Comparison: Michigan Imputation Server / Shapeit+Minimac3 (Chromosome 8)

Ignacio Tolosana

Imputation

Imputed data exploration

```
michigan_chr8
```

```
## 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...
```

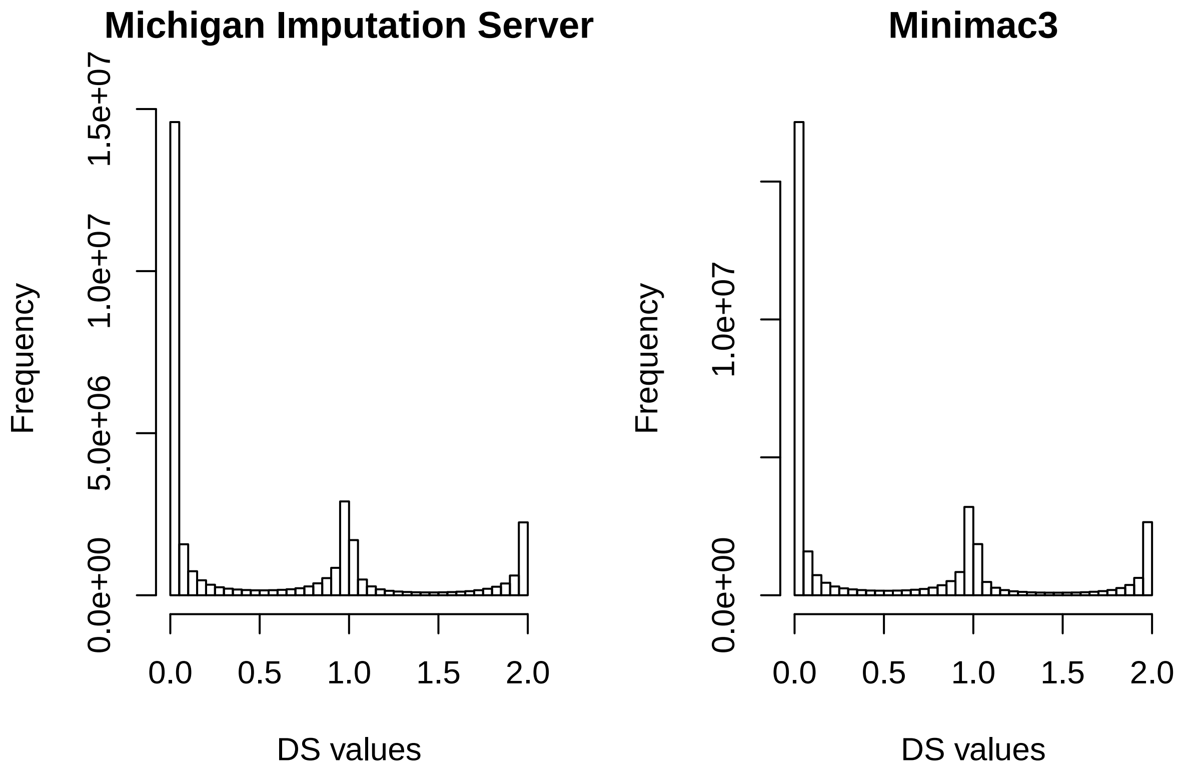
```
minimac_chr8
```

```
## 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="DS values")
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
```

