Introduction to the 'AFLP' package (version 0.4.0-71)

Thierry Onkelinx

April 23, 2014

Chapter 1

Terminology

specimen the sample as recieved in the lab. Each sample requires a unique code. It will be divided into one or more replicates. Specimens with multiple replicates will be used to asses the repeatability.

group the a priori clustering of specimens e.g. per location, per species, ...

replicate a subsample of a specimen. The DNA extraction, PCR reaction and fluorescence measurement are done at this level. Thus replicates are unique.

plate a batch of replicates with simultaneous PCR reaction. Each cell in a plate is referenced by a lane and a capilar

capilar label of the capilar. The number of capilars depend on the lab equiment. E.g. an ABI 3500 uses 8 capilars, a LiCor slabgel is regarded as one capilar.

lane on a LiCor slabgel: the position of the replicates. On a capilar system: the number of the run within the plate.

fluorescence

marker

normalisation

classification

repeatability

Chapter 2

Reading in data

2.1 The design

First we must define the design: the position of each replicate on the plates, the link between replicate and specimen. The *Group* column is optional and only relevant if some a priori clustering is assumed. The grouping in the Tilia dataset is the field determination of the species (see §A). All columns, exect *Group*, present in the TiliaDesign `data.frame` are mandatory. Extra columns will be appended to the AFLP object, but ignored during analysis.

When the data.frame is proper formatted, you can transform it into an `AFLP` object using the **as.AFLP** function. All mandatory columns will be converted to factors.

```
> data(TiliaDesign)
> str(TiliaDesign)
```

```
'data.frame':        512 obs. of  6 variables:
 $ Replicate: Factor w/ 512 levels "C.09.1744","C.09.1745"..
 $ Plate    : Factor w/ 8 levels "1","2","3","4",..: 1 1 1..
 $ Lane     : Factor w/ 64 levels "A1","A2","A3",..: 1 2 3..
 $ Specimen : Factor w/ 423 levels "11201","11202",..: 166..
 $ Group    : Factor w/ 4 levels "Tilia cordata",..: 1 3 4..
 $ Capilar  : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 ...
```

```
> summary(TiliaDesign)
```

```
     Replicate       Plate          Lane
 C.09.1744:  1   1      : 64   A1     :  8
 C.09.1745:  1   2      : 64   A2     :  8
 C.09.1746:  1   3      : 64   A3     :  8
 C.09.1747:  1   4      : 64   A4     :  8
 C.09.1748:  1   5      : 64   A5     :  8
 C.09.1749:  1   6      : 64   A6     :  8
```

```
 (Other)  :506   (Other):128   (Other):464
     Specimen                   Group     Capilar
 QC method:  8   Tilia cordata     :102   1:512
 7109     :  3   Tilia platyphyllos:148
 7203     :  3   Tilia europea (x) :163
 C/07/3323:  3   Unknown           : 99
 C/07/3327:  3
 C/07/3331:  3
 (Other)  :489
```

```
> Tilia <- as.AFLP(TiliaDesign)
```

2.2 The fluorescence data

In the example the AFLP analysis was run on a LiCor slabgel and the fluorescence was measured using SAGA software. It is easy to add the text file output from SAGA to an `AFLP` object using the **readSAGA** function. Be carefull with the names of the replicates. The names in SAGA file and the `AFLP` object must be indentical. Keep in mind that the replicate names are read as header by **readSAGA** and thus all rules for the names of `data.frame` apply. We recommend to start names of replicates with a letter and to use only letters, numbers and points. Please note that R is case-sensitive. The *textclean* argument can be used to pass a user-defined function to do some cleaning on the replicate names.

```
> Tilia <- readSAGA(
+   system.file("extdata", "Tilia_bandvaluespc1", package = "AFLP"),
+   add.to = Tilia)
> str(fluorescence(Tilia))
```

```
'data.frame':        13824 obs. of  6 variables:
 $ PC          : Factor w/ 1 level "PC1": 1 1 1 1 1 1 1 1 ..
 $ Replicate   : Factor w/ 512 levels "C.09.1744","C.09.1"..
 $ Fluorescence: num  1419 702 786 1261 NA ...
 $ Marker      : num  676 621 592 539 528 480 449 431 420 ..
 $ Normalised  : logi  NA NA NA NA NA NA ...
 $ Score       : logi  NA NA NA NA NA NA ...
```

```
> summary(fluorescence(Tilia))
```

```
   PC            Replicate      Fluorescence
 PC1:13824   C.09.1744:   27   Min.   :    1.5
             C.09.1745:   27   1st Qu.: 1259.0
             C.09.1746:   27   Median : 4845.3
             C.09.1747:   27   Mean   :10187.9
             C.09.1748:   27   3rd Qu.:15206.0
```

```
            C.09.1749:   27   Max.   :59779.3
            (Other)  :13662   NA's   :2889
    Marker      Normalised      Score
Min.   : 97.0   Mode:logical   Mode:logical
1st Qu.:171.0   NA's:13824     NA's:13824
Median :288.0
Mean   :325.3
3rd Qu.:449.0
Max.   :676.0
```

