Comparison: Michigan Imputation Server / Shapeit+Minimac3 (Chromosome 7)

Ignacio Tolosana

Imputation

Imputed data exploration

```
michigan_chr7
```

```
## 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...
```

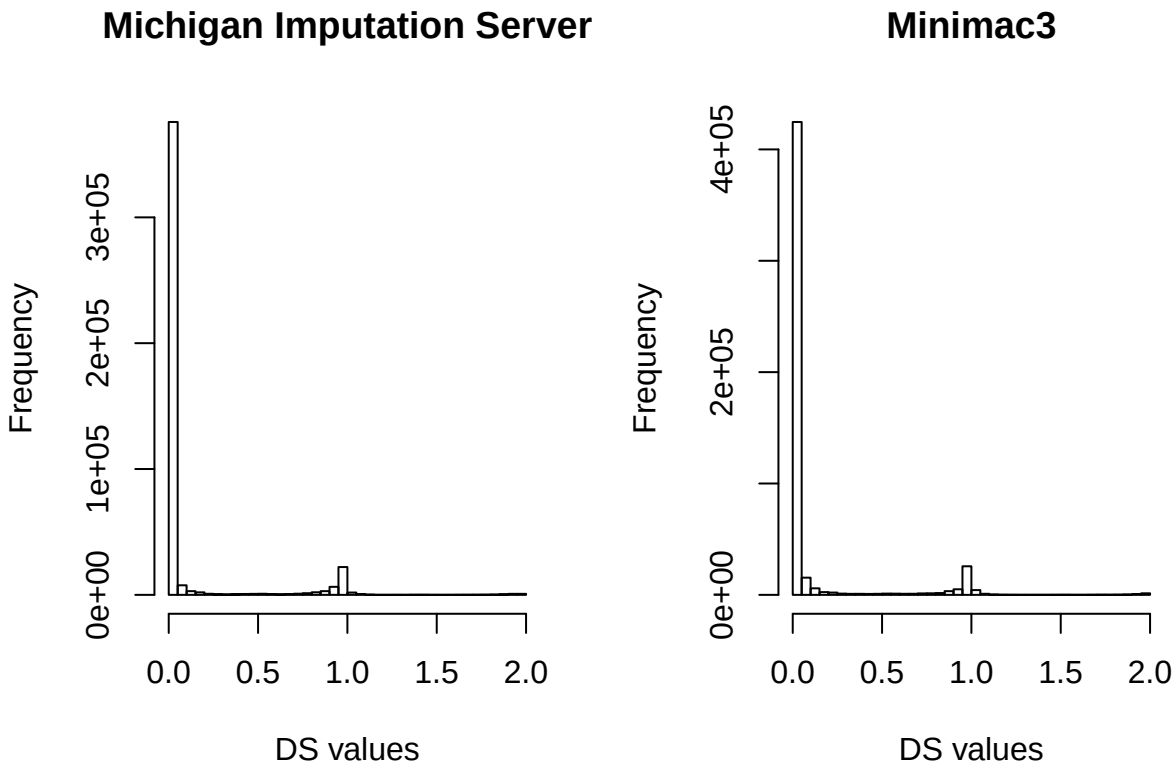
```
minimac_chr7
```

```
## 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="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] 0.4930321
```

```
max(cor_by_ind)
```

