Comparison: Michigan Imputation Server / Shapeit+Minimac3 (Chromosome 17)

Ignacio Tolosana

Imputation

```
## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind,
##     colMeans, colnames, colSums, dirname, do.call, duplicated,
##     eval, evalq, Filter, Find, get, grep, grepl, intersect,
##     is.unsorted, lapply, lengths, Map, mapply, match, mget, order,
##     paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind,
##     Reduce, rowMeans, rownames, rowSums, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which, which.max,
##     which.min

## Loading required package: GenomeInfoDb

## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:base':
##
##     expand.grid
```

```
## Loading required package: IRanges

## Loading required package: GenomicRanges

## Loading required package: SummarizedExperiment

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

## Loading required package: DelayedArray

## Loading required package: matrixStats

##
## Attaching package: 'matrixStats'

## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians

## Loading required package: BiocParallel

##
## Attaching package: 'DelayedArray'

## The following objects are masked from 'package:matrixStats':
##
##     colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges

## The following objects are masked from 'package:base':
##
##     aperm, apply

## Loading required package: Rsamtools

## Loading required package: Biostrings

## Loading required package: XVector

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:DelayedArray':
##
##     type
```

```
## The following object is masked from 'package:base':
##
##     strsplit

##
## Attaching package: 'VariantAnnotation'

## The following object is masked from 'package:base':
##
##     tabulate

## Loading required package: survival

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following object is masked from 'package:VariantAnnotation':
##
##     expand

## The following object is masked from 'package:S4Vectors':
##
##     expand

## Loading required package: magrittr
```

Imputed data exploration

```
michigan_chr17
```

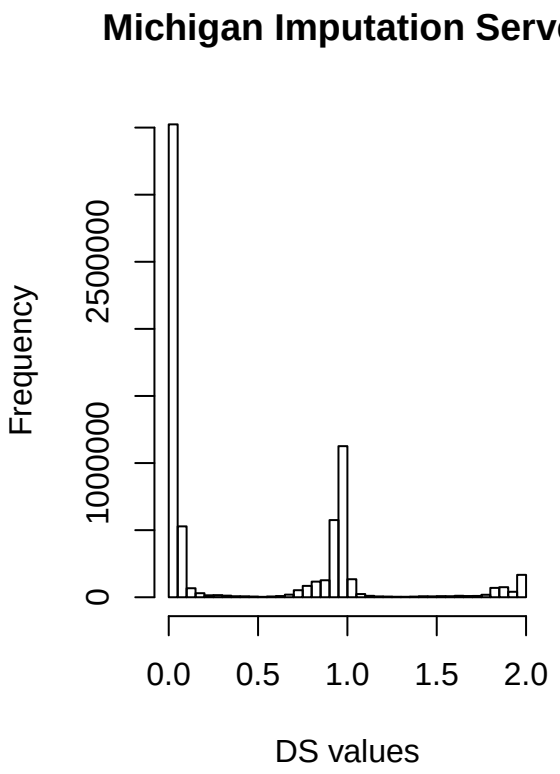
```
## class: CollapsedVCF
## dim: 3041 2286
## rowRanges(vcf):
##   GRanges with 5 metadata columns: paramRangeID, REF, ALT, QUAL, FILTER
## info(vcf):
##   DataFrame with 4 columns: AF, MAF, R2, ER2
## info(header(vcf)):
##        Number Type  Description
##    AF  1      Float Estimated Alternate 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...
```

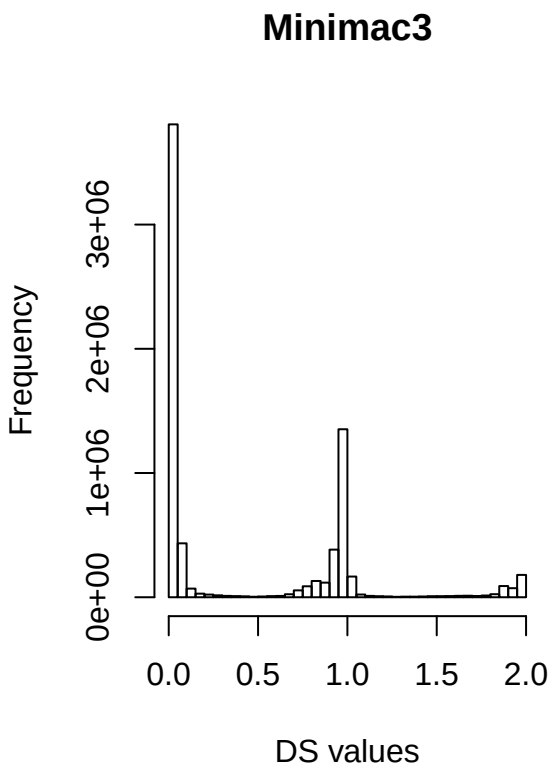
```
minimac_chr17
```

```
## 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="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] 0.8202466
```

```
max(cor_by_ind)
```

