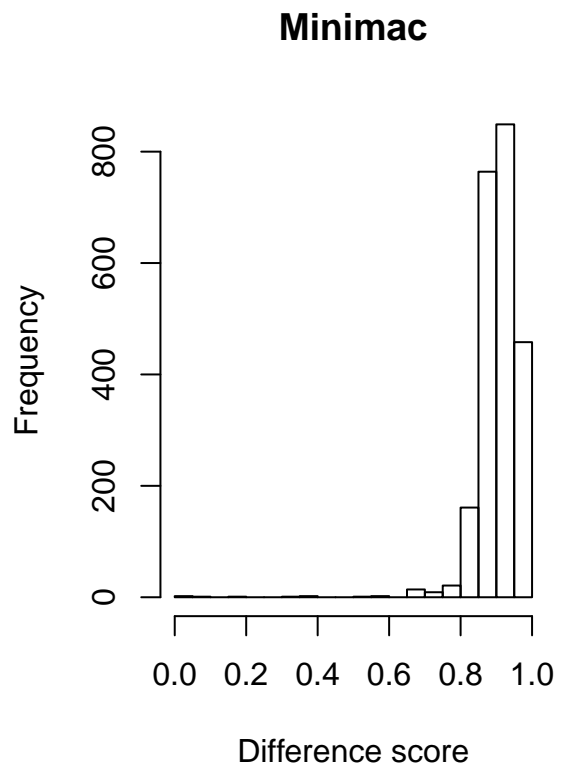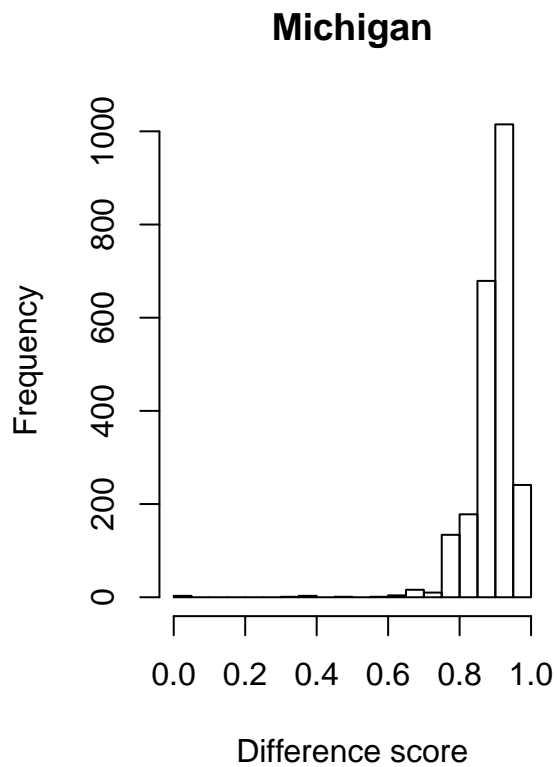
Michigan **Minimac**

Numbers of scores used

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

```
## [1] 2225
```

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

```
## [1] 2225
```

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

```
## [1] 2225
```

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

```
## [1] 2254
```

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

```
## [1] 2254
```

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

[1] 2254

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

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

[1] 2286

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

[1] 2283

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

[1] 0.9986877

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

[1] 2286

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

[1] 2283
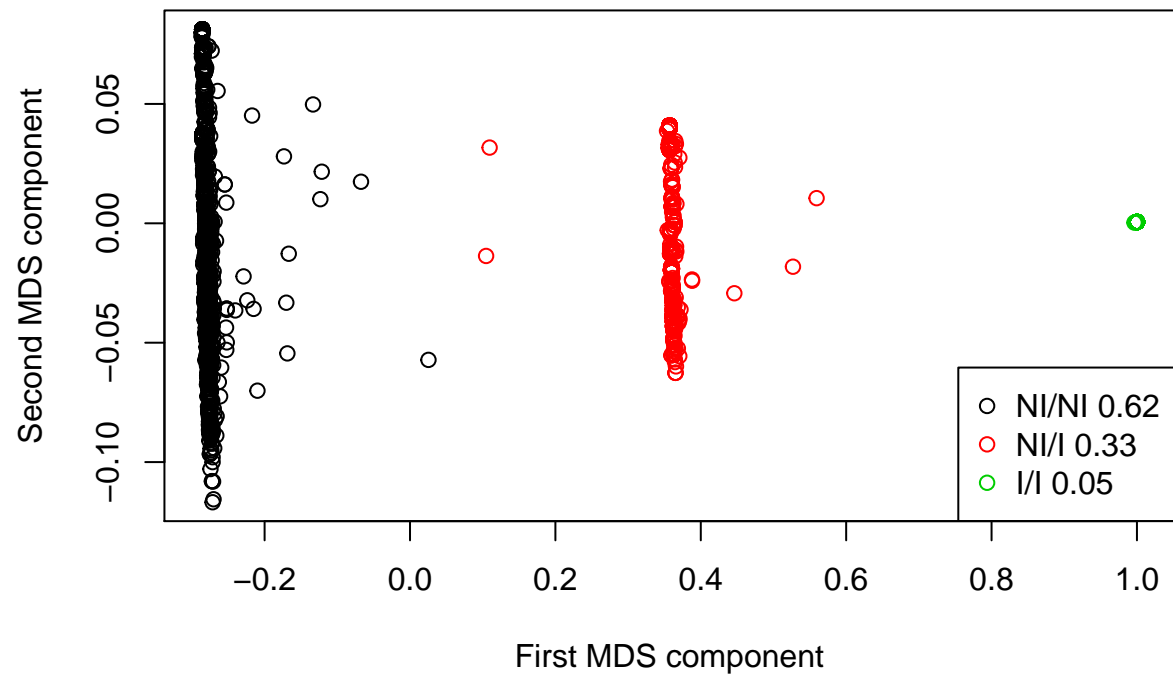
```r
length(classification(minimac_inv, minDiff = 0.1, callRate = 0.9))/length(classification(minimac_inv))
```

[1] 0.9986877

Plots with invClust

Michigan:

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

Minimac:

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