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

```r
mean(cor_by_ind)
```

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

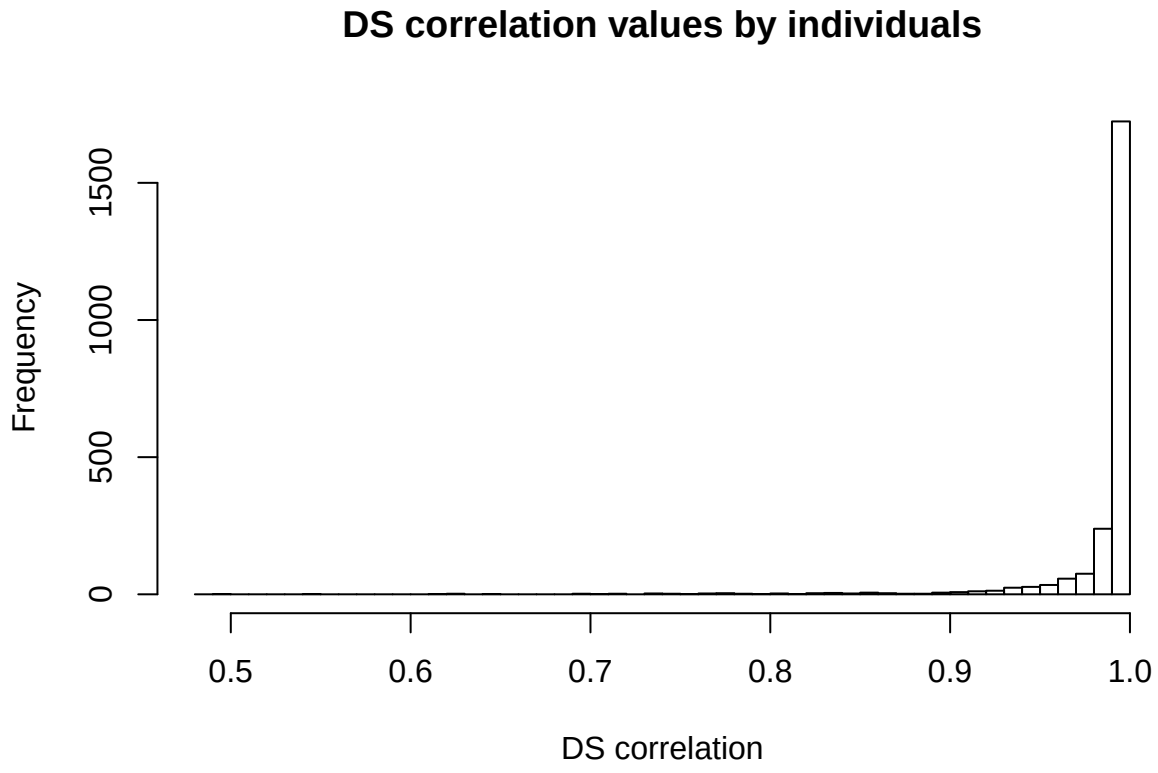
```r
mean(cor_by_ind > 0.95)
```

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

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

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

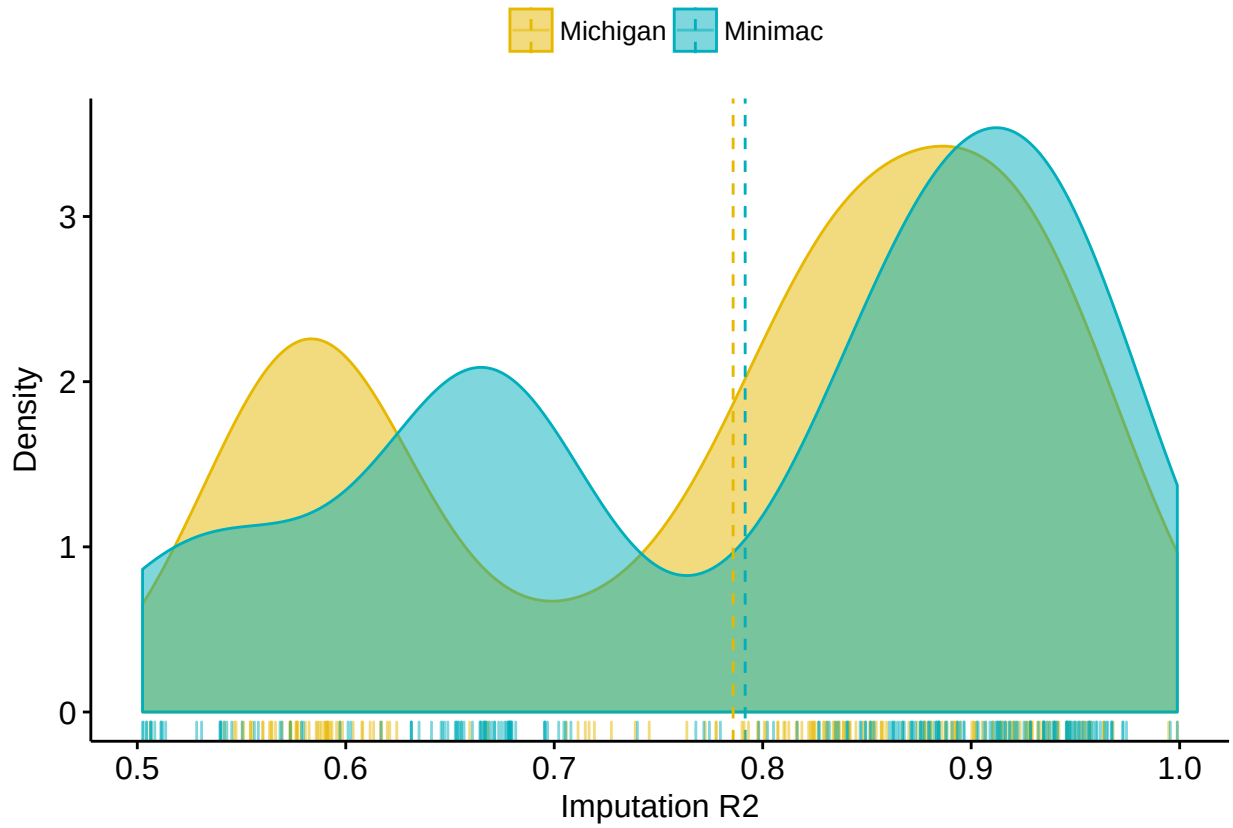
```r
# Histogram of the DS correlation values by individuals
par(mfrow=c(1,1))
CORR_HIST <- hist(cor_by_ind, breaks=seq(0.48, 1, by=0.01),
                  main="DS correlation values by individuals", xlab="DS correlation")
```