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

```
mean(cor_by_ind)
```

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

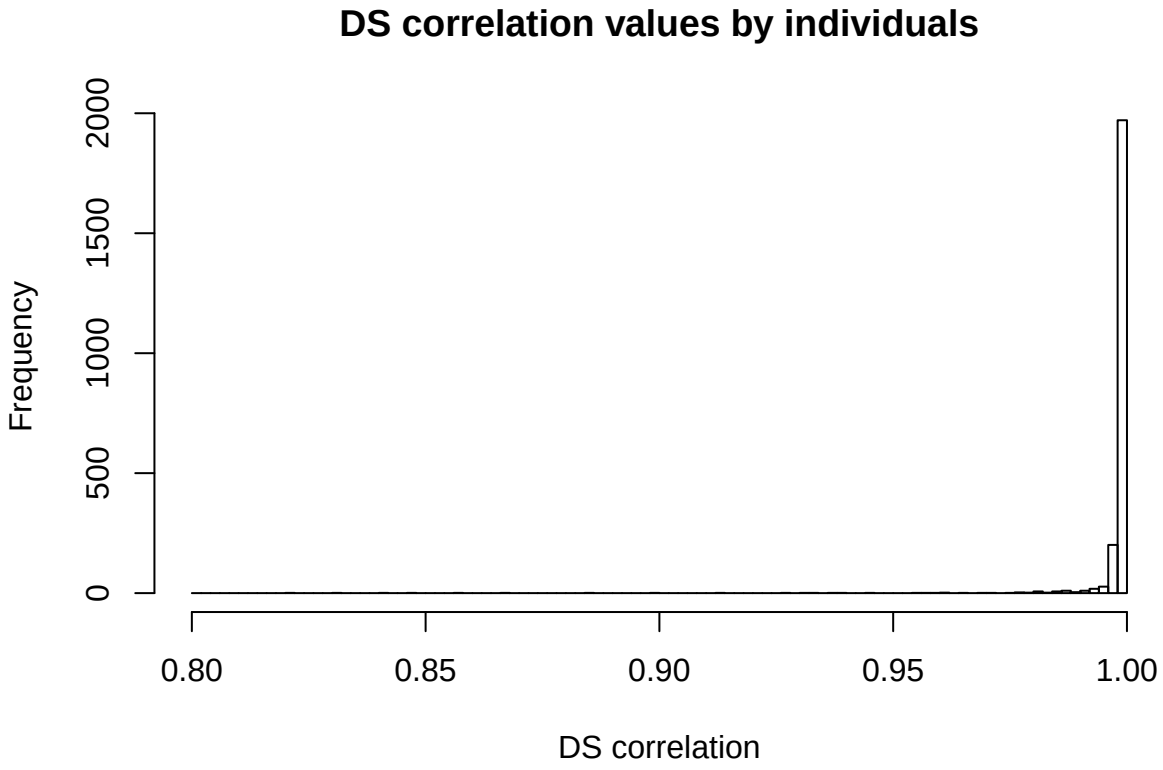
```
mean(cor_by_ind > 0.95)
```

```
## [1] 0.9934383
```

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

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

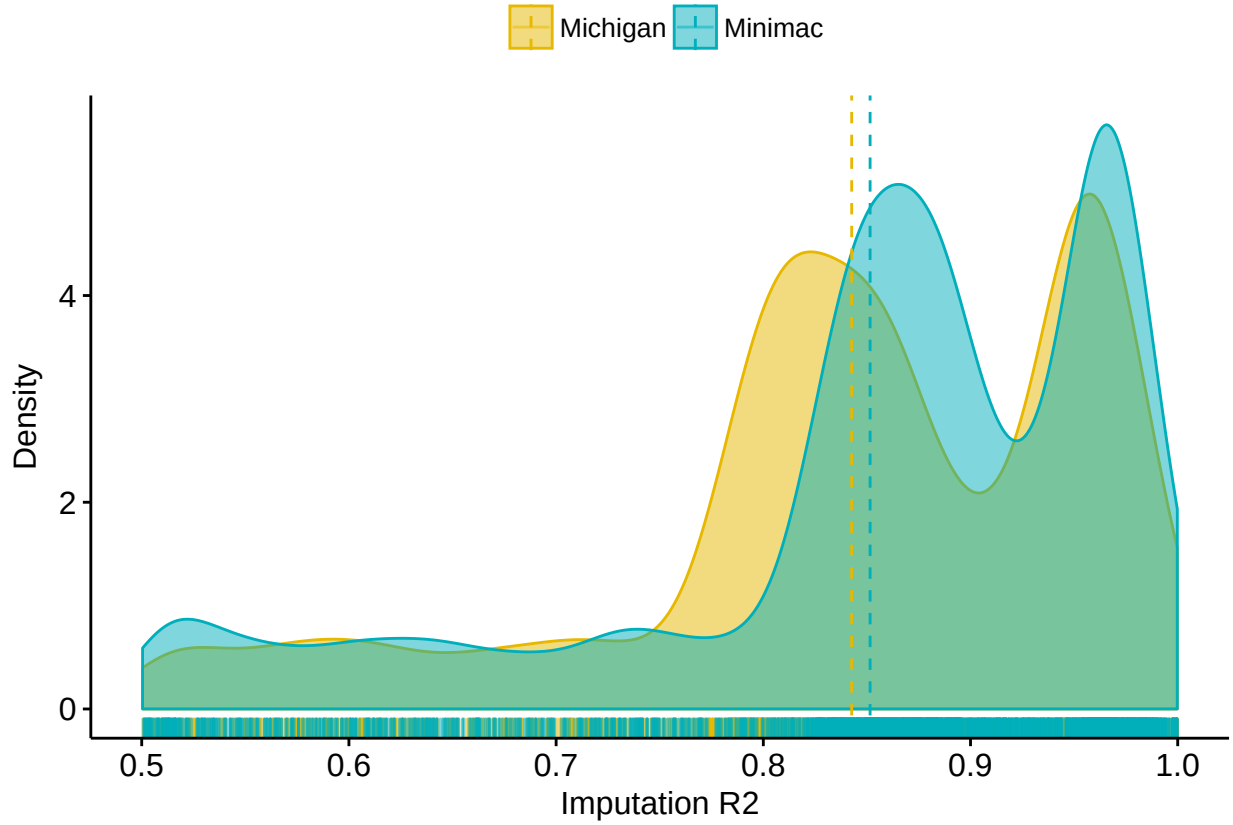
```r
# Histogram of the DS correlation values by individuals
par(mfrow=c(1,1))
CORR_HIST <- hist(cor_by_ind, breaks=seq(0.8, 1, by=0.002),
                  main="DS correlation values by individuals", xlab="DS correlation")
```



