NR Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Preprocessing of gene-expression data related to breast cancer diagnosis

Note

Note no.	SAMBA/35/14
Authors	Clara-Cecilie Günther, Marit Holden, Lars Holden
Date	28. okt. 2014

NR Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Norsk Regnesentral
Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Title	**Preprocessing of gene-expression data related to breast cancer diagnosis**
Authors	**Clara-Cecilie Günther, Marit Holden, Lars Holden**
Date	28. okt. 2014
Year	2014
Publication number	SAMBA/35/14

Abstract

The work is performed in close cooperation with the University of Tromsø and professor Eiliv Lund and is financed by the ERC TICE project. This note describes the preprocessing steps of gene expression data and focuses particularly on the filtering and normalization steps as the choices made here greatly affects the set of probes used in later analyses. In the filtering step, two parameters are set. Firstly, a cut-off for the detection p-value for each probe is set, and a probe is present in a given sample if its detection p-value is smaller than this cut-off. Secondly, the present limit is set. It is used to decide in how many samples a probe has to be present in order to be included in the dataset. The results show that a p-value cut-off at 0.01 and a present limit at 0.01 are reasonable choices. After filtering, the data can be normalized. Four different approaches are evaluated, and for the available dataset, quantile normalization of the data on original scale gives the most stable results.

Keywords	Gene expression data; Breast cancer; Filtering; Normalization
Availability	Open
Project number	220641
Research field	MMH
Number of pages	41
© Copyright	Norsk Regnesentral

Table of Content

1 Introduction

This note describes the pre-processing steps of gene expression data related to breast cancer diagnosis. The data have been provided by the group of professor Eiliv Lund at the University of Tromsø. The hypothesis is that genes related to the stages of cancer development could be differentially expressed over time, perhaps in a small but consistent manner. Marit and Lars Holden [3] have started developing methods for testing whether there is such a development in time, and for identifying groups of genes with similar behavior, or functional form, the last years before diagnosis. Hence, they are looking for weak signals from a large number of genes in contrast to stronger signals from a few genes. Through their analysis they have found that as the data are updated and with slightly different choices in the preprocessing steps, the subsets of genes that are selected, differ for the different versions of the dataset. To better understand the importance of the choices in the preprocessing steps, we have looked into what choices are being made and the effect of each of these on the resulting set of genes.

The dataset consists of a data matrix of gene expression values in blood cells, a data matrix with negative controls and background information that contains clinical and technical data for each sample. Each case sample is matched to its corresponding control through the case/control identifier given in the background information.

The main steps in the data cleaning process are:

1. Remove probes for genes involved with blood type.

2. Remove case-control pairs when either the case or control is an outlier with respect to quality, and remove the pairs of the following samples:

- Cases without breast cancer
- Cases with other types of cancer
- Cases with previous types of breast cancer
- Controls with breast cancer

3. Perform background correction using negative control probes.

4. Filter out probes that are not detectable or sufficiently present and translate from probes to genes.

The outlier detection in step 2 is done by using density plots, principal component analysis and inspection of laboratory quality measures. The background correction is done by using negative control probes, i.e. probes that should not be expressed. An offset is added to the background corrected data, but changing this offset does not affect the succeeding steps. Step 1-3 are described in more detail in Section 5. In the remainder of this report we will focus on step 4. Section 2 describes the parameters in this step and Section 3 describes the effect of varying the different filtering parameters and the normalization methods.

2 Filtering and mapping to genes

To decide which probes should be filtered out, three aspects must be considered.

Firstly, it is determined which probes are detectable. Each probe has a corresponding p-value for each sample that indicates whether it is detectable or not for that sample. A p-value cut-off is chosen, and the probes with detection p-values less than or equal to this cut-off are detectable. A higher p-value cut-off results in more probes being detectable. This cut-off will be denoted the p-value cut-off in the remainder of this report. In current analyses it is set to 0.01.

Secondly, detectable probes should be present in a certain number of samples to be included in further analysis. The present limit is the percentage of samples for which a probe is detectable. A lower present limit results in less probes being removed. Currently this limit is set at 0.01, i.e. a probe has to be detectable in 1% of the samples to be present. This threshold will be denoted the present limit in the remainder of this report.

Thirdly, each probe is mapped to a unique Illumina ID, which is a necessary step before mapping to genes, as different chips have different probe ids. Probes with poor quality of this mapping are filtered out.

The final analysis is done on gene level. A gene can be represented by several probes, and one probe has to be chosen for each gene. In a set of probes that belong to the same gene, the interquartile range (IQR) value is calculated for each probe across all samples. The probe with the highest IQR value is chosen, and the other probes are removed. It is possible to select another measure than the IQR value.

The prospective study consists of three runs. Until now the mapping to genes has been done for each run separately. However, for a specific gene, different probes may be chosen in each run. Even though the analysis is done on gene level, the identifiers in the data are still probes. When combining datasets from different runs, the probes that are not present in all runs will be removed, and this means many probes may be lost that otherwise could have been kept. This mapping should therefore be done after the runs have been combined. It is also possible to do the analysis on probe level, however the interpretation of the results will then be more difficult.

In practice, one would do the filtering before mapping probes to genes. Mapping probes to genes more than halves the number of probes. In order to study the direct effect of changing the filtering parameters, we therefore perform the mapping before the filtering in the evaluations in this report.

3 Effects of changing filtering parameters

There are several settings that can be changed in the filtering process. We want to evaluate how these settings affects the set of genes that are used in the final analysis.

The outlier removal steps are assumed to be done appropriately, and we will not change anything in this part of the preprocessing. The background correction also seems reasonable, there is no need to use another method, and changing the offset (the only parameter in this step) does not affect the filtering process. Therefore we only consider the filtering step.

In this evaluation we focus on the dataset from run 1 of the prospective study. After outlier removal, there are 588 samples left, i.e. 294 case-control pairs. The dataset contains 39 388 probes. The data are background corrected as described in Section 5.7, probes with bad quality Illumina mapping are removed, and the rest of the probes are mapped to genes which results in a set of 17 411 probes/genes. We also consider the datasets from run 2 and 3. They contain 234 samples (132 case-control pairs) and 270 samples (135 case-control pairs), respectively. The number of probes after mapping to genes is 17 411 also for these datasets.

3.1 Vary p-value cut-off and present limit

There are two parameters to investigate, the detection p-value cut-off and the present limit. Table 1 -Table 3 show the number of probes that remains in the dataset for different settings of these parameters in run 1 - 3. The numbers are quite similar for the three runs. The number of probes left increases with increasing p-value cut-off and decreases with increasing present limit as one would expect. If the present limit is 0, a probe does not have to be present in any samples in order to be included in the final dataset, thus all probes are kept, independently of the p-value cut-off. A present limit of 1, which means that the probes have to be present in all samples, greatly reduces the number of probes, and would be a very strict and limiting requirement. In the applied R-function, p-value cut-offs greater than 0.5 are treated as 1 minus the cut-off since expression values with detection p-values close to 1 are supposed to be detectable. In the last column of Table 1, the numbers shown are given that the value 1 has been used as a an actual cut-off.

		p-value cut-off						
		0	0.01	0.05	0.1	0.3	0.5	1
Present limit	0	17 411	17 411	17 411	17 411	17 411	17 411	17 411
	0.01	8 004	10 418	13 537	15 614	17 324	17 407	17 411
	0.05	7 223	9 450	11 792	13 710	16 856	17 340	17 411
	0.1	6 825	9 001	10 953	12 687	16 215	17 214	17 411
	0.3	5 882	8 063	9 700	10 854	14 102	16 190	17 411
	0.5	5 237	7 398	8 887	9 897	12 490	14 720	17 411
	1	1 973	3 097	3 881	4 309	5 379	6 197	17 411

Table 1 Number of probes that remains after filtering for various values of the p-value cut-off and present limit in run 1.

		p-value cut-off						
		0	0.01	0.05	0.1	0.3	0.5	1
Present limit	0	17 411	17 411	17 411	17 411	17 411	17 411	17 411
	0.01	8 734	11 550	13 764	14 421	15 054	15 632	17 411
	0.05	7 627	9 608	11 782	13 046	14 509	14 948	17 411
	0.1	7 226	8 973	10 882	12 157	14 159	14 739	17 411
	0.3	6 245	7 811	9 011	10 028	12 852	14 114	17 411
	0.5	5 431	7 032	8 000	8 733	11 338	13 414	17 411
	1	1 897	2 893	3 465	3 778	4 730	5 916	17 411

Table 2 Number of probes that remains after filtering for various values of the p-value cut-off and present limit in run 2.

		p-value cut-off						
		0	0.01	0.05	0.1	0.3	0.5	1
Present limit	0	17 411	17 411	17 411	17 411	17 411	17 411	17 411
	0.01	8 716	11 152	13 413	14 729	16 308	17 003	17 411
	0.05	7 842	10 033	11 834	13 188	15 527	16 417	17 411
	0.1	7 409	9 558	11 155	12 413	15 034	16 059	17 411
	0.3	6 335	8 616	9 898	10 866	13 501	15 130	17 411
	0.5	5 534	8 016	9 147	9 949	12 265	14 184	17 411
	1	1 921	4 648	5 348	5 832	6 892	7 833	17 411

Table 3 Number of probes that remains after filtering for various values of the p-value cut-off and present limit in run 3.

Figure 1 shows the densities for the mean of the log2 expression of all samples in run 1 for each probe divided by the corresponding standard deviation of the log2 expression, denoted mean standard deviation ratio, at different p-value cut-offs and a fixed present limit at 0.01. The red curves are the densities for the probes that are removed, and the black curves are the densities of the probes that are kept in the dataset. The figure shows that there is a clear separation in the distribution of the mean standard deviation ratio for p-value cut-off at 0.01 with a mean at about 10 and the probes above this cut off with a mean at about 15. This implies that when the p-value cut-off is increased, the variation in the distribution increases. The number of probes kept in each of the small figures is shown in the second row in Table 1.

A similar pattern as in Figure 1 can be seen in Figure 2 when increasing the present limit with a fixed p-value cut-off at 0.01. Here, there are separate distributions for the mean standard deviation ratio for all present limits above 0. For probes with present value above 0.01 the mean is about 10, while for probes with present value below 0.01, the mean is about 15. The number of probes kept in each of the small figures are shown in the second column in Table 1, ranging from 7 398 to 10 418.

Figure 3 shows simultaneous variation of the p-value cut-off and present limit. Notice that we get about the same separation of the mean standard deviation ratio in two distributions in the left column and bottom row (number of probes between 7 398 and 10 418), while the two distributions are more merged when approaching the upper right corner with 15 614 probes. Hence, we can maintain the separation with p-value cut-off at 0.01 and present value at 0.5 and then either increase the p-value cut-off or decrease the present limit. If both variables are changed, then the kept and removed distributions will be more overlapping.

Figure 4 shows the distribution of the mean standard deviation ratio of the probes in run 1 as in Figure 1. The red line is the distribution of the probes removed when the p-value cut-off is 0.5. The orange, purple, green and blue lines are the distributions of the additional probes removed as the p-value cut-off is successively decreasing to 0.01, and the black line is the distribution for the probes kept with a p-value cut-off of 0.01. This gives separate distributions for the different intervals for p-value cut-off with small variation for each interval. The mean value in the distributions increases for increasing p-value cut-off. However, each of the curves for intermediate values of the p-value cut-off are based on quite few data points.

