EPIC Pipeline

10 April, 2019

Contents

This is a pipeline for preprocessing EPIC-Methylation data using R.

PART 1: Data and parameters

```
## Warning in DataFrame(sampleNames = c(colnames(rgSet),
## colnames(referenceRGset)), : 'stringsAsFactors' is ignored
```

Read data and parameters

We are working with data from directory data/ studies/ 13_Collaborations/ 07_Flotho_JMML/ 00_data/ 00_raw_data/ 01_idats_epic which contains 91 idat files. The annotationfile used is data/ programs/ pipelines/ CPACOR-EPIC_pipeline/ annotationfileB4_2017-09-15.csv - if problems occur with annotation, please have a look at Illumina downloads *Infimum Methylation EPIC Product files.*

Output is directed to data/ studies/ 13_Collaborations/ 07_Flotho_JMML/ 01_analysis/ 02_output. We use samples listed in data/ studies/ 13_Collaborations/ 07_Flotho_JMML/ 01_analysis/ 01_input/ samplefile_JMML_epic.txt
for quality control.

As given in *parameterfile.R*, the following parameters were used:

parameter	value
arraytype	IlluminaHumanMethylationEPIC
detPthreshold	10^{-16}
callrate.thres	0.95
filterOutlierCtrlQC	TRUE
QuantileNormalize	TRUE
InterQuartileRangeCalculation	FALSE
estimateWBCs	TRUE
extractSNPs	FALSE

Further we interpret the values for the gender in the samplesfile as **1=female** and **0=male**.

```
## The samples listed in
##  data/ studies/ 13_Collaborations/ 07_Flotho_JMML/ 01_analysis/ 01_input/ samplefile_JMML_epic_final
##  are used for quantile normalization and calculation of outliers regarding inter-quartile-range.
```

When reading the data using the minfi-package we apply Illumina Background correction. Within this process we also

- extract control-probe information.
- calculate detection p-values.
- estimate the white blood cell distribution assuming whole blood samples using minfi.
- separate the data by channel (red / green) and Infinium I / II type.

We use this detection p-values and control probe information for high-level quality control and the white blood cell estimations for further processing as phenodata.

White Blood Cell estimation

If the switch estimateWBCs is set to TRUE in the parameterfile, white blood cell distributions are estimated assuming measurements from whole blood. The estimation of White Blood Cells results in a data.frame est.wbc.minfi for further use as part of the phenodata.
Notice: Dependencies of the results on the sample selection is possible. To avoid that the WBC estimation of fine runs is disturbed by problematic measurements, the estimation is based on samples from the final sample file only.

```
##                            CD8T      CD4T            NK      Bcell
## 202292320102_R01C01 0.16345548 0.1277519  1.464051e-01 0.29630221
## 202292320102_R03C01 0.12180568 0.2564297  1.347788e-02 0.22908237
## 202292320102_R04C01 0.15847149 0.3435667 -1.387779e-17 0.08136514
## 202292320102_R05C01 0.08551308 0.2723701  0.000000e+00 0.07181416
## 202292320102_R06C01 0.10548865 0.1900578  7.979642e-02 0.23385023
## 202292320102_R07C01 0.02926605 0.2065018  9.692872e-02 0.12954703
##                            Mono      Gran
## 202292320102_R01C01 0.17573070 0.2834445
## 202292320102_R03C01 0.31661405 0.1006506
## 202292320102_R04C01 0.09608631 0.3495113
## 202292320102_R05C01 0.20664404 0.3607841
## 202292320102_R06C01 0.18650585 0.2925847
## 202292320102_R07C01 0.14706265 0.3687178
```

SNP data extraction

If the switch extractSNPs is set to TRUE in the parameterfile, the SNP information from the RGset is extracted and exported to the file /data/studies/13_Collaborations/07_Flotho_JMML/01_analysis/02_output/extracted-snps_JMML_EPIC-2019-04-10-15h57m.csv. This information can be used to detect mismatches in labeling of samples and often explains most of the sex mismatches. Homozygotes should have values around 0 and 1, whereas heterozygotes have values close to 0.5. Please consider this when comparing to the genotyping from other sources which may be coded 0 - 1 - 2.

Data preparation

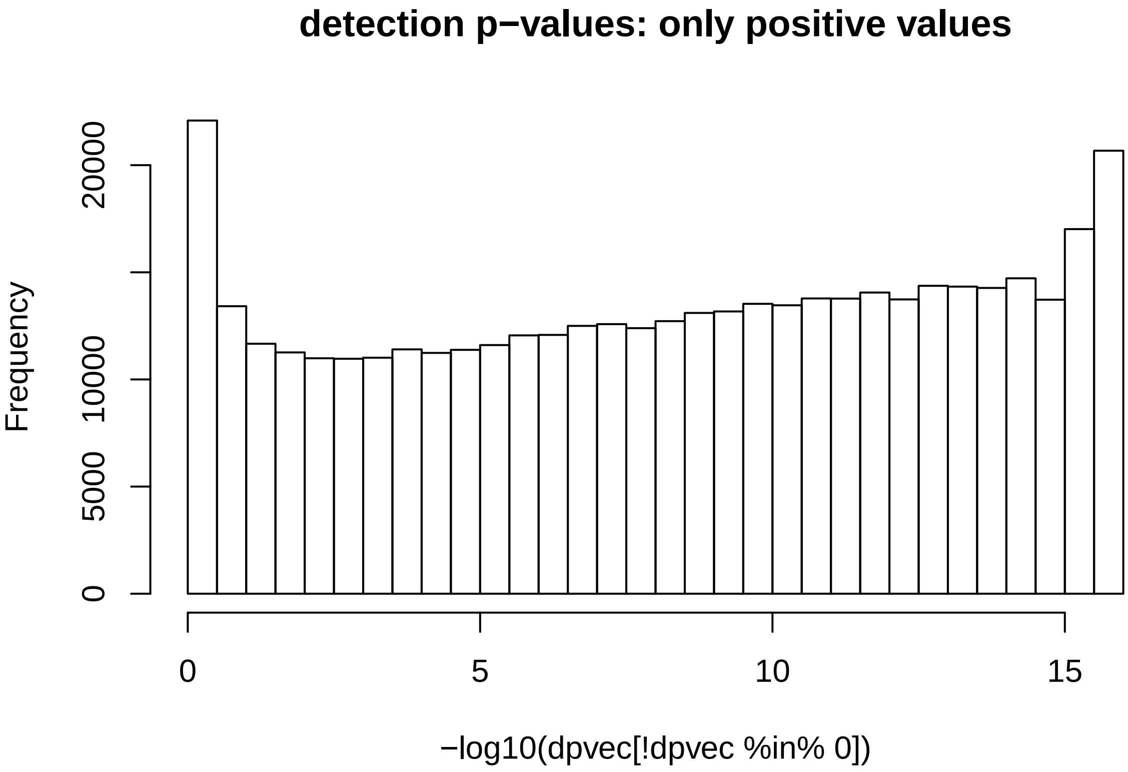
The probes are divided by chromosome type: autosomal probes, chromosome X probes and chromosome Y probes. For this step we need the annotationfile.

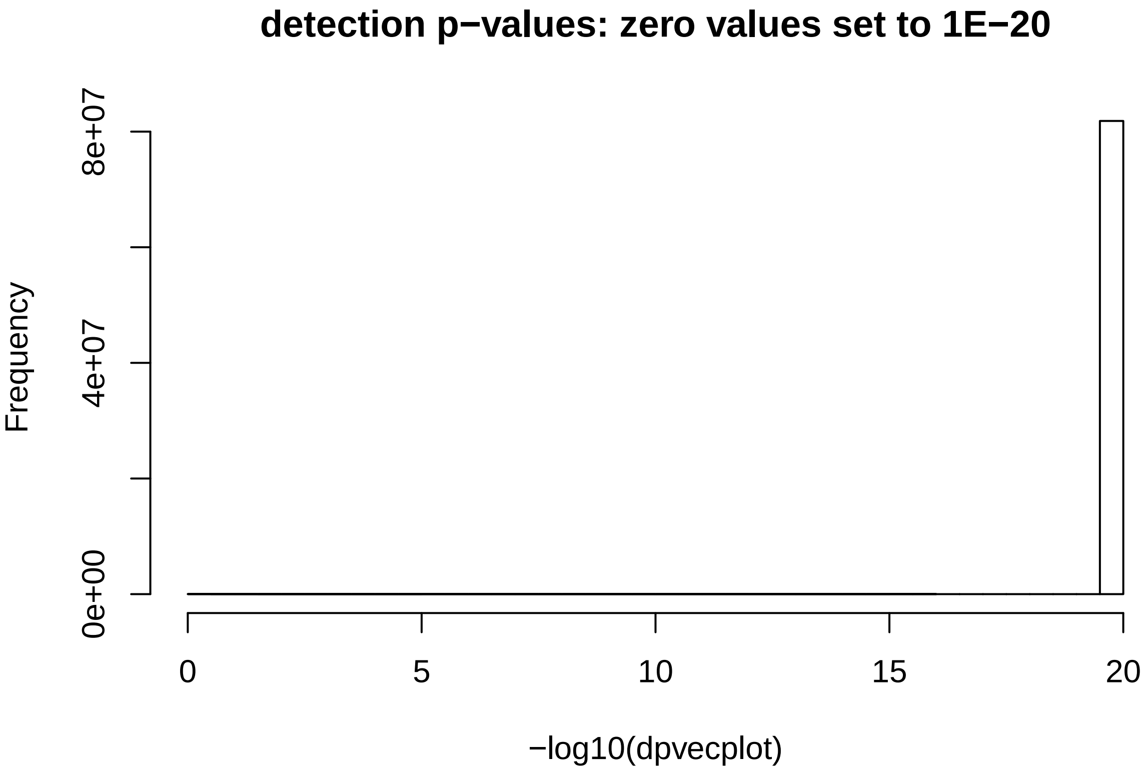
High-level quality control

High level quality control includes detection p-value filter, restriction the the samples listed in the samplesfile and call rate filtering.

We calculate raw beta values for both autosomal and sex chromosome data. In our further calculations beta value data always is processed separately for autosomes and gametes.

- Detection p-values are illustrated in the following plots. Low p-values indicate that the signal is unlikely to be background noise.




detection p−values: zero values set to 1E−20
Frequency
0e+00
4e+07
8e+07
0
5
10
15
20
−log10(dpvecplot)

The following table summarizes how many detection p-values are smaller than the threshold 10^{-16} given in the *parameterfile*, or 0.01:

threshold	count	percentage
10^{-16}	81849573	0.9947851
0.01	82220217	0.9992899

0 detection p-values are missing.

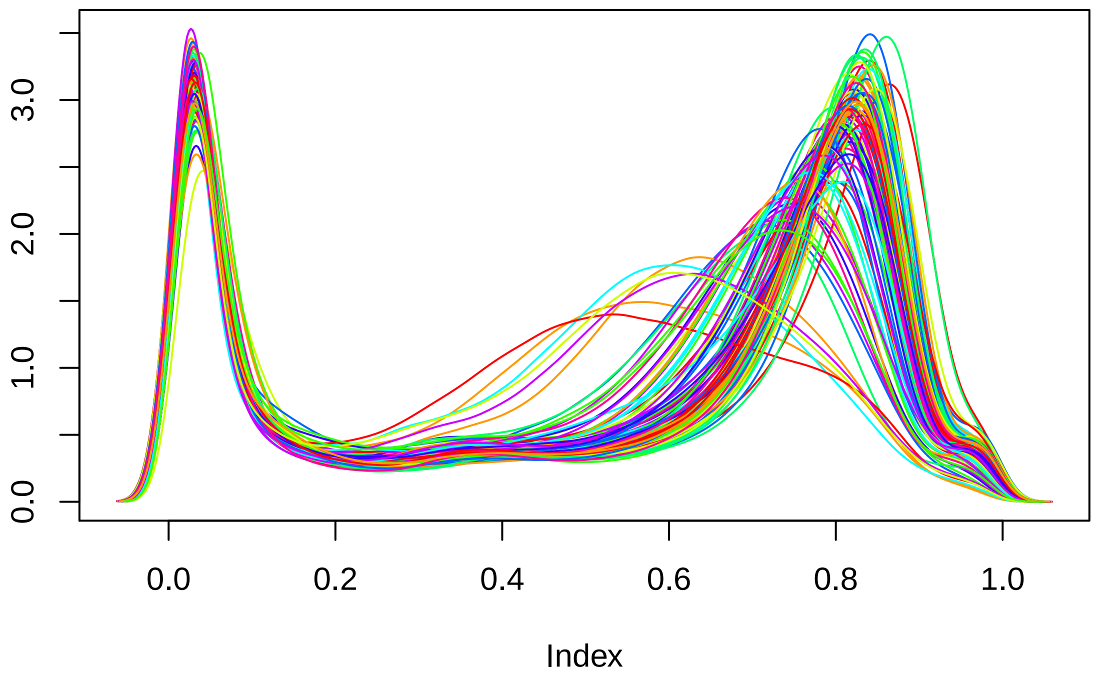
429072 (0.005 %) measurements are excluded because their detection p-value is bigger than 10^{-16}. Only values with a detection p-value strictly smaller the threshold are kept. To skip this filtering, set the parameter *detPthreshold* to a value **strictly** bigger than 1 in the *parameterfile.R*.

- Identified by the samplefile, 93 samples are included in the analysis.

beta values:	autosomes	sex chromosomes
dimension:	$846232, 93$	$19627, 93$

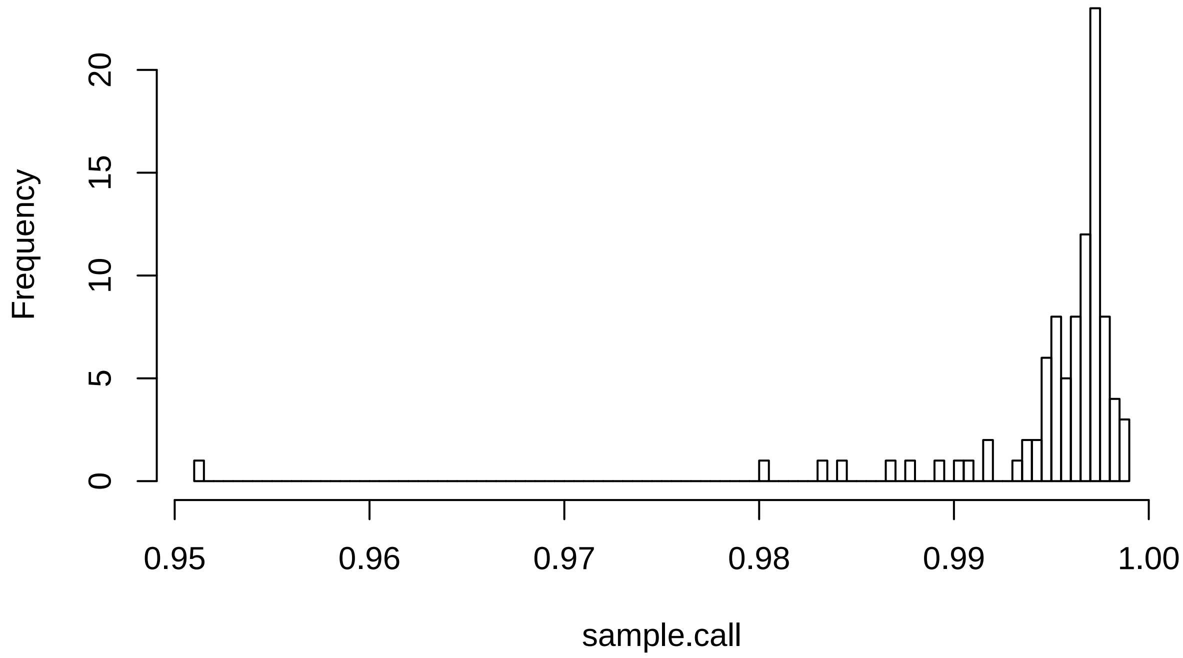
```
##
## The sample names of the included samples:

##  [1] "A027"  "A040"  "A088"  "A118"  "A132"  "B011"  "B034"  "B057"
##  [9] "CH064" "CZ022" "CZ045" "CZ118" "D1004" "D1059" "D1070" "D1077"
## [17] "D1102" "D1167" "D1173" "D1240" "D1241" "D1257" "D1276" "D129"
## [25] "D1292" "D1295" "D155"  "D217"  "D284"  "D310"  "D311"  "D316"
## [33] "D361"  "D378"  "D422"  "D454"  "D456"  "D530"  "D576"  "D614"
## [41] "D675"  "D712"  "D738"  "D763"  "D865"  "D915"  "D943"  "D953"
## [49] "D960"  "D989"  "D994"  "E016"  "E030"  "E041"  "E051"  "H014"
## [57] "H034"  "H036"  "I109"  "I160"  "I215"  "I244"  "I282"  "I283"
## [65] "I289"  "I323"  "I334"  "I338"  "I340"  "IR012" "NL045" "NL056"
## [73] "NL079" "NL095" "NL105" "PL010" "PL080" "SC044" "SC049" "SC070"
## [81] "SC103" "SC182" "SC192" "SC199" "SC201" "SC204" "SC207" "SC214"
## [89] "SC215" "SC218" "SK001" "SK002" "SK008"
```