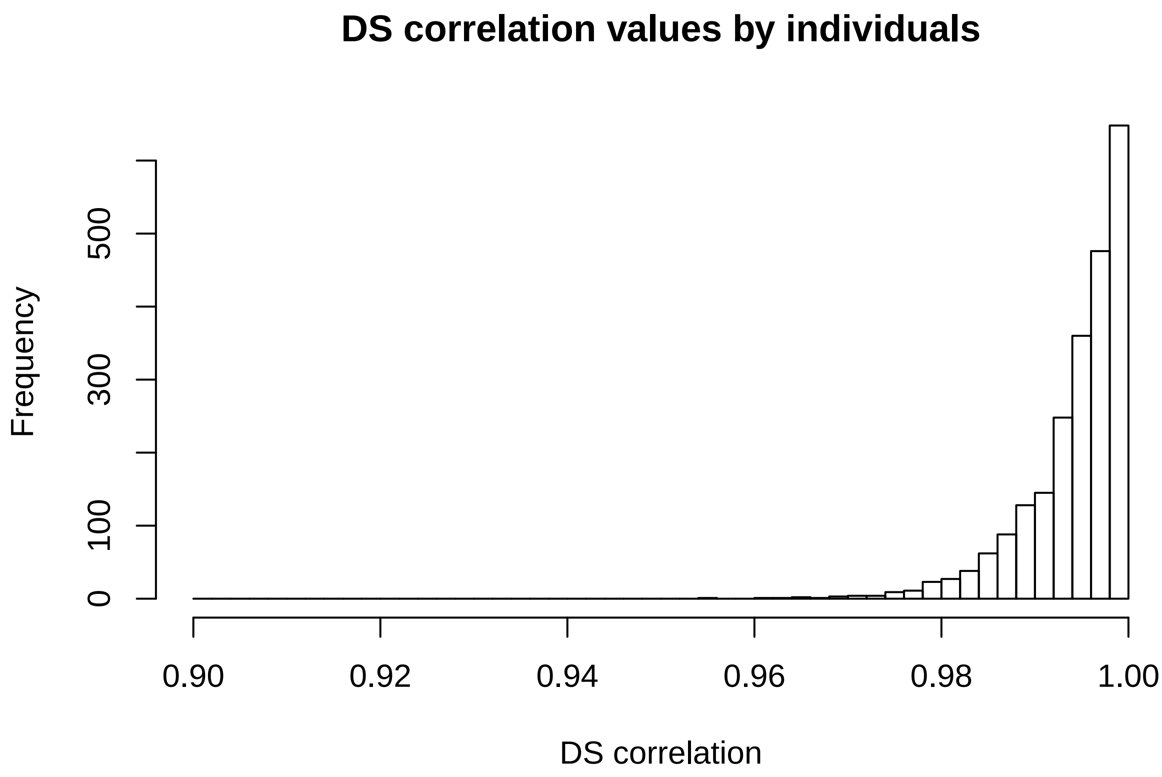
```r
mean(cor_by_ind > 0.95)
```

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

```r
mean(cor_by_ind > 0.99)
```

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

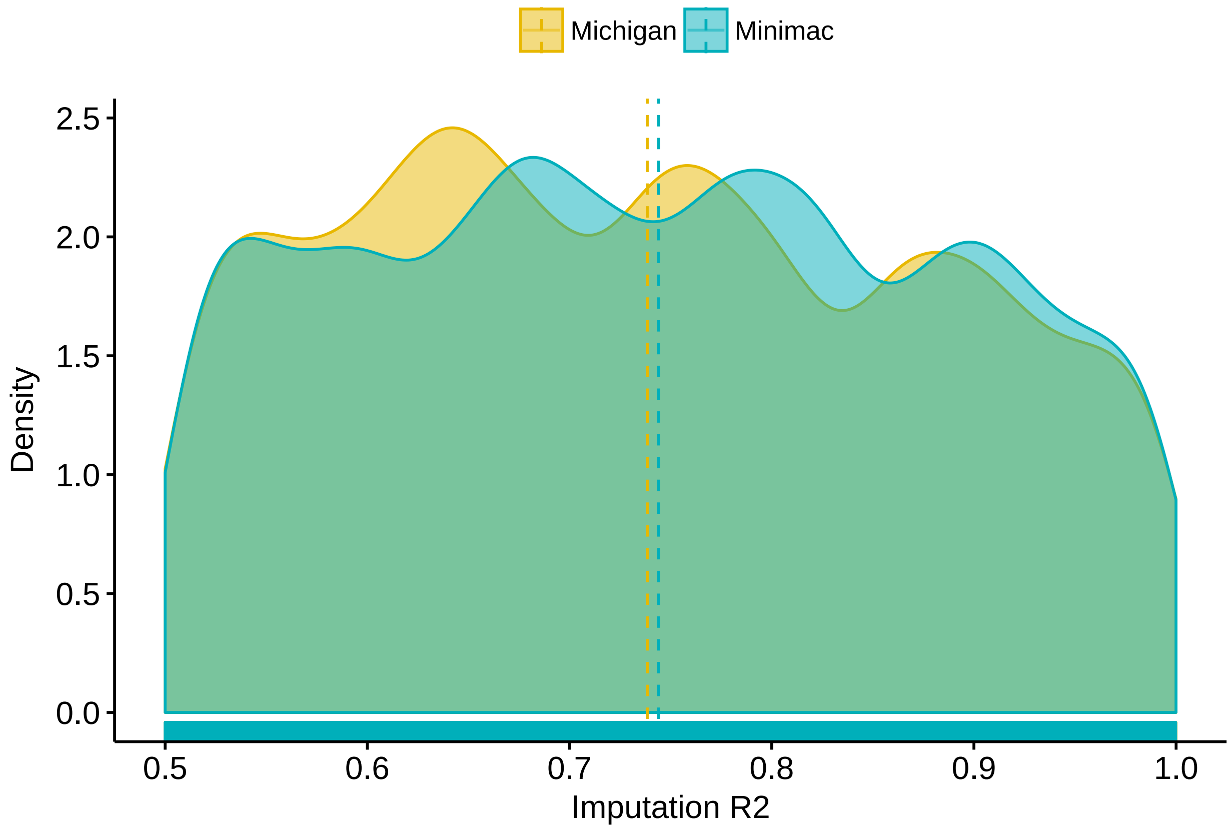
```r
# Histogram of the DS correlation values by individuals
par(mfrow=c(1,1))
CORR_HIST <- hist(cor_by_ind, breaks=seq(0.9, 1, by=0.002),
                  main="DS correlation values by individuals", xlab="DS correlation")
```