Figure 5 shows the distribution of the mean standard deviation ratio as in Figure 2. The red line is the distribution of the probes removed with a present limit at 0.01. The blue, green, purple, orange and light blue lines correspond to the distributions for the additional probes removed as the present limit is successively increased to 1. The black line is the distribution of the probes kept with a present limit of 1, this corresponds to the minimal set of probes that will be kept, independently of the present limit. The distributions for succeeding intervals are overlapping, and the mean of the distributions is decreasing as the present limit increases. Each of the curves for intermediate values of the present limit are based on quite few data points.

The same figures as in Figure 1Figure 5 for run 1 are given in Section 6 for run 2 and 3. The mean standard deviation distributions are quite different for run 2 and 3 compared to run 1. For run 2, the distribution has two peaks, one high and one small in the right-hand tail. The distribution for run 3 has a heavy long right-hand tail. The distribution of the probes that are removed in run 2 overlaps the distribution of the probes that are kept, but the smaller peak is removed, even for less strict values of the parameters. In run 3, the distributions are less overlapping, and more similar to run 1. The probes removed are mainly from the tail. There were several technical problems associated with run 2, and in this respect, we would trust the results from run 1 and 3 more.

Figure 1 - 5 show that there is a narrow distribution for the mean standard deviation ratio when we have strict limitations on the p-value cut-off and present limit. If we relax on these conditions, we get larger values for the mean standard deviation ratio. It is difficult to interpret this result. A p-value cut-off at 0.01 seems to be a reasonable choice, but how to choose the appropriate value of the present limit is more unclear. We discuss this matter further in Section 3.3.

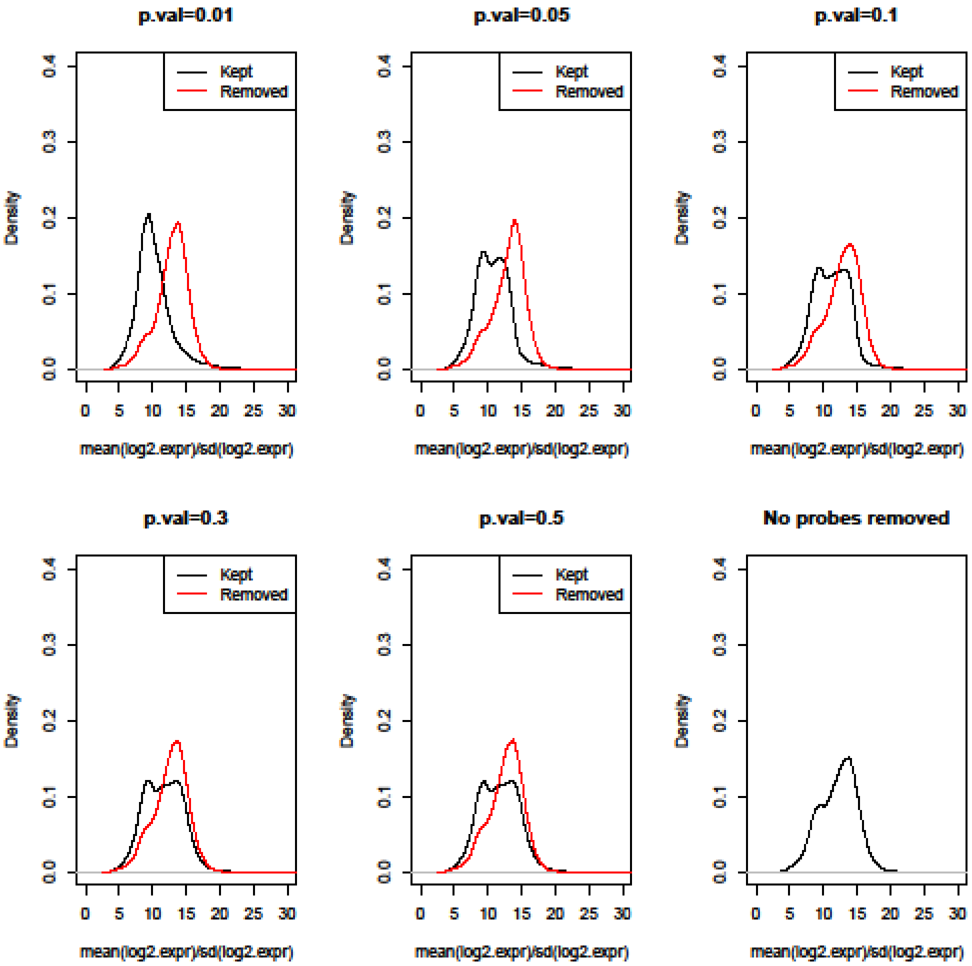

Figure 1 Density of mean(log2 expr)/sd(log2 expr) for probes that are removed and kept for different p-value cut-offs. The present limit is set to 0.01.

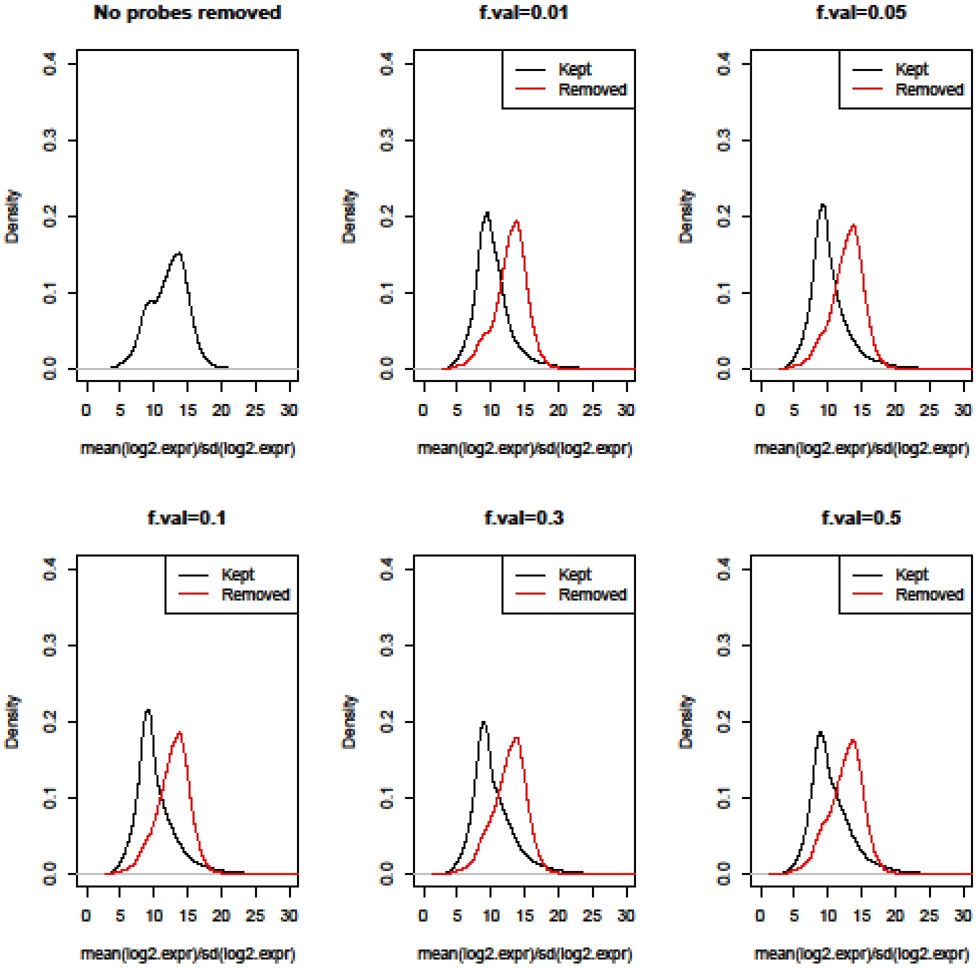

Figure 2 Density of mean(log2 expr)/sd(log2 expr) for probes that are removed and kept for different present limits (denoted by f.val). The p-value cut-off is set to 0.01.

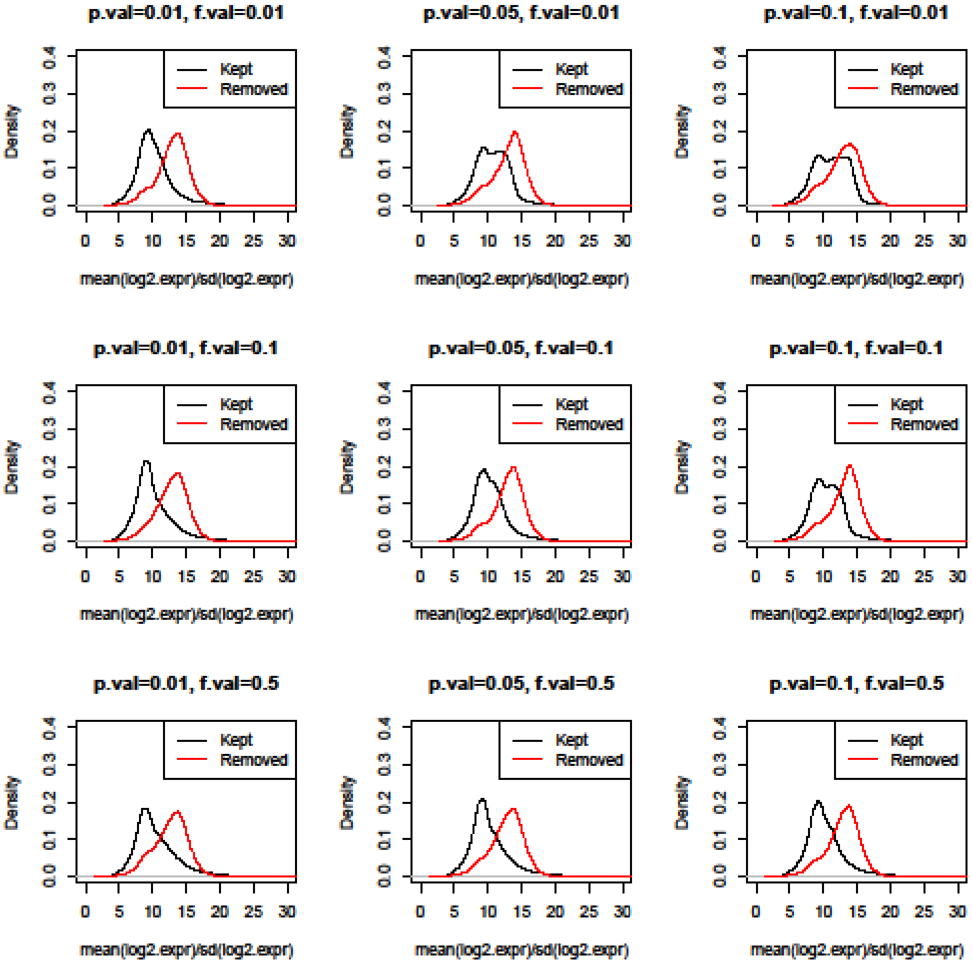

Figure 3 Density of mean(log2 expr)/sd(log2 expr) for probes that are removed and kept for different p-value cut-offs and present limits. The f-val denotes the present limit.