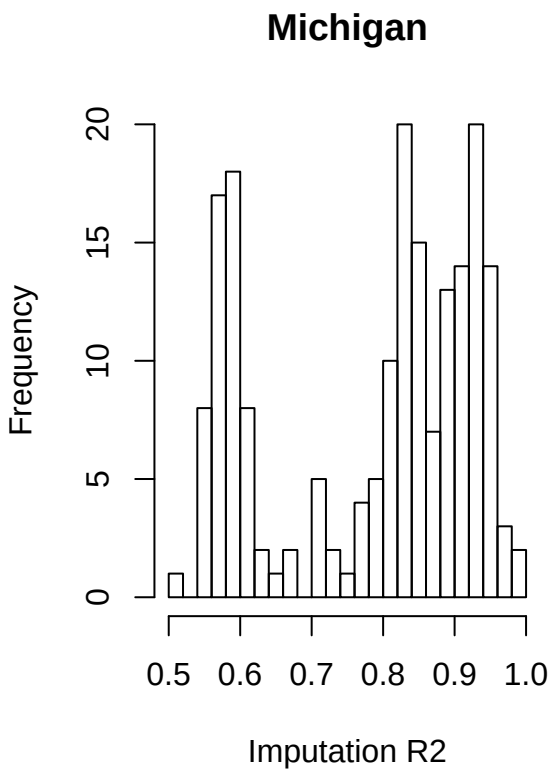
R^2

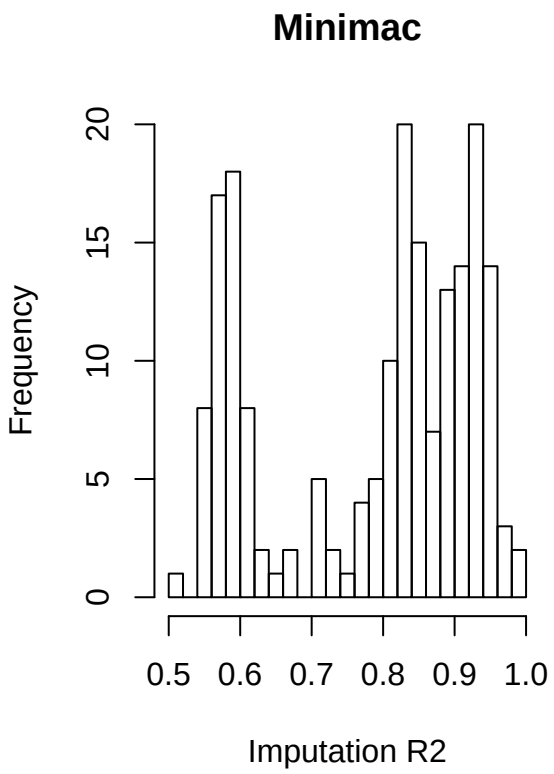
Density and histogram plots comparing the RS2 values in both methods ("ismich = TRUE" indicates the values for the Michigan imputation, whereas "ismich = FALSE" shows the values for the Minimac imputation)

```
ggdensity(comparison, x = "rsq",
          add = "mean", rug = TRUE,
          color = "ismich", fill = "ismich",
          palette = c("#E7B800", "#00AFBB"),
          legend.title = c(""),
          xlab = ("Imputation R2"),
          ylab = ("Density"))
```




```
par(mfrow=c(1,2))
HIST_R2_MICHIGAN <- hist(info(michigan_chr7)$R2, breaks=seq(0.5, 1, by=0.02),
                         main="Michigan", xlab="Imputation R2")
HIST_R2_MINIMAC <- hist(info(michigan_chr7)$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] 0.7864583
```

```
max(perc_by_ind)
```

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

```
mean(perc_by_ind)
```

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

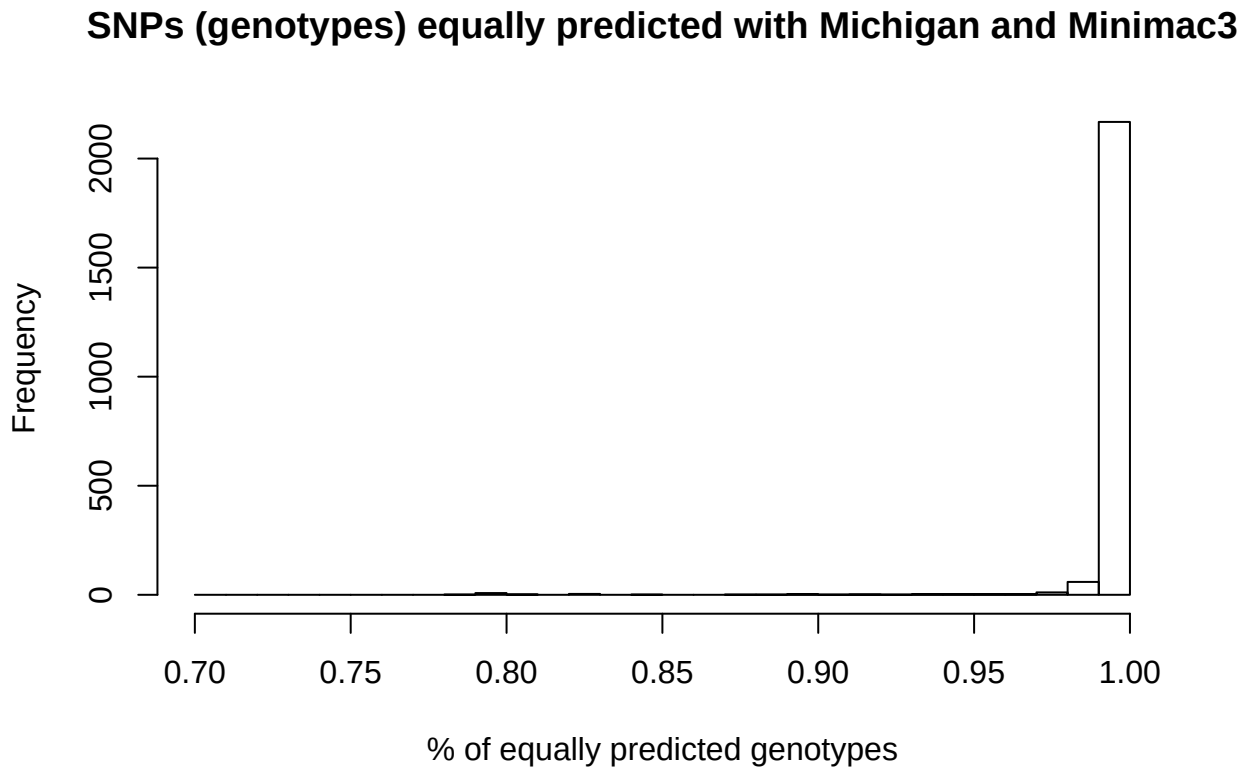
```r
mean(perc_by_ind > 0.95)
```

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

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

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

```r
# 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(0.7, 1, by=0.01),
                  main="SNPs (genotypes) equally predicted with Michigan and Minimac3",
                  xlab="% of equally predicted genotypes")
```



Inversion prediction

Predicted inversions with scoreInvHap

```r
michigan_inv_chr7
```

```
## 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_chr7
```

```
## 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%
```

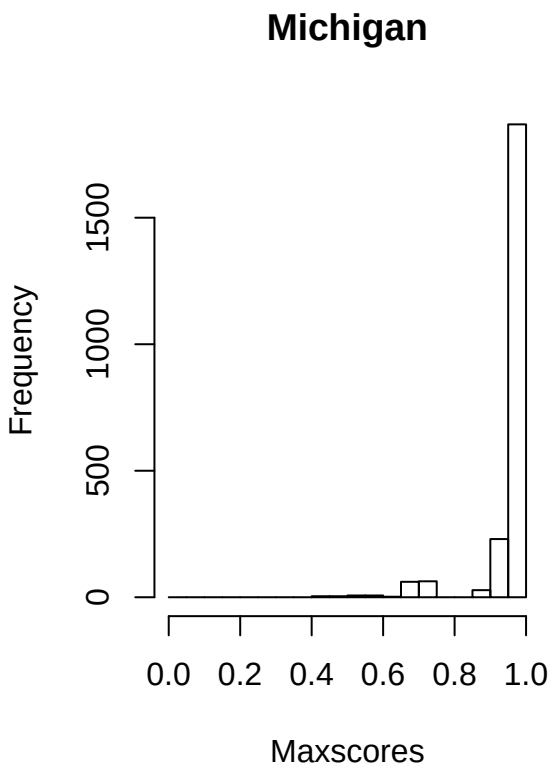
```
# Comparison table
scoreinvhap_table
```