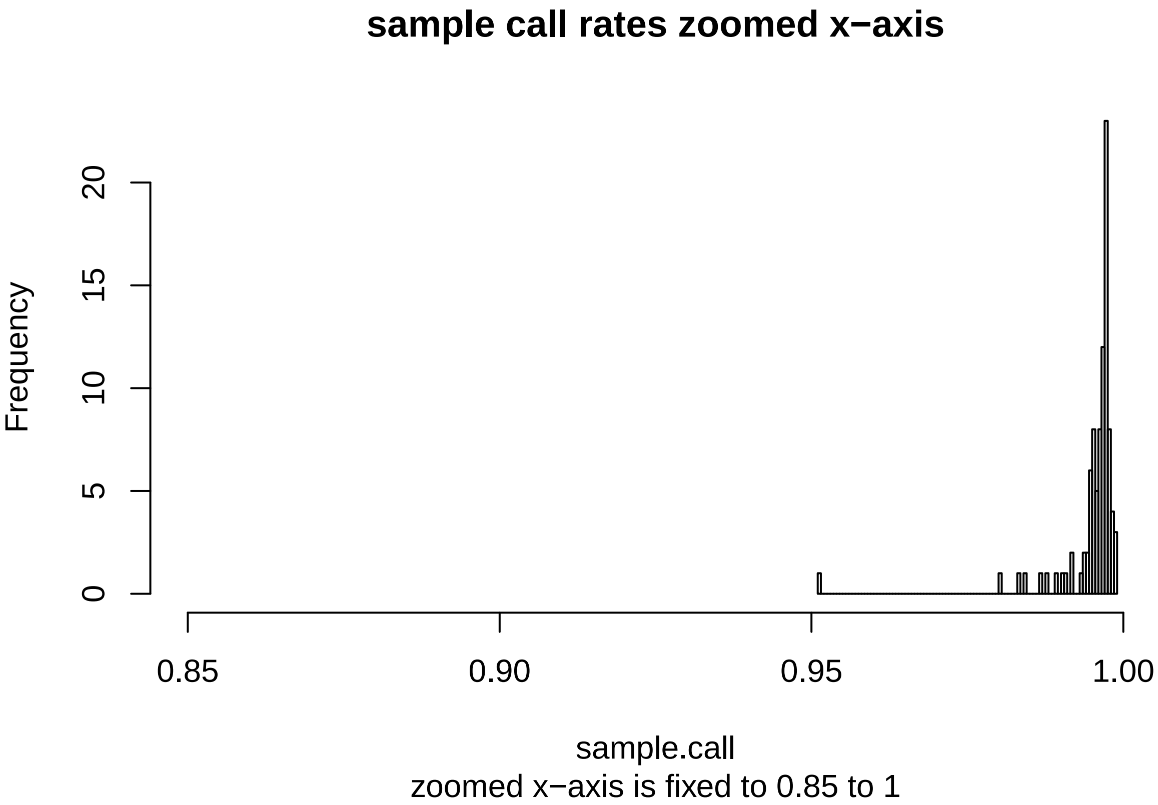
Densities of raw autosomal beta values per sample

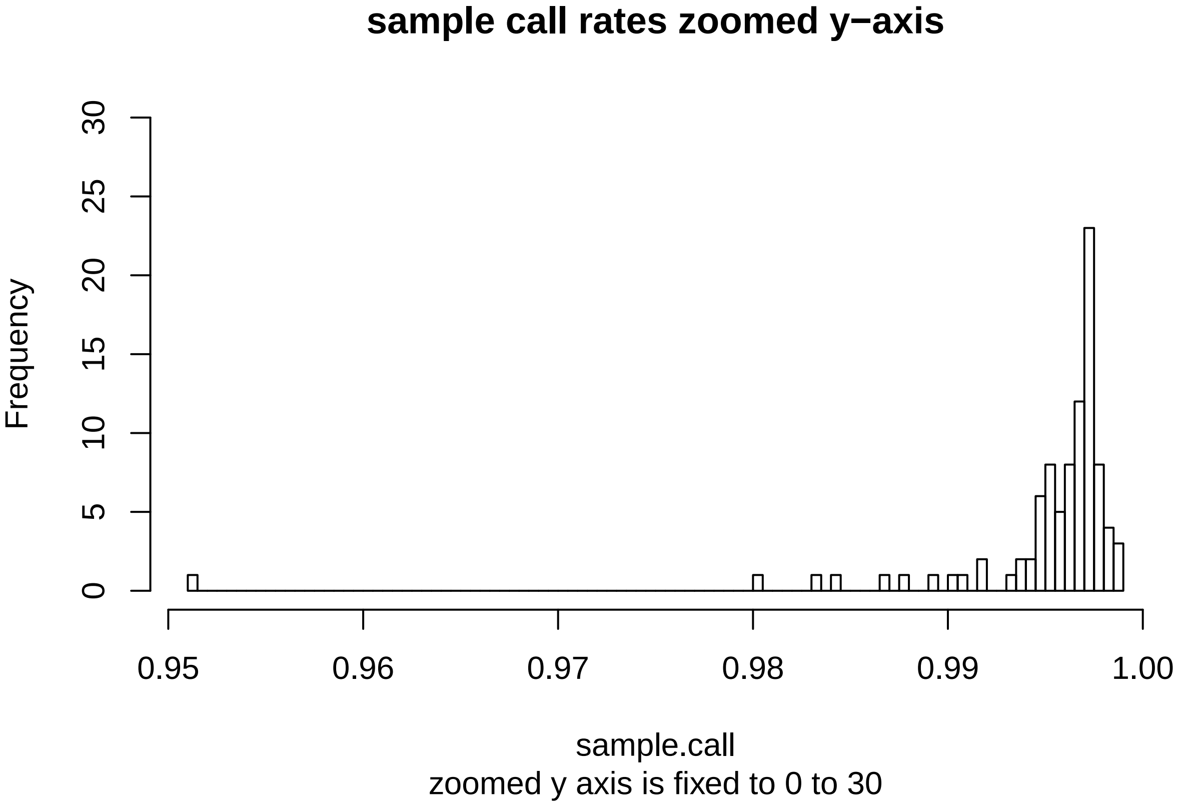


- There is call-rate filtering with threshold 0.95.
- 0 samples were tagged for exclusion because the call-rate was below the threshold 0.95.

sample call rates

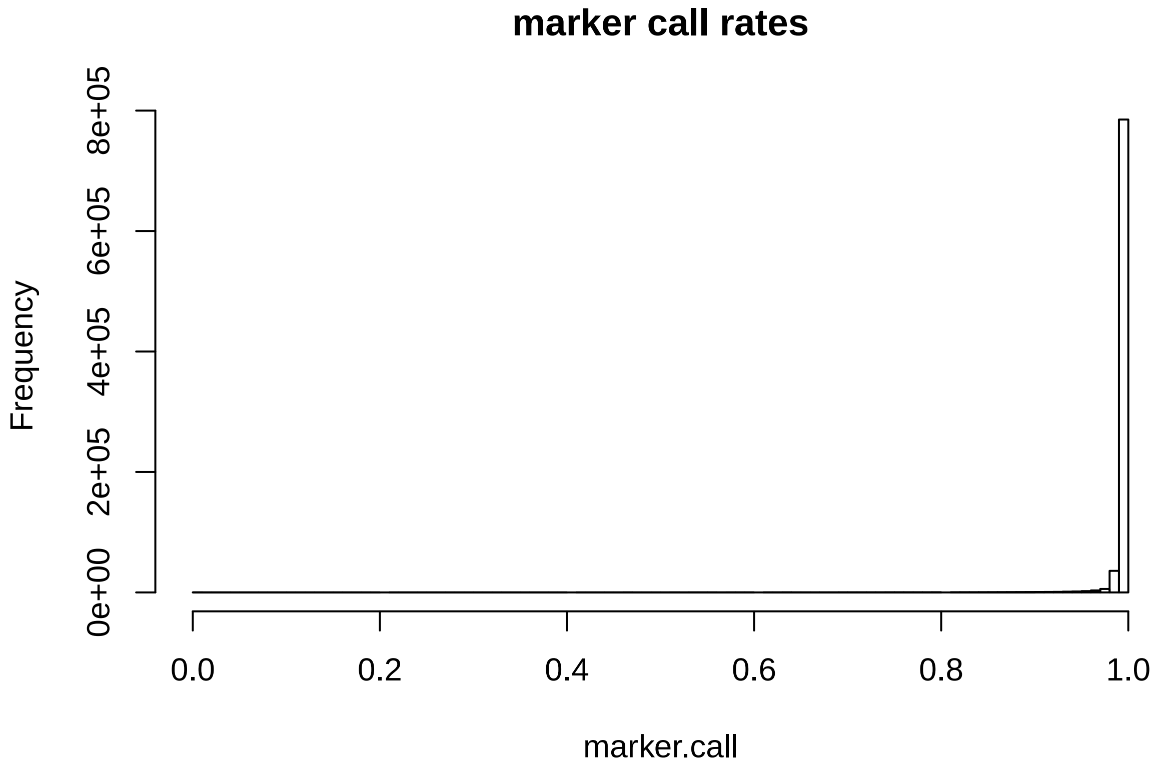


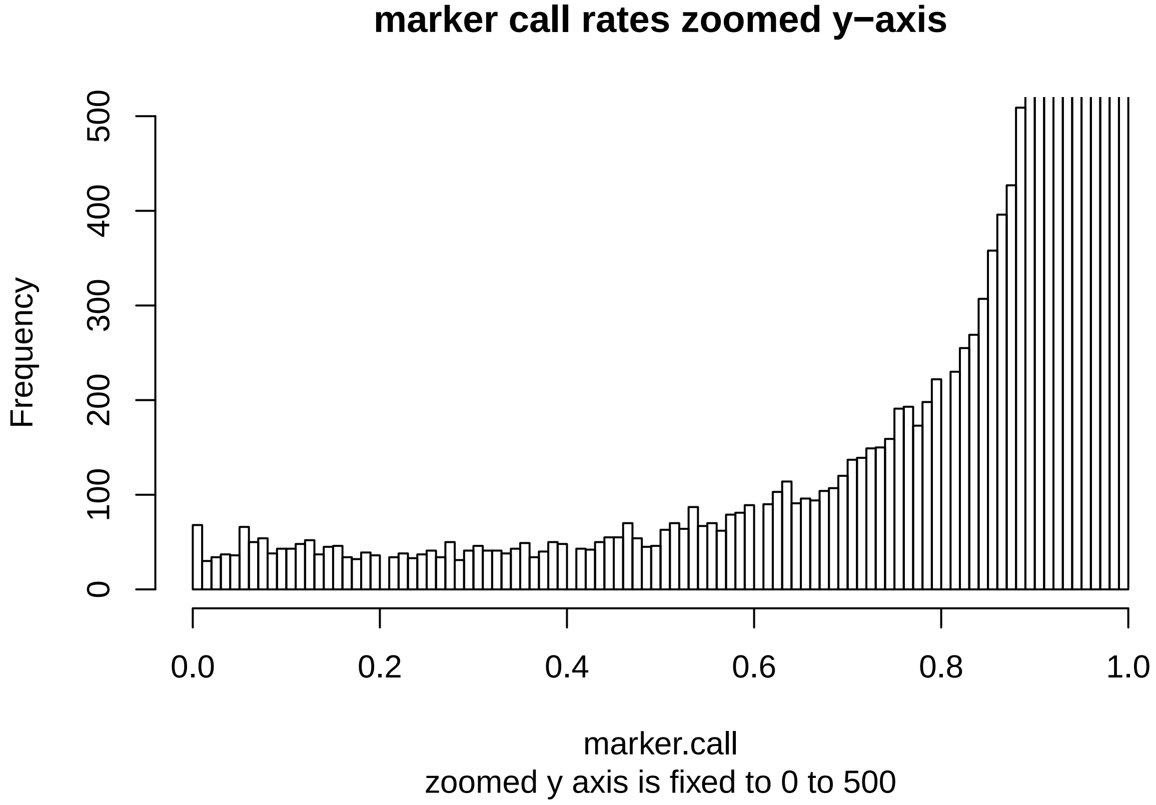




1 of all sample call rates are lower than 0.98,
and 0 are lower than the threshold 0.95.

We have a look at the marker call rates as well:





2.5224×10^4 of all marker call rates are lower than 0.98,
and 13991 are lower than the threshold 0.95.

The results of the sample call rate filter are included in the export file **samples__filtered.csv** which also documents the following control-probe based quality control.

PART 2: Low-Level Quality Control

The quality control consists of two parts:

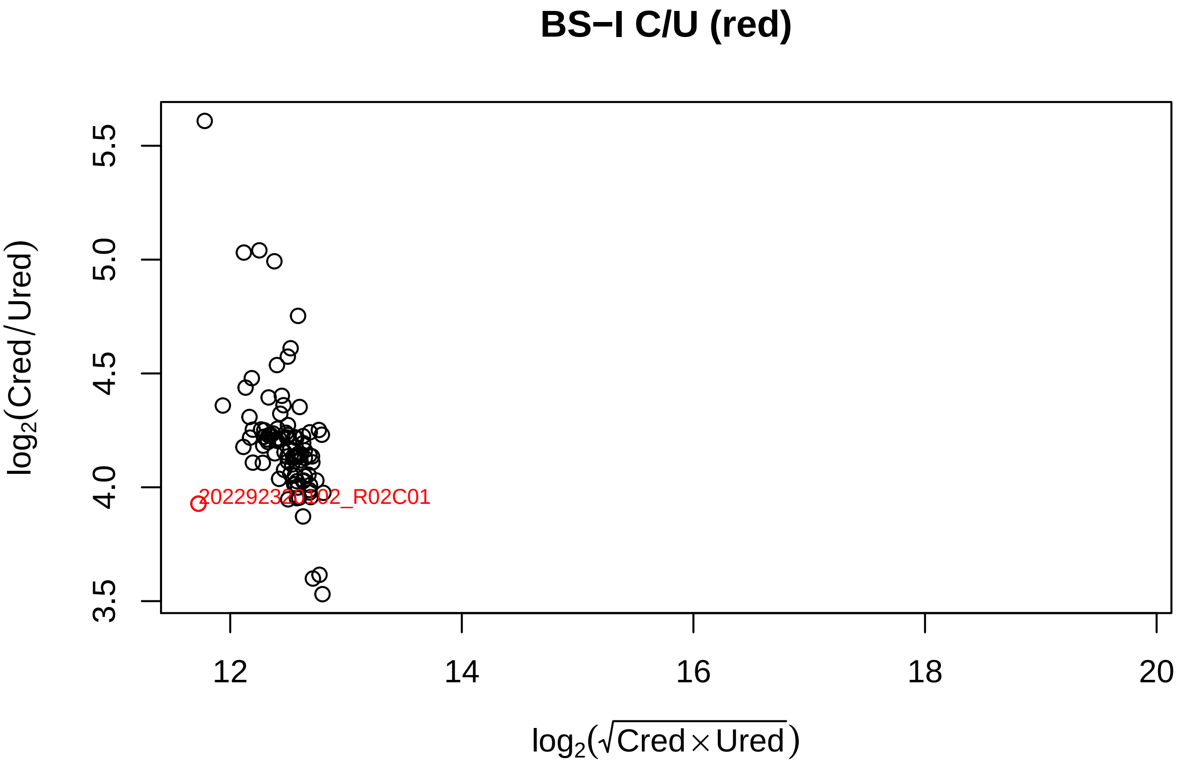
1. The first part is based on control probes. Details are given in the ILMN HD methylation assay protocol guide (15019519).
2. Secondly, data is checked for sex mismatch.