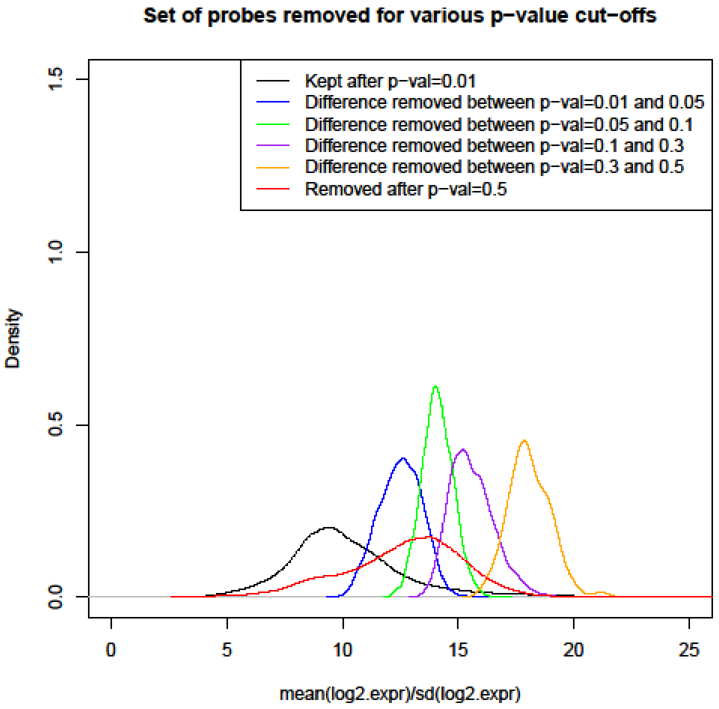

Figure 4 Density of mean(log2 expr)/sd(log2 expr). The black line represents the probes that are the kept when the p-value cut-off is 0.01. The red line represents the probes that always removed, i.e. the ones that are removed when the cut-off is 0.5. The blue, green, purple and orange lines represent the additional probes that are removed when the cut-off is decreased from 0.5 to 0.01. The present limit is set to 0.01.

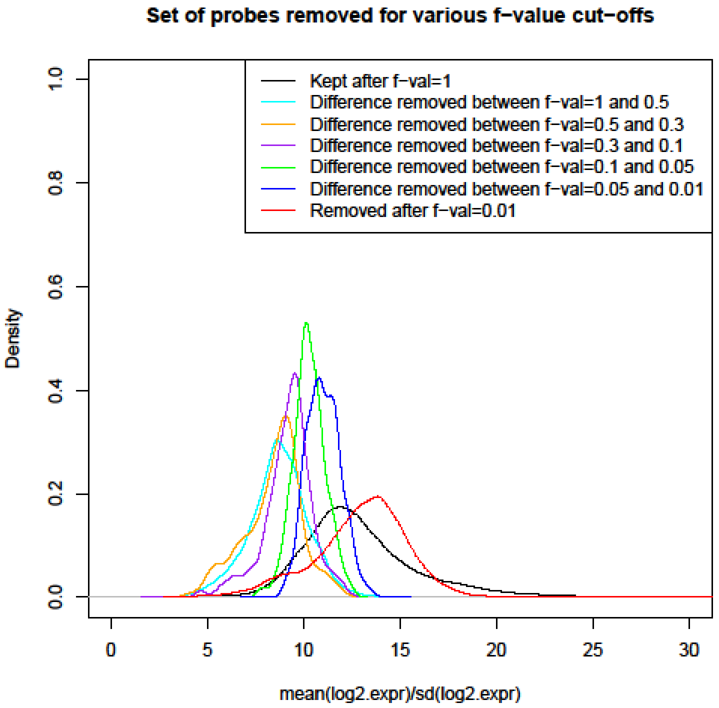

Figure 5 Density of mean(log2 expr)/sd(log2 expr). The black line represents the probes that in the minimal set of kept probes, i.e. when the present limit is 1 and the probes have to be present in all samples. The red line represents the probes that removed when the present limit is 0.01. The light blue, orange, purple, green and blue lines represent the additional probes that are removed when the present limit is increased from 0.01 to 1. The p-value cut-off is set to 0.01.

3.2 Comparing different normalization methods

The aim of the final statistical analyses is to compare the gene expression patterns between different groups (strata). The choice of parameters in the preprocessing affects the results of these analyses, i.e. the set of genes that have significantly different patterns. In standard gene expression analysis, the gene expression values are normalized after the preprocessing, as this removes unwanted variation due to technical differences between the samples. For gene expression values, quantile normalization is frequently used, this results in all samples having the same distribution across the probes. In our setting, it is not clear which normalization method should be used.

To study the effects of the preprocessing parameters and normalization methods, first a set of 45 hypothesis tests are performed for all probes . We are here comparing the values from each pair of six different strata (screening/interval/outside x spread/not spread) and three different time periods giving 3x$\binom{6}{2} = 45$ combinations. It is important to note that we do not know whether there are any significant differences between the strata in the data set. We therefore expect a priori that all the null hypotheses are correct, i.e. that no genes are differentially expressed between different strata. With this assumption, we expect to reject a proportion of the null hypotheses equal to the chosen significance level of the hypothesis tests. The tests are done on log2 expression background corrected data that are not filtered, and the tests are based on calculating the t-statistics for comparing the two groups of data. For a given p-value cut-off for the detection filtering, the mean number of significantly differentially expressed probes at each possible present limit is calculated. This is done the following way: For each probe, find the proportion of samples for which the probe is present. Then for each probe, find its proportion of significant tests among the 45 tests. For each possible value of proportion of present probes, calculate the mean proportion of significant tests among the probes with the given present proportion. The proportion of present probes is plotted against the proportion of significant tests for a chosen set of significance levels shown in Figure 6 - Figure 9. We have applied 0.01 as the detection p-value cut-off. To smooth the resulting curve, at each present value, at least 3 000 probes should contribute to the mean. For those values that have less than 3 000 probes, the probes from the nearest values are included until this number is reached. The curve is plotted for different significance levels and for four different normalization methods that are described in Table 4.

Figure 6 - Figure 9 show curves for the significance levels 0.0001, 0.001, 0.05 and 0.10. In general not normalizing results in far less probes being significant than expected, except when the probes are present in all samples. It is difficult to interpret why we get less significance than if the data was only noise. The variation in the data may be so large that the assumptions in a t-test is violated.

For all three normalization methods, the proportion of significant tests is quite close to the given significance level. For lower present proportions, fewer probes are declared significant. The signal in these probes is probably noisy, and we would not expect to find many significant probes here. When the present proportion increases, the proportion of significant tests also increases and is above the significance level. In particular, when all probes are present the proportion of significant tests is clearly higher than the significance level. The relative increase compared to significance level is highest for the strictest significance levels. This indicates that there are significant differences in the sample and that this is more visible when all probes are present and for the strictest significant levels.

Normalization methods 2 and 3 produce almost identical results. Normalization method 1 differ from method 2 and 3 by detecting fewer probes for low proportion of present samples, and more for higher proportions of present samples.

The distribution of the proportion of present probes is not uniform, as shown in Figure 10. There are two peaks, with 16 864 probes not being present in any samples and 4 075 probes being present in all[1]. These two peaks are clearly visible in the plots, and the proportion of significant tests differ for these values compared to the rest. When all probes are present, the proportion of significant probes is much higher than when fewer probes are present, in particular for normalization method 1. With this respect, method 2 and 3 seem more stable, especially for lower significance levels.

Normalization method	Input data for normalization	Type of normalization
0	-	None
1	log2 expression case – log2 expression controls	Quantile
2	log2 expression	Quantile
3	expression	Quantile

Table 4 Overview of normalization methods.

[1] These numbers correspond to the full dataset of 39 388 probes, and are therefore not comparable to the numbers in Table 1.

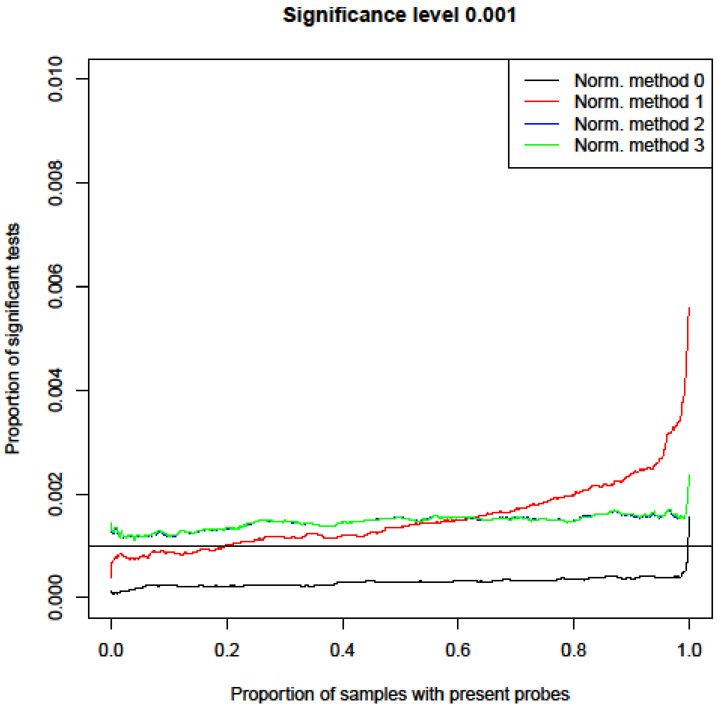

Figure 6 Mean proportion of significant probes for 45 tests at significance level 0.001 against present proportions of the samples.

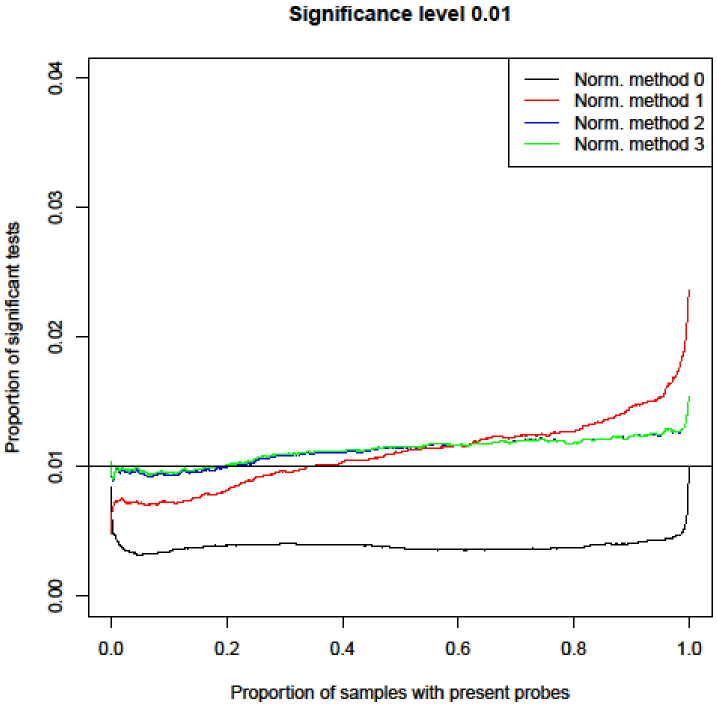

Figure 7 Mean proportion of significant probes for 45 tests at significance level 0.01 against present proportions of the samples.