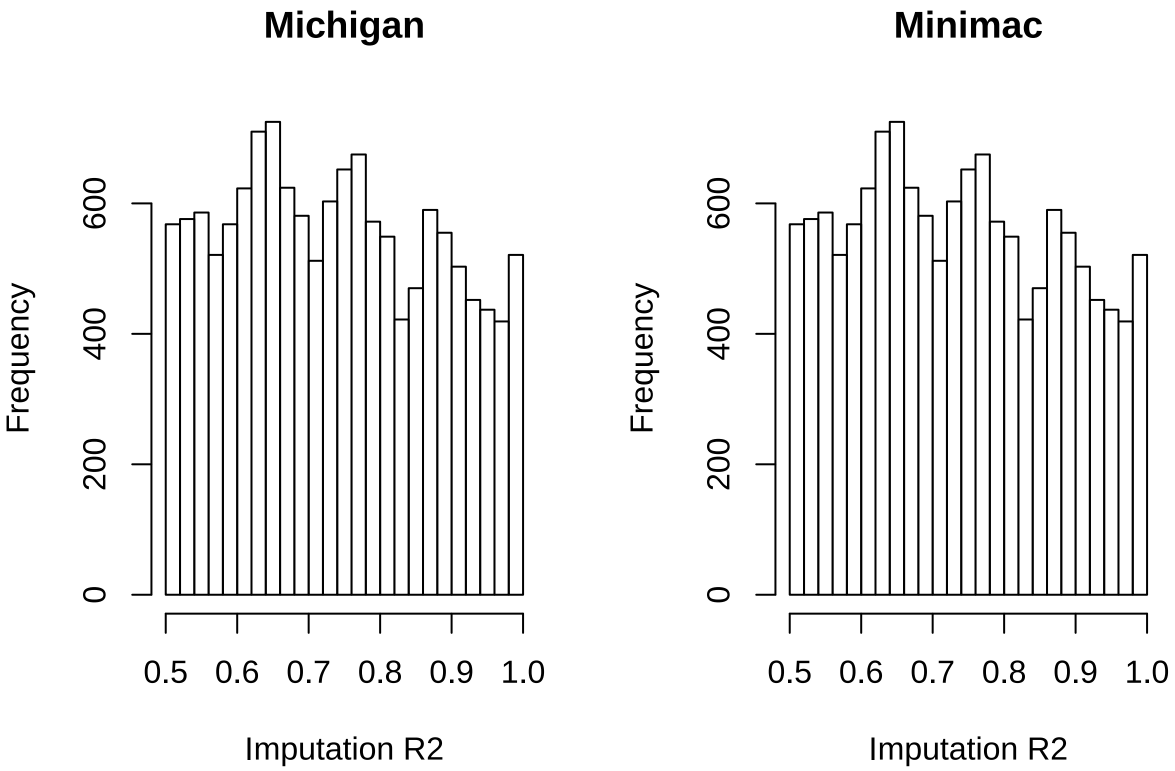
R^2

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)

```r
ggdensity(comparison, x = "rsq",
          add = "mean", rug = TRUE,
          color = "ismich", fill = "ismich",
          palette = c("#E8B800", "#00AFBB"),
          legend.title = c(""),
          xlab = ("Imputation R2"),
          ylab = ("Density"))
```



```r
par(mfrow=c(1,2))
HIST_R2_MICHIGAN <- hist(info(michigan_chr8)$R2, breaks=seq(0.5, 1, by=0.02),
                         main="Michigan", xlab="Imputation R2")
HIST_R2_MINIMAC <- hist(info(michigan_chr8)$R2, breaks=seq(0.5, 1, by=0.02),
                        main="Minimac", xlab="Imputation R2")
```



Genotype predictions

```
## non-single nucleotide variations are set to NA
## non-single nucleotide variations are set to NA
```

Compare the genotype predictions (BestGuess) 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
```

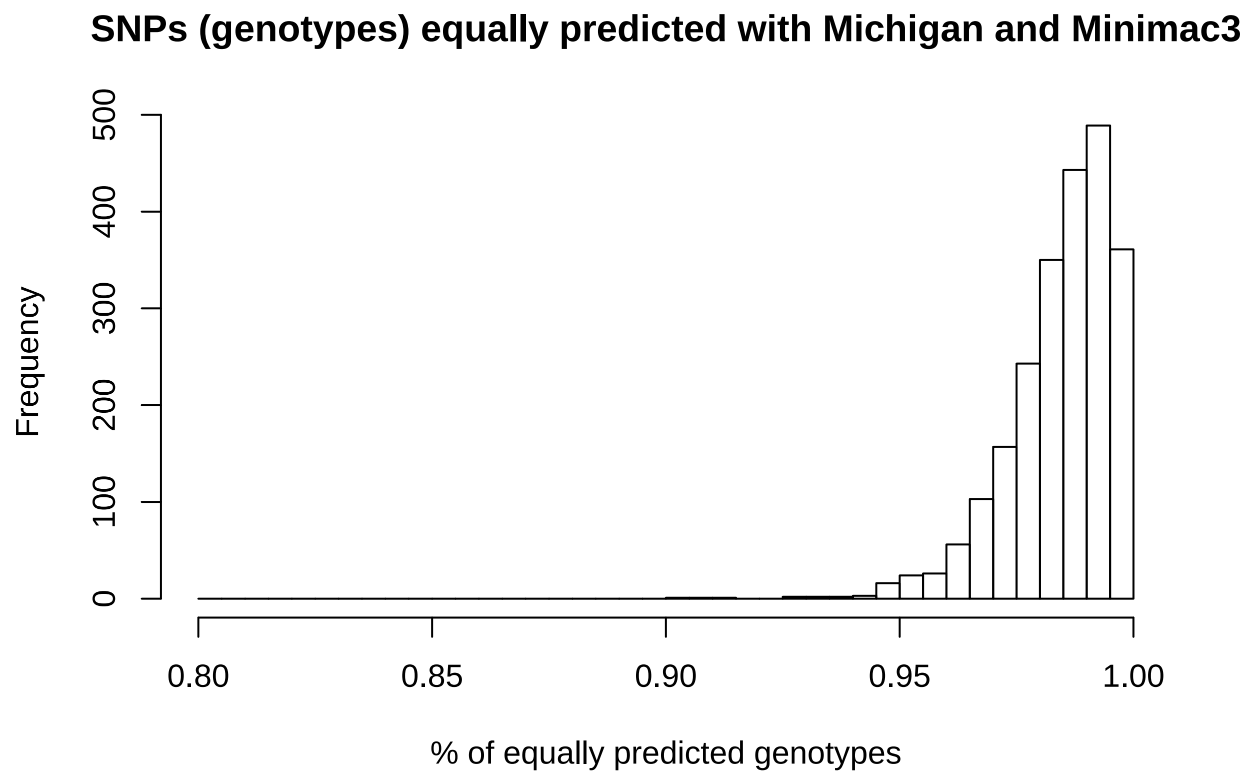
```
mean(perc_by_ind > 0.95)
```

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