The fluorescence can be added manually as well. In this case you need to prepare a `data.frame` with 6 columns: *PC* (a factor indicating the primer combination), *Replicate* (a factor with the replicate ID), *Fluorescence* (the measured fluorescence), *Marker* (the size of the marker in basepairs), *Normalised* (NA , will hold the normalised fluorescence) and *Score* (NA , will hold the classification).

```
> fluorescence(Tilia) <- Your.data.frame
```

Chapter 3

Normalising the raw fluorescence

3.1 clean and normalise

Prior to the normalisation you should use the **clean** function. This will do some sanity checking on the `AFLP` object, especially on missing data.

```
> Tilia <- clean(Tilia)
> output <- normalise(Tilia, output = "none")
```

The normalisation estimates the effects of replicate, plate, lane, capilar and marker on the average fluorescence. The algorithm selects the appropriate combination of effects based on the design and the number of markers. The user has to decide which transformation to use. One can choose among `transformation = "log"` (default), `transformation = "logit"` and `transformation = "none"`.

The user has three options for *output*: `"screen"` (default), `"tex"` and `"none"`. `"screen"` does the normalisation and displays the model that is used in the normalisation and some standard graphs and tables to asses potential problems with the data. `"tex"` is equal to `"screen"` except that the output generates LaTeXcode for include the graphs and table in a document. It generates a section for each primer combination and a subsection for each random effect. `"none"` does the normalisation without generating tables and graphs. Therefore it is the fastest option.

The graphs consist of QQ plots for the best unbiased lineair predictions (BLUPs) of the random effects and QQ plots of the residuals. As simple linear model is fitted for each QQ plot, quantifying the linear relationship between the observed and the theoretical values. The predictions of this model and their $100 \times level\%$ prediction intervals and added to the QQ plots. *level* is an argument of **normalise** and defaults to `0.99`. Observed values outside these

prediction interval are marked as "possible outliers" and tabulated. See §**??** for more information on outlier detection.

```
> output <- normalise(Tilia, output = "tex", device = "png")
```

3.2 PC1

```
Linear mixed model fit by REML ['lmerMod']
Formula:
log(Fluorescence) ~ 1 + (1 | Plate) + (1 | Replicate) + Marker +
    (1 | fMarker)
   Data: z[z$UseIt, ]

REML criterion at convergence: 30445.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.4733 -0.5905  0.0376  0.6129  3.6111

Random effects:
 Groups    Name        Variance Std.Dev.
 Replicate (Intercept) 0.1184   0.3441
 fMarker   (Intercept) 0.4591   0.6775
 Plate     (Intercept) 0.0799   0.2827
 Residual              0.8751   0.9355
Number of obs: 10935, groups: Replicate, 512; fMarker, 27; Plate, 8

Fixed effects:
             Estimate Std. Error t value
(Intercept) 10.4420586  0.3029761   34.46
Marker      -0.0060394  0.0007803   -7.74

Correlation of Fixed Effects:
       (Intr)
Marker -0.838
```

3.2.1 Replicate

Label	Observed
C.09.2245	-1.741
C.09.1915	-1.560
C.09.1952	-1.500
C.09.1911	-1.154
C.09.2006	-1.135
C.09.2220	-1.086
C.09.1944	-1.051
C.09.1945	-1.014
C.09.1953	-1.012
C.09.2133	-0.984
C.09.2042	-0.861
C.09.2205	-0.855

C.09.2011	-0.853
C.09.2184	0.572

Table 3.1: QQ-plot of the random effects at the level Replicate for primer combination PC1