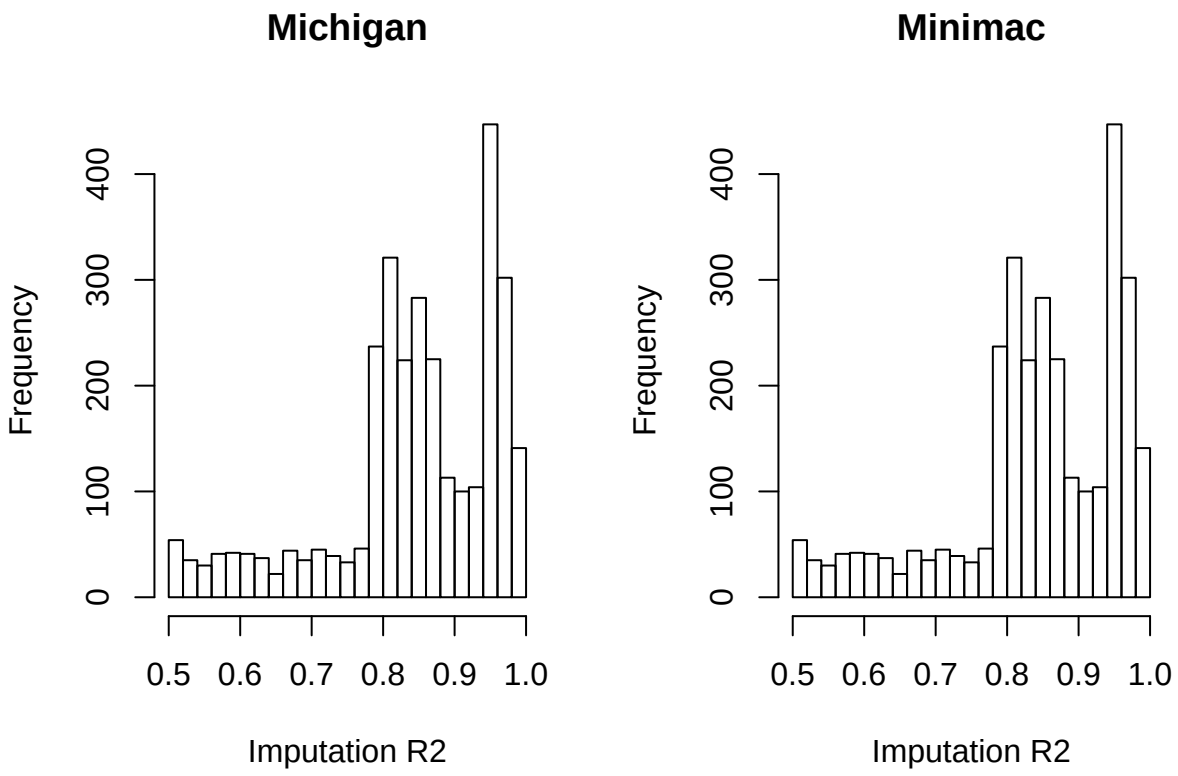
R^2

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("#E7B800", "#00AFBB"),
          legend.title = c(""),
          xlab = ("Imputation R2"),
          ylab = ("Density"))
```




```
par(mfrow=c(1,2))
HIST_R2_MICHIGAN <- hist(info(michigan_chr17)$R2, breaks=seq(0.5, 1, by=0.02),
                         main="Michigan", xlab="Imputation R2")
HIST_R2_MINIMAC <- hist(info(michigan_chr17)$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] 0.8913759
```