```
mean(perc_by_ind > 0.99)
```

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

```
# Plot histogram of the genotypes equally predicted (BestGuess) by individuals
# in both methods
par(mfrow=c(1,1))
GENO_HIST <- hist(perc_by_ind, breaks=seq(0.8, 1, by=0.005),
                  main="SNPs (genotypes) equally predicted with Michigan and Minimac3",
                  xlab="% of equally predicted genotypes")
```



Inversion prediction

Predicted inversions with scoreInvHap

```
michigan_inv_chr8
```

```
## 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_chr8
```

```
## 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%
```

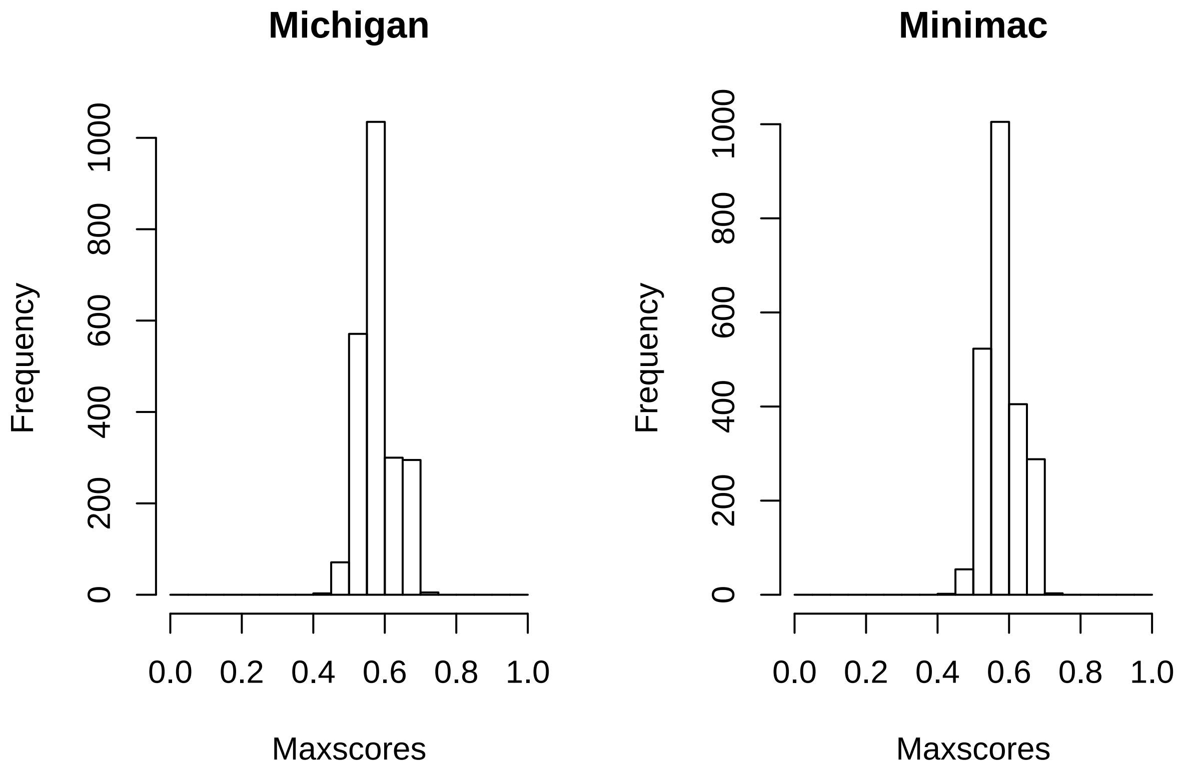
```
# Comparison table
scoreinvhap_table
```