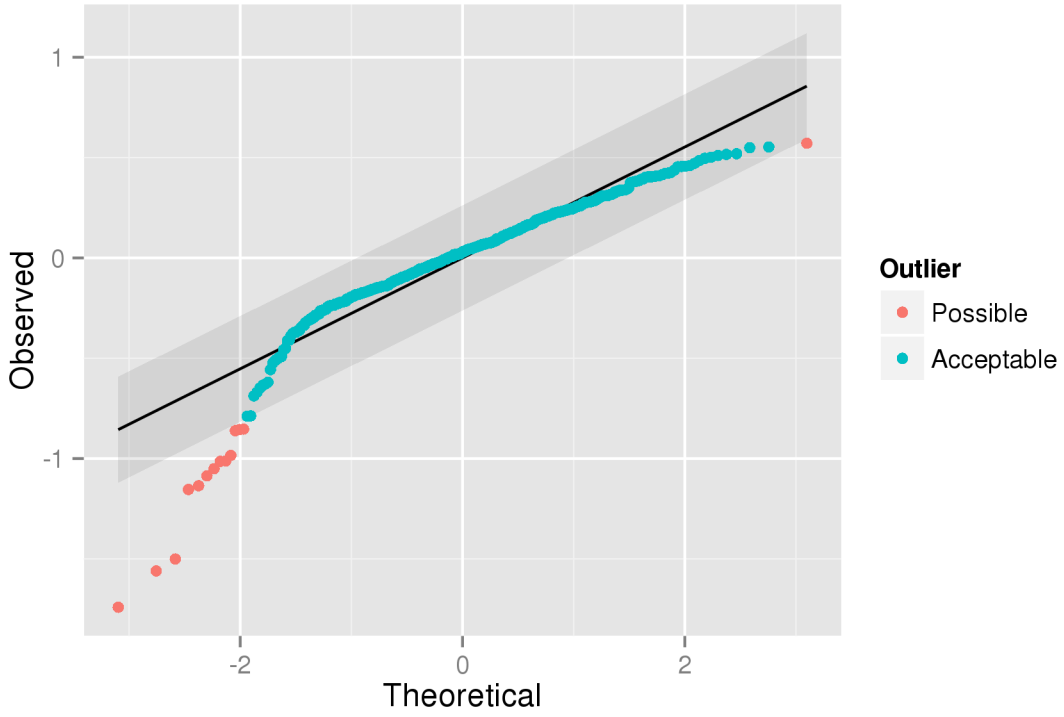

Figure 3.1: QQ-plot of the random effects at the level Replicate for primer combination PC1

3.2.2 fMarker

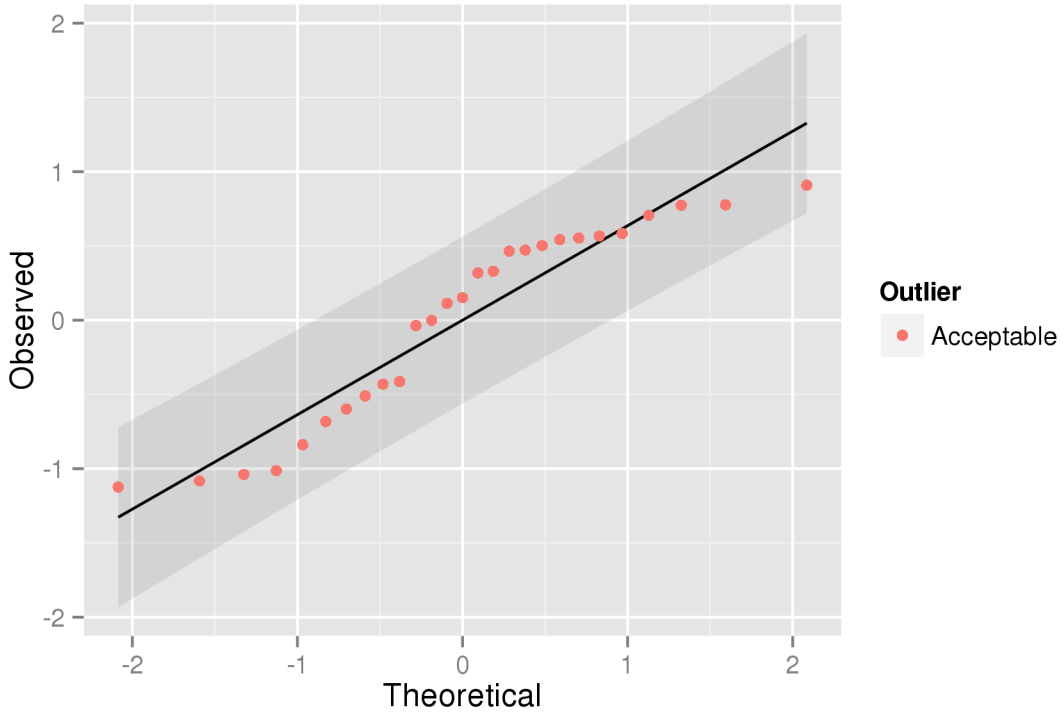

Figure 3.2: QQ-plot of the random effects at the level fMarker for primer combination PC1