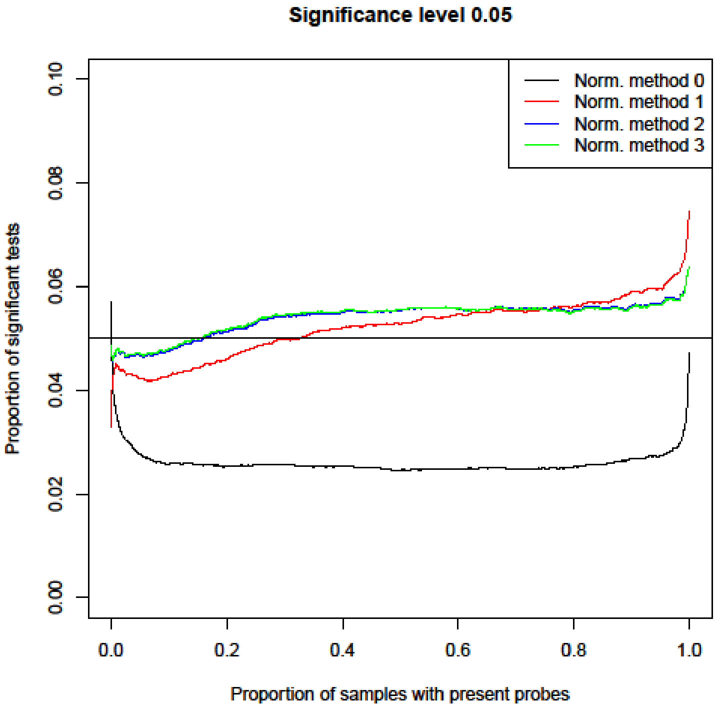

Figure 8 Mean proportion of significant probes for 45 tests at significance level 0.05 against present proportions of the samples.

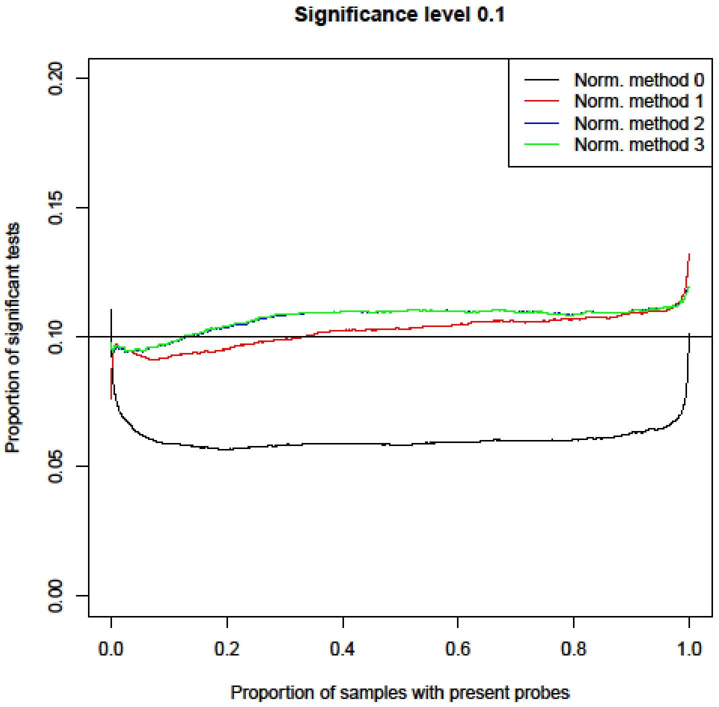

Figure 9 Mean proportion of significant probes for 45 tests at significance level 0.1 against present proportions of the samples.

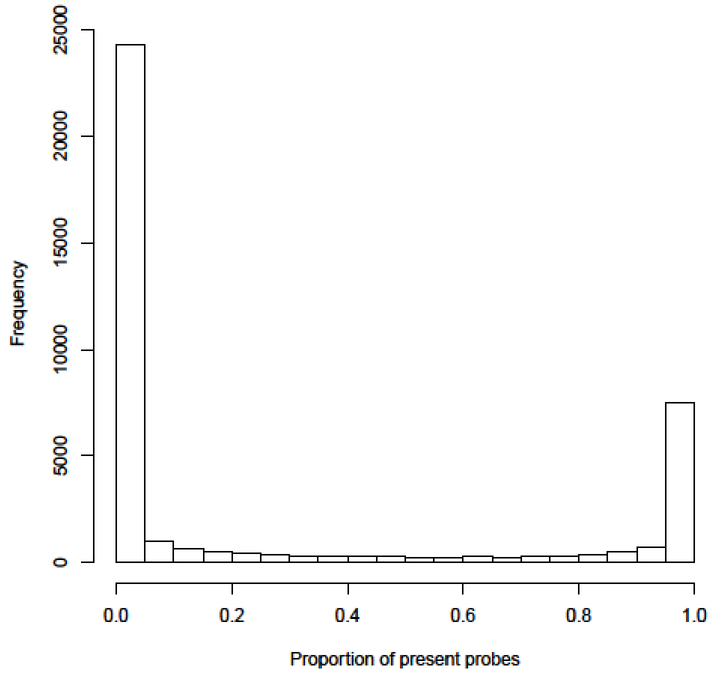

Figure 10 Histogram of the proportion of present probes in the dataset, the p-value cut-off is set to 0.01.

The interpretation of Figure 6 - Figure 9 is not straightforward, in particular because we are not sure what to expect in each situation. The figures are based on 45 tests, which are a mix of tests expected to be true and not true. To obtain a clearer picture, we consider two single tests separately, one where we expect a difference between the groups and one where we do not expect a difference. The description of these tests and the detailed results are given in the long version of this report [1]. The results can be summarized as follows. Not normalizing results in too many positive findings, showing that normalization is necessary. One should be careful using normalization method 1 if there are few samples present. Normalization method 3 seems slightly better than normalization method 2 and considerably better than normalization method 1. Normalization method 3 gives more significant tests when we expect significant differences and closer to the significance level when we expect no difference. Normalization method 2 and 3 give more significant values also when there are fewer probes present, while normalization method 1 needed almost all probes present in order to give significant results.

3.3 Conclusion on filtering parameters and normalization

The results in Section 3.1 and 3.2 shows that the p-value cut-off, the present limit and the normalization methods are critical for the statistical analysis. The results in Section 3.1 indicate that the strength of the signal differ when we vary the two filtering parameters. Less strict parameter choices results in less probes being removed from the dataset, however the signal in probes that are not detectable or only present in a few samples may be noisy and disturb the analysis. Some filtering is therefore necessary. The p-value cut-off has currently been set at 0.01, and from the results in Section 3.1 this seems reasonable, we therefore recommend that in order for a probe to be counted as present, its detection p-value should be less than 0.01. The choice of present limit seems less critical, given a low p-value cut-off, and the results in Section 3.2 indicate that we are able to find significant results also when many of the probes are not present. A lower present limit can therefore be used, in particular if normalization method 3 is used.

The tests of the different normalization methods indicate that method 3 is best suited for this data set and give quite strong evidence that it is necessary with normalization. Normalization method 3 gave most significant results where we expected this and close to the significance level where we know the data is only noise.

In this report, the outlier detection, background correction, filtering and normalization have been done simultaneously for all samples in the given run. The analysis of differential expression is however done on subsets corresponding to different strata. The question is whether filtering and/or normalization should be done on the whole set or on these subsets.

Intuitively, it seems reasonable to do the normalization on the full dataset. This ensures data are comparable across the various subsets. If several comparisons are described in the same paper, it will probably be confusing if the data do not contain the same number of probes in all analyses. There are however reasons to perform the filtering on each subset separately. If a present limit is applied to the full dataset, there might be probes that would be included if this limit was applied to the subset, but not to the full set. The probes to genes mapping may also result in different probes being chosen if applied to the subsets instead of the full dataset.

A compromise would be to do the filtering on the full dataset, but to use a liberal present limit, so that the set of probes included is as large as possible. We therefore recommend to use a present limit at 0.01.

In the final analyses, three different runs are combined. Each run is now background corrected separately, and probes that have poor mapping to Illumina ids are removed. Then the runs are combined, and the filtering and mapping to genes are done on the combined dataset. This way it is ensured that probes are not lost because the mapping was done differently for each run. The normalization is done on the combined dataset.

We conclude that the appropriate choices should be a p-value cut-off at 0.01, a present limit at 0.01 and normalization method 3.

3.4 Suggested procedure for preprocessing breast cancer data

After having concluded what values should be used for the parameters, we have decided on the following preprocessing procedure. Outlier removal, background correction and normalization should always be done for the complete dataset. When separate analyses are performed on subdatasets, the filtering and mapping of probes to genes will be done on the subdatasets, not the complete dataset.

1. Remove probes for genes involved with blood type.

2. Remove the case-control pairs when either the case or control is an outlier with respect to quality measures.

Remove the pairs of the following samples:

- Cases without breast cancer
- Cases with other types of cancer
- Cases with previous types of breast cancer
- Controls with breast cancer

Remove case-control pairs when either the case or control has missing values of the EN variable or M values not equal to 0.

4. Perform background correction for complete dataset using negative control probes.

5. Normalize complete dataset using quantile normalization on original scale level.

6. Split dataset into subdatasets that will be used in different analyses.

7. For each subdataset, filter out probes that are not detectable or sufficiently present using a p-value cut-off of 0.01 and a suitable present limit.

8. For each subdataset, map probes to genes using normalized data.

4 Appendix: Technical details on access to stallo

Knut Hansen (knut.hansen@uit.no) can be contacted for help with access to stallo.

First, the form on this page has to be filled out:

http://uit.no/ansatte/organisasjon/artikkel?p_document_id=299809&p_dimension_id=88223&p_menu=49281

The necessary information is:

"Account Responsible" is you, mobile phone and e-mail are mandatory.

"Project Title" is "Tice"

"Project manager" is Eiliv Lund

Start og stopp can be skipped, but you still have to write something there, I wrote NA

Choose No for Multi-User Account

Skip, i.e. NA, for CPU and Disk usage

Choose Yes for previously granted quotas

Choose username

Write R under Software/Public Domain

Refer to previously submitted description of the Tice project

You will then receive an e-mail with username and a temporary password. Once your user account is activated, you can log onto stallo by ssh:

```
ssh stallo-ism.uit.no
```

or with ssh username@stallo.uit.no if your username is not the same as your username at NR.

To start R at stallo, it first has to be loaded:

```
module load R
```

Then it is started with the command

```
R
```

and works as normal, and you can load packages, run scripts etc. Our files are located in the folder /project/data1/tice2/Clara_Cecilie

5 Appendix: Details on the pre-processing steps

5.1 Data object

The raw data are by Knut Hansen, he receives all the data from NTNU and put it together. The data are saved in a *lumi* object. In this report we will use the run 1 data from the prospective study as an example. The data are stored in the *run1-bc-366cc-jan-31-2014.Rdata* file that we load into R. This file contains (for run 1) the following R objects:

- exprs – matrix of gene expression, dimension: 732 x 39388
- data – the lumi object
- negCtrl – matrix of negative controls, dimension: 732 x 728
- background10 – dataframe with background patient information like case/control status, birth year, date when blood sample is taken, date at the time of diagnosis, dimension : 732x125
- labInfo – dataframe with laboratory information, like RNA quality measures, plate and chip information, dimension: 732x21

The expression matrix can also be extracted from the lumi object by the command

t(exprs(data))

The probe ids are extracted by the command

featureNames(data)

and the detection p-values are extracted from

data@assayData$detection

5.2 Remove probes for genes with blood type

Probes that are related to blood should be removed because they should not be related to cancer. This is done by importing a list of 38 IlluminaID-values (hla_hist.txt). They first have to mapped to nuID, this is done with the function *IlluminaID2nuID.* The list was originally supplied by Arnar Flatberg at NTNU. For the prospective data, this has been done before creating the data objects we work on.

5.3 Checking case and control status

This is done in SAS, and the individual ID is written down, so that the corresponding samples can be identified and removed. This has to be done by someone in Tromsø, probably either the person who replaces Nicolle or Marita Melhus. The list must then be sent to us.