```
##         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_chr8), breaks=seq(0, 1, by=0.05), main="Michigan", xlab="Maxscores")
hist(maxscores(minimac_inv_chr8), 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
```

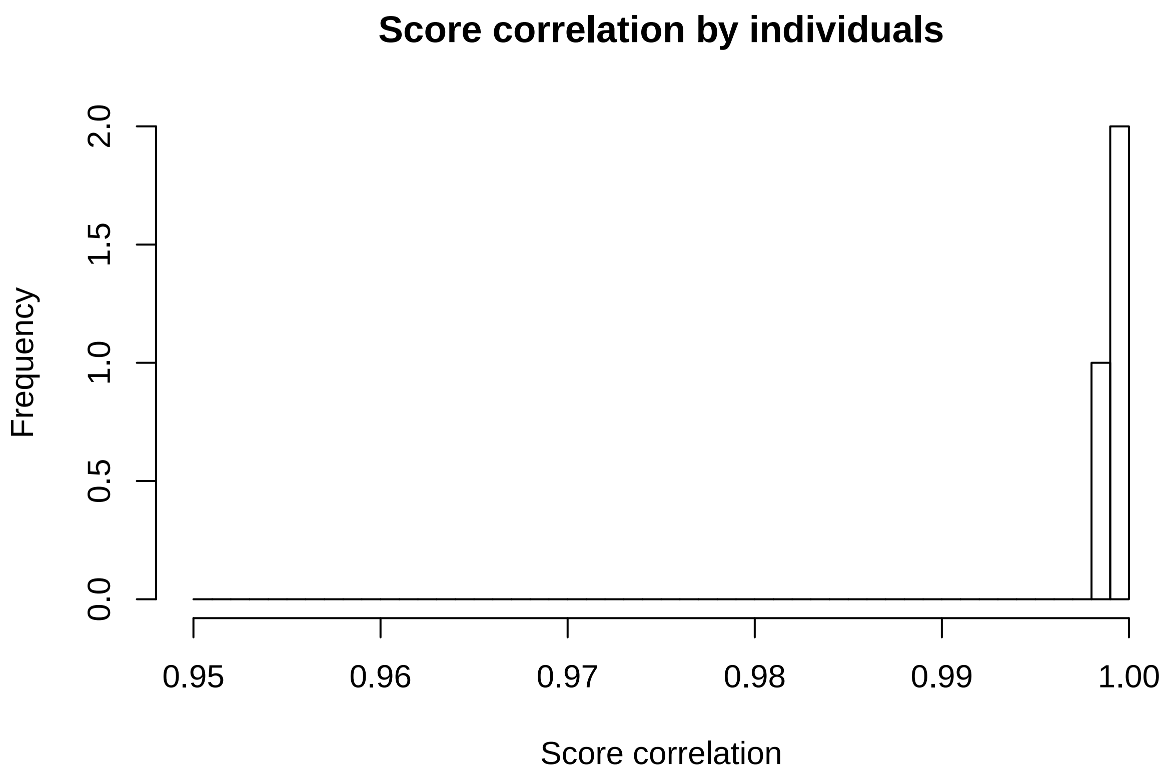
```
max(score_corr)
```

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

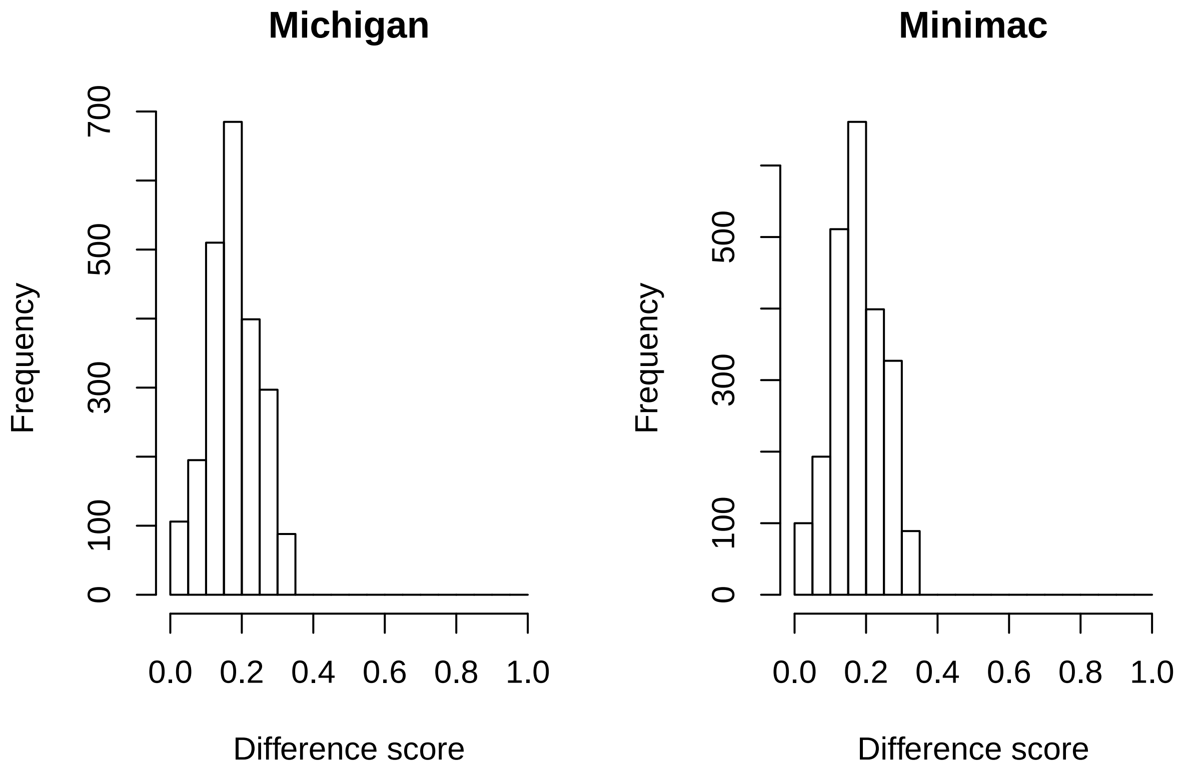
```
mean(score_corr)
```

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