3.2.3 Plate

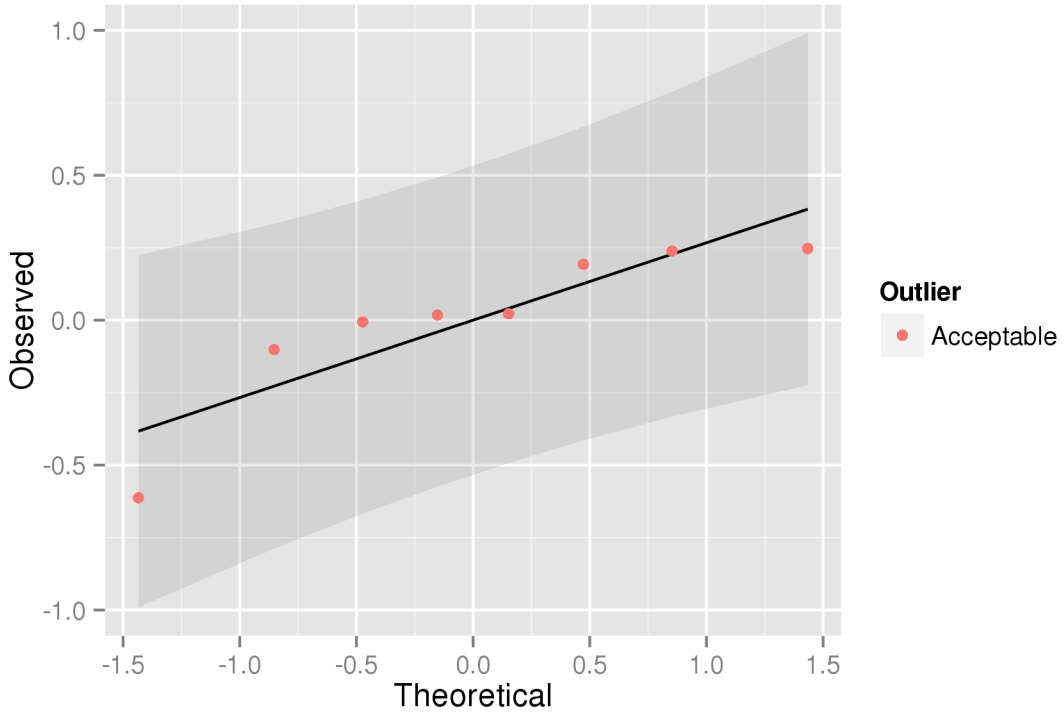

Figure 3.3: QQ-plot of the random effects at the level Plate for primer combination PC1

3.2.4 Globale outliers

Replicate	Marker	Observed
C.09.1863	431.000	-6.991
C.09.1915	281.000	-6.592
C.09.1940	411.000	-6.321
C.09.1952	387.000	-5.679
C.09.2176	431.000	-5.254
C.09.2040	420.000	-5.032
C.09.1953	347.000	-5.001
C.09.2183	539.000	-4.966
C.09.2245	431.000	-4.915
C.09.2245	420.000	-4.699
C.09.1952	431.000	-4.617
C.09.2011	480.000	-4.566
C.09.1911	387.000	-4.333
C.09.1915	387.000	-4.283
C.09.1997	387.000	-4.262
C.09.1915	431.000	-4.085
C.09.1952	420.000	-4.062
C.09.2182	431.000	-3.985
C.09.1944	420.000	-3.840
C.09.2042	420.000	-3.737
C.09.2178	420.000	-3.712
C.09.2011	431.000	-3.703
C.09.1915	347.000	-3.689
C.09.2055	539.000	-3.648
C.09.2027	539.000	-3.584
C.09.2006	480.000	-3.544
C.09.2183	411.000	-3.494

C.09.1915	420.000	-3.440
C.09.2166	420.000	-3.306
C.09.1911	431.000	-3.301
C.09.1927	431.000	-3.282
C.09.1997	411.000	-3.265
C.09.2241	411.000	-3.222
C.09.2025	449.000	-3.217
C.09.2178	411.000	-3.184
C.09.1911	420.000	-3.183
C.09.2245	411.000	-3.168
C.09.2042	480.000	-3.160
C.09.1935	411.000	-3.160
C.09.1944	411.000	-3.119
C.09.1928	411.000	-3.116
C.09.2023	539.000	-3.091
C.09.2024	431.000	-3.077
C.09.1945	411.000	-3.054
C.09.1911	411.000	-3.012
C.09.1805	297.000	-3.001
C.09.2006	387.000	-2.984
C.09.2040	431.000	-2.970
C.09.2025	480.000	-2.954
C.09.1952	281.000	-2.954
C.09.1911	528.000	-2.950
C.09.2049	480.000	-2.934
C.09.1811	411.000	-2.880
C.09.2011	420.000	-2.879
C.09.1772	347.000	-2.842
C.09.2000	431.000	-2.811
C.09.1944	431.000	-2.801
C.09.1945	250.000	-2.801
C.09.2245	281.000	-2.796
C.09.1859	431.000	-2.747
C.09.1940	420.000	-2.742
C.09.2241	420.000	-2.704
C.09.2040	411.000	-2.682
C.09.1796	250.000	-2.649
C.09.2133	250.000	-2.615
C.09.1911	347.000	-2.611
C.09.2025	431.000	-2.606
C.09.2242	411.000	-2.605
C.09.1809	347.000	-2.595
C.09.2081	387.000	-2.592
C.09.2006	431.000	-2.570
C.09.2177	431.000	-2.567
C.09.1811	387.000	-2.543
C.09.2220	178.000	-2.540