Samples that are removed are:

- Cases without breast cancer
- Cases with other types of cancer
- Cases with previous types of breast cancer
- Controls with breast cancer

The samples removed are removed from background data, the expression matrix, two sets of laboratory information, negative controls and data. The laboratory information is collected by the lab at NTNU. We believe this has been done prior to creating our data objects.

5.4 Quality evaluation and outlier detection – first round

The *lumiQ*-function from the *lumi* package is used for initial quality evaluation. This function performs quality control evaluation of a LumiBatch object and returns a summary of the data. This quality evaluation is done on the original scale, not on log2 data.

The summary of the data contains the mean, standard deviation, detection rate and distance to sample mean for each array in the dataset. The detection rate is printed for a given detection threshold, the default is 0.01 and corresponds to the detection p-value, i.e. probes with a detection p-value of less than 0.01 are supposed to be detectable. The returned rate is the detectable probe ratio for the sample.

There are six types of quality control plots for a lumiQ-object. These are

- **Density plot of the chips (what=”density”)**
- Boxplot of the chip intensities (what=”boxplot”)
- Hierarchical tree to illustrate the correlation among chips (what=”pair”)
- MA-plot between chips (what=”chips”)
- **Sample relations (what=”sampleRelation”)**
- **Distance to the center to be used for outlier detection (what=”outlier”)**
- Density of the coefficients of variance of the chips (what=”cv”)

For our data either only detection of outliers have been used, see Figure 1, or sometimes also the sample relations and density plots are made too. Note: This is done on the original data, not the log2-transformed data. The outliers are detected based on a distance to the cluster center, it performs hierarchical clustering with eucledian distance measure. The outliers are defined as having distance to the center larger than a threshold x the median distance.

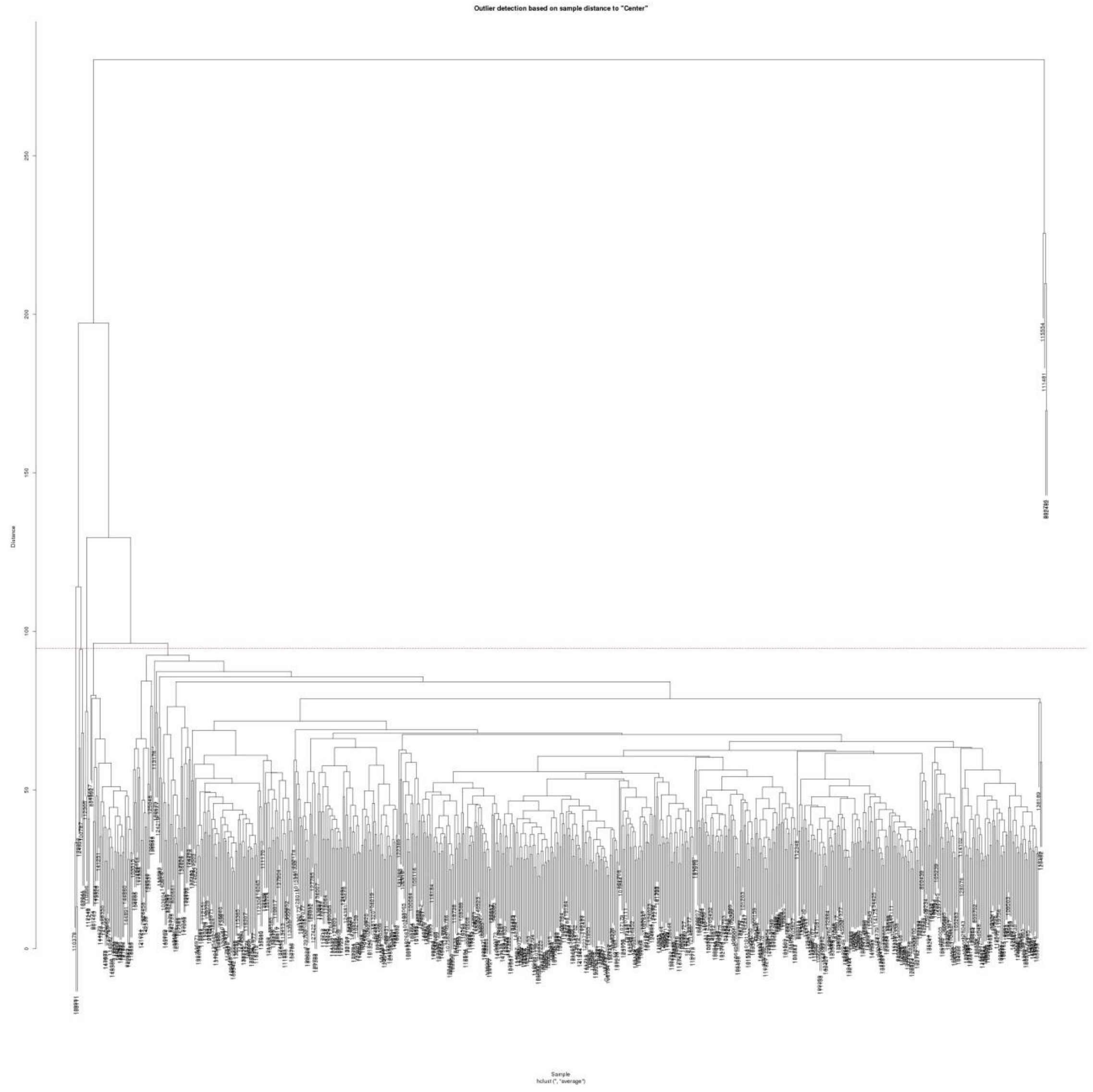

Figure 11 Dendrogram for outlier detection.

The identification of outliers is done by principal component analysis (PCA) and the dendrogram from the quality evaluation. For some datasets, AQM identification is used too. The outliers that are identified are then evaluated by density plots, median gene plots, laboratory measures and days in the mail. Finally, differences between the plates are investigated.

Figure 11 shows the resulting dendrogram from the quality evaluation. The potential outliers are the ones above the red line.

Principal component analysis is performed on the log2-transformed data. The percentage of explanation for each component is calculated by dividing the square of the standard deviation of each component by the sum of squared standard deviations for all components.

ggplot is used to plot the first two components of the PCA-object. By visual inspection, a cut-off is chosen to create a set of potential outliers. For these data, a cut-off was set at 200 for the absolute value of the first component, see Figure 12.

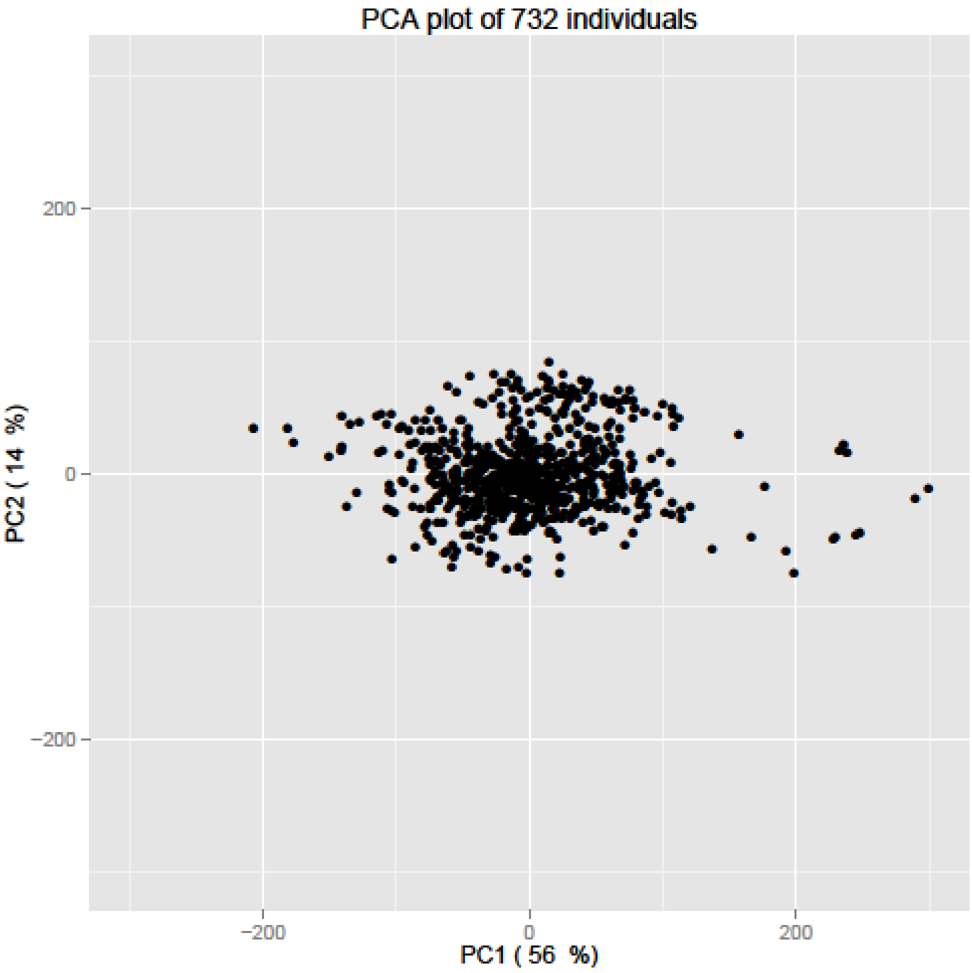

Figure 12 PCA plot of the first two components.

In some dataset the AQM method is used to detect outliers. This approach uses the *arrayQualityMetrics* package [2]. In order to use this package, an *eset* object must be made, and through the *prepdata*-function, the data are log2-transformed. Three plots are made initially, heatmap, MA-plot and boxplot. It is checked whether the identified outliers are the same in all three plots, but the outliers are also chosen if they are only found in two of the plots. This is done manually. The MA-plot takes a long time to run, and is therefore excluded in some scripts.

To evaluate the identified outliers, the outliers from the dendrogram (and AQM) are also marked in the PCA plot, and in this case they are among the ones above the PCA cut-off. Density plots are made for each plate, with the identified outliers colored in red. Median gene plots are made as well, where for each individual, log2 expression values for all the probes are plotted against the median log2 expression value of all individuals.

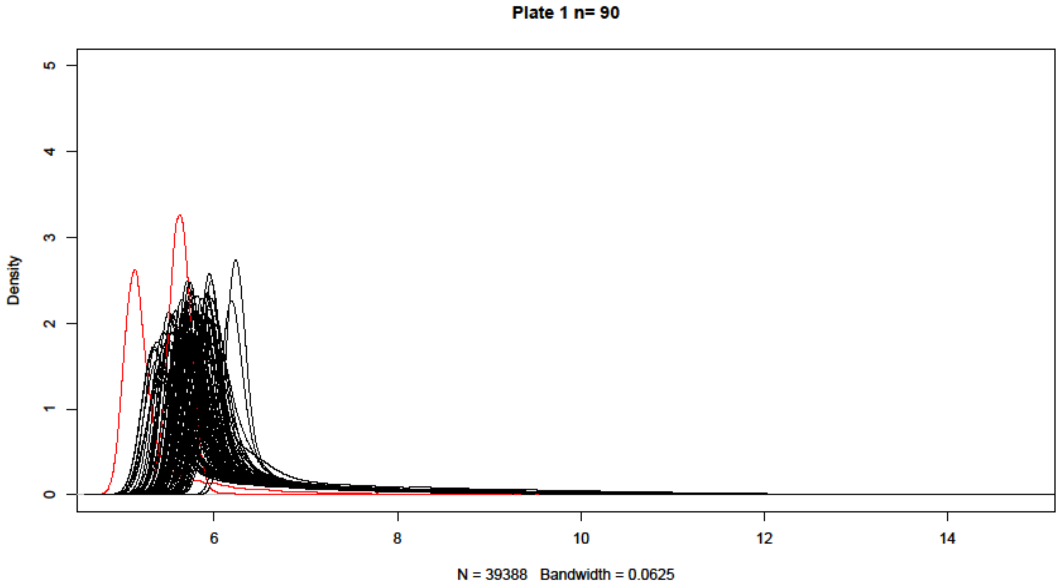

Figure 13 Density plot for all the samples on a plate.