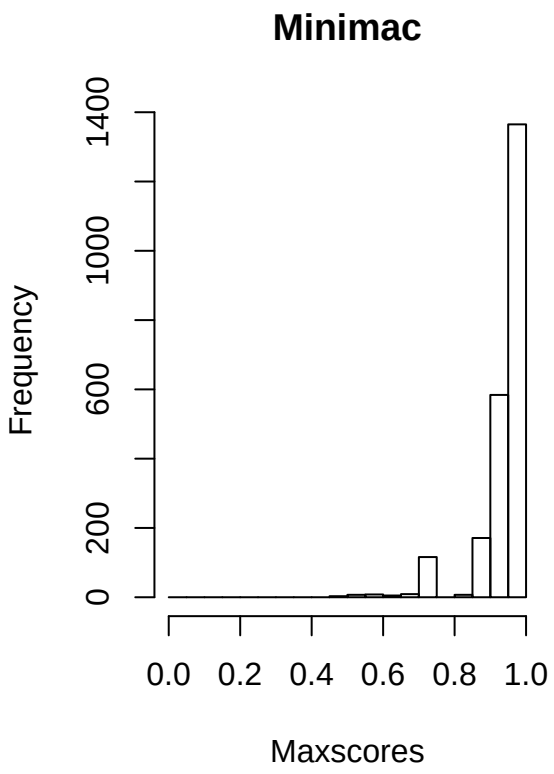
```
##         Minimac
## Michigan IaIa IaIb IbIb NaIa NaIb NaNa NaNb NbIa NbIb NbNb
##     IaIa   89    0    0    0    0    0    0    0    0    0
##     IaIb    0    4    0    1    0    0    0    0    0    0
##     IbIb    0    0  344    0    0   23    0    0    0    0
##     NaIa    0   66    1  564    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_chr7), breaks=seq(0, 1, by=0.05), main="Michigan", xlab="Maxscores")
hist(maxscores(minimac_inv_chr7), 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
```