Table 3.2: QQ-plot of the residuals for primer combination PC1

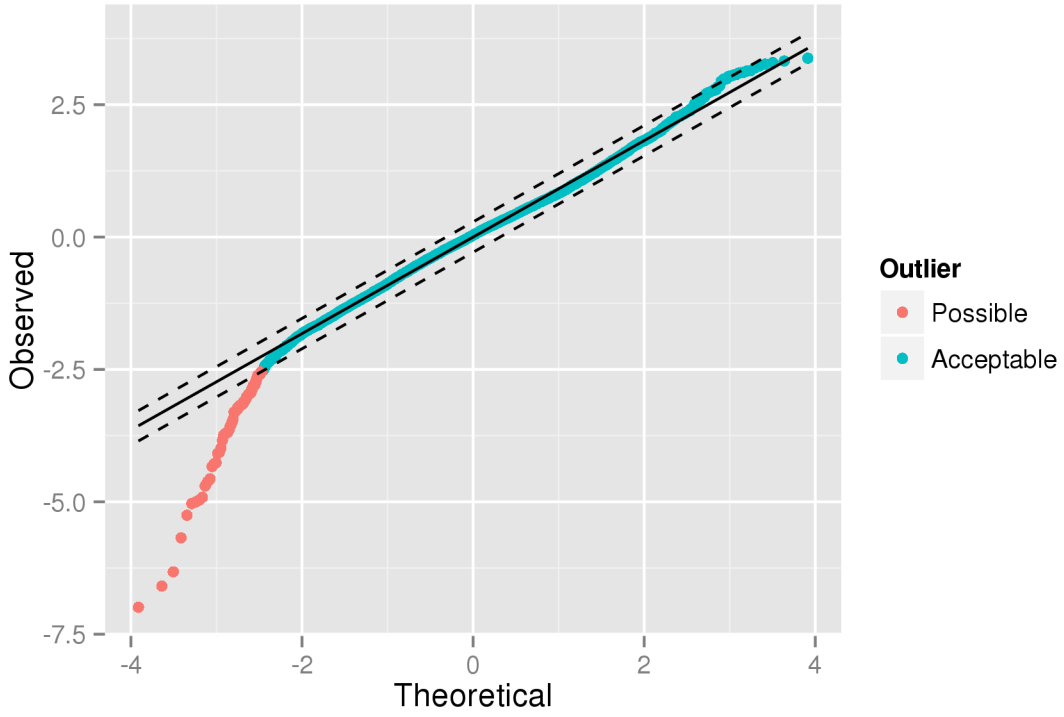

Figure 3.4: QQ-plot of the residuals for primer combination PC1

3.3 Detecting and removing outliers

The result of **normalise** is a `list` with two objects: *data* contains an `AFLP` object and *outliers* contains an `AFLP.outlier` object. The `AFLP` object in the output of **normalise** is the one passed to the *data* argument with updated the `model` and `Normalised` values. All other information remains unchanged. The `AFLP.outlier` object contains all tabulated outliers (see §3.1). Note than although the option `output = "none"` will no display outliers, it will add them to the `AFLP.outlier` object. However we recommend to use the graphs for outlier inspection.

Before taking decissions on removing outliers, we need to know how to interpet the QQ-plots. We take the replicates as an example. An observed value of 0 indicates that signal of that replicate is as strong as the signal of an average replicate. Positive observed values thus indicate stronger signals and negative values weaker signals. A values of -0.5 with the `"log"` transformation can be interpreted as the strength of the signal of this replicate is $\exp(-0.5) = 0.607$ times the strength of an average replicate. Very weak signals can be due to

failed amplification. Very strong signals can be due to contamination. We can give a similar interpretation to the effects of plate, capilar, marker and residual. Strong effects for plate or capilar indicate that possibly something when wrong in the lab. Strong marker effects indicate differences in amplification among markers. Outlying residuals are indications of problems with measuring the fluorescence.

Ideally all points on the QQ-plot form a more or less continuous pattern, which does not need to be a straight line. Isolated points at both ends of the QQ-plot are the most important points to look for. If they are present (e.g. the three lowest points and maybe the highest point in the QQ-plot for the replicates), we recommend to have a look at the lab data to see if there is a problem. Keep the point if the lab data has no indications for problems, otherwise remove it.

We recommend to check the outliers in a stepwise fashion. First start with the worst possible lab problems: entire plates which are problematic. Then check for smaller lab problems: problems at the level of individual replicates. Then we look at the level of the markers. And finaly at the residuals. When outliers are removing for a given step, then one should run **normalise** again and restart the checking at the level of the plates.

In primer combination 1 of the `Tilia` dataset we find no outliers at the plates levels. The inspection of the replicate level highlights 3 low effects and 1 high effect. We descid to remove only the 3 low replicate after inspection of the slab gels.

```
> #extra the outliers for the replicates from the
> #AFLP.outlier object in the output
> repOutliers <- replicates(output$outliers)
> #select the 3 lowest effects for PC1
> toRemove <- head(subset(repOutliers, PC == "PC1"), 3)
> toRemove

        PC Replicate  Observed
PC1.14 PC1 C.09.2245 -1.740983
PC1.2  PC1 C.09.1915 -1.560076
PC1.5  PC1 C.09.1952 -1.500320

> #no outliers in the Tilia dataset
> replicates(outliers(Tilia))

[1] PC        Replicate Observed
<0 rows> (or 0-length row.names)

> Tilia <- addOutliers(Tilia, toRemove)
> #the outliers are added
> replicates(outliers(Tilia))

        PC Replicate  Observed
PC1.14 PC1 C.09.2245 -1.740983
```

```
PC1.2  PC1 C.09.1915 -1.560076
PC1.5  PC1 C.09.1952 -1.500320
```