```
par(mfrow=c(1,1))
SCORE_CORR_HIST <- hist(score_corr, breaks=seq(0.95, 1, by=0.001),
                        main="Score correlation by individuals",
                        xlab="Score correlation")
```



```
# Difference score between the highest similarity score and the second highest,
# in both imputation methods
par(mfrow=c(1,2))
hist(diffscores(michigan_inv_chr8), breaks=seq(0, 1, by=0.05),
     main="Michigan", xlab="Difference score")
hist(diffscores(minimac_inv_chr8), breaks=seq(0, 1, by=0.05),
     main="Minimac", xlab="Difference score")
```



```
# Numbers of scores used
mean(numSNPs(michigan_inv_chr8))
```

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

```
mean(numSNPs(minimac_inv_chr8))
```

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

```
# Number of samples in both imputation methods before and after QC filtering
length(classification(michigan_inv_chr8))
```

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

```
length(classification(michigan_inv_chr8, minDiff = 0.1, callRate = 0.9))
```

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

```
length(classification(michigan_inv_chr8, minDiff = 0.1, callRate = 0.9))/
  length(classification(michigan_inv_chr8))
```

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

```r
length(classification(minimac_inv_chr8))
```

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

```r
length(classification(minimac_inv_chr8, minDiff = 0.1, callRate = 0.9))
```

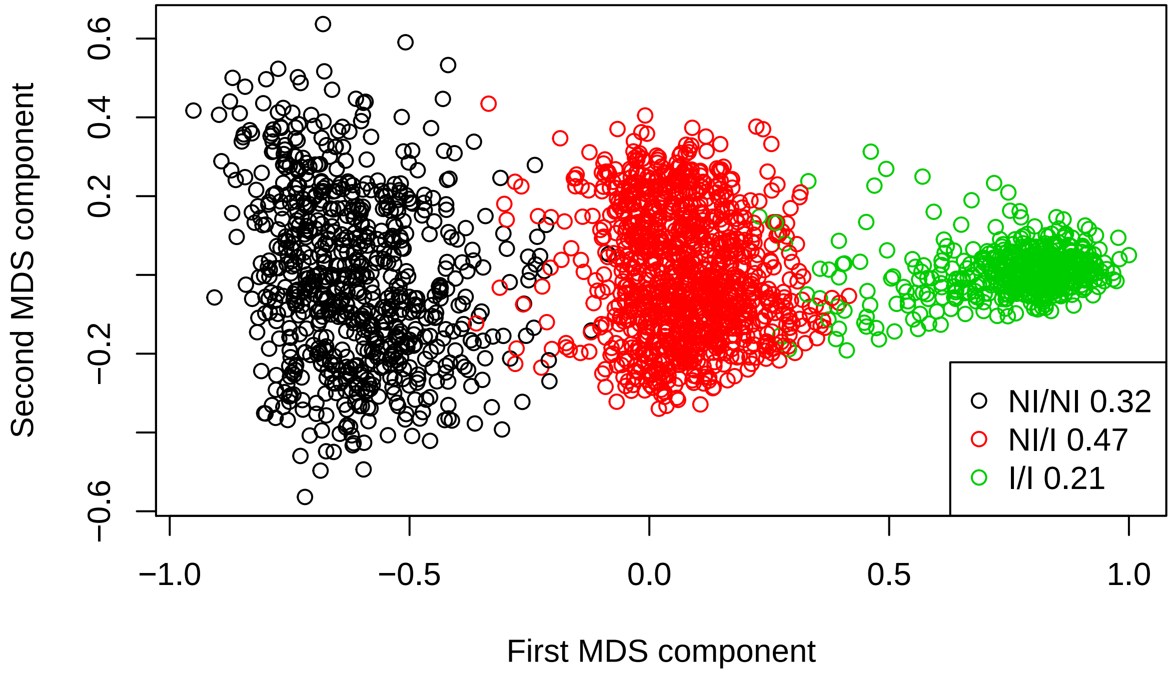
```
## [1] 1987
```