```
max(perc_by_ind)
```

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

```
mean(perc_by_ind)
```

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

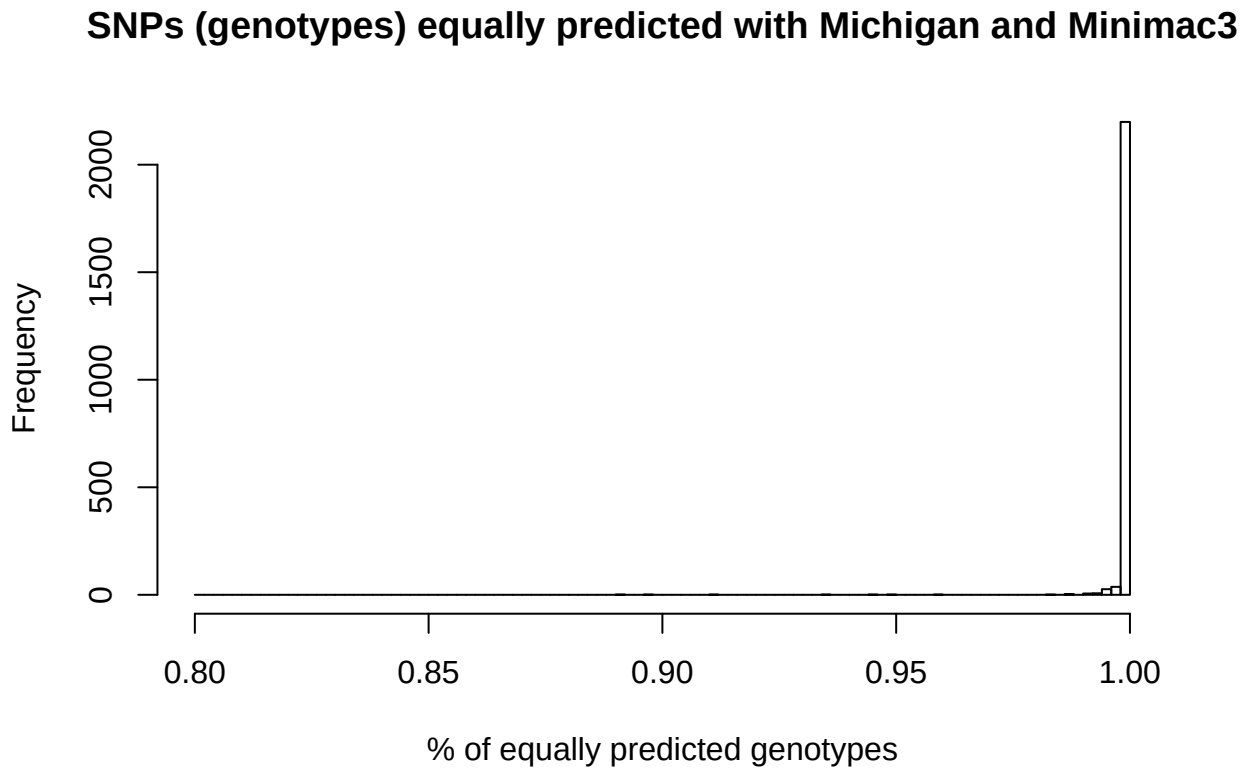
```r
mean(perc_by_ind > 0.95)
```

```
## [1] 0.9973753
```

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

```
## [1] 0.9951881
```

```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(0.8, 1, by=0.002),
                  main="SNPs (genotypes) equally predicted with Michigan and Minimac3",
                  xlab="% of equally predicted genotypes")
```



Inversion prediction

Predicted inversions with scoreInvHap

```r
michigan_inv_chr17
```

```
## 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_chr17
```

```
## 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%
```

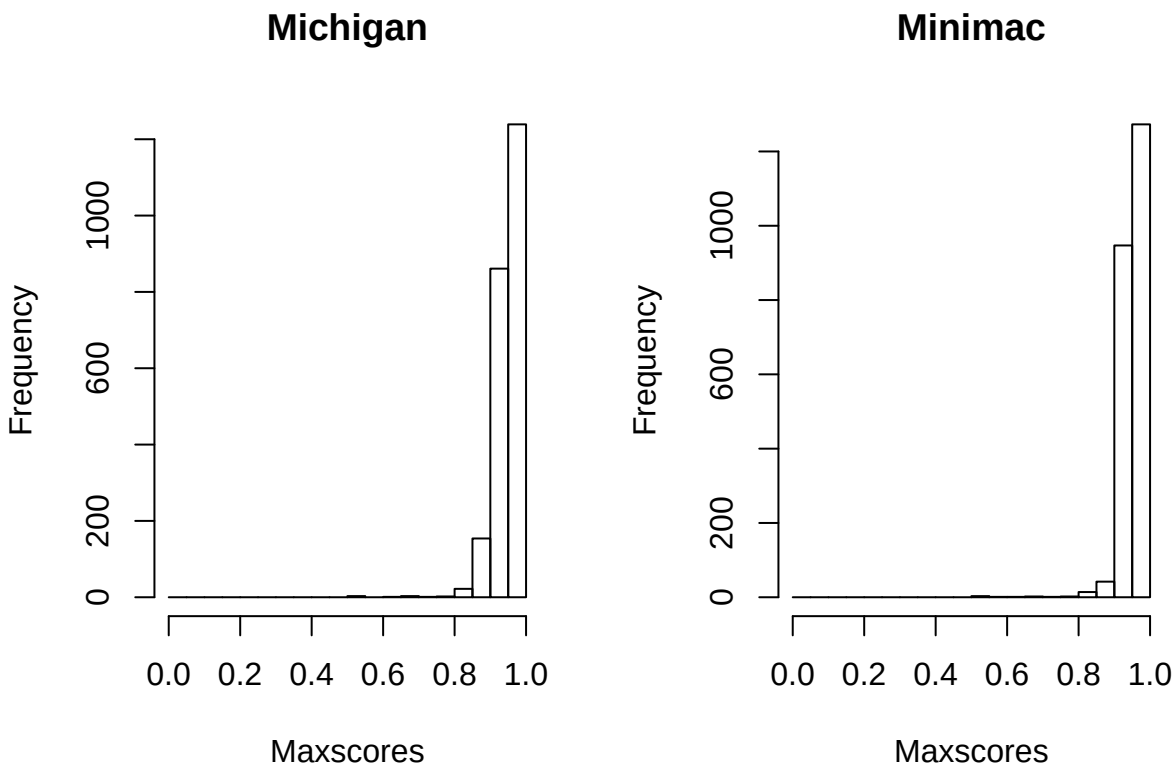
```
# Comparison table
scoreinvhap_table
```