Now we have to rerun the normalisation.

```
> output <- normalise(Tilia, output = "tex", device = "png")
```

3.4 PC1

```
Linear mixed model fit by REML ['lmerMod']
Formula:
log(Fluorescence) ~ 1 + (1 | Plate) + (1 | Replicate) + Marker +
    (1 | fMarker)
   Data: z[z$UseIt, ]

REML criterion at convergence: 29852.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.6213 -0.6028  0.0407  0.6200  3.6353

Random effects:
 Groups    Name        Variance Std.Dev.
 Replicate (Intercept) 0.09163  0.3027
 fMarker   (Intercept) 0.45774  0.6766
 Plate     (Intercept) 0.08270  0.2876
 Residual              0.84770  0.9207
Number of obs: 10875, groups: Replicate, 509; fMarker, 27; Plate, 8

Fixed effects:
             Estimate Std. Error t value
(Intercept) 10.4474323  0.3030585   34.47
Marker      -0.0060180  0.0007791   -7.72

Correlation of Fixed Effects:
       (Intr)
Marker -0.837
```

3.4.1 Replicate

Label	Observed
C.09.1911	-1.110
C.09.2006	-1.065
C.09.2220	-1.045
C.09.1944	-1.001
C.09.1945	-0.970
C.09.1953	-0.968
C.09.2133	-0.915
C.09.2205	-0.828
C.09.2042	-0.808

C.09.2011	-0.800
C.09.1937	-0.760
C.09.2025	-0.739
C.09.2184	0.530

Table 3.3: QQ-plot of the random effects at the level Replicate for primer combination PC1

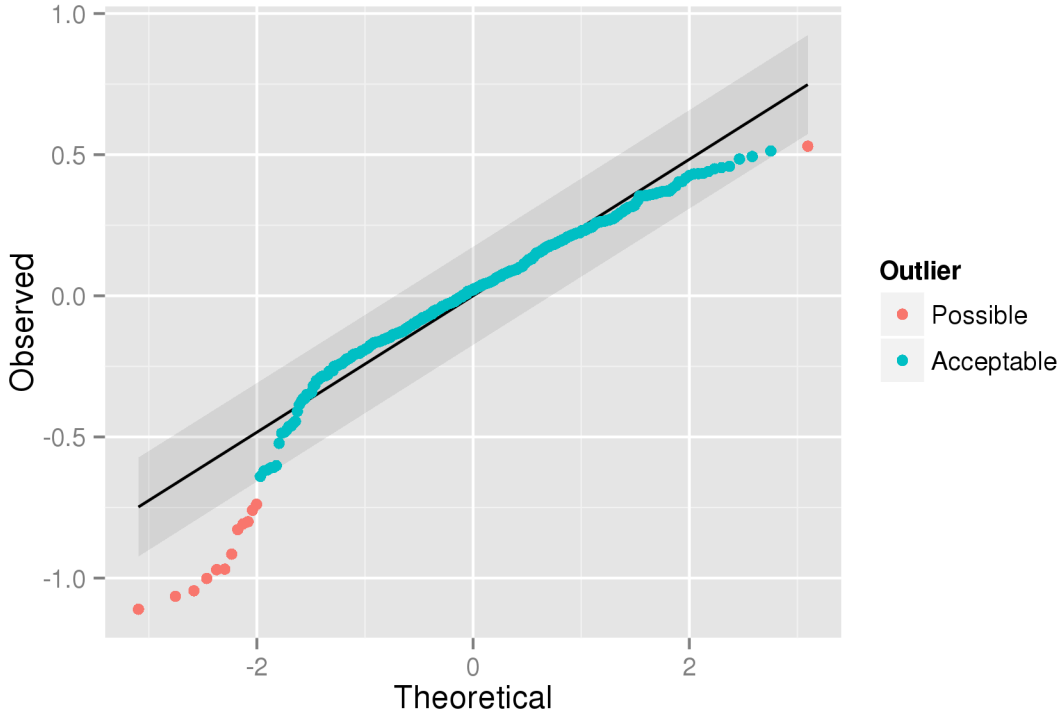

Figure 3.5: QQ-plot of the random effects at the level Replicate for primer combination PC1