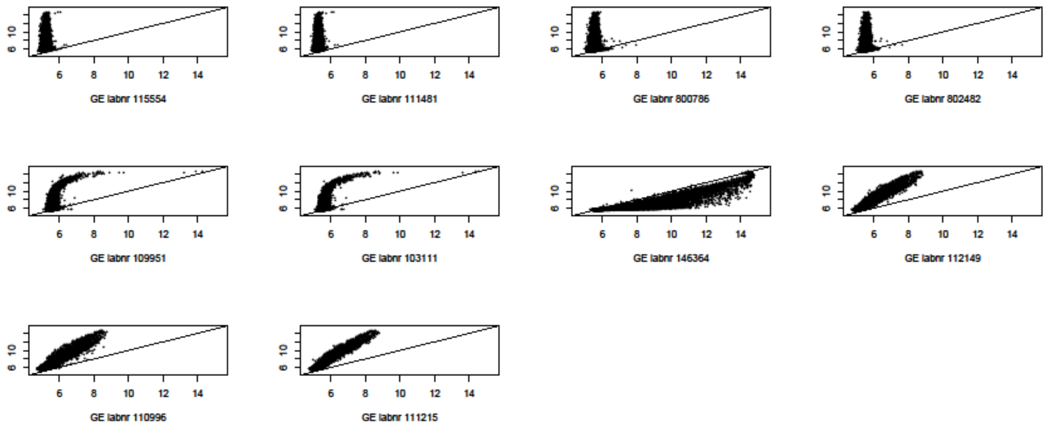

Figure 14 Median gene plots for possible outlier samples

Laboratory measures are assessed for the potential outliers. Good quality indicators are

- 50 < RNA < 500
- RIN>7
- 260/280 > 2
- 260/230 > 1.7

The time in the mail, i.e. the time between DATOINN and PROVEDATO is checked, for the detected outliers, in case any of them could be due to a delay between when the blood sample is taken and when it arrived.

In this dataset, nine outliers are detected and since pairs will be removed, not only individuals, in total 18 samples are removed. After this round of outlier removal, we are left with 714 individuals.

5.5 Quality evaluation and outlier detection – second round

In the second round of outlier identification, the same steps as in Section 5.4 are repeated. Seven individuals, and thus seven pairs are removed, and the resulting dataset now consists of 700 individuals (350 pairs). This is the final round of outlier detection, and the data are saved in the R object *run1-bc-350cc-feb-03-2014.Rdata*

5.6 Comments regarding the removal of outliers

The detection is done on the individual level, before the whole pair has to be removed. The final decision of which outliers to remove is subjective, however the PCA plot, the dendrogram and the median plots are the ones trusted the most. The approach is quite conservative, the samples should not be removed unless they consistently look strange.

5.7 Background correction

The background correction is done by the use of the *nec* function in *limma*. There are 728 negative control probes. These are probes that should not be expressed, and can therefore be used to determine the background level. The common approach is to assume that the gene expression is exponentially distributed and the background is normally distributed, which leads to a normal+exponential convolution model. Let X_i be the observed intensity for gene *i*, and let

$$X_i = S_i + B_i,$$

where S_i is the signal intensity and $B_i$is the noise intensity for gene *i*. Furthermore,

$$S_i \sim \mathrm{Exp}(\alpha)$$

and

$$B_i \sim N(\mu, \sigma^2).$$

The parameters can be estimated by a non-parametric estimator derived by Xie et al [4]. Then

$$\tilde{\alpha} = \bar{X} - \bar{X}_0,$$

$$\bar{\mu} = \bar{X}_0$$

and

$$\sigma^2 = \frac{\sum_{j=1}^{J} (X_{0j} - \bar{X}_0)^2}{J-1},$$

which can be inserted into the formula for the expected true signal:

$$E(S_i|X_i = x_i) = a_i + b\frac{\varphi(\frac{a_i}{b})}{\Phi(\frac{a_i}{b})},$$

where $a_i = x_i - \mu - \sigma^2/\alpha$ and $b = \sigma$.

For each gene there is in indicator of whether the gene is a negative control or not, and this the input to the *nec* function. An offset is added to the background corrected data, the default value is 16. Changing the offset does not affect the filtering in the next step.

5.8 Additional details of the filtering step

The filtering step is described in Section 2. This section includes some additional details.

Each probe is mapped to a gene through the *nuID2IlluminaID* function. The quality of the mapping can be assessed by the *illuminaHumanv3PROBEQUALITY* function. Each probe has a quality grade assigned to it, which is either “Perfect”, “Good”, “Bad” or “No Match”. The probes with quality “Bad” or “No Match” are filtered out. The probes that have poor quality are removed. These are removed before filtering on the present proportion and detections values.

We are then left with a reduced set of probes. The final analysis will however be done on the gene level. A gene can be represented by several probes, and we therefore have to choose only one probe for each gene. This is done by the *nsFilter* function. The Entrez id is used as a gene identifier. As described in Section 2, when there are several probes that maps to the same gene, the probe with the highest IQR value is chosen.

5.9 Removing of samples based on clinical information

Samples that have missing values of the EN variable or M values not equal to 0 are identified and the case-control pair is removed. This results in 112 samples being removed, and the final dataset consists of 588 samples or 294 pairs.

We believe this is an additional removal step, after step 5.3.

5.10 Defining new clinical variables and create final dataset to analyse

In this step, the node_spread, insitu and follow_up_time are defined from the background data. The data from the blood questionnaire are added, they are imported from the file MHolden_extravars_07jul2014.csv. (We don’t have this file.) These data include the variables

- smoke
- hrt
- RMENS

- age_at_dx
- age_at_blood
- status

Finally, the data are combined in a data frame with the log2 gene expressions (one column per gene) , node spread, follow up time, case/control indicator, patient id, insitu indicator, smoke indicator, hrt measurement, BMENS, age at diagnosis, age when blood sample was taken and status.

6 Appendix: Plots for run 2 and 3

This appendix gives the same figures for run 2 and 3 as Figure 1 -Figure 5 for run 1. The comparison of the three runs is given in Section 3.1.

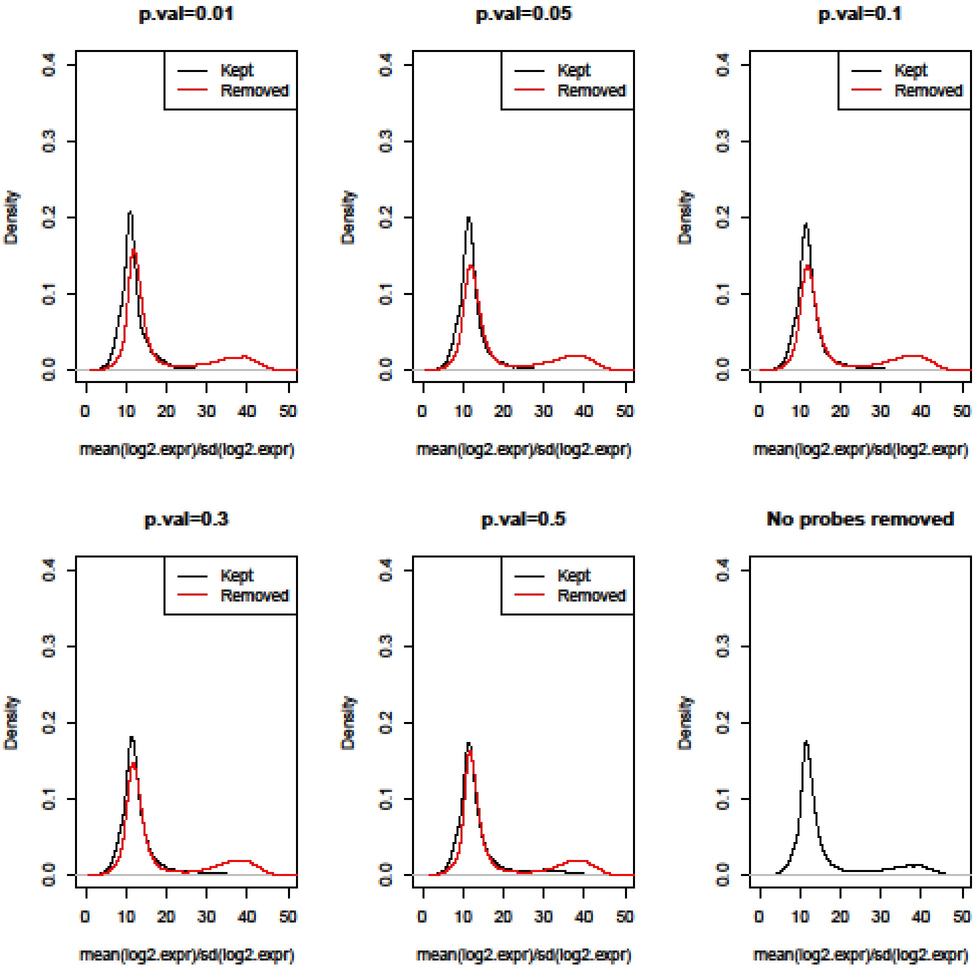

Figure 15 Density of mean(log2 expr)/sd(log2 expr) for probes in run 2 that are removed and kept for different p-value cut-offs. The present limit is set to 0.01.

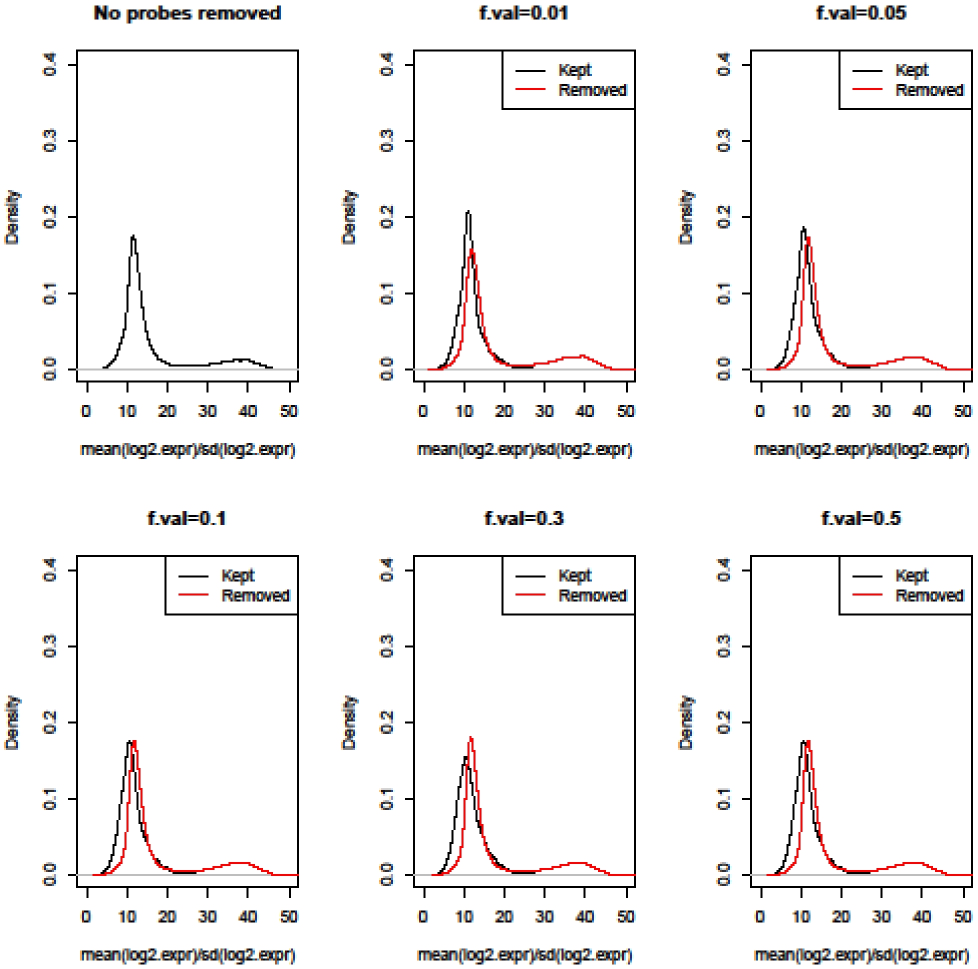

Figure 16 Density of mean(log2 expr)/sd(log2 expr) for probes in run 2 that are removed and kept for different present limits. The p-value cut-off is setl to 0.01.