```
##         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_chr17), breaks=seq(0, 1, by=0.05), main="Michigan", xlab="Maxscores")
hist(maxscores(minimac_inv_chr17), breaks=seq(0, 1, by=0.05), main="Minimac", xlab="Maxscores")
```




```
# Score correlation by individuals between both imputation methods
min(score_corr)
```

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

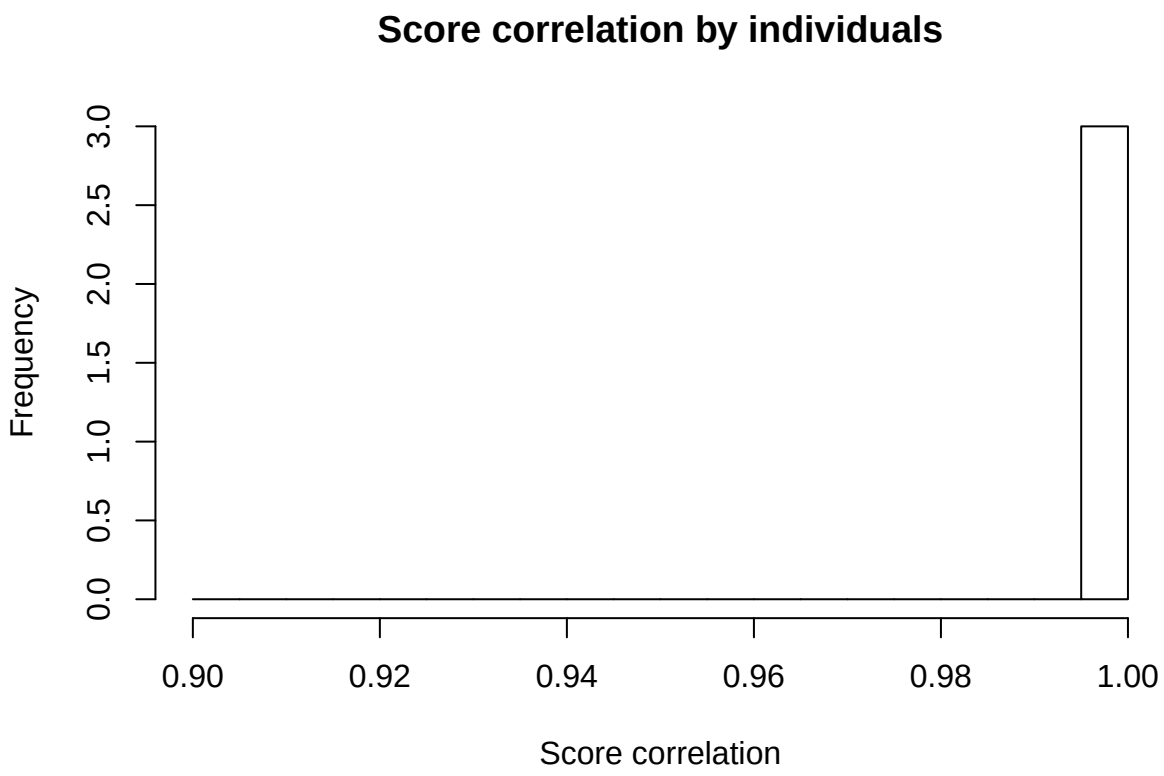
```
max(score_corr)
```

```
## [1] 0.9999092
```

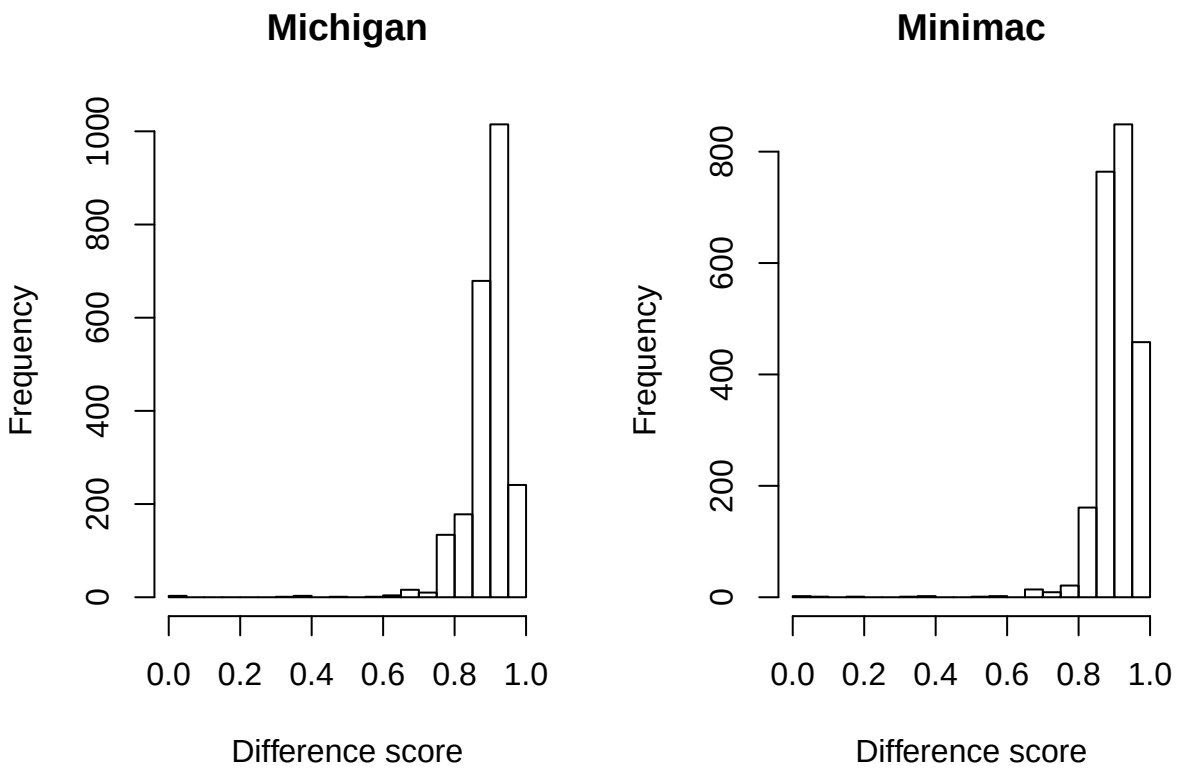
```
mean(score_corr)
```

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