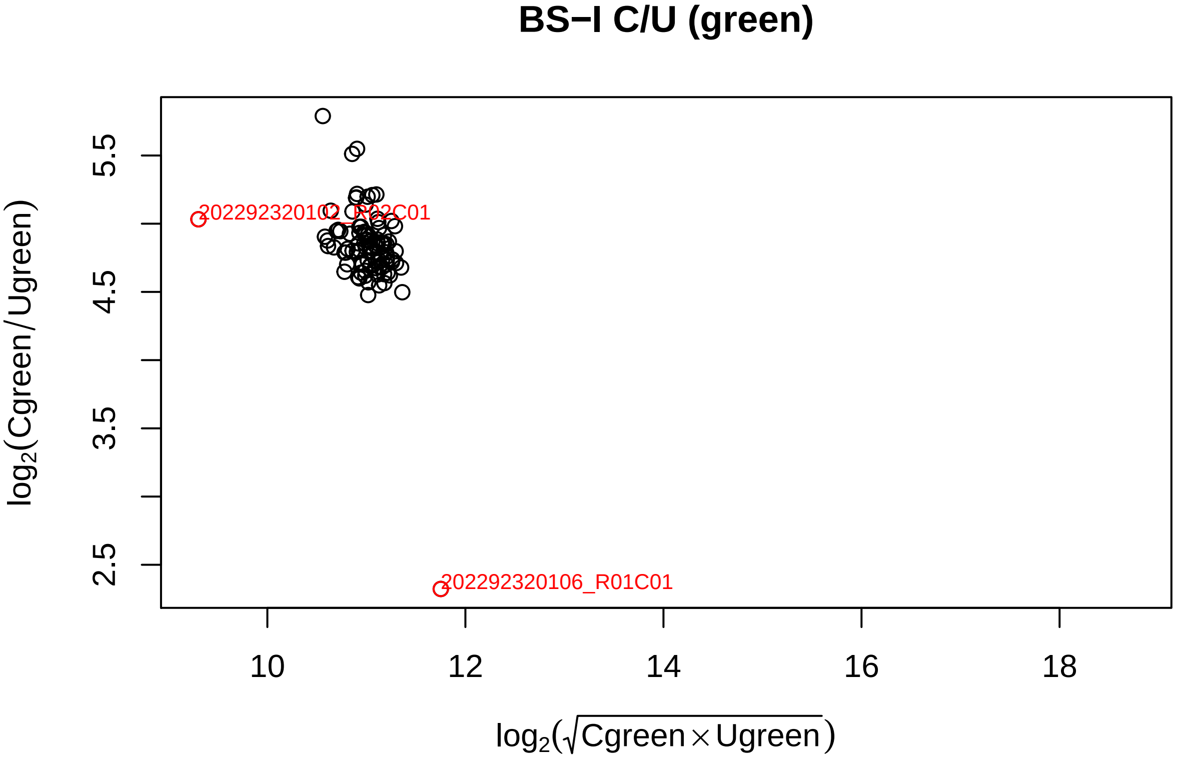
- By conducting a PCA on the control-probe information we obtain controlprobe scores.
- Then we look at the controls.

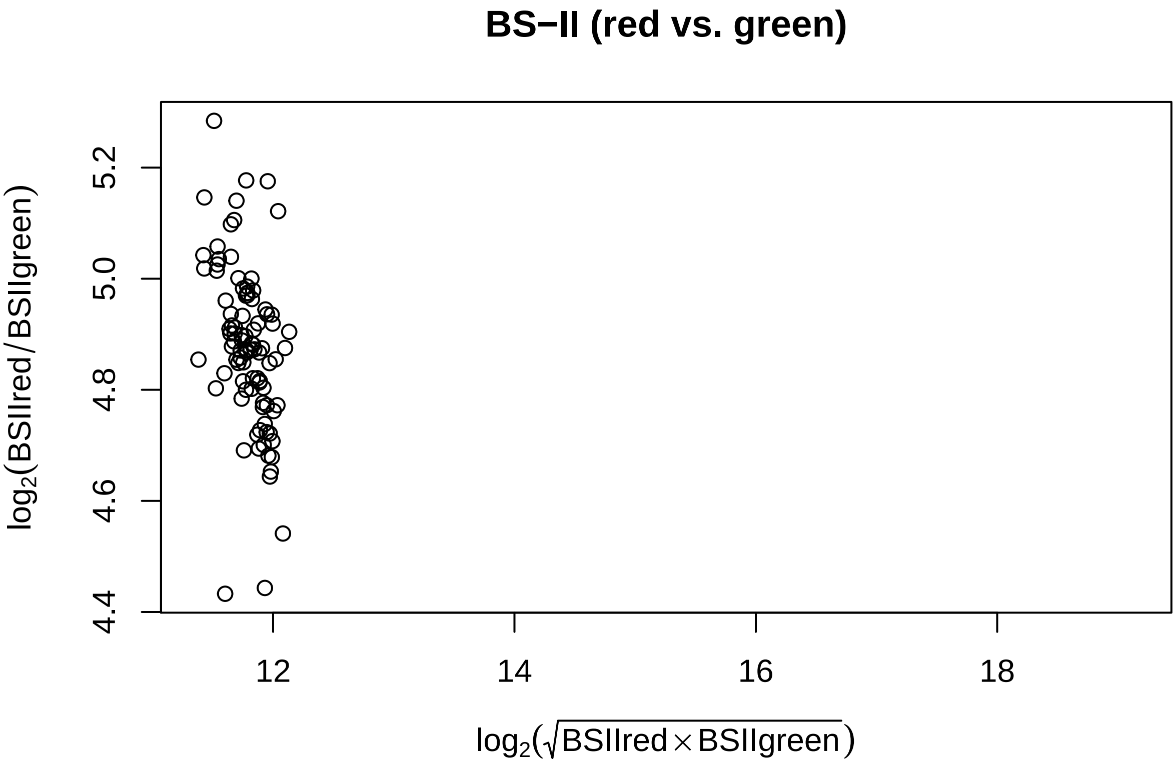
The first 3 rows of control probes information from QC contain the following information:

```
##                      Cgreen     Cred Ugreen     Ured BSIIgreen  BSIIred
## 202292320102_R01C01 8369.0 17752.00  293.0  865.000    496.00 14348.25
## 202292320102_R02C01 3616.5 13212.33  110.5  868.000    669.25 14454.25
## 202292320102_R03C01 9138.0 20669.33  351.0 1199.333    484.50 15703.00
##                       HybH HybL   TR SpecIPMred SpecIPMgreen SpecIMMred
## 202292320102_R01C01 16922 4416 89.5   13169.00     4481.000   543.6667
## 202292320102_R02C01 18684 4906 72.5   12637.67     2766.667   635.0000
## 202292320102_R03C01 18851 5029 85.0   14329.00     4562.333   724.0000
##                     SpecIMMgreen SpecIIspec SpecIIunspec   ExtCG   ExtAT
## 202292320102_R01C01     79.00000   15957.00     177.3333 19053.0 34341.5
## 202292320102_R02C01     55.00000   16805.67     191.6667 20722.5 33854.5
## 202292320102_R03C01     59.33333   18127.33     182.0000 21522.0 36808.0
##                     StainingRedH StainingGreenH StainingRedB
## 202292320102_R01C01        32542          18281          711
## 202292320102_R02C01        37527          15979          938
## 202292320102_R03C01        37571          17246          913
##                     StainingGreenB
## 202292320102_R01C01            140
## 202292320102_R02C01            135
## 202292320102_R03C01            171
```

In the following control probes are checked. For a more detailed description see e.g. the ILMN HD methylation assay protocol guide (15019519) or the Illumina BeadArray Controls Reporter Software Guide,pages 6-8. Probes are evaluated by MA plots. BS-I and BS-II control probes check the DNA bisulfite conversion step.

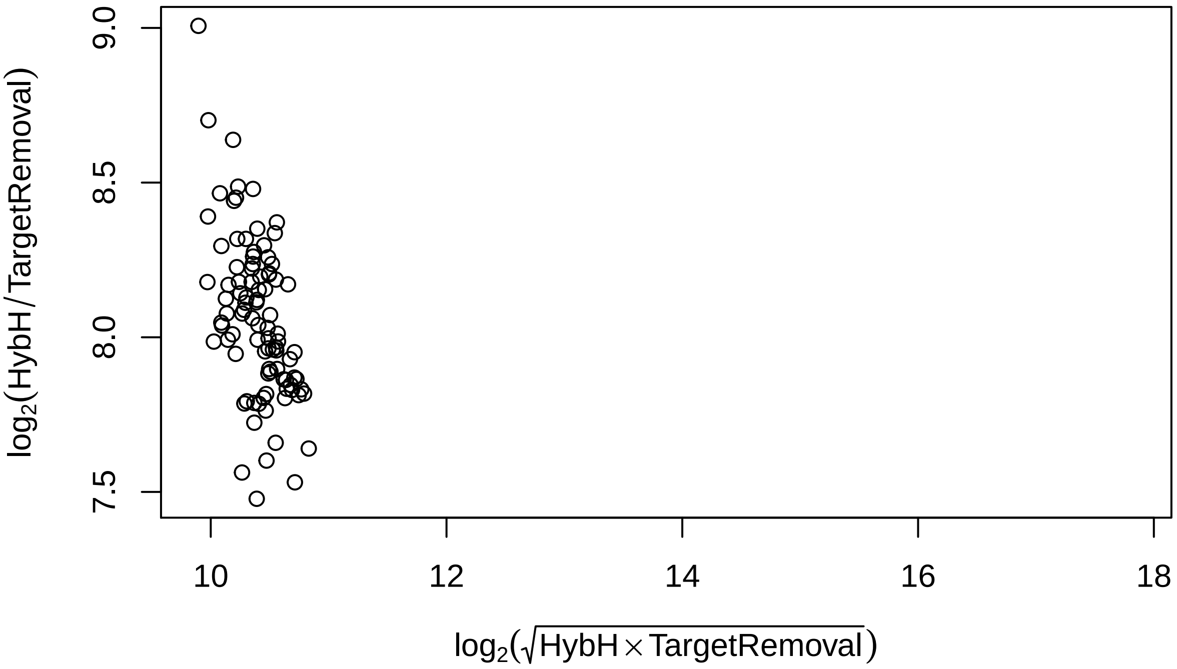

BS−I C/U (red)
$\log_2(\text{Cred}/\text{Ured})$
5.5
5.0
4.5
4.0
3.5
202292320102_R02C01
12
14
16
18
20
$\log_2(\sqrt{\text{Cred} \times \text{Ured}})$


BS−I C/U (green)
$\log_2(\text{Cgreen}/\text{Ugreen})$
5.5
4.5
3.5
2.5
202292320102_R02C01
202292320106_R01C01
10
12
14
16
18
$\log_2(\sqrt{\text{Cgreen} \times \text{Ugreen}})$

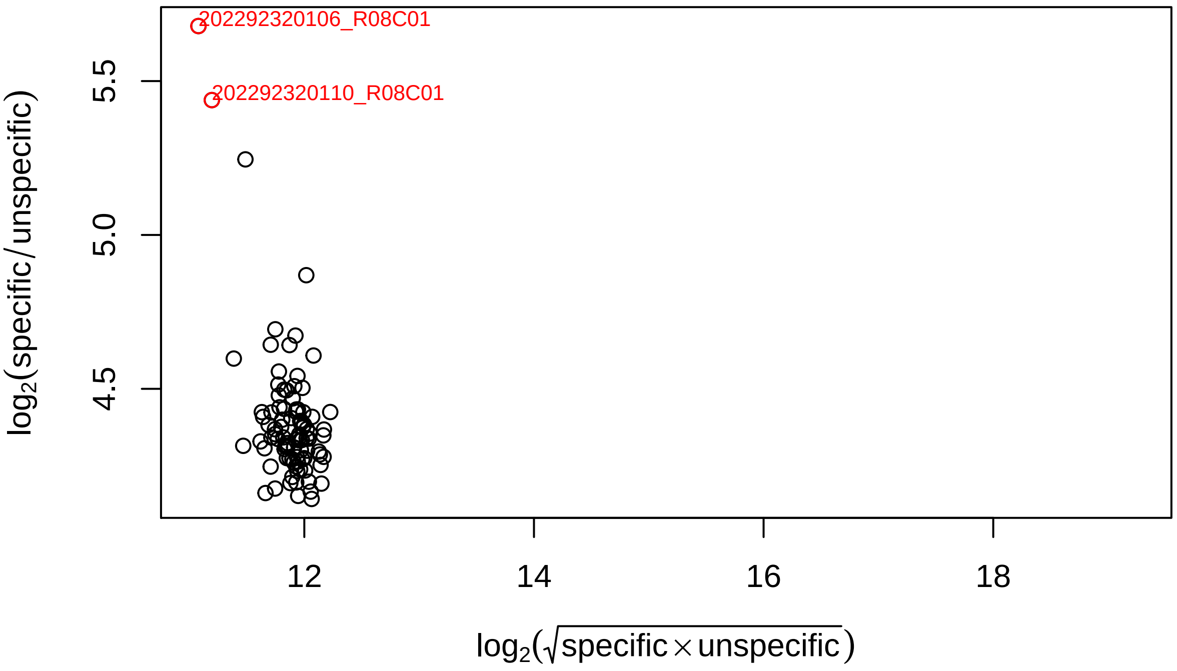


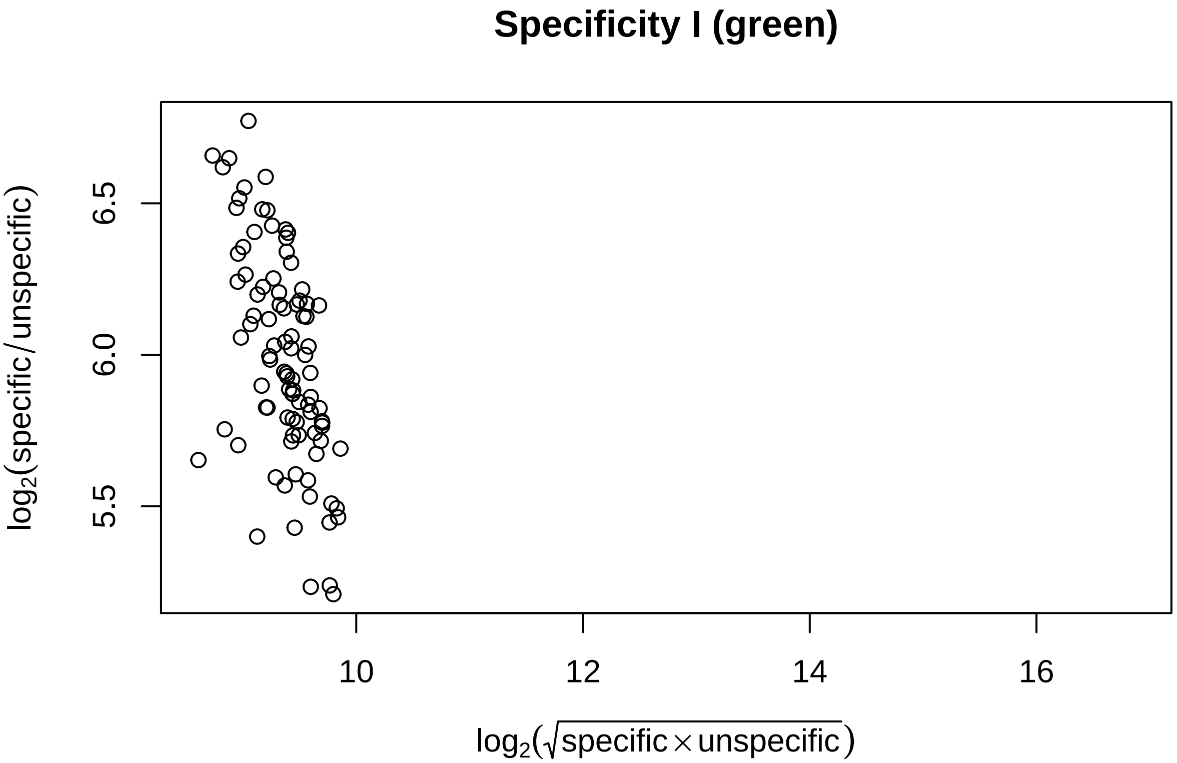
We also check the Hybridisation of the amplified DNA to the array:

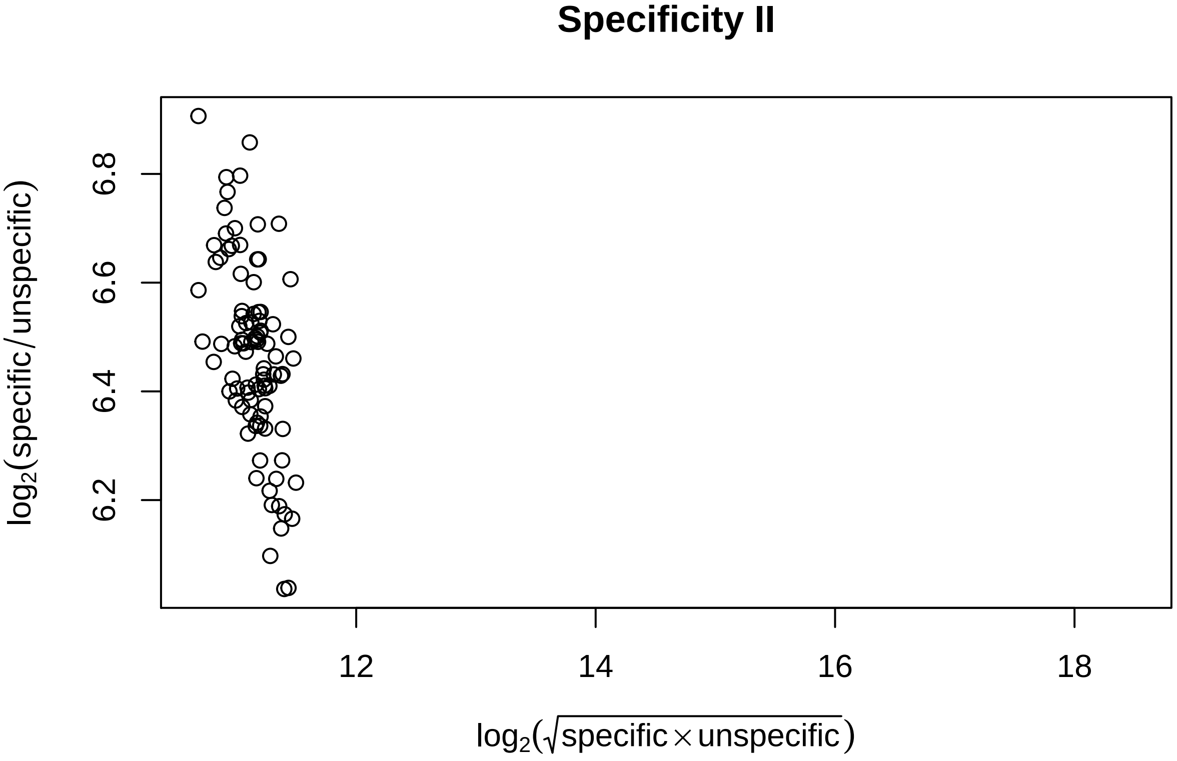
Hybridization (green)

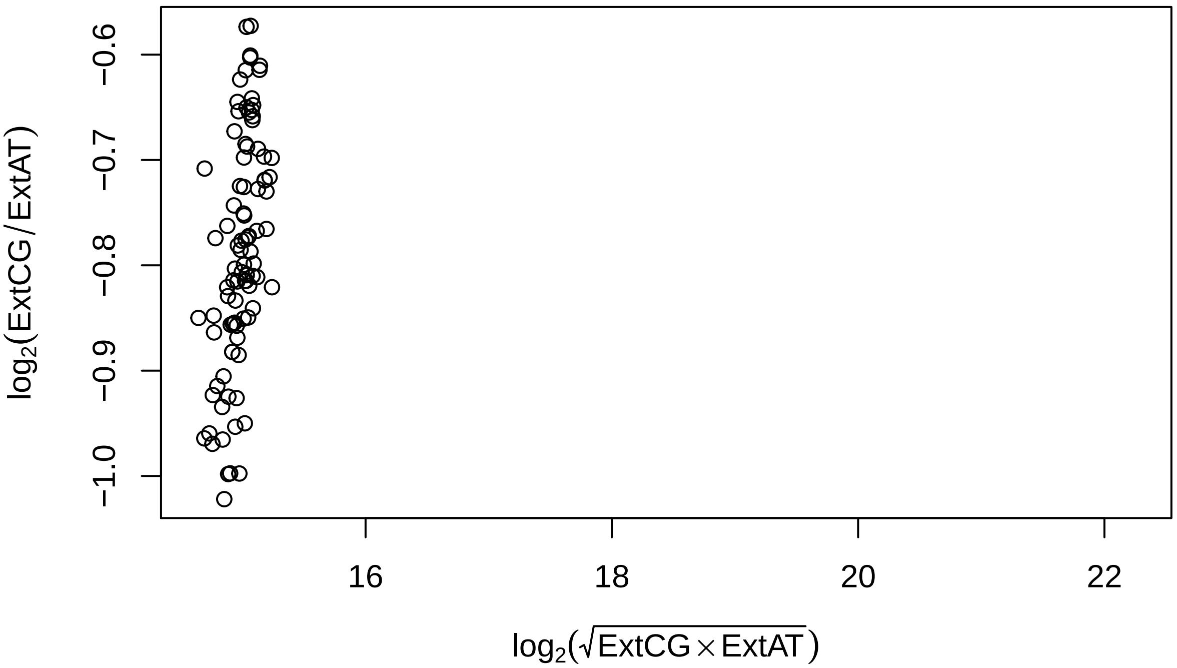


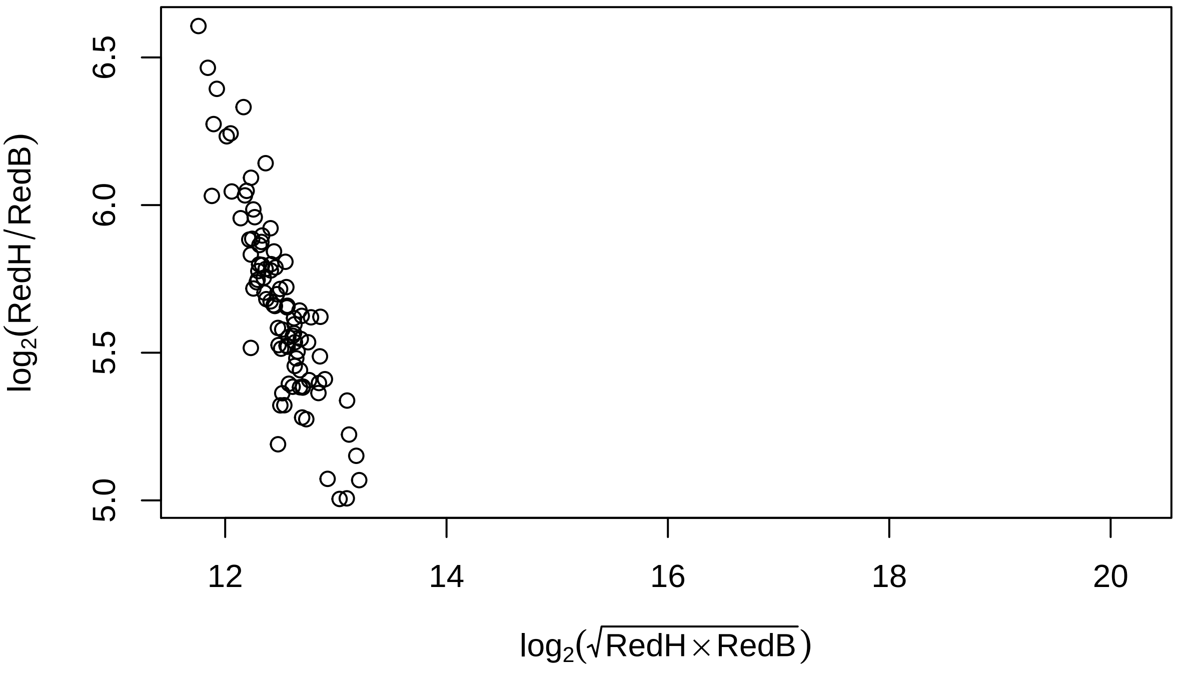
Specificity I (red)

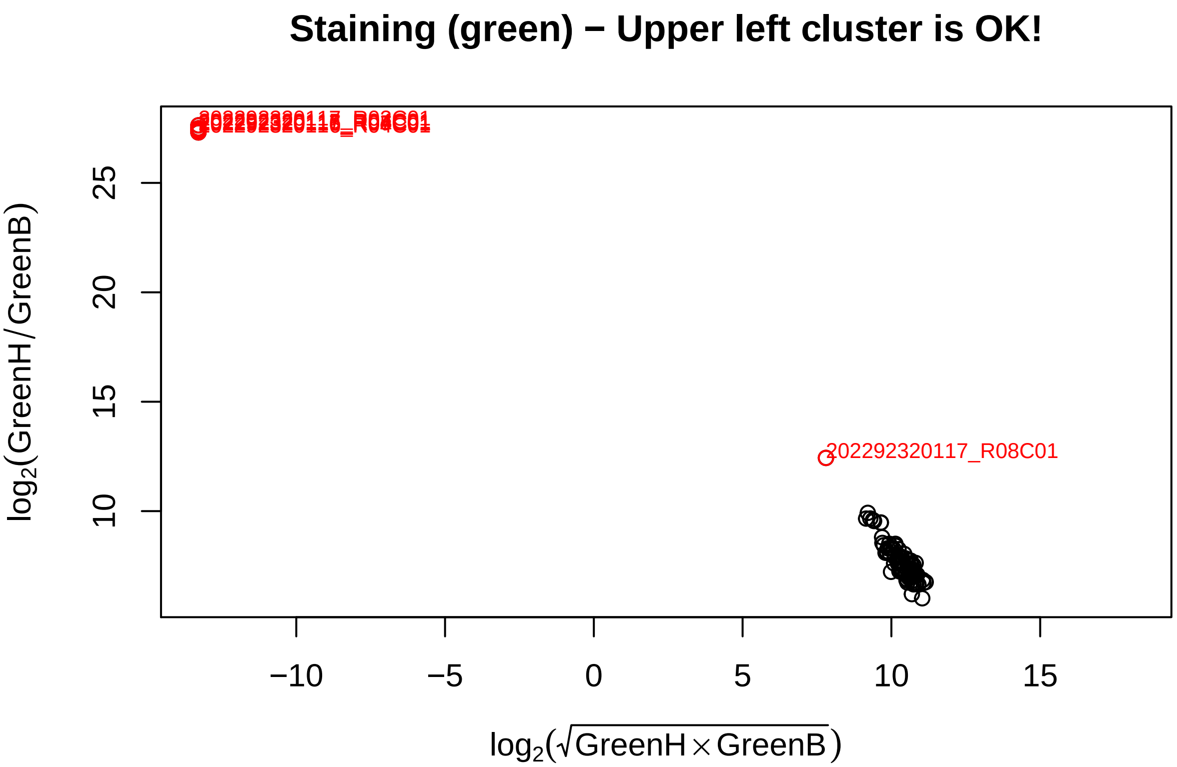



Specificity I (green)
6.5
6.0
5.5
$\log_2(\text{specific}/\text{unspecific})$
10
12
14
16
$\log_2(\sqrt{\text{specific} \times \text{unspecific}})$

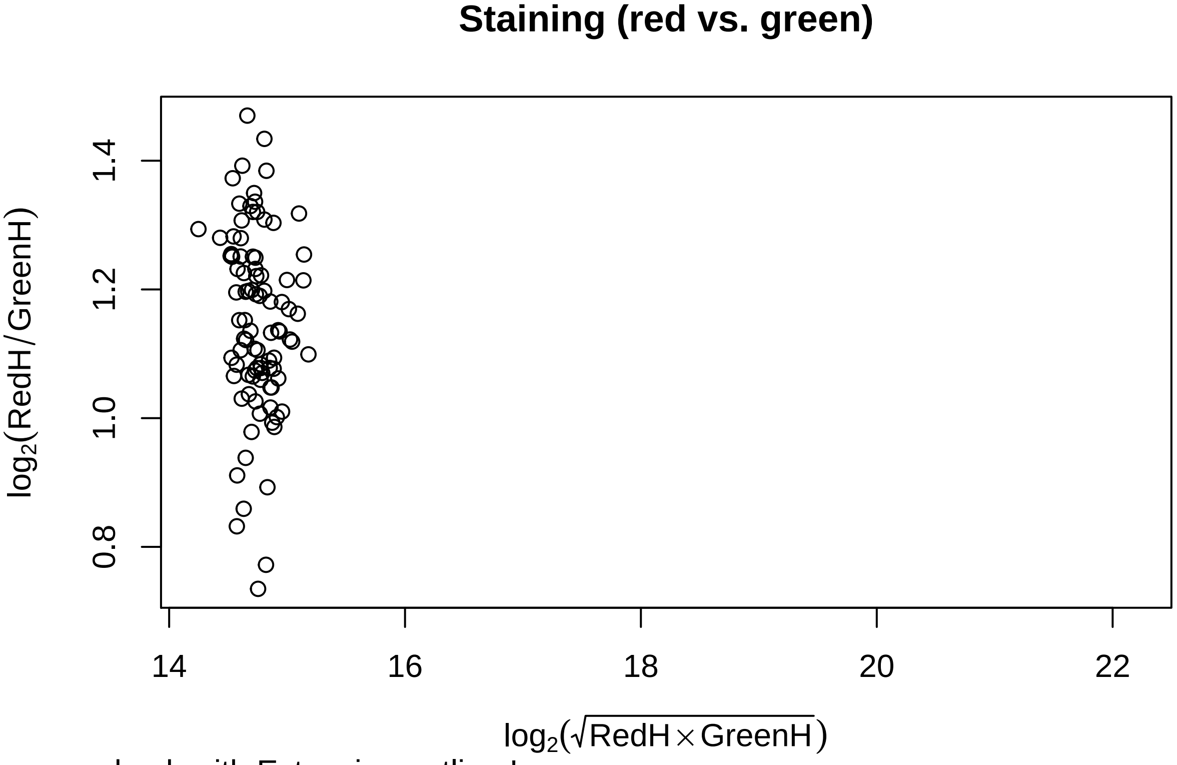

Specificity II
6.8
6.6
6.4
6.2
$\log_2(\text{specific}/\text{unspecific})$
12
14
16
18
$\log_2(\sqrt{\text{specific} \times \text{unspecific}})$


Extensions


Staining (red) − Upper left cluster is OK!



Staining (green) − Upper left cluster is OK!

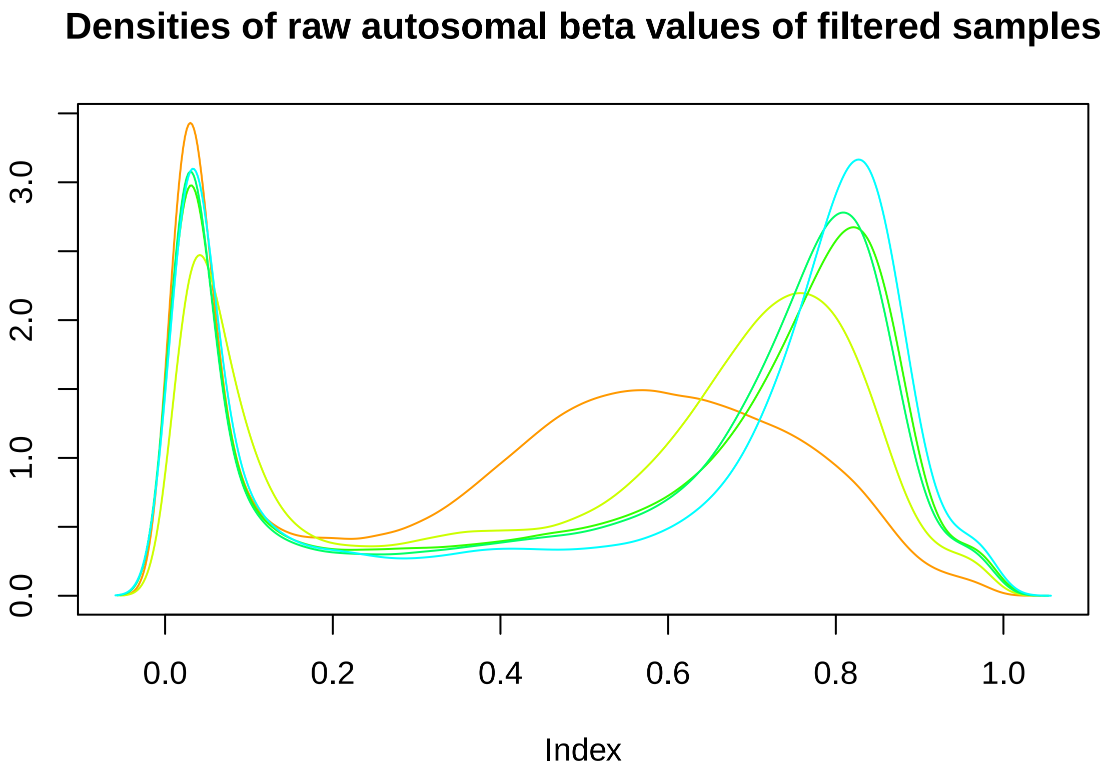


Staining (red vs. green)

cross−check with Extension outliers!