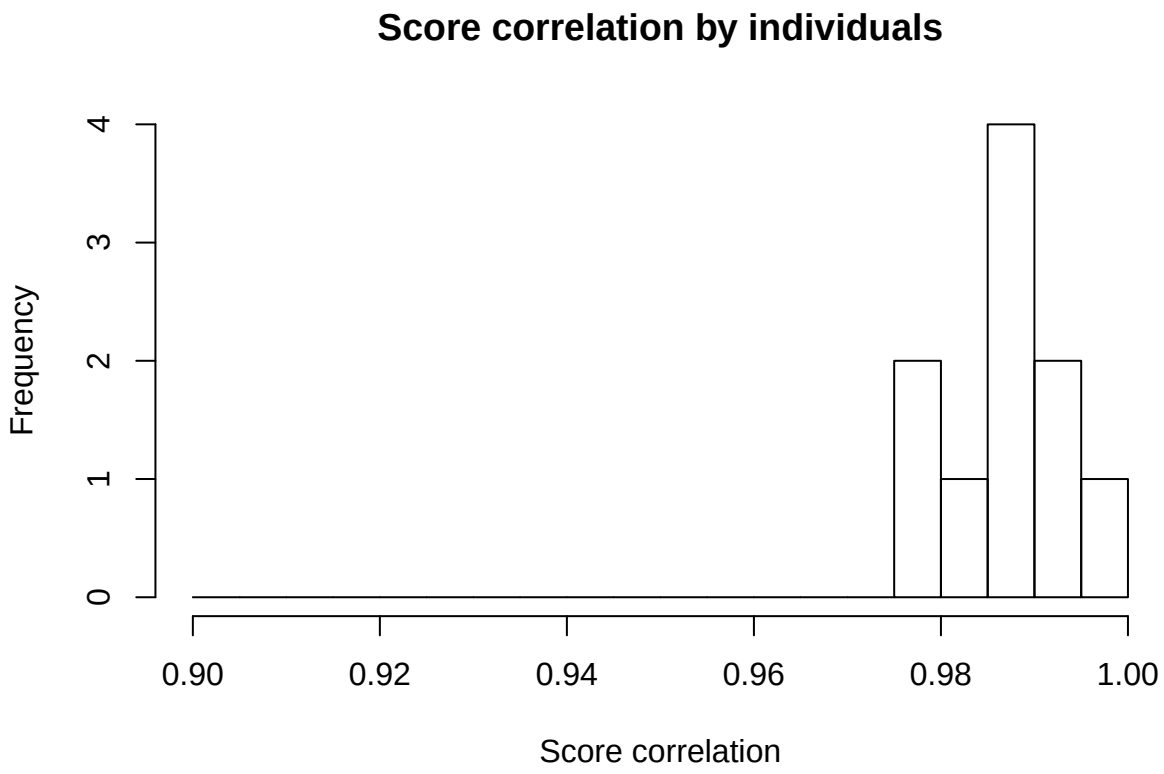
```
max(score_corr)
```

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

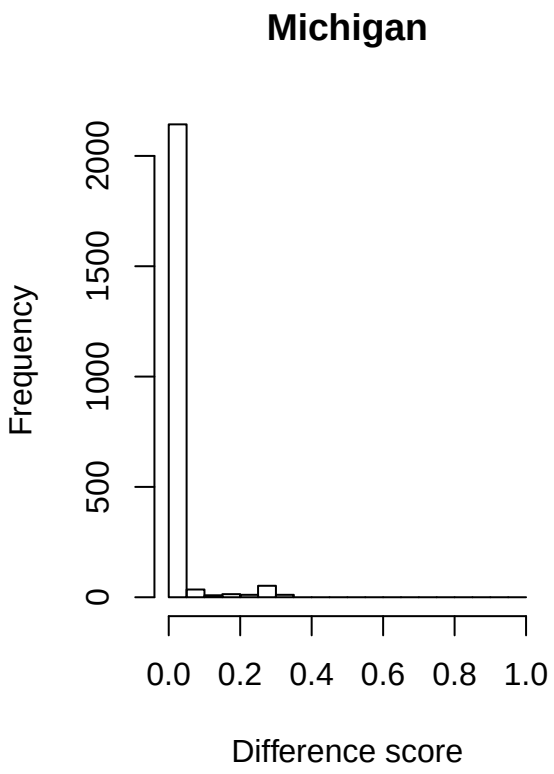
```
mean(score_corr)
```

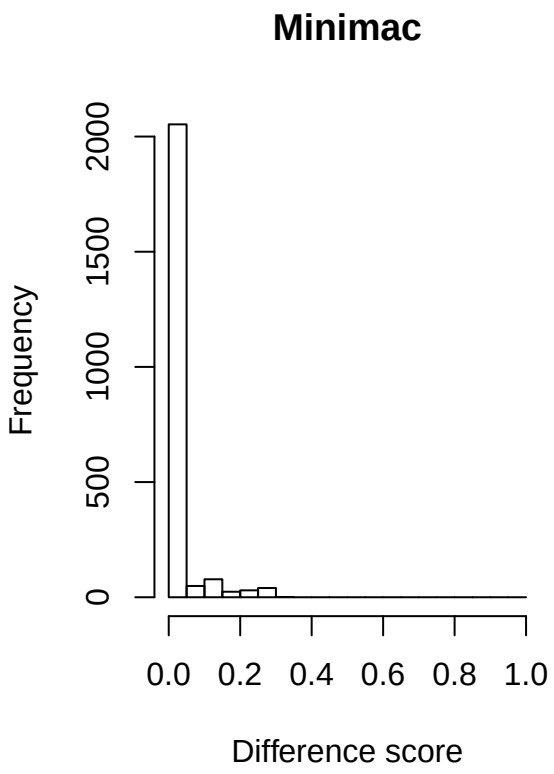
```
## [1] 0.9866348
```

```
par(mfrow=c(1,1))
SCORE_CORR_HIST <- hist(score_corr, breaks=seq(0.9, 1, by=0.005),
                        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_chr7), breaks=seq(0, 1, by=0.05),
     main="Michigan", xlab="Difference score")
hist(diffscores(minimac_inv_chr7), breaks=seq(0, 1, by=0.05),
     main="Minimac", xlab="Difference score")
```






```
# Numbers of scores used
mean(numSNPs(michigan_inv_chr7))
```

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

```
mean(numSNPs(minimac_inv_chr7))
```

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

```
# Number of samples in both imputation methods before and after QC filtering
length(classification(michigan_inv_chr7))
```

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

```
length(classification(michigan_inv_chr7, minDiff = 0.1, callRate = 0.9))
```

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

```
length(classification(michigan_inv_chr7, minDiff = 0.1, callRate = 0.9))/
  length(classification(michigan_inv_chr7))
```

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

```
length(classification(minimac_inv_chr7))
```

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

```
length(classification(minimac_inv_chr7, minDiff = 0.1, callRate = 0.9))
```

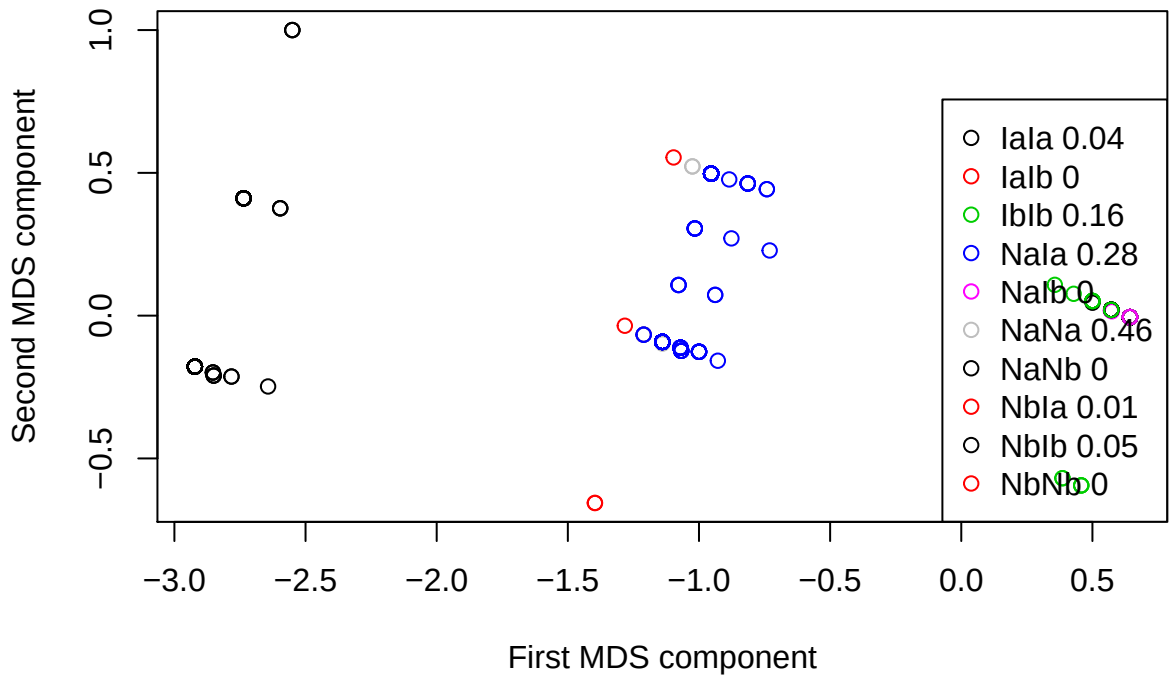
```
## [1] 173
```

```
length(classification(minimac_inv_chr7, minDiff = 0.1, callRate = 0.9))/
  length(classification(minimac_inv_chr7))
```