```
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_chr17), breaks=seq(0, 1, by=0.05),
     main="Michigan", xlab="Difference score")
hist(diffscores(minimac_inv_chr17), breaks=seq(0, 1, by=0.05),
     main="Minimac", xlab="Difference score")
```




```
# Numbers of scores used
mean(numSNPs(michigan_inv_chr17))
```

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

```
mean(numSNPs(minimac_inv_chr17))
```

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

```
# Number of samples in both imputation methods before and after QC filtering
length(classification(michigan_inv_chr17))
```

```
## [1] 2286
```

```
length(classification(michigan_inv_chr17, minDiff = 0.1, callRate = 0.9))
```

```
## [1] 2283
```

```
length(classification(michigan_inv_chr17, minDiff = 0.1, callRate = 0.9))/
  length(classification(michigan_inv_chr17))
```

```
## [1] 0.9986877
```

```
length(classification(minimac_inv_chr17))
```

```
## [1] 2286
```

```
length(classification(minimac_inv_chr17, minDiff = 0.1, callRate = 0.9))
```

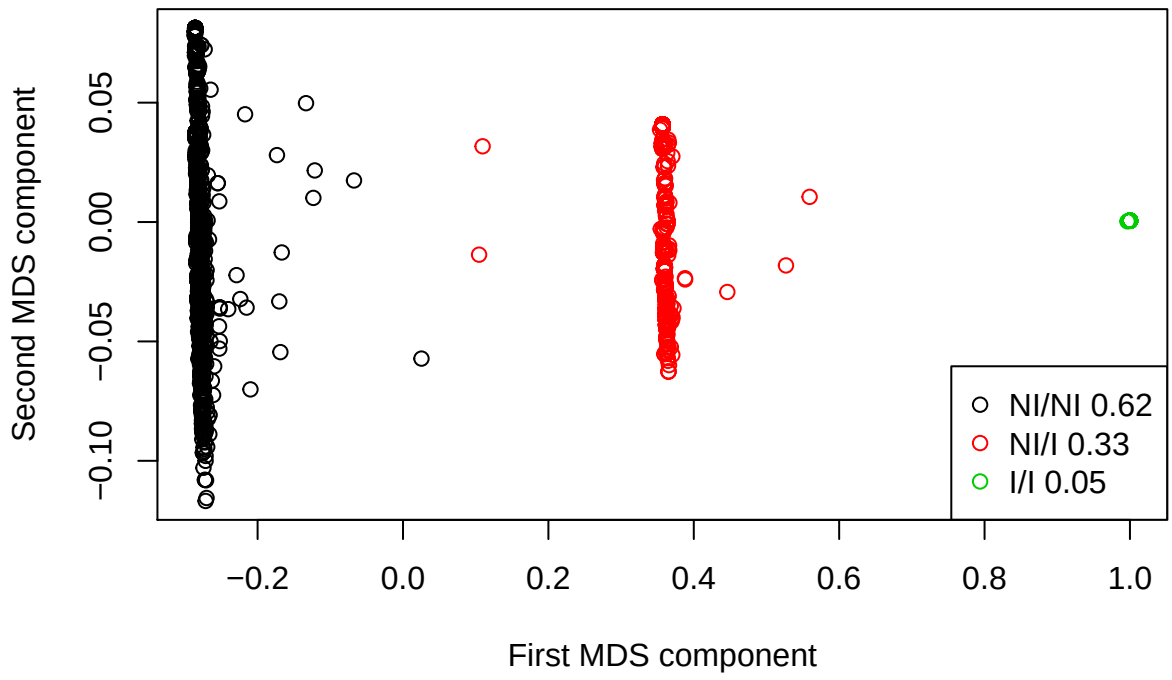
```
## [1] 2283
```