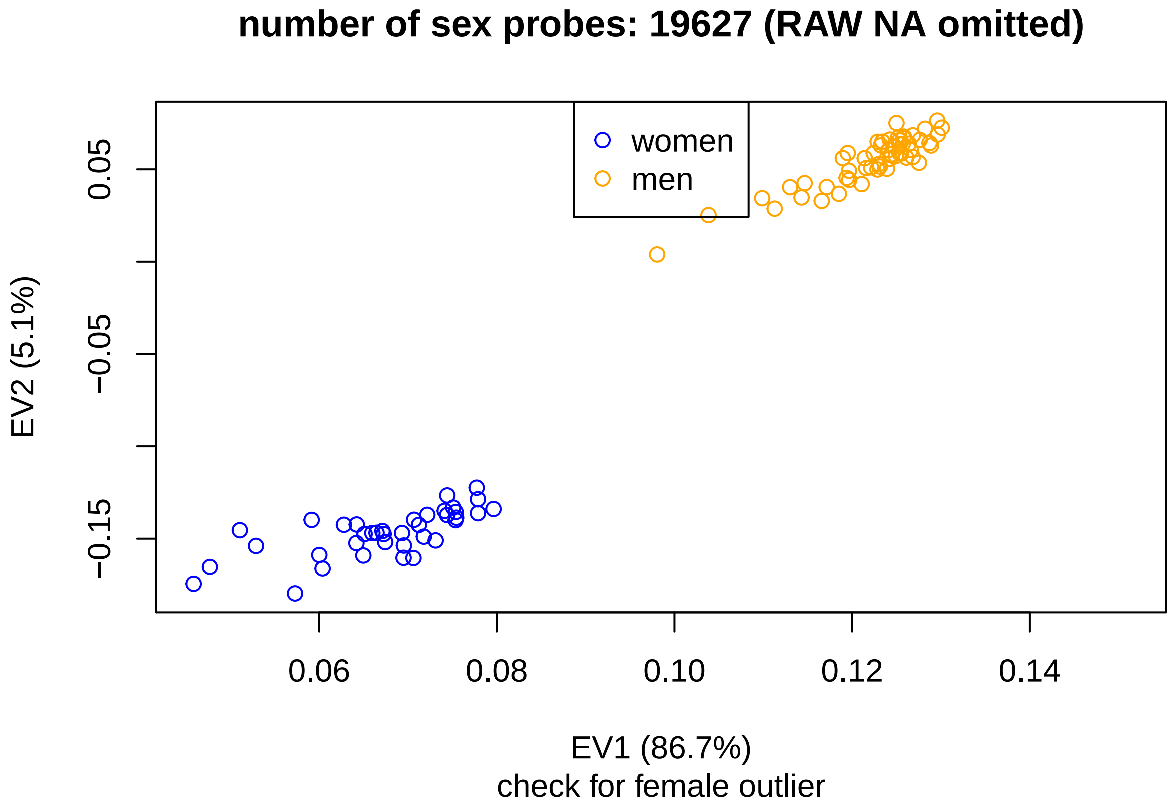
The following table lists the detected outliers identified by the quality control and can be found in the file **samples-filtered.csv** in a slightly expanded version.

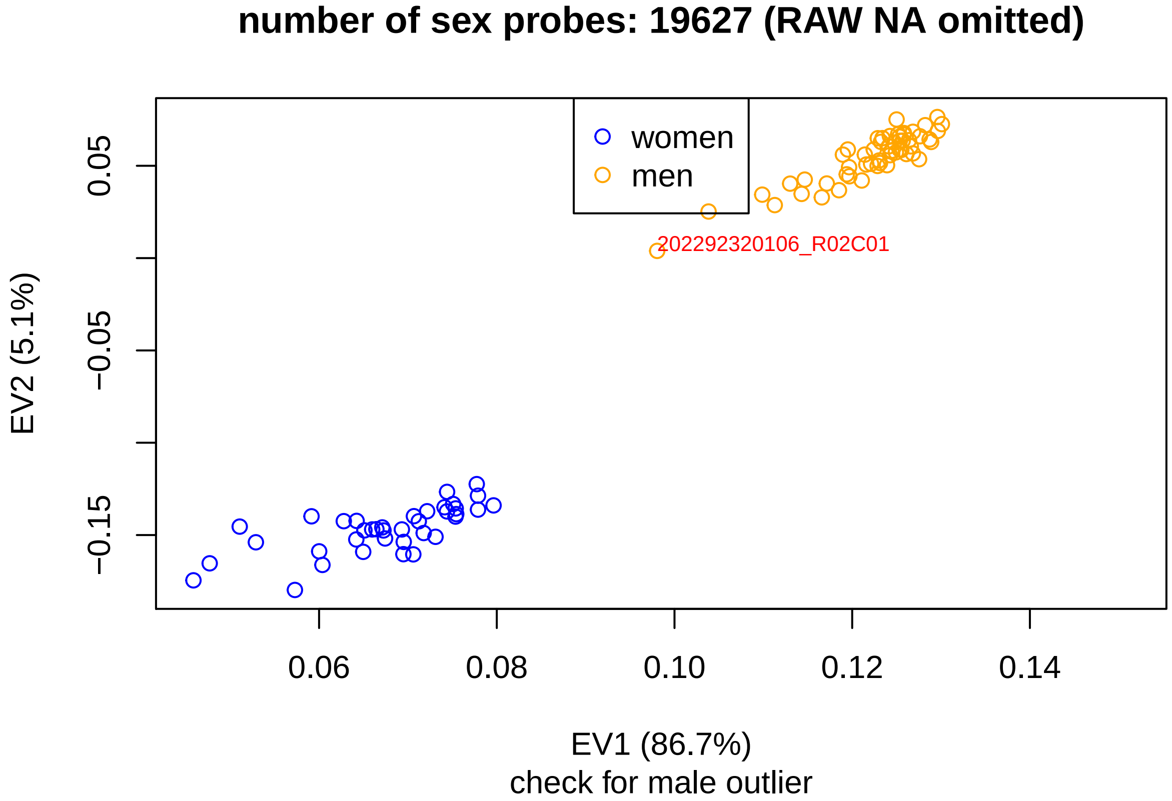
```
##   Sample_Name                                                    filter
## 1        A040                                              BS I-C (red)
## 2        A040                                        BS I-C (green)
## 3        D129                                        BS I-C (green)
## 4        D310                                   Specificity I (red)
## 5        E041                                   Specificity I (red)
## 6       SC204 Staining Green (cross-check with Extension outliers!)
##                sample         x         y Sample_Well Sample_Plate
## 1 202292320102_R02C01 11.725574  3.928046        <NA>       epic_1
## 2 202292320102_R02C01  9.304141  5.032475        <NA>       epic_1
## 3 202292320106_R01C01 11.751504  2.322770        <NA>       epic_1
## 4 202292320106_R08C01 11.077842  5.678874        <NA>       epic_1
## 5 202292320110_R08C01 11.194841  5.438215        <NA>       epic_1
## 6 202292320117_R08C01  7.800843 12.431711        <NA>       epic_1
##   Sample_Group Pool_ID           Sample_ID Gender MFGender  callrate
## 1         <NA>    <NA> 202292320102_R02C01      1        F 0.9512982
## 2         <NA>    <NA> 202292320102_R02C01      1        F 0.9512982
## 3         <NA>    <NA> 202292320106_R01C01      0        M 0.9905723
## 4         <NA>    <NA> 202292320106_R08C01      0        M 0.9802064
## 5         <NA>    <NA> 202292320110_R08C01      1        F 0.9867507
## 6         <NA>    <NA> 202292320117_R08C01      1        F 0.9918403
```



Sex mismatch

Gamete methylation can be used to check sex mismatches. First, we see the loadings of the PCA on markers in the space of samples. This loadings plot shows the eigenvectors (EV) multiplied with the eigenvalues. The loading indicates the importance of the variables for the PCs.

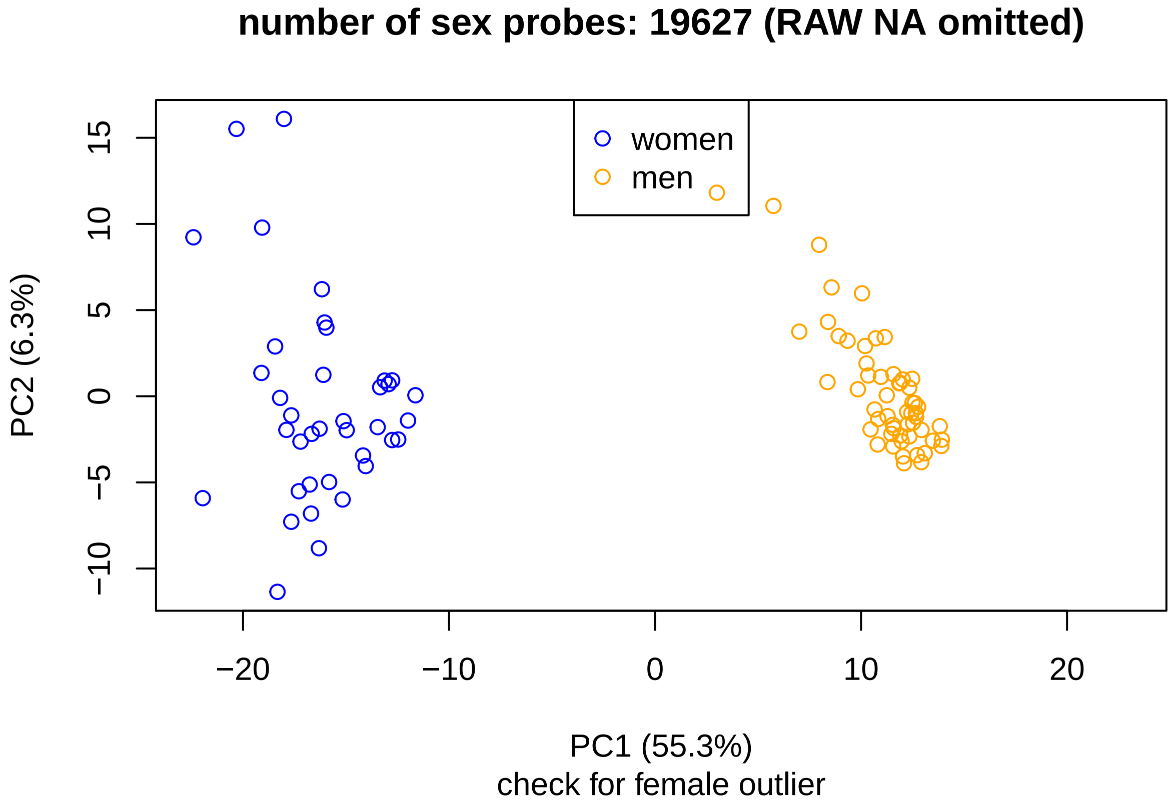


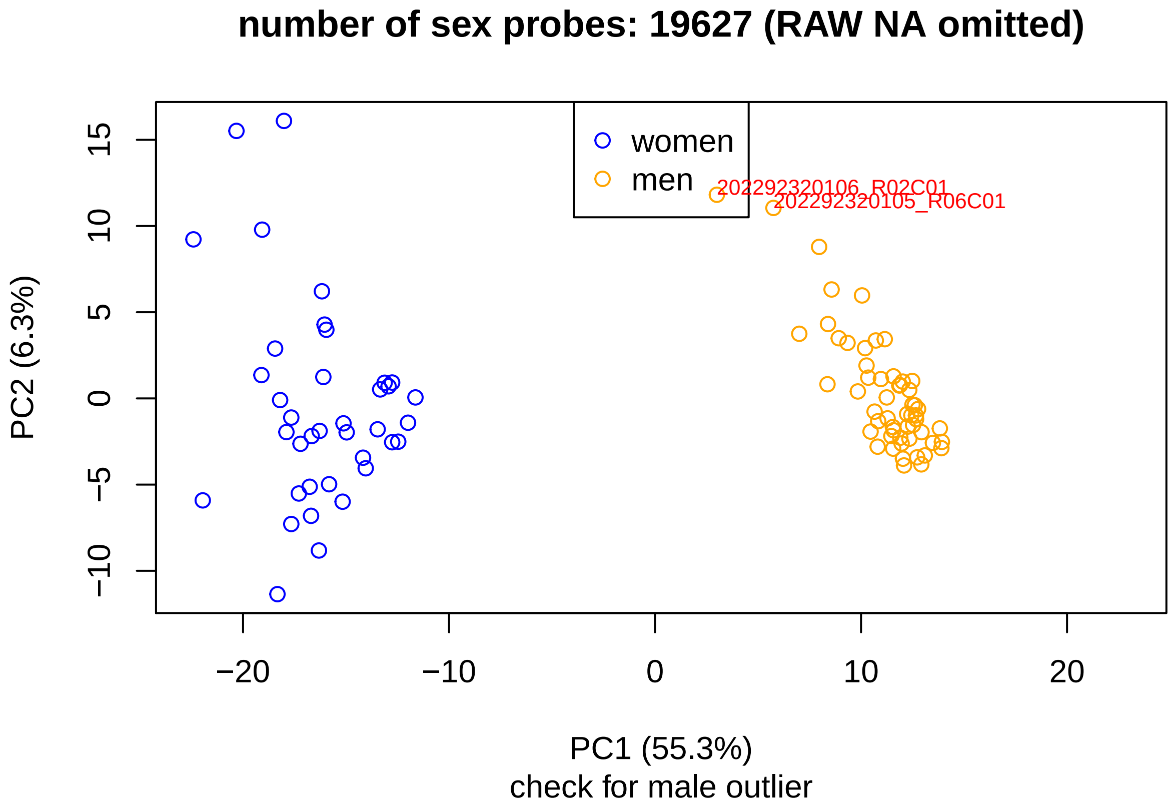


The samples whose loadings are more than three standard deviations from the mean of the loadings with same assigned sex according to the first two principal components:

```
## [1] "202292320106_R02C01"
```

The second plot shows the principle components of the PCA on the samples in the space of samples (a standard PCA plot).





The samples whose values are more than three standard deviations from the mean of the samples with same assigned sex according to the first two principal components:

```
## [1] "202292320106_R02C01" "202292320105_R06C01"
```

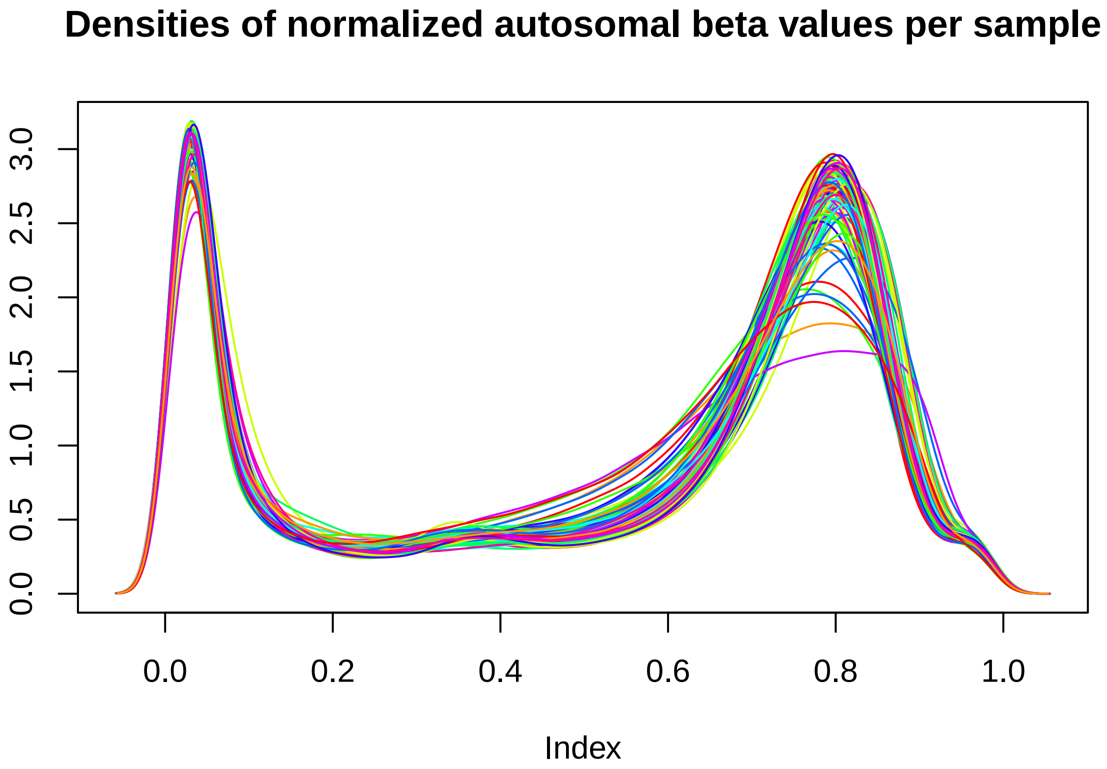
To further investigate possible sex mismatches, it might be useful to compare SNP alleles. Some of these are recorded on the methylation array for this purpose and can be obtained using the R-package `minfi`: First read the rgSet from the concerned samples and then run the function **getSnpBeta** from minfi to get a list of rs-identifiers and corresponding SNP information. These can be compared with genotype information from other sources to clarify whether samples were mixed up.