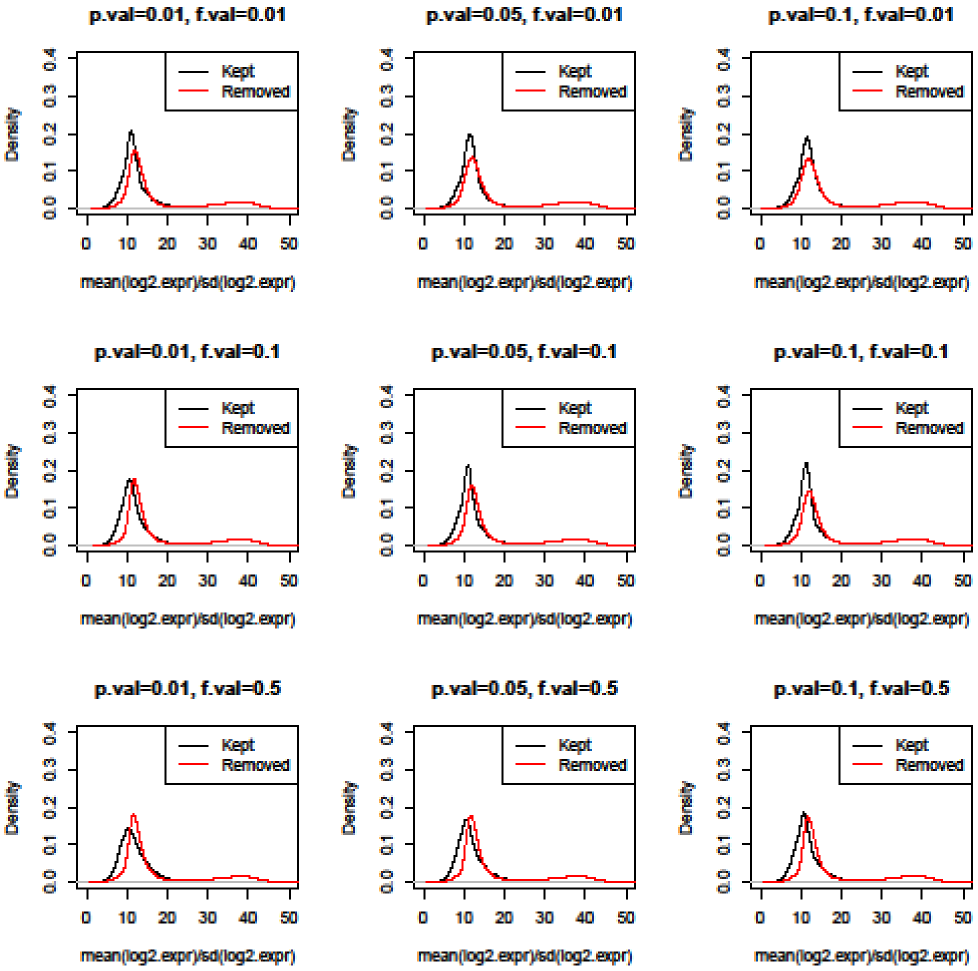

Figure 17 Density of mean(log2 expr)/sd(log2 expr) for probes in run 2 that are removed and kept for different p-value cut-offs and present limits.

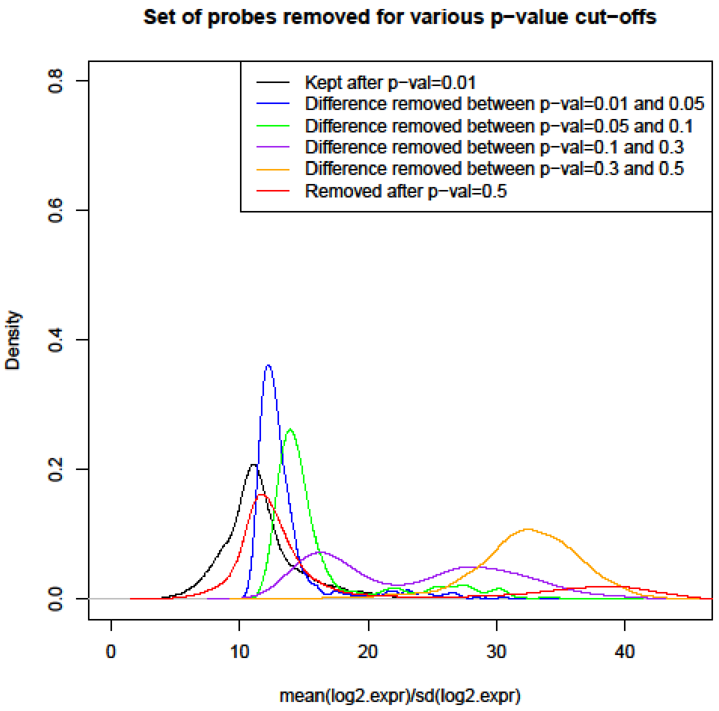

Figure 18 Density of mean(log2 expr)/sd(log2 expr). The black line represents the probes that are the kept when the p-value cut-off is 0.01. The red line represents the probes that always removed, i.e. the ones that are removed when the cut-off is 0.5. The blue, green, purple and orange lines represent the additional probes that are removed when the cut-off is decreased from 0.5 to 0.01. The present limit is set to 0.01.

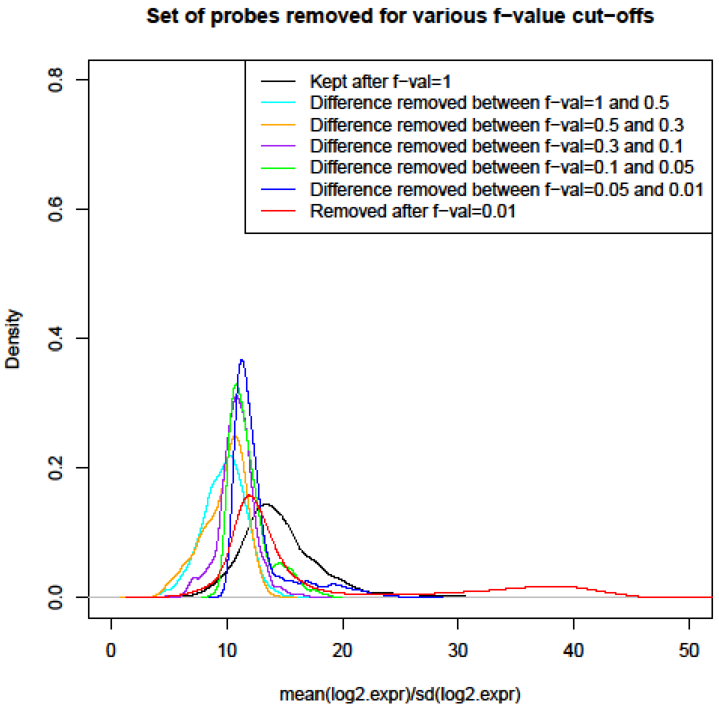

Figure 19 Density of mean(log2 expr)/sd(log2 expr). The black line represents the probes that in the minimal set of kept probes, i.e. when the present limit is 1 and the probes have to be present in all samples. The red line represents the probes that removed when the present limit is 0.01. The light blue, orange, purple, green and blue lines represent the additional probes that are removed when the present limit is increased from 0.01 to 1. The p-value cut-off is set to 0.01.

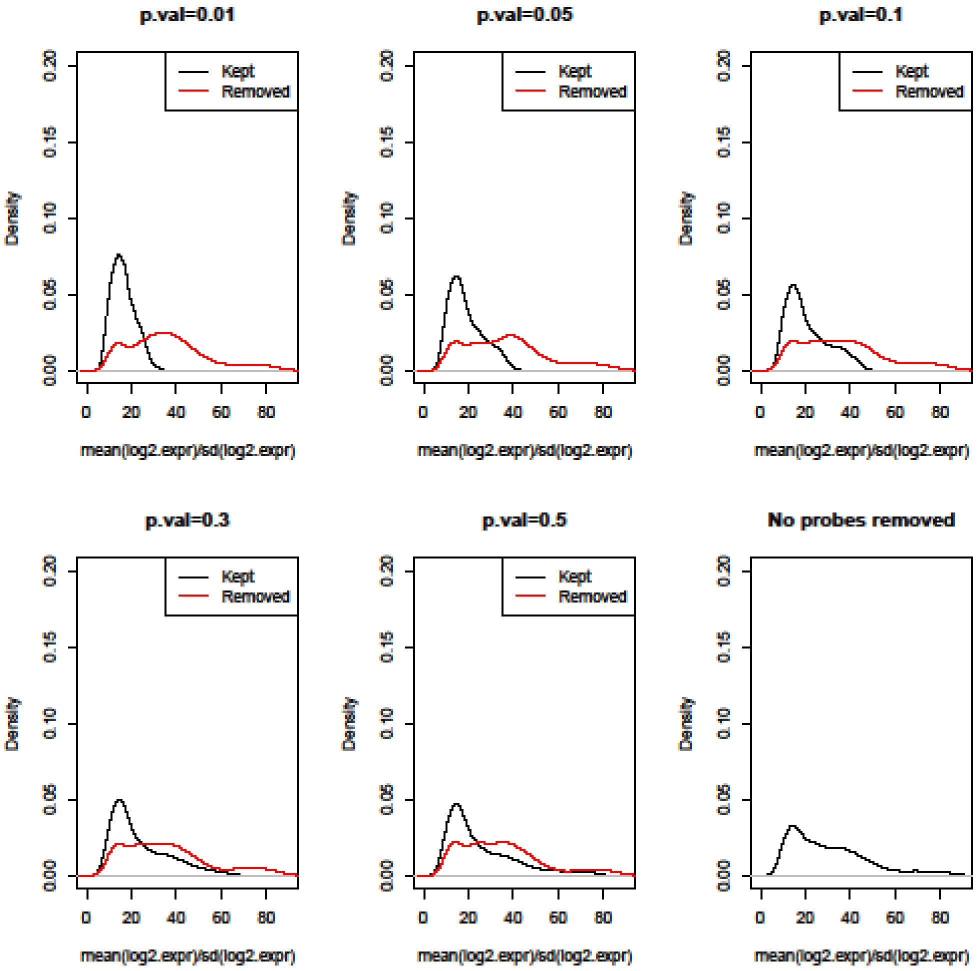

Figure 20 Density of mean(log2 expr)/sd(log2 expr) for probes in run 3 that are removed and kept for different p-value cut-offs. The present limit is equal to 0.01.