3.4.2 fMarker

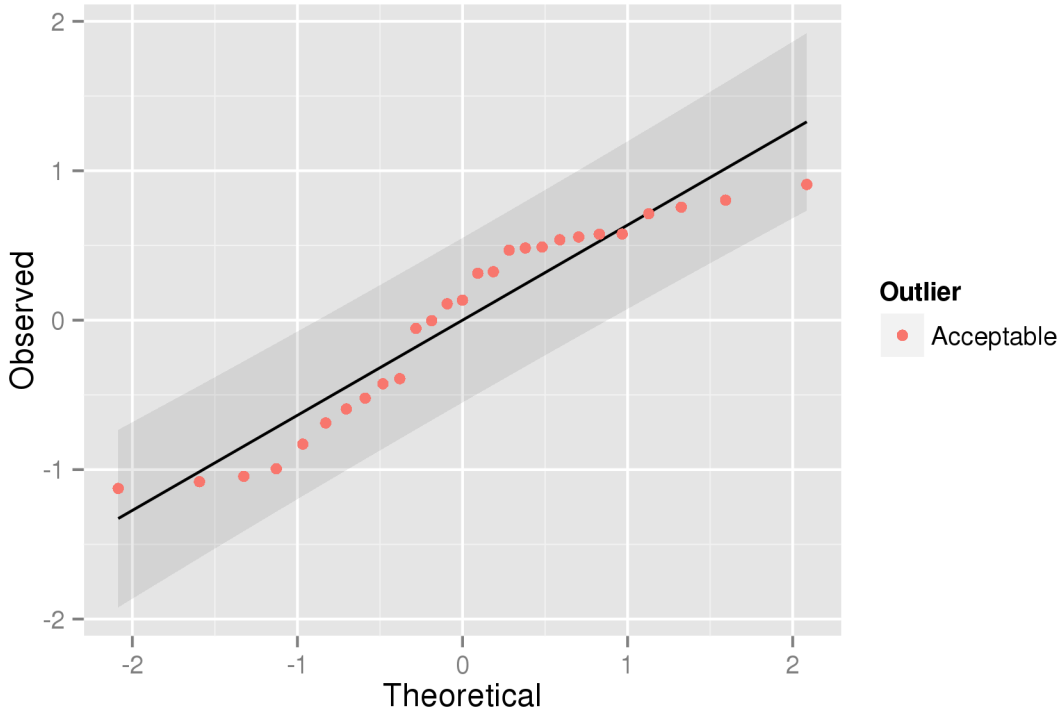

Figure 3.6: QQ-plot of the random effects at the level fMarker for primer combination PC1

3.4.3 Plate

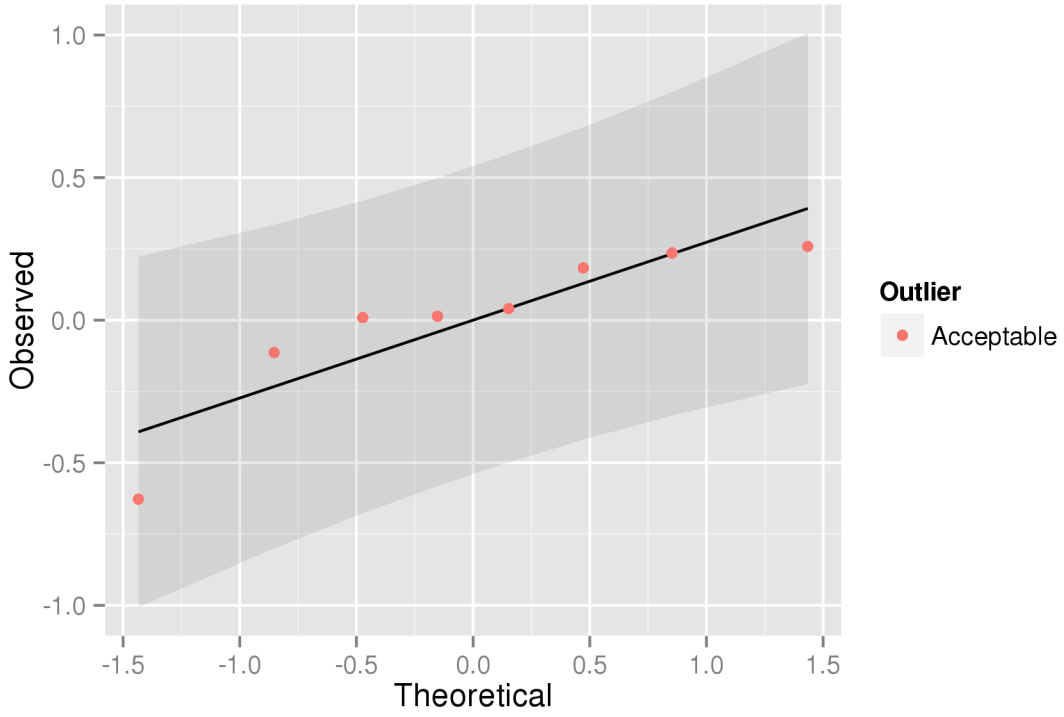

Figure 3.7: QQ-plot of the random effects at the level Plate for primer combination PC1