```
length(classification(minimac_inv_chr17, minDiff = 0.1, callRate = 0.9))/
  length(classification(minimac_inv_chr17))
```

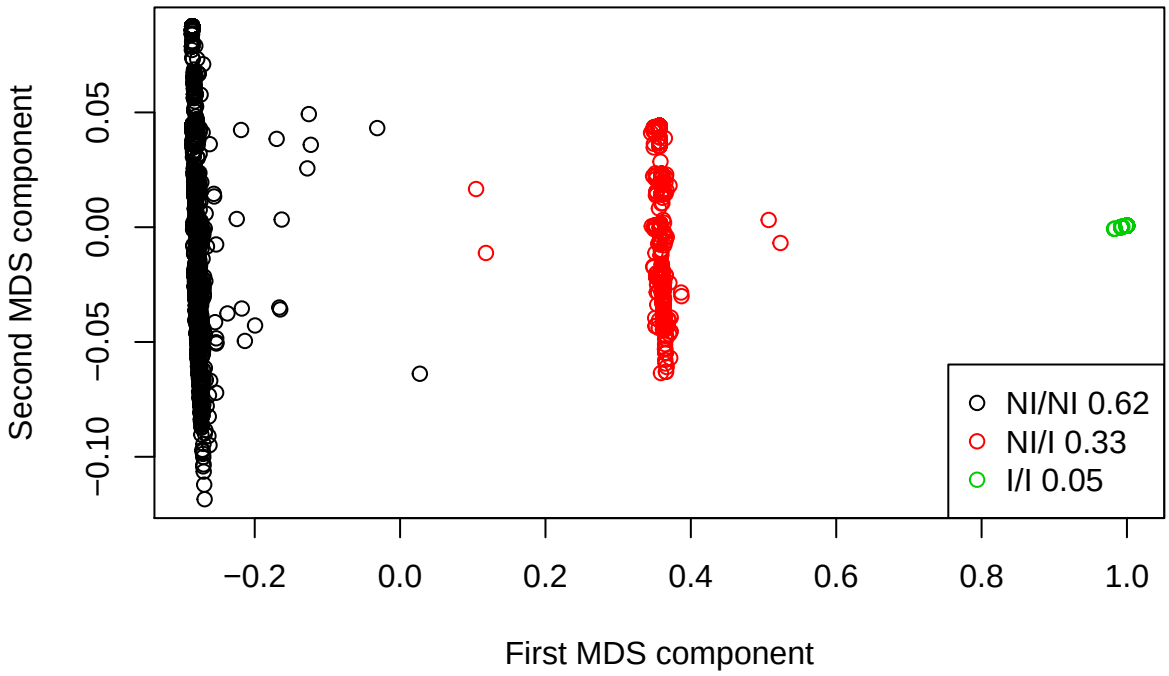
```
## [1] 0.9986877
```

Plots with invClust

```
# Michigan
par(mfrow=c(1,1))
plotInv(michigan_invclust_chr17, classification = classification(michigan_inv_chr17))
```



```
# Minimac
plotInv(minimac_invclust_chr17, classification = classification(minimac_inv_chr17))
```



No filtered imputed data

```
nofilter_minimac_17
```

```
## class: CollapsedVCF
## dim: 12717 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)):
```

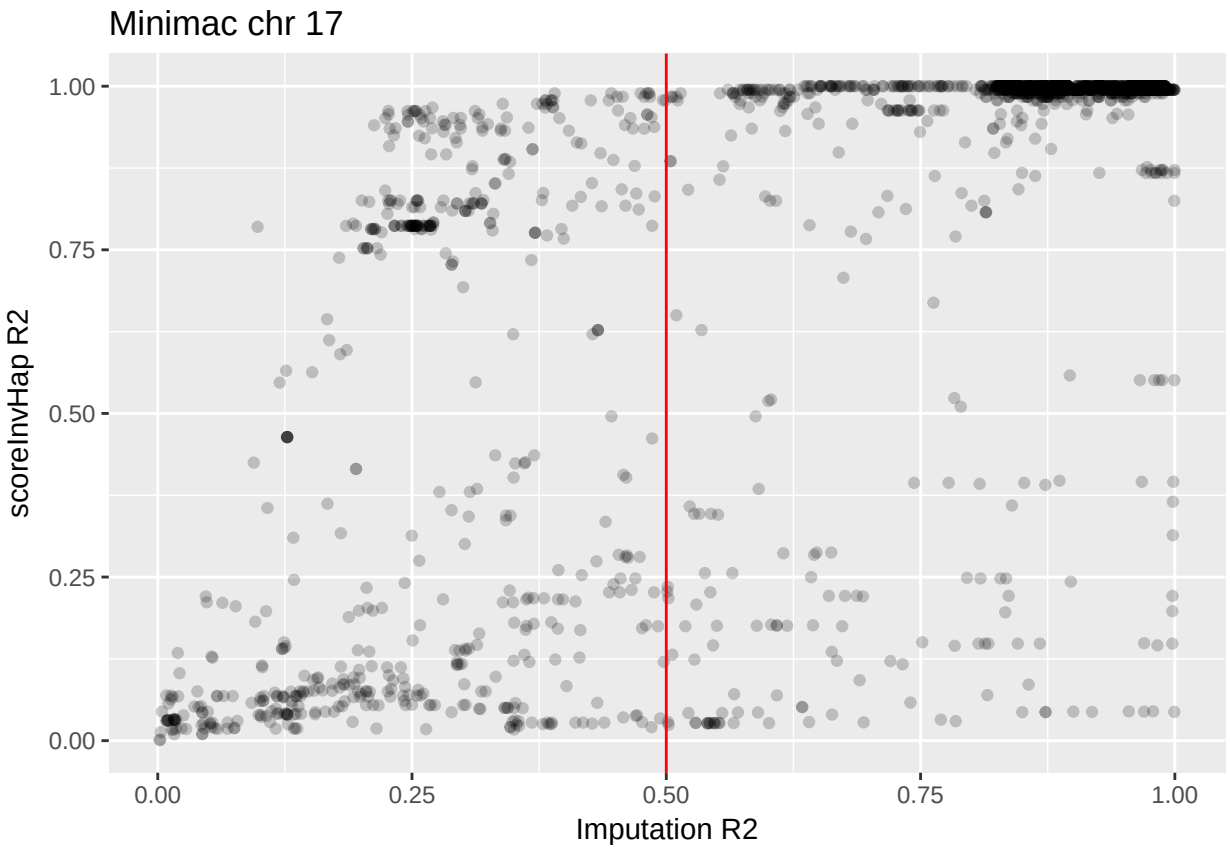
```
##       Number Type   Description
##    GT 1      String Genotype
##    DS 1      Float  Estimated Alternate Allele Dosage : [P(0/1)+2*P(1/1)]
##    GP 3      Float  Estimated Posterior Probabilities for Genotypes 0/0...
```