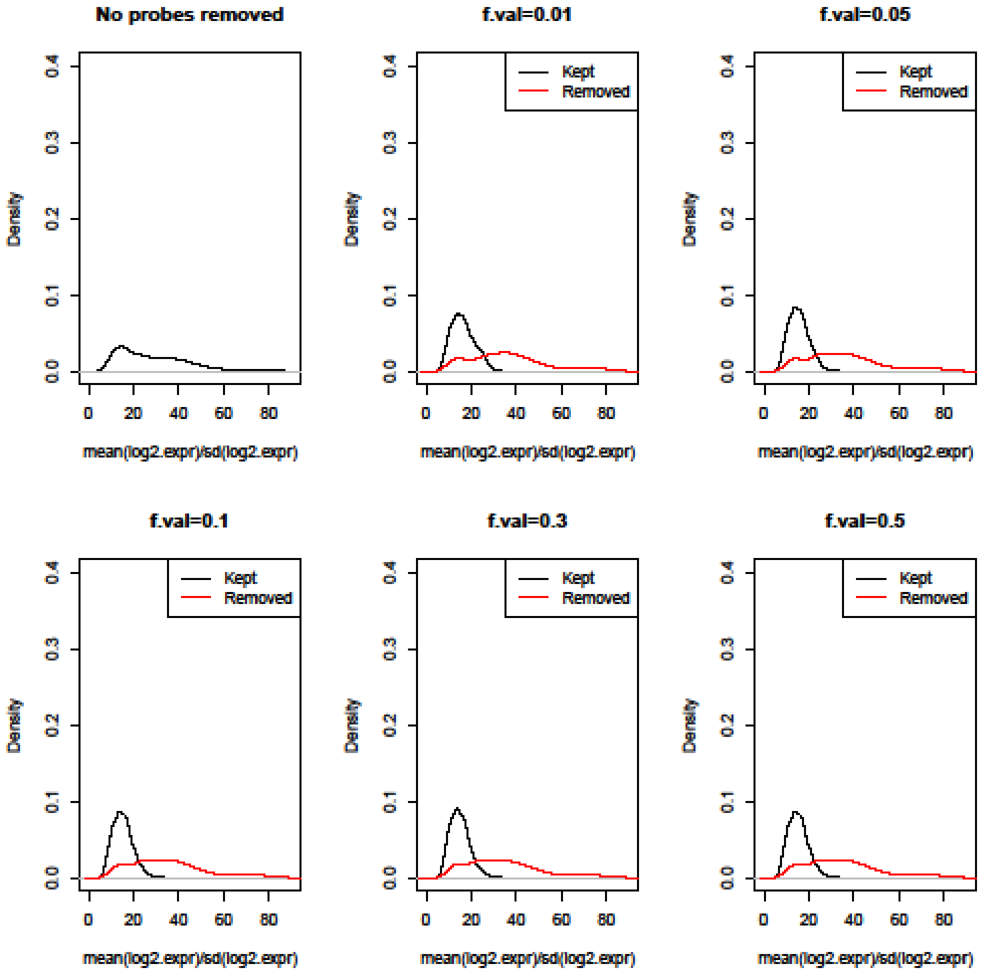

Figure 21 Density of mean(log2 expr)/sd(log2 expr) for probes in run 3 that are removed and kept for different present limits. The p-value cut-off is equal to 0.01.

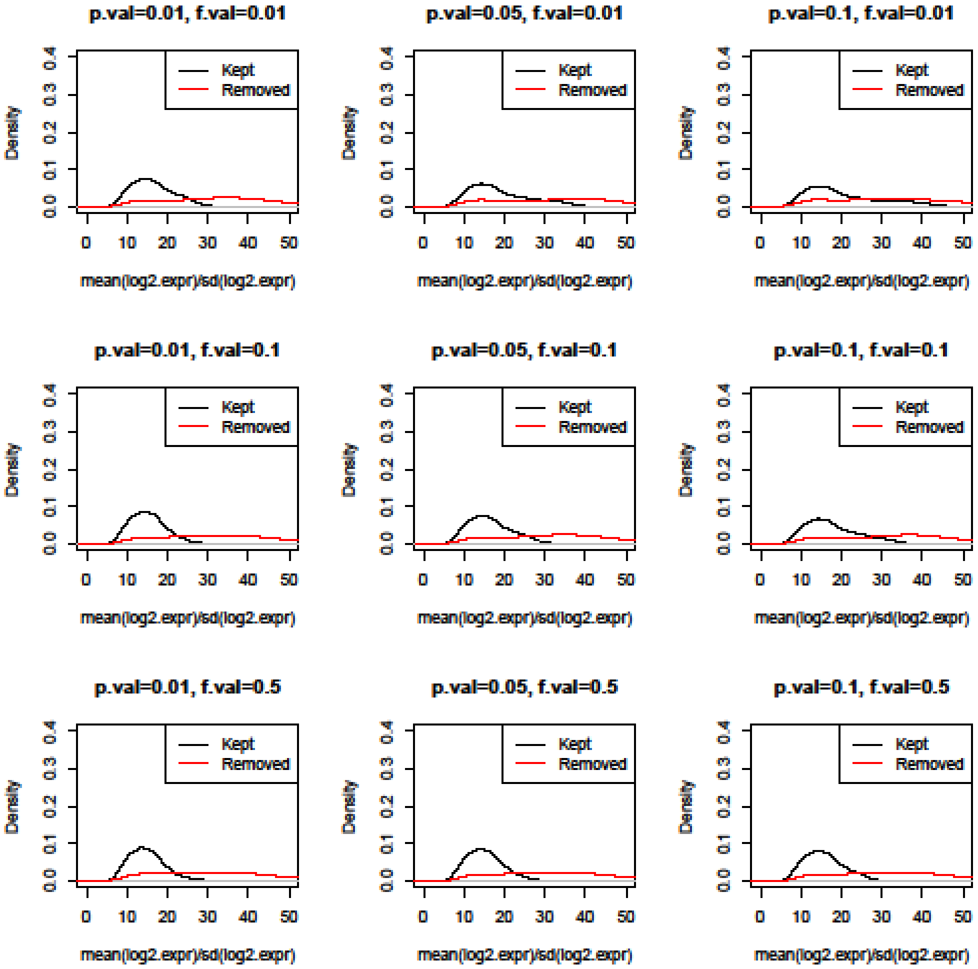

Figure 22 Density of mean(log2 expr)/sd(log2 expr) for probes in run 3 that are removed and kept for different p-value cut-offs and present limits.

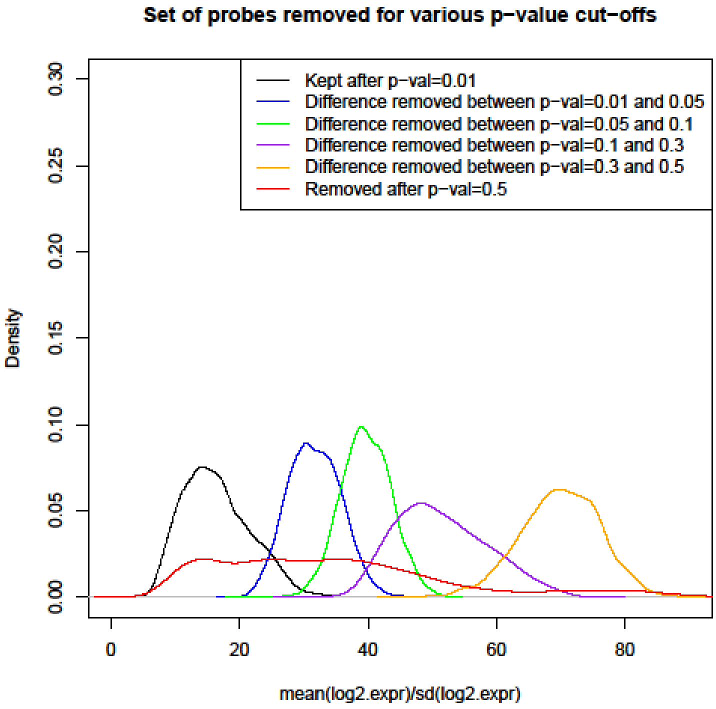

Figure 23 Density of mean(log2 expr)/sd(log2 expr). The black line represents the probes that are the kept when the p-value cut-off is 0.01. The red line represents the probes that always removed, i.e. the ones that are removed when the cut-off is 0.5. The blue, green, purple and orange lines represent the additional probes that are removed when the cut-off is decreased from 0.5 to 0.01. The present limit is set to 0.01.

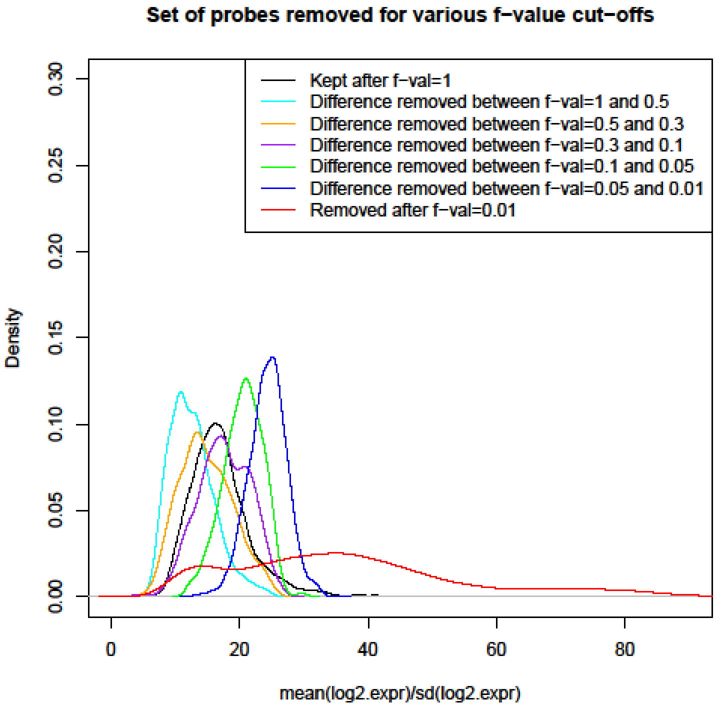

Figure 24 Density of mean(log2 expr)/sd(log2 expr). The black line represents the probes that in the minimal set of kept probes, i.e. when the present limit is 1 and the probes have to be present in all samples. The red line represents the probes that removed when the present limit is 0.01. The light blue, orange, purple, green and blue lines represent the additional probes that are removed when the present limit is increased from 0.01 to 1. The p-value cut-off is set to 0.01.

7 References

1. Günther CC, Holden M and Holden L (2014). “Preprocessing of gene expression data related to breast cancer diagnosis.” NR-note SAMBA/32/14.

2. Kauffmann A, Gentleman R and Huber W (2009). “arrayQualityMetrics–a bioconductor package for quality assessment of microarray data.” *Bioinformatics*, **25**(3), pp. 415–6.

3. Holden M and Holden L (2014). “Statistical analysis of gene expression data related to breast cancer diagnosis.” NR-note SAMBA/19/14.

4. Xie Y, Wang X and Story, M (2009). “Statistical methods of background correction for Illumina BeadArray data.” *Bioinformactics*, **25**(6), pp, 751-757.