PART 3: Quantile Normalisation

The previous calculations provide all information needed to filter the samples and make a tab-separated file *samplesfilefinal* for further use. For this analysis, data/ studies/ 13_Collaborations/ 07_Flotho_JMML/ 01_analysis/ 01_input/ samplefile_JMML_epic_final.txt was used as list of samples for the final preprocessing steps.

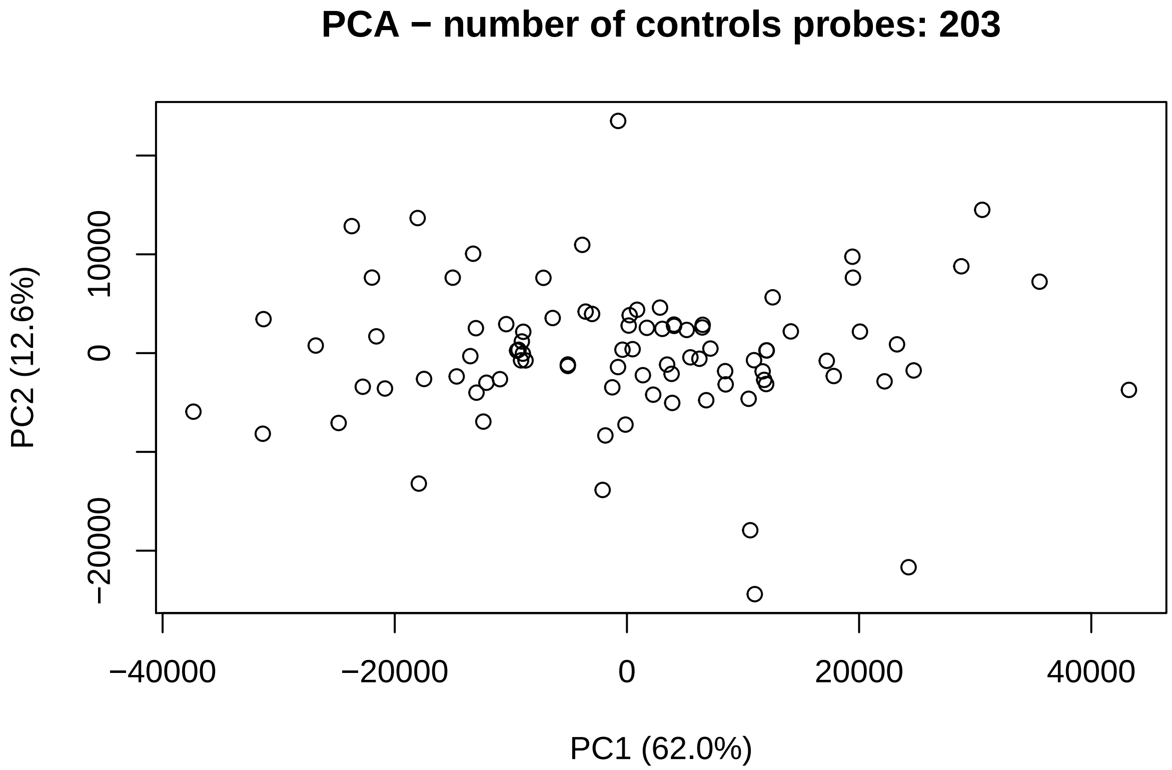
91 samples are included in the analysis, identified by the *samplesfilefinal* and existing sample calls.

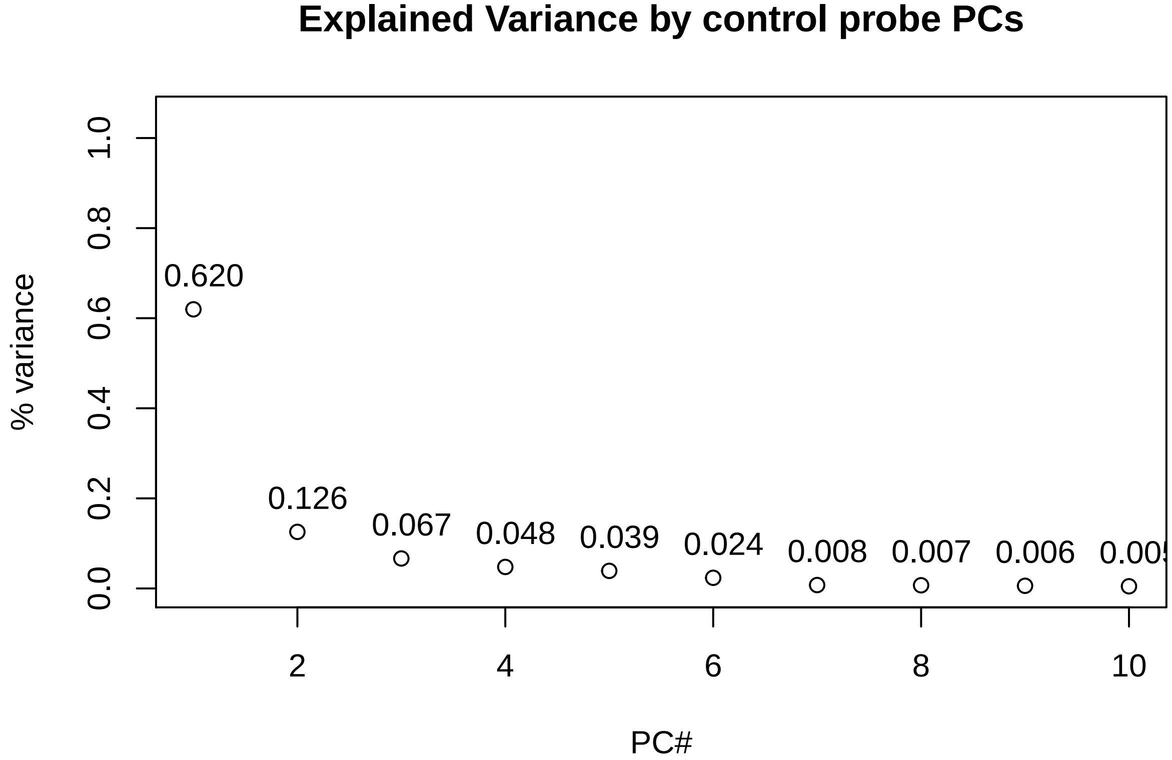
Of the persons from whom samples are processed, 37 are women and 54 men.



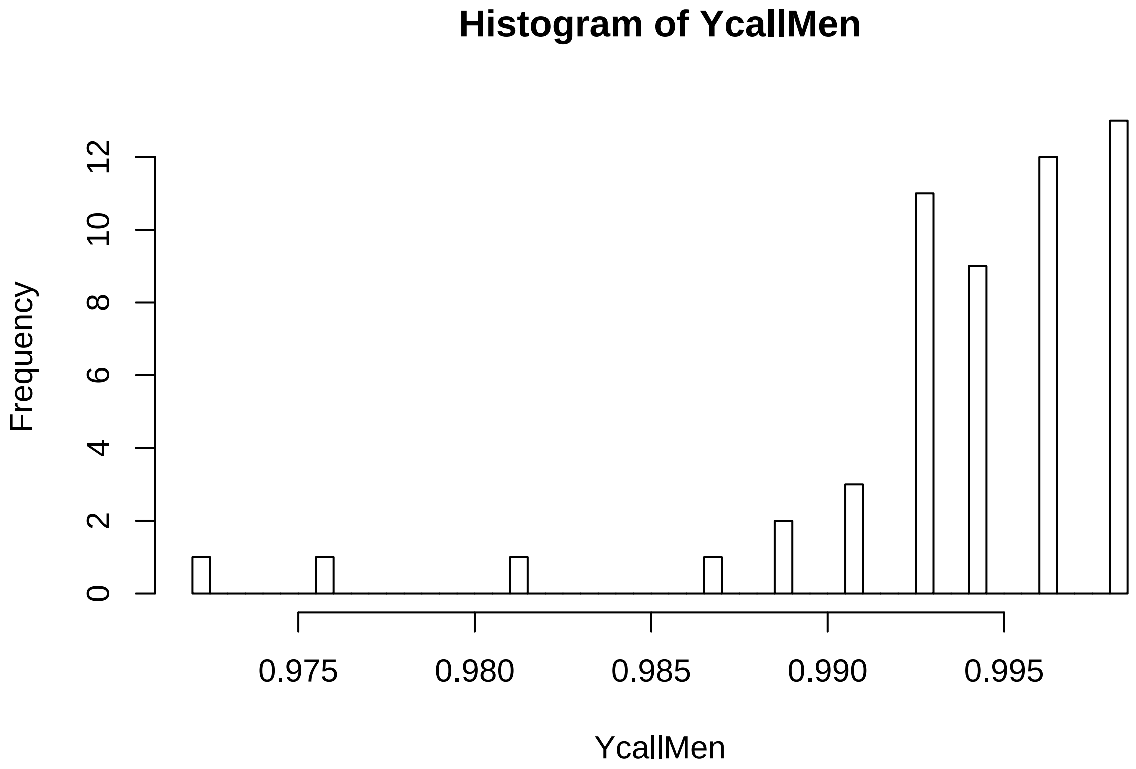
PART 4: Principal Component Checks

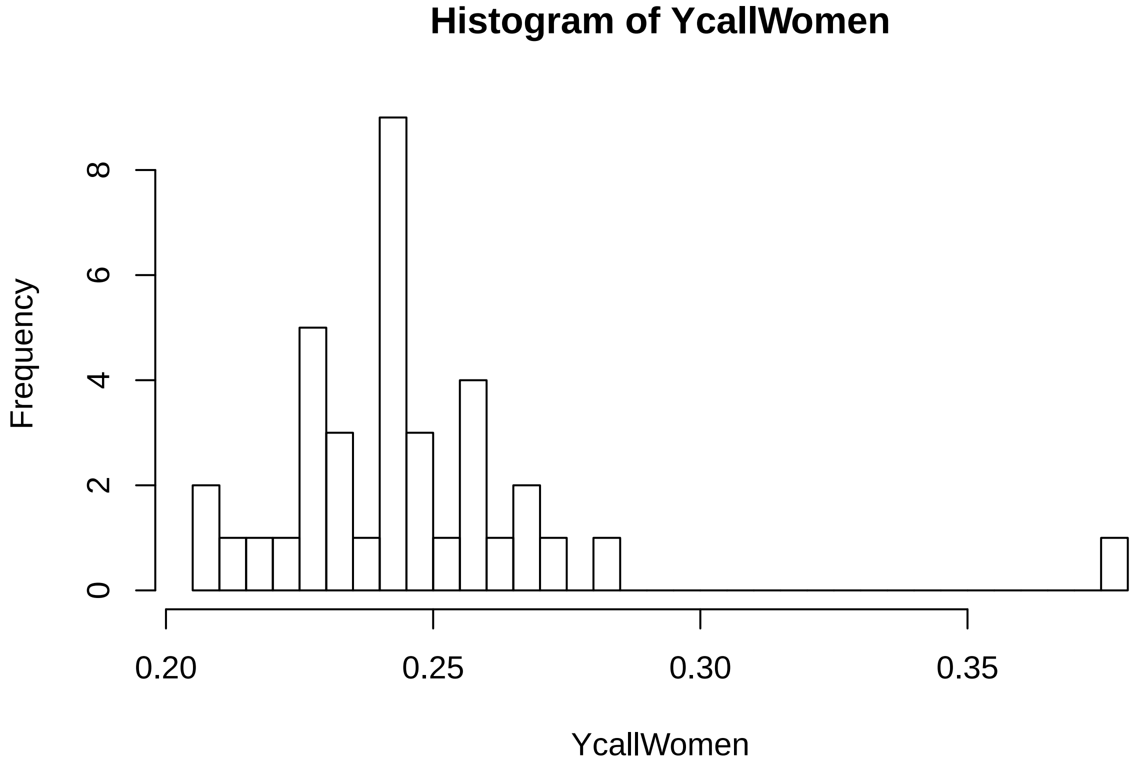
Control probes track technical bias between batches. To adjust for these in the analysis while minimizing the number of variables due to convergence reasons, we calculate Principal Components and show how much variance they explain. Markers which had at least one cpg site missing were excluded for the following PCA.