```
nofilter_minimac_inv_17
```

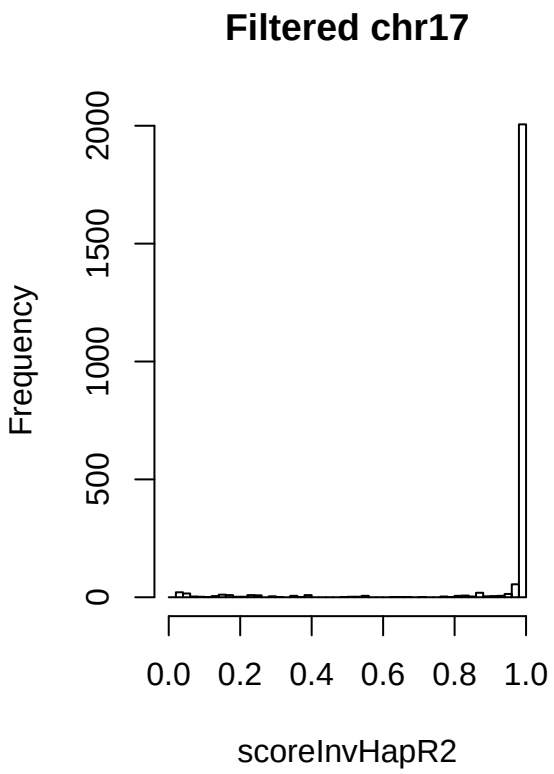
```
## scoreInvHapRes
## Samples:  2286
## Genotypes' table:
##  NI/NI    NI/I    I/I
##  1407     760     119
## - Inversion genotypes' table:
##  NN   NI  II
##  1407     760     119
## - Inversion frequency: 21.83%
```

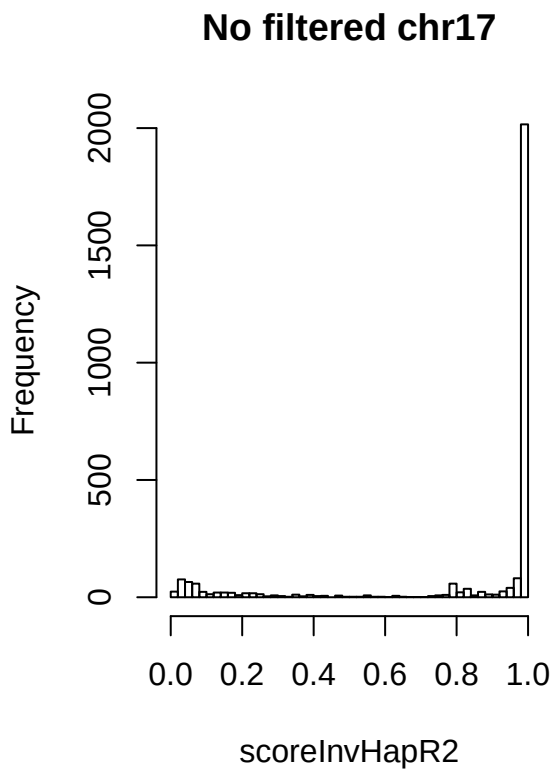
```
# Select SNPs in both elements to represent them in the plot
snps_minimac_17 <- intersect(rownames(info(nofilter_minimac_17)), names(SNPsR2$inv17q21.31))

# Plot Imputation R2 vs scoreInvHap R2 (red line = filter in the previous data)
ggplot() +
  geom_point(aes(x = info(nofilter_minimac_17)[snps_minimac_17,]$R2,
                 y = SNPsR2$inv17q21.31[snps_minimac_17]),
                alpha = 0.2) +
  geom_vline(aes(xintercept=0.5), colour="red") +
  ggtitle("Minimac chr 17") +
  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$inv17q21.31[rownames(minimac_chr17)], breaks=seq(0, 1, by=0.02),
     main="Filtered chr17", xlab="scoreInvHapR2")
hist(SNPsR2$inv17q21.31[rownames(nofilter_minimac_17)], breaks=seq(0, 1, by=0.02),
     main="No filtered chr17", xlab="scoreInvHapR2")
```






```
#Correlation between Imputation R2 and scoreInvHap R2 (NO filtered data)
cor(info(nofilter_minimac_17)[snps_minimac_17,]$R2, SNPsR2$inv17q21.31[snps_minimac_17])
```

```
## [1] 0.7349726
```

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

```
##            Filtered
## No_filtered 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_filt))/sum(scoreinvhap_table_filt)
```

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