```r
length(classification(minimac_inv_chr8, minDiff = 0.1, callRate = 0.9))/
  length(classification(minimac_inv_chr8))
```

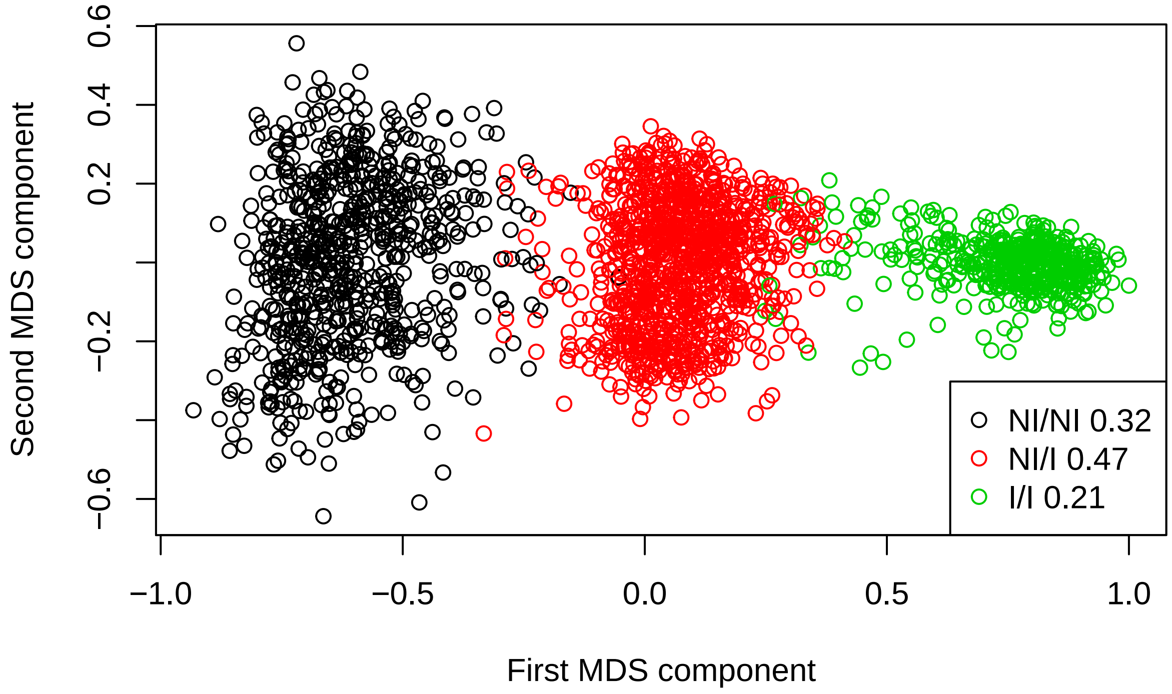
```
## [1] 0.8714912
```

Plots with invClust

```r
# Michigan
par(mfrow=c(1,1))
plotInv(michigan_invclust_chr8, classification = classification(michigan_inv_chr8))
```



```
# Minimac
plotInv(minimac_invclust_chr8, classification = classification(minimac_inv_chr8))
```



No filtered imputed data

```
nofilter_minimac_8
```

```
## class: CollapsedVCF
## dim: 98521 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)):
```

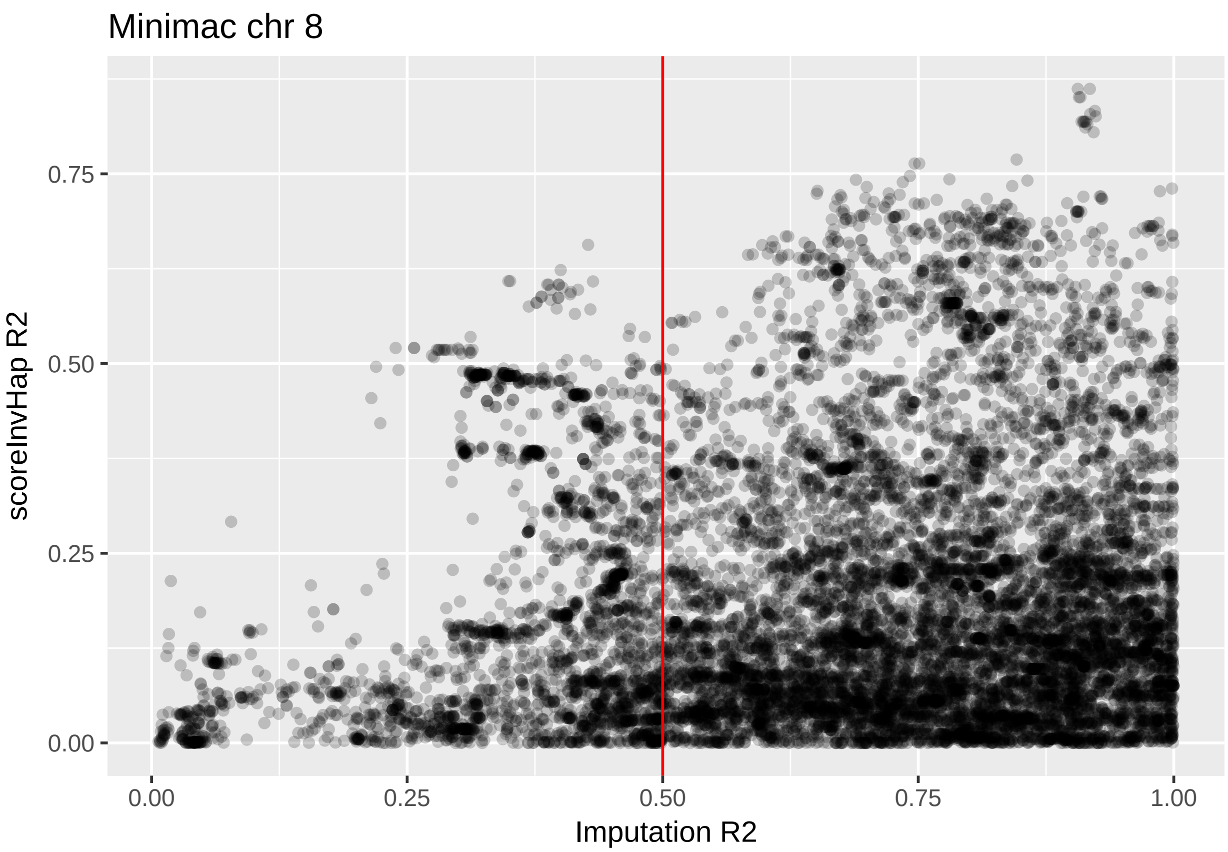
```
##       Number Type   Description
##    GT 1      String Genotype
##    DS 1      Float  Estimated Alternate Allele Dosage : [P(0/1)+2*P(1/1)]
##    GP 3      Float  Estimated Posterior Probabilities for Genotypes 0/0...
```