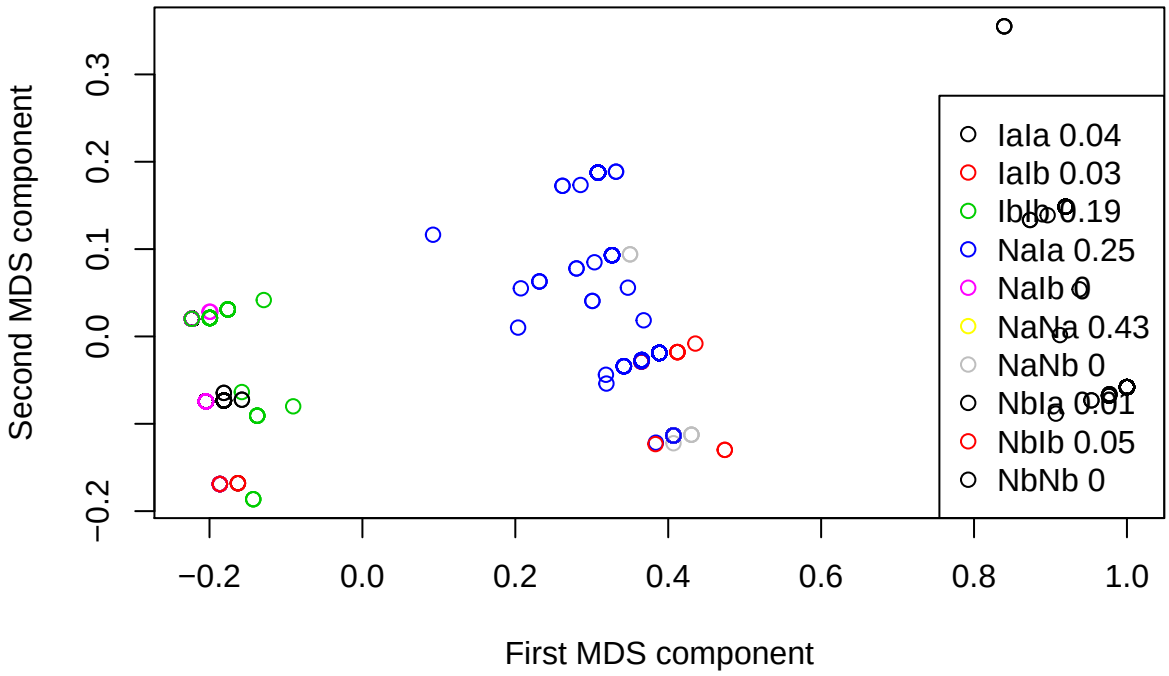
```
## [1] 0.07604396
```

Plots with invClust

```
# Michigan
par(mfrow=c(1,1))
plotInv(michigan_invclust_chr7, classification = classification(michigan_inv_chr7))
```



```
# Minimac
plotInv(minimac_invclust_chr7, classification = classification(minimac_inv_chr7))
```



No filtered imputed data

```
nofilter_michigan_inv_7
```

```
## scoreInvHapRes
## Samples:  2275
## Genotypes' table:
##  IaIa    IaIb    IbIb    NaIa    NaIb    NaNa    NaNb    NbIa    NbIb    NbNb
##  89   40  536     595     0   866     1   22  121     5
## - Inversion genotypes' table:
##  NN   NI  II
##  872  738     665
## - Inversion frequency: 45.45%
```

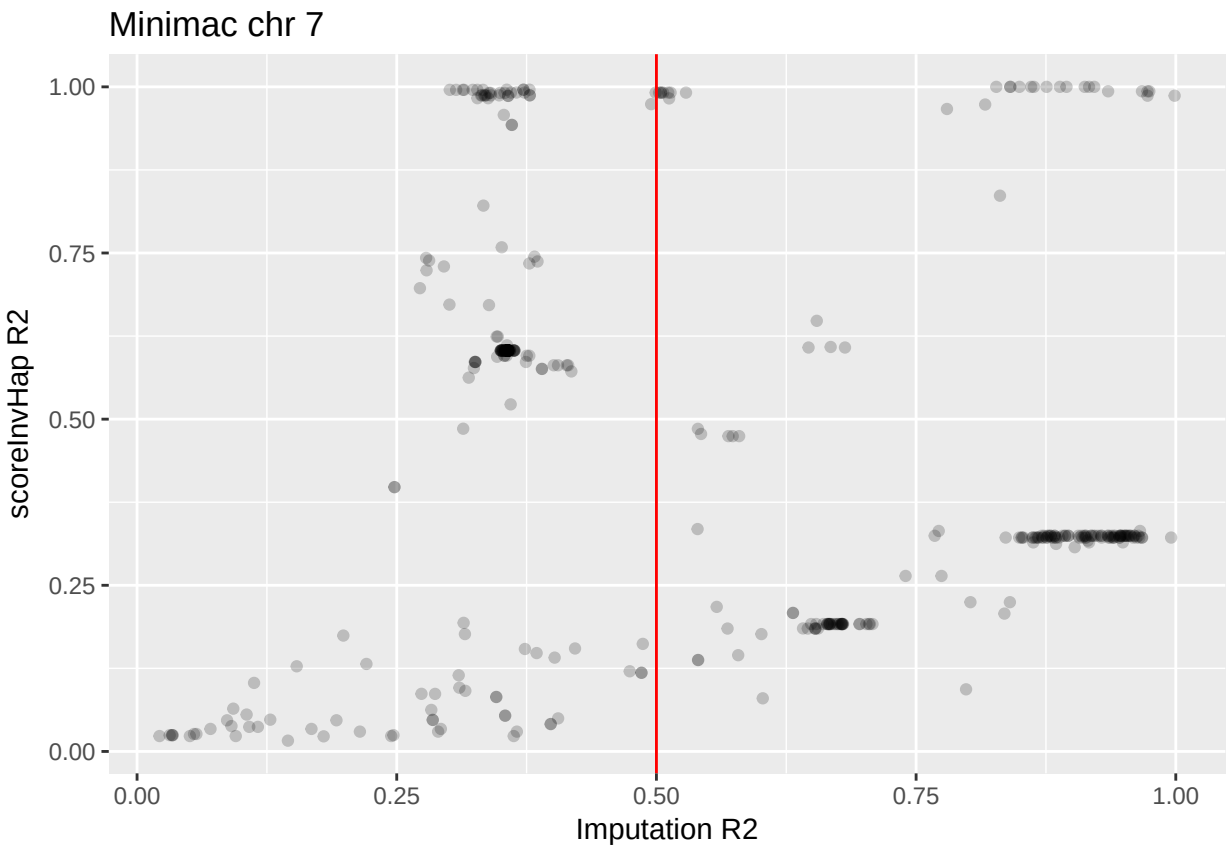
```
nofilter_minimac_inv_7
```

```
## scoreInvHapRes
## Samples:  2275
```

```
## Genotypes' table:
##  IaIa    IaIb    IbIb    NaIa    NaIb    NaNa    NaNb    NbIa    NbIb    NbNb
##  89   261     185     368     710     508     2   27  117     8
## - Inversion genotypes' table:
##  NN   NI  II
##  518  1222    535
## - Inversion frequency: 50.37%
```