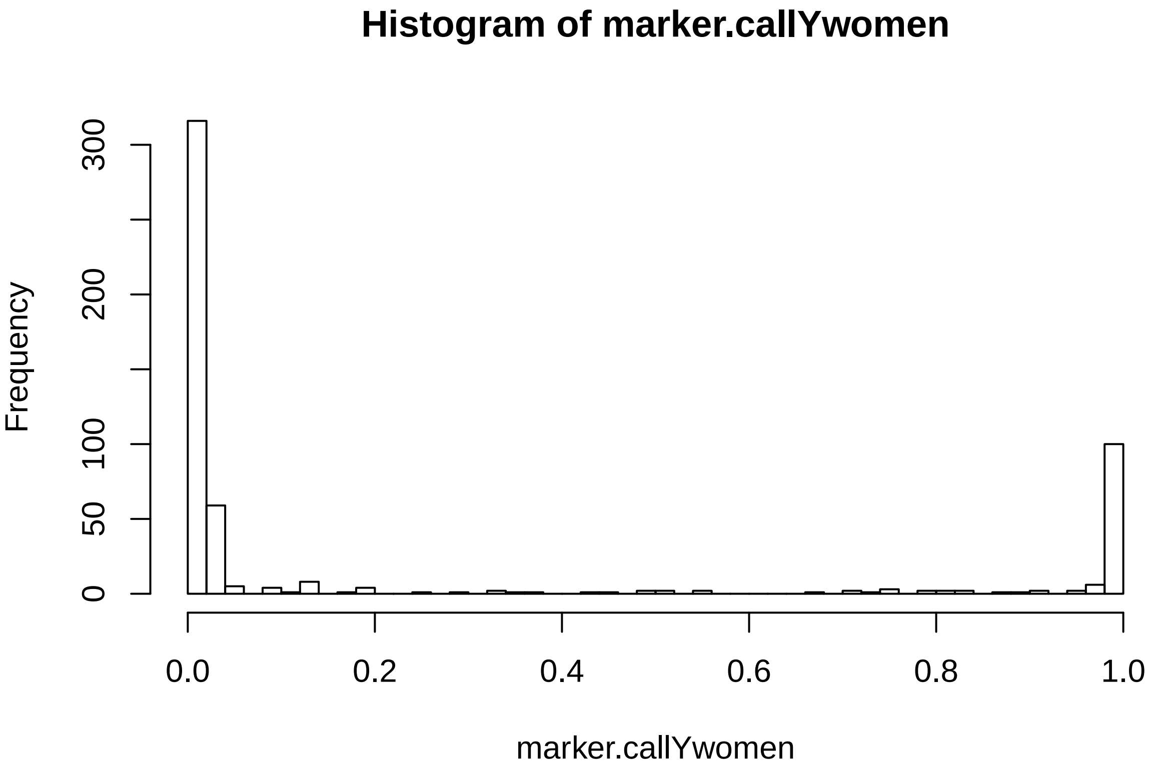




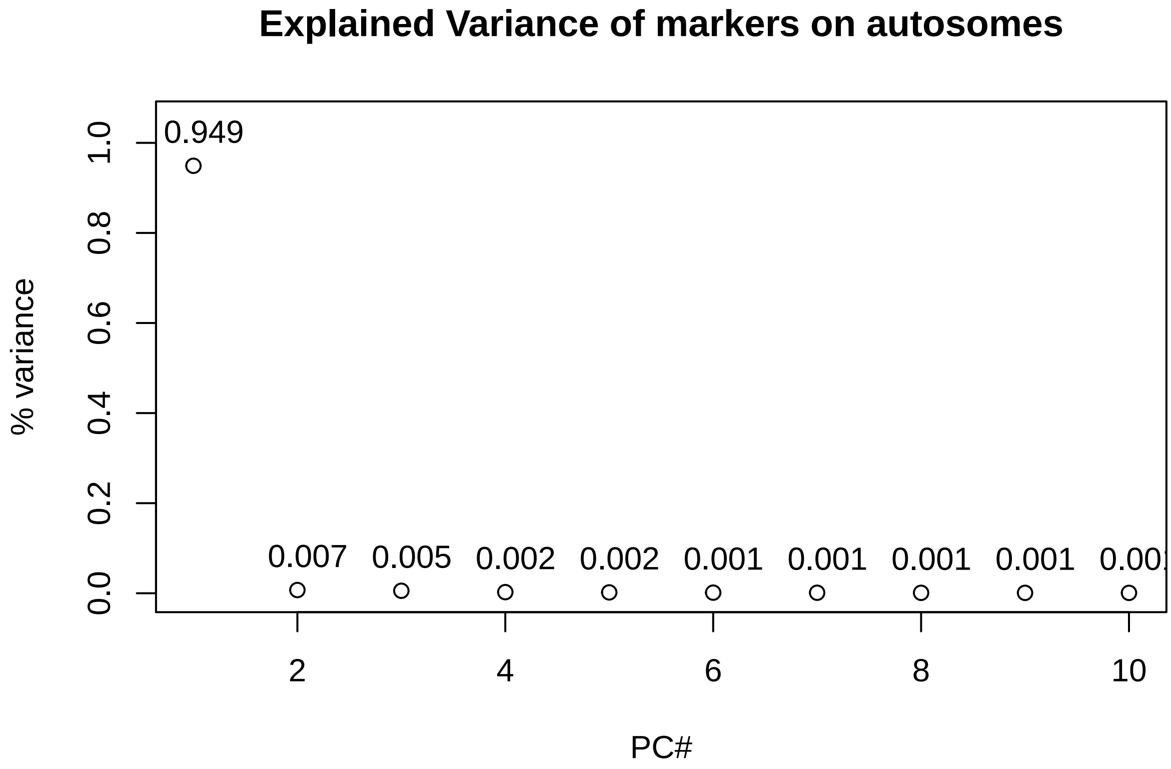
The following plots show the call rates of the sex chromosomes stratified per sex.


Histogram of YcallMen
Frequency
0
2
4
6
8
10
12
0.975
0.980
0.985
0.990
0.995
YcallMen


Histogram of YcallWomen
Frequency
0
2
4
6
8
0.20
0.25
0.30
0.35
YcallWomen

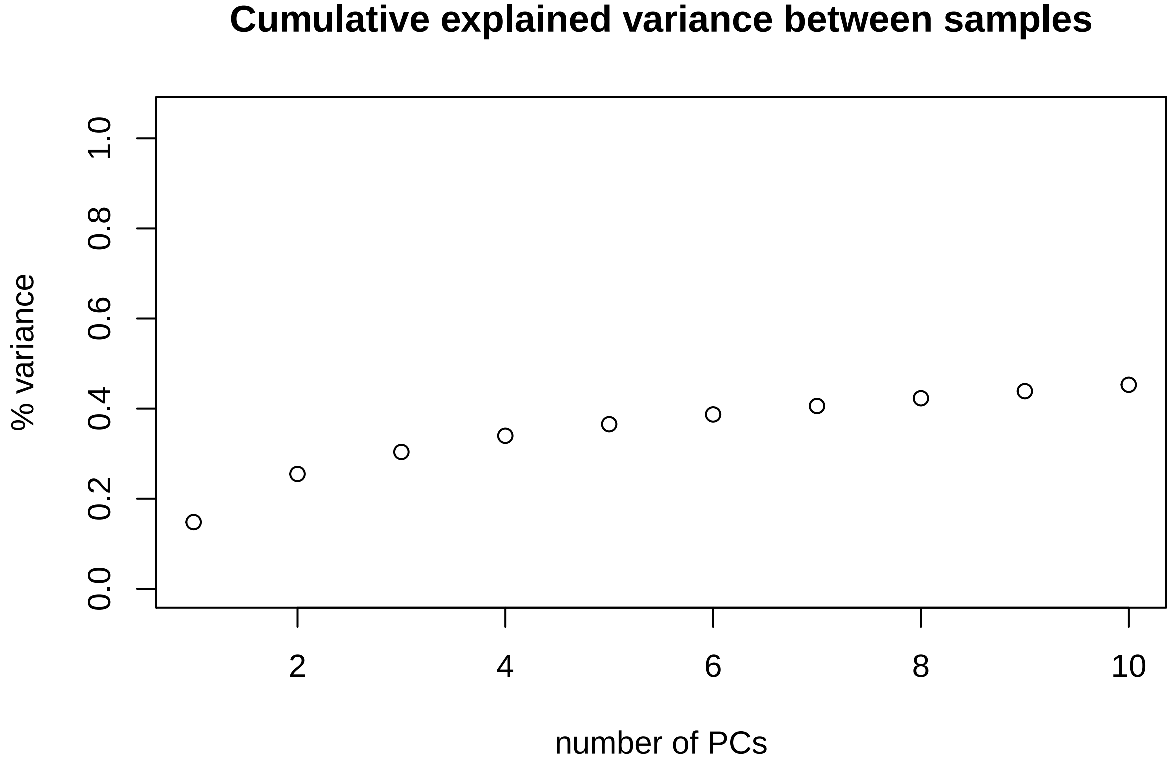


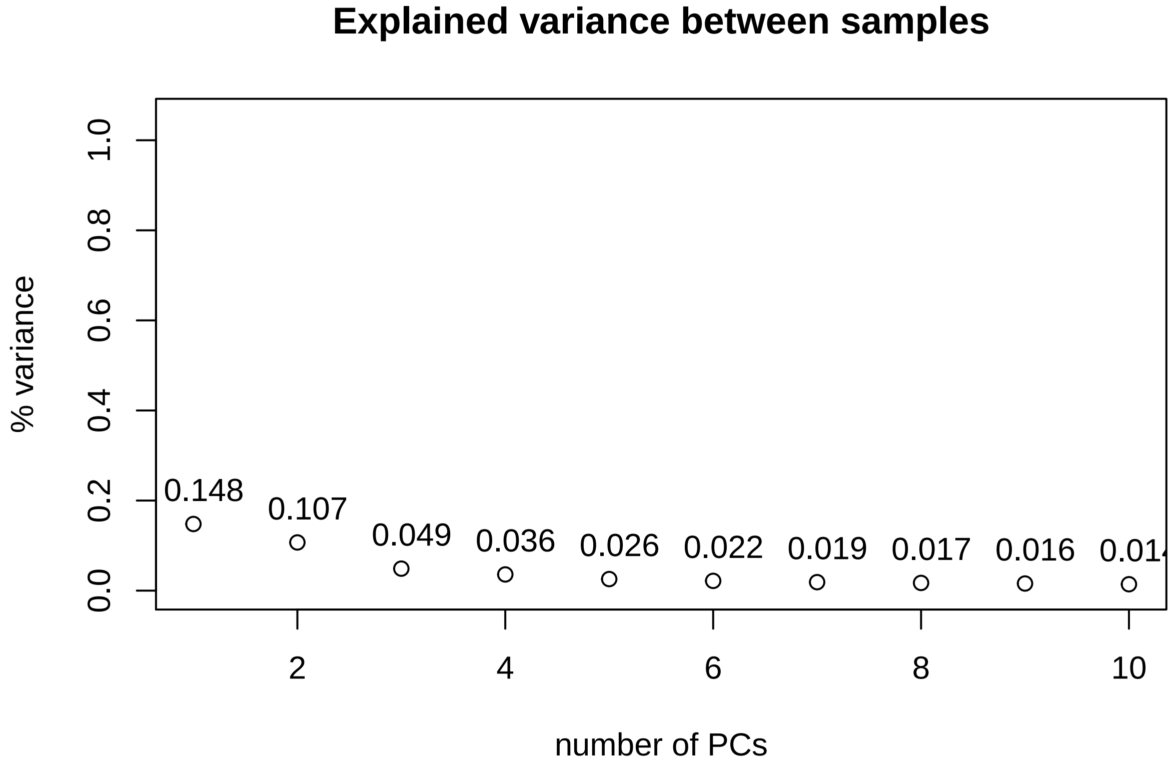
Now a PCA will be conducted on autosomal beta values after quantile normalisation. Here we investigate how well the markers can be separated by choosing different base vectors. Rows are observations, here markers.



The next PCA is also calculated on the quantile normalized beta values (omitting missing values), but now such that the first principal component explains most of the variation of the autosomal sample data. This is the same data as used in the plot before, but with transposed data (rows are observations, thus samples).

So the variation which is explained by the first pricipal component indicates how well the samples can be separated by choosing a new base vector.

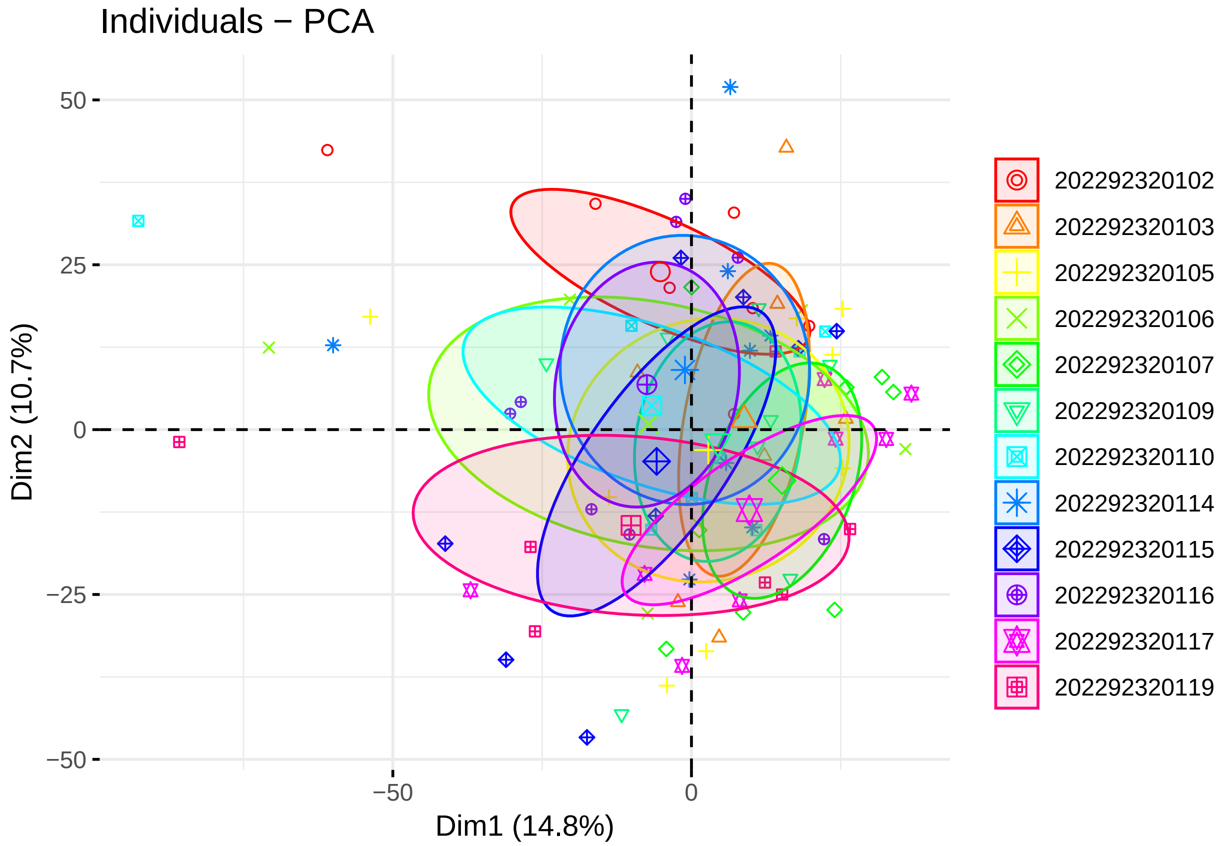


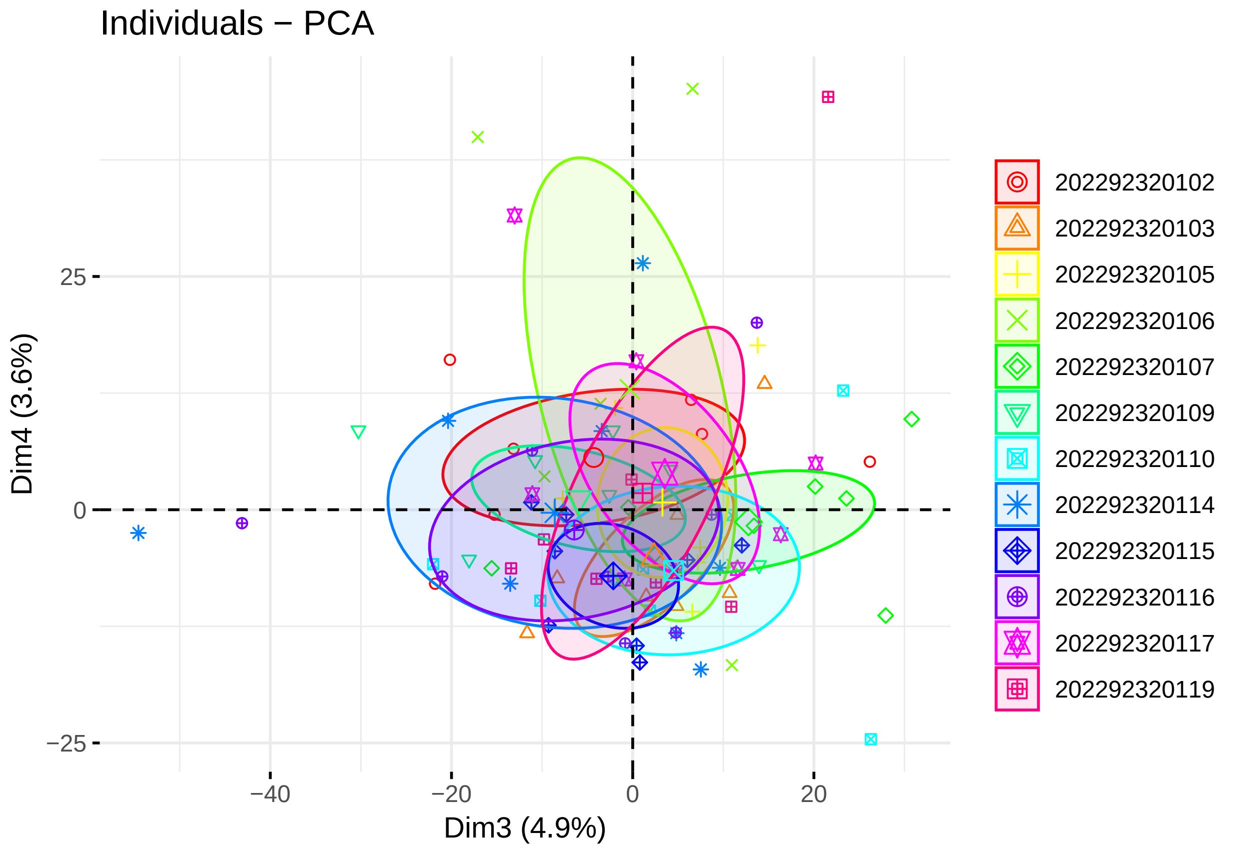


The following plots show these principal components coloured by Sentrix_ID and Sample_Plate. Because color coding for more than 25 groups is problematic, we split the batch variable in disjoint random groups of maximum 24 members and plot each group together with a 25th member who summarizes the other batches.