```
nofilter_minimac_inv_8
```

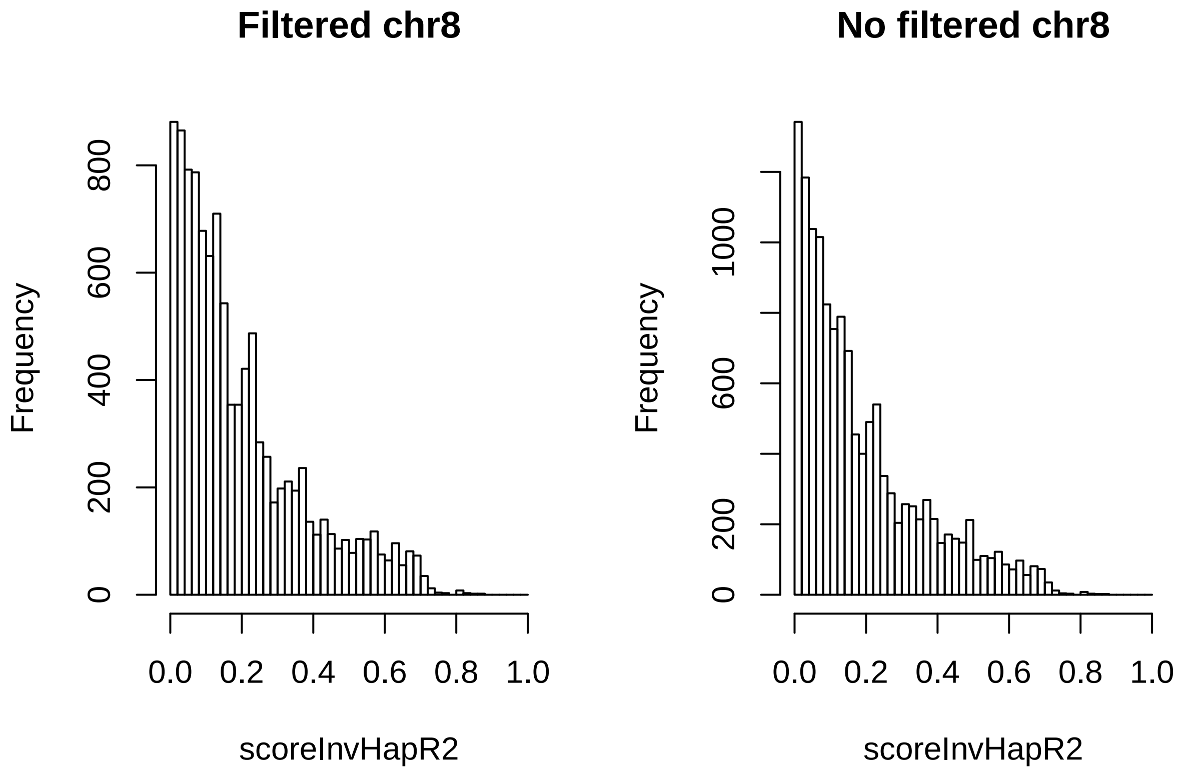
```
## scoreInvHapRes
## Samples:  2280
## Genotypes' table:
##  NI/NI    NI/I    I/I
##  732  1065    483
## - Inversion genotypes' table:
##  NN   NI  II
##  732  1065    483
## - Inversion frequency: 44.54%
```

```
# Select SNPs in both elements to represent them in the plot
snps_minimac_8 <- intersect(rownames(info(nofilter_minimac_8)), names(SNPsR2$inv8p23.1))

# Plot Imputation R2 vs scoreInvHap R2 (red line = filter in the previous data)
ggplot() +
  geom_point(aes(x = info(nofilter_minimac_8)[snps_minimac_8,]$R2,
                 y = SNPsR2$inv8p23.1[snps_minimac_8]),
                alpha = 0.2) +
  geom_vline(aes(xintercept=0.5), colour="red") +
  ggtitle("Minimac chr 8") +
  xlab("Imputation R2") +
  ylab("scoreInvHap R2")
```



```
# Histograms of scoreInvHap R2 in the filtered imputed data and in the NO filtered imputed data
par(mfrow=c(1,2))
hist(SNPsR2$inv8p23.1[rownames(minimac_chr8)], breaks=seq(0, 1, by=0.02),
     main="Filtered chr8", xlab="scoreInvHapR2")
hist(SNPsR2$inv8p23.1[rownames(nofilter_minimac_8)], breaks=seq(0, 1, by=0.02),
     main="No filtered chr8", xlab="scoreInvHapR2")
```



```
#Correlation between Imputation R2 and scoreInvHap R2 (NO filtered data)
cor(info(nofilter_minimac_8)[snps_minimac_8,]$R2, SNPsR2$inv8p23.1[snps_minimac_8])
```

```
## [1] 0.1257115
```

```
# Comparison table scoreInvHap with filtered and no filtered data
scoreinvhap_table_filt
```

```
##            Filtered
## No_filtered NI/NI NI/I  I/I
##       NI/NI   730    2    0
##       NI/I      1 1061    3
##       I/I       0    0  483
```

```
sum(diag(scoreinvhap_table_filt))/sum(scoreinvhap_table_filt)
```

```
## [1] 0.9973684
```