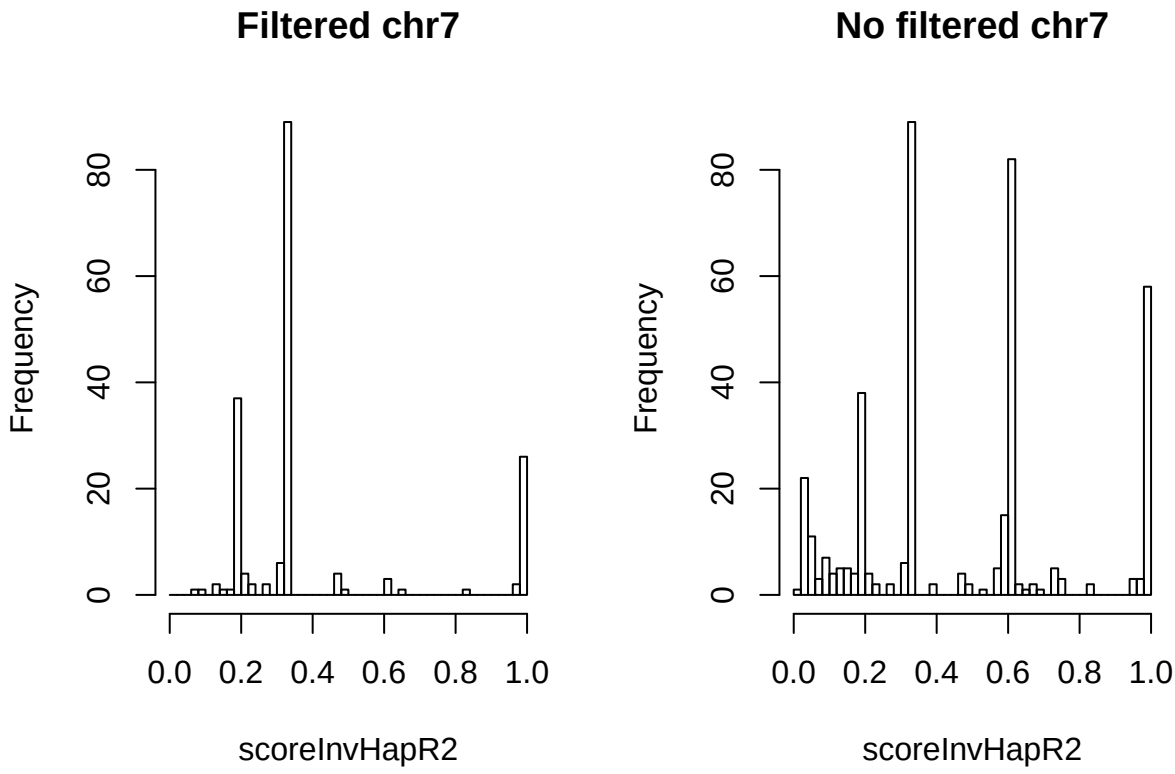
```r
# Select SNPs in both elements to represent them in the plot
snps_minimac_7 <- intersect(rownames(info(nofilter_minimac_7)), names(SNPsR2$inv7p11.2))

# Plot Imputation R2 vs scoreInvHap R2 (red line = filter in the previous data)
ggplot() +
  geom_point(aes(x = info(nofilter_minimac_7)[snps_minimac_7,]$R2,
                 y = SNPsR2$inv7p11.2[snps_minimac_7]),
                alpha = 0.2) +
  geom_vline(aes(xintercept=0.5), colour="red") +
  ggtitle("Minimac chr 7") +
  xlab("Imputation R2") +
  ylab("scoreInvHap R2")
```




```r
# Histograms of scoreInvHap R2 in the filtered imputed data and in the NO filtered imputed data
par(mfrow=c(1,2))
hist(SNPsR2$inv7p11.2[rownames(minimac_chr7)], breaks=seq(0, 1, by=0.02),
     main="Filtered chr7", xlab="scoreInvHapR2")
```