3.4.4 Globale outliers

Replicate	Marker	Observed
C.09.1863	431.000	-7.017
C.09.1940	411.000	-6.371
C.09.2176	431.000	-5.289
C.09.2040	420.000	-5.095
C.09.1953	347.000	-5.081
C.09.2183	539.000	-4.993
C.09.2011	480.000	-4.621
C.09.1911	387.000	-4.428
C.09.1997	387.000	-4.331
C.09.2182	431.000	-3.993
C.09.1944	420.000	-3.942
C.09.2042	420.000	-3.812
C.09.2011	431.000	-3.780
C.09.2178	420.000	-3.772
C.09.2055	539.000	-3.660
C.09.2006	480.000	-3.618
C.09.2027	539.000	-3.594
C.09.2183	411.000	-3.533
C.09.1911	431.000	-3.401
C.09.2166	420.000	-3.343
C.09.1927	431.000	-3.331
C.09.1997	411.000	-3.326
C.09.1911	420.000	-3.280
C.09.2025	449.000	-3.266
C.09.2241	411.000	-3.257
C.09.2178	411.000	-3.233
C.09.2042	480.000	-3.217

C.09.1944	411.000	-3.211
C.09.1935	411.000	-3.175
C.09.1945	411.000	-3.141
C.09.1928	411.000	-3.139
C.09.2024	431.000	-3.117
C.09.2023	539.000	-3.107
C.09.1911	411.000	-3.099
C.09.2006	387.000	-3.075
C.09.1911	528.000	-3.039
C.09.2040	431.000	-3.035
C.09.1805	297.000	-3.010
C.09.2025	480.000	-3.005
C.09.2049	480.000	-2.971
C.09.2011	420.000	-2.953
C.09.1811	411.000	-2.911
C.09.1944	431.000	-2.906
C.09.1945	250.000	-2.877
C.09.1772	347.000	-2.845
C.09.2000	431.000	-2.837
C.09.1940	420.000	-2.802
C.09.1859	431.000	-2.762
C.09.2241	420.000	-2.749
C.09.2040	411.000	-2.735
C.09.1911	347.000	-2.691
C.09.2133	250.000	-2.687
C.09.2025	431.000	-2.679
C.09.2006	431.000	-2.665
C.09.2242	411.000	-2.647
C.09.2081	387.000	-2.641
C.09.1796	250.000	-2.641
C.09.2220	178.000	-2.608
C.09.1809	347.000	-2.596
C.09.2177	431.000	-2.593
C.09.1911	592.000	-2.584
C.09.1811	387.000	-2.583
C.09.1944	250.000	-2.572
C.09.2179	539.000	-2.543
C.09.1848	347.000	-2.535
C.09.1794	232.000	-2.517
C.09.1770	159.000	-2.512
C.09.1909	281.000	-2.505
C.09.1926	288.000	-2.490

Table 3.4: QQ-plot of the residuals for primer combination PC1

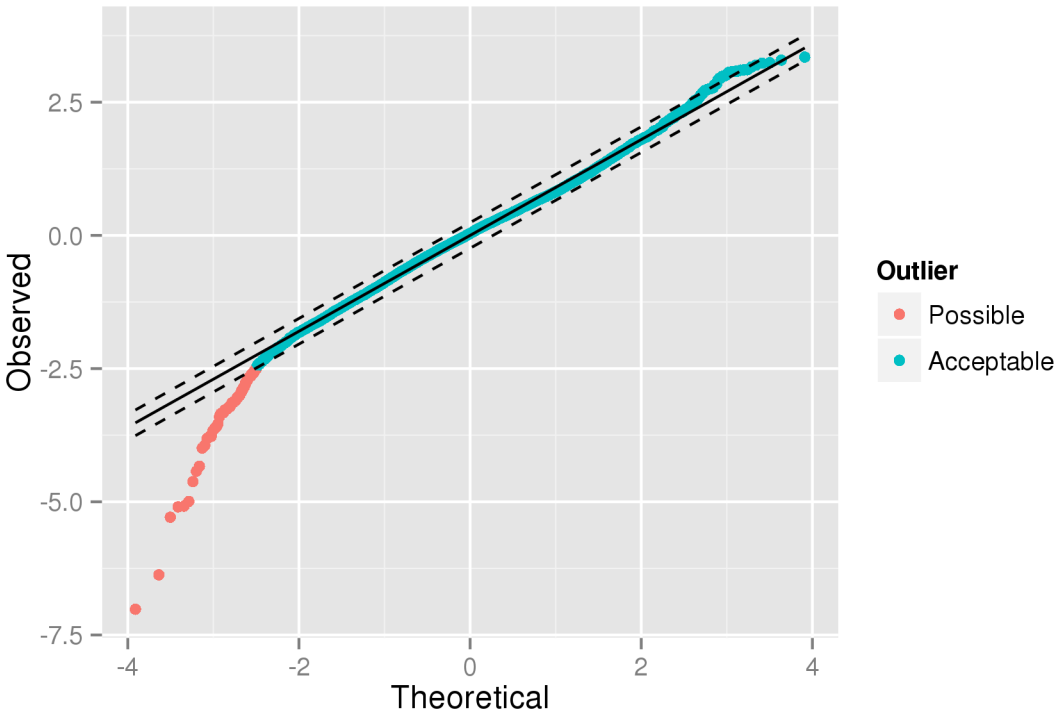

Figure 3.8