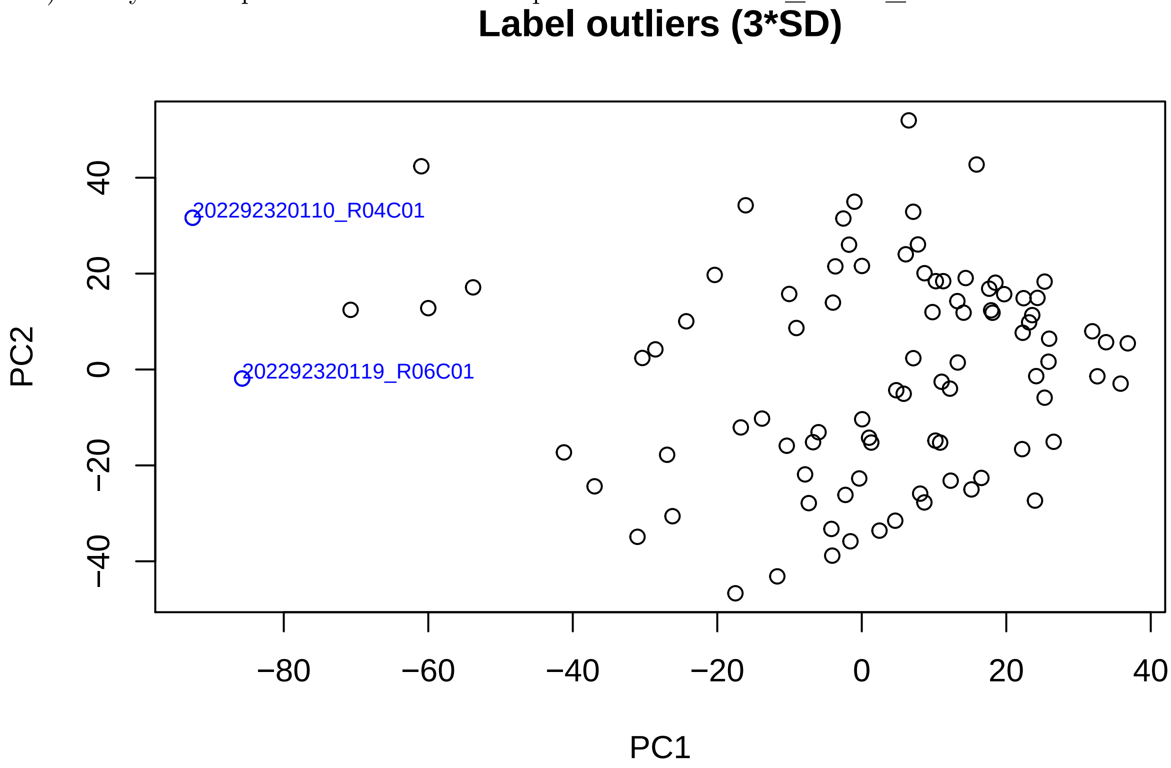
For each group of 24 batches, we plot the PCs 1&2 and 3&4.

First, we plot regarding *Sentrix_ID*:

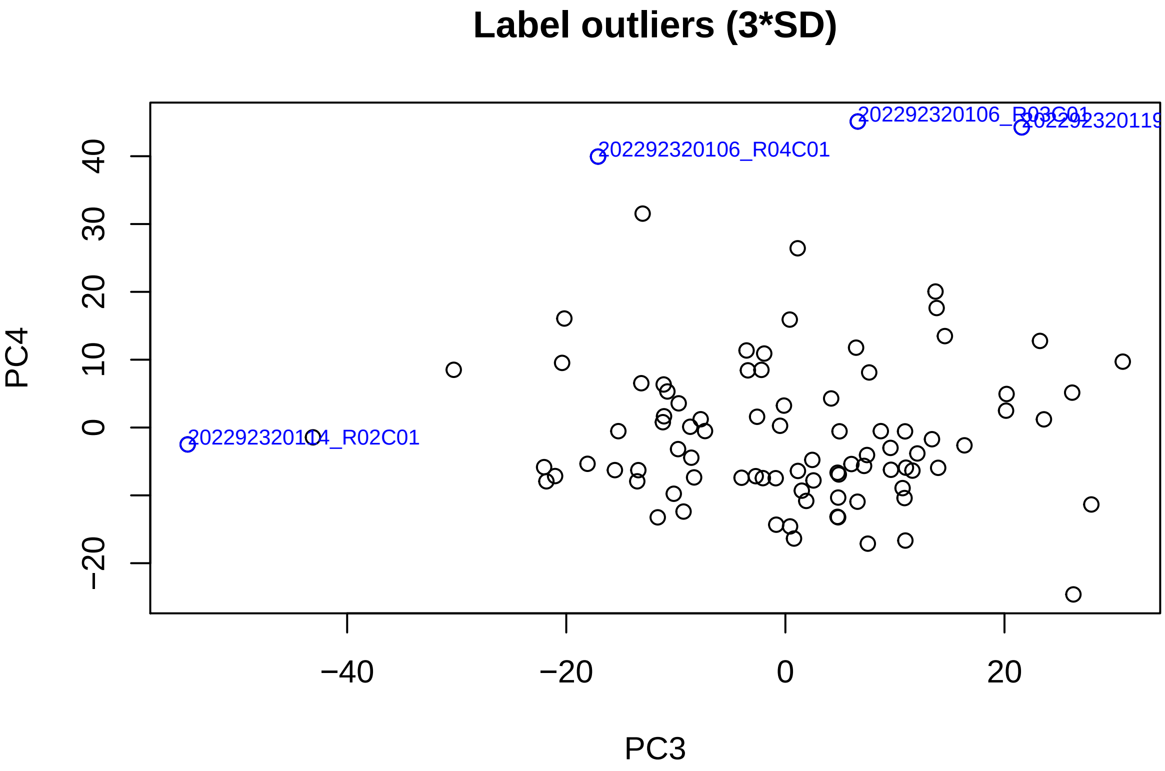




Additionally we supply the Sentrix_IDs of outliers (defined by three times the standard deviation per PC). They are exported to the file samples-with-batch-effect__JMML__EPIC-2019-04-10-15h57m.csv.



```
## [1] "202292320110_R04C01" "202292320119_R06C01"
```



```
## [1] "202292320114_R02C01" "202292320106_R03C01" "202292320106_R04C01"
## [4] "202292320119_R07C01"
```

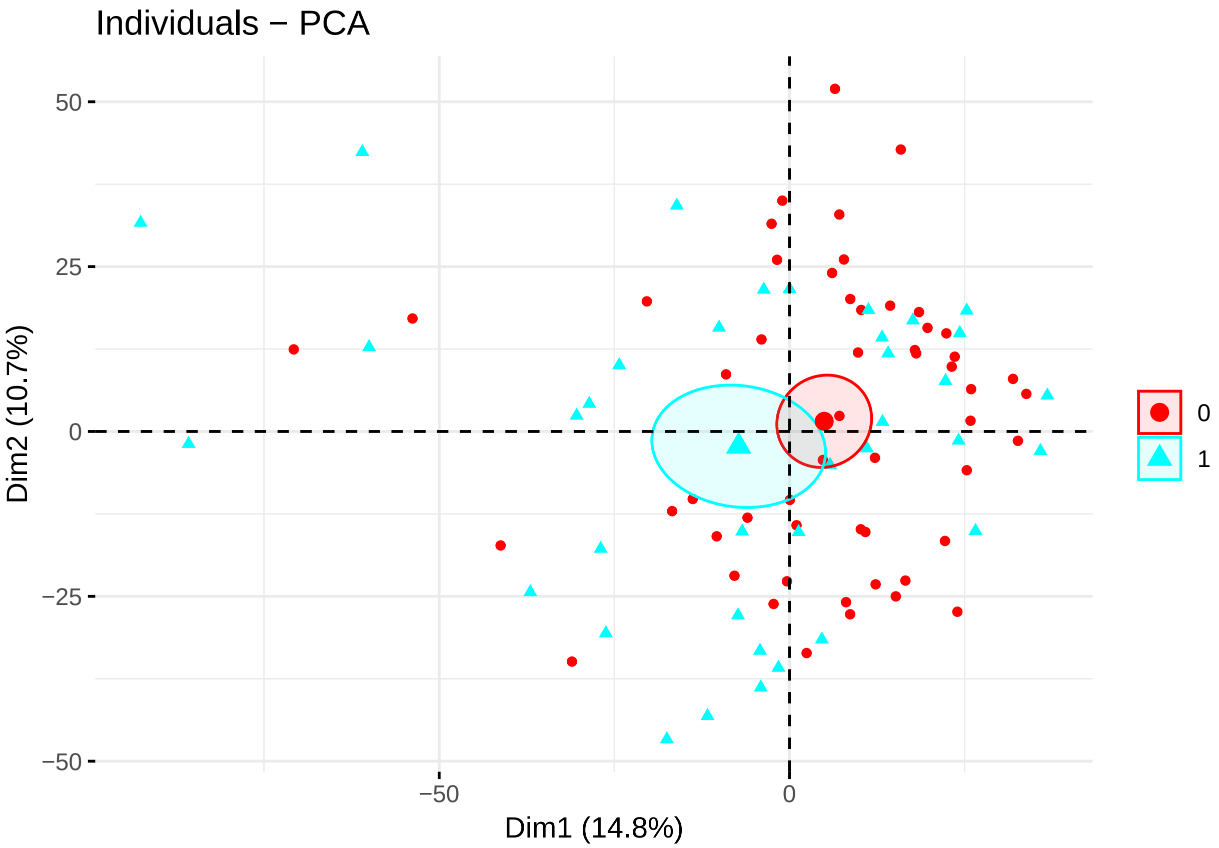
Next, we plot with respect to *Sample_Plate*:

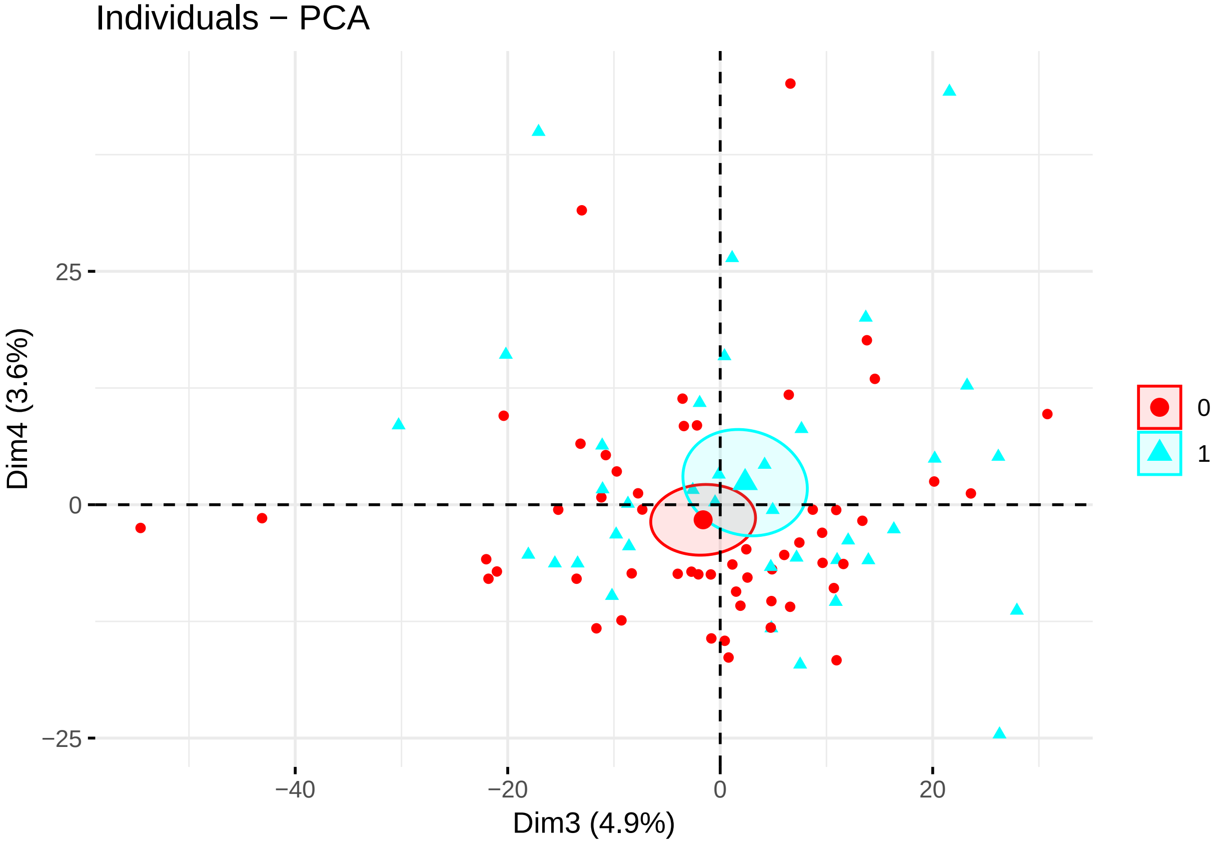
```
##
## There is no point in investigating batch effects with respect to  Sample_Plate
##  as there are less than two of these batches.
```

To better understand these results, compare the **number of samples per Plate**.

```
## epic_1
##     91
```

```
## Warning in read.table(BatchVariablesFile, header = FALSE, stringsAsFactors
## = F): incomplete final line found by readTableHeader on '/data/
## studies/13_Collaborations/07_Flotho_JMML/01_analysis/01_input/
## additionalBatchVariables.tsv'
```


Individuals − PCA
Dim2 (10.7%)
50
25
0
−25
−50
−50
0
Dim1 (14.8%)
● 0
▲ 1


Individuals − PCA
Dim4 (3.6%)
25
0
−25
−40
−20
0
20
Dim3 (4.9%)
● 0
▲ 1

Output

1. The following data is included in the Rdata-file QC_data_JMML_EPIC_2019-04-10-15h57m.Rdata:

   ctrl.all, ctrl.complete.Red.all, ctrl.complete.Green.all, control.info control probe data

   ctrlprobes.scores rotated control probe intensities into the coordinates given by principal component analysis

   dp.all detection p-values of unfiltered idat data

   TypeII.Red.All etc list the corresponding intensity values

   TypeII.Red.All.d etc list the corresponding intensity values where detection p-values smaller than threshold 10^{-16} are set to missing (NA).

   sample.call the sample call rates

   marker.call the marker call rates

   beta.raw the β values calculated for autosomal probes with detection p value filter only

   beta.raw.sex the β values calculated for sex chromosome probes with detection p value filter only

   pcaAuto output of function prcomp conducting PCA of **betaQN**

   sample.callY, sample.callX sample call rates for chromosomes X and Y for quantile normalized filtered samples

   marker.call.all marker callrate of all quantile normalized samples incl. the ones excluded by IQR-filter

2. The data needed for further analysis is saved as analysis_ready_JMML_EPIC_2019-04-10-15h57m.Rdata:

   pcaControls output of function prcomp conducting the PCA of control probes

   betaQN β values calculated for autosomal probes of filtered samples with quantile normalization applied

   betaQN.sex β values calculated for sex chromosome probes of filtered samples with quantile normalization applied

   betaQN.all combines **betaQN** and **betaQN.sex**

   est.wbc.minfi the white-blood-cell estimations by minfi method based on RGsets

Memory load and processing time

The maximum memory load in this run was 1.84865×10^4 Mb . It took 22.64703 mins of processing time.

Methods draft

For processing and quality control of the raw methylation data, a customized version of the CPACOR pipeline < PMID: 25853392> was used for quality control, data normalization and calculation of beta values, calculating principal components of the control probes for adjustment and exclusion of outliers based on the Inter-Quartile-Range. The threshold for the sample call rate was set to CHANGE. White blood cell type (WBC) sub-populations were estimated based on 100 CpG sites by the Houseman method REFERENCE as implemented in the minfi R package REFERENCE. XXXXX samples discordant for reported and genetic sex, based on CpGs on the X- and Y-chromosome, was excluded from analyses. Additionally, quality control based on principal component analyses of the control probes was conducted to detect samples with measurement failures.

Credits

The code for the CPACOR analysis pipeline was adapted from Lehne et. al. (Genome Biology, 2015) which was developed and written by Benjamin Lehne (Imperial College London) and Alexander Drong (Oxford University). The code for the low-level quality control was developed and written by Alexander Teumer (University Medicine Greifswald/ Erasmus MC Rotterdam). The code was combined into the current pipeline by Pascal Schlosser and Franziska Grundner-Culemann. It is available at https://github.com/genepi-freiburg/Infinium-preprocessing.

See *A coherent approach for analysis of the Illumina HumanMethylation450 Bead Chip improves quality and performance in epigenome-wide association studies* by Lehne et. al., Genome Biology (2015) for the basic idea. The method was then extended to EPIC arrays. Please cite this article in your publication.