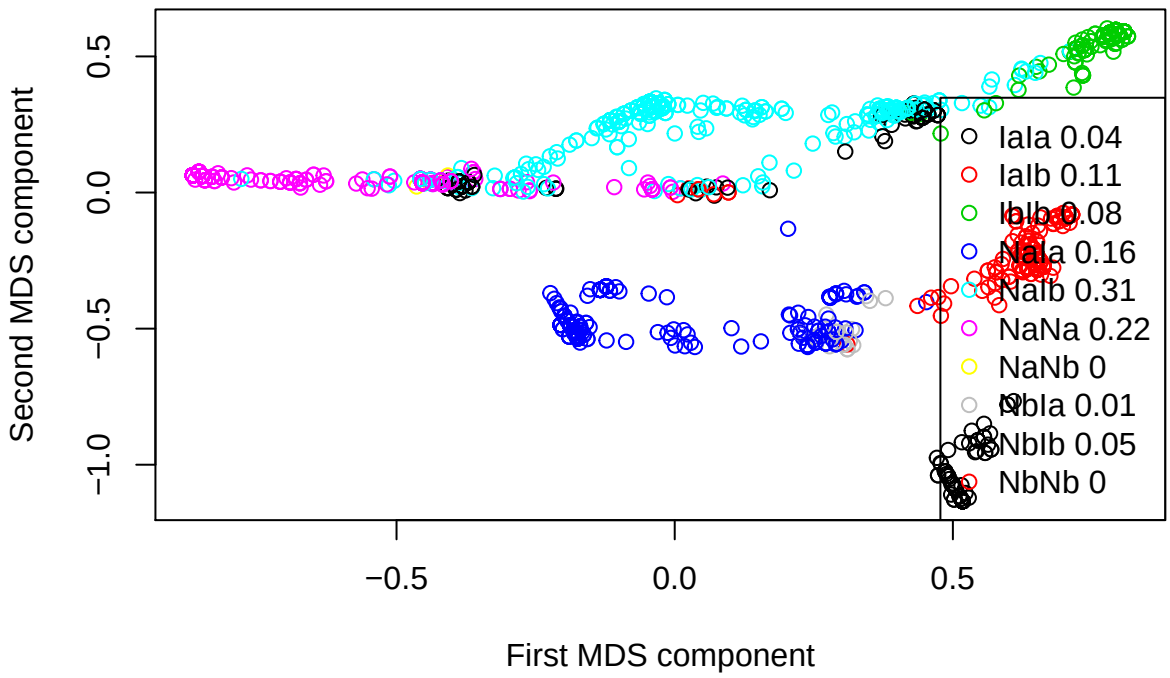
```
hist(SNPsR2$inv7p11.2[rownames(nofilter_minimac_7)], breaks=seq(0, 1, by=0.02),
     main="No filtered chr7", xlab="scoreInvHapR2")
```



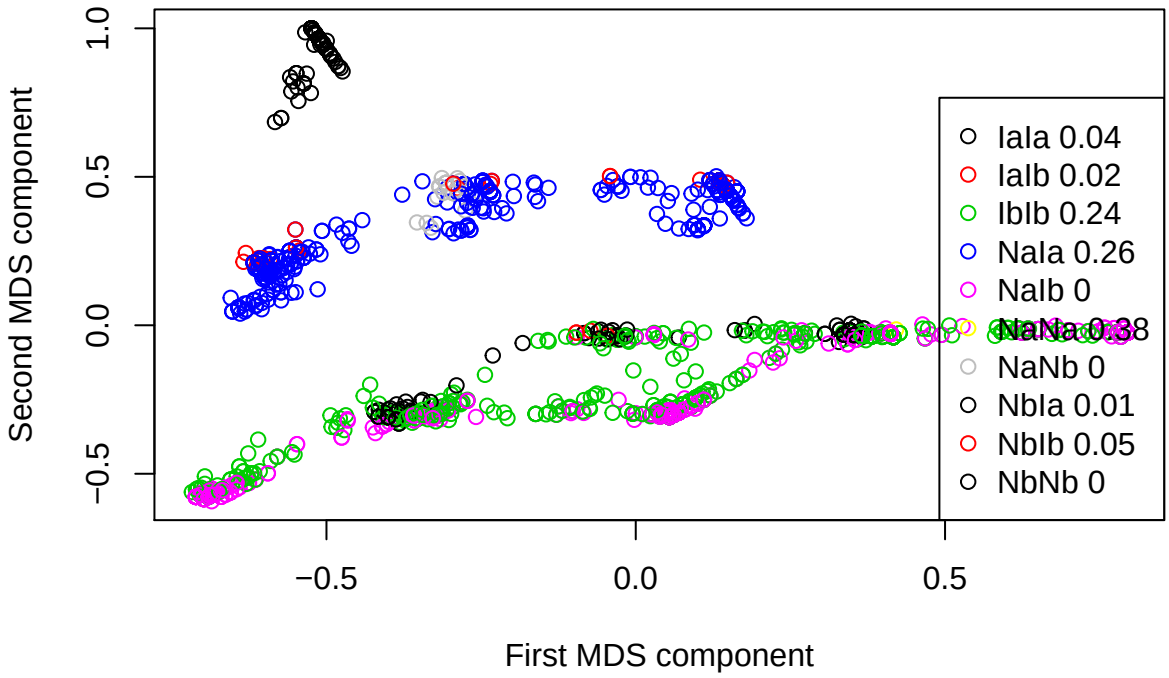

```
#Correlation between Imputation R2 and scoreInvHap R2 (NO filtered data)
cor(info(nofilter_minimac_7)[snps_minimac_7,]$R2, SNPsR2$inv7p11.2[snps_minimac_7])
```

```
## [1] -0.08281116
```

```
# invClust plot of the NO filtered imputed data (minimac)
plotInv(nofilter_minimac_invclust_7,
        classification = classification(nofilter_minimac_inv_7))
```




```
# invClust plot of the NO filtered imputed data (michigan)
# (invClust only used 96 SNPs in this case)
plotInv(nofilter_michigan_invclust_7,
        classification = classification(nofilter_michigan_inv_7))
```


Second MDS component
1.0
0.5
0.0
−0.5
−0.5
0.0
0.5
First MDS component
Ia|a 0.04
Ia|b 0.02
Ib|b 0.24
Na|a 0.26
Na|b 0
Na|Na 0.38
Na|Nb 0
Nb|a 0.01
Nb|b 0.05
Nb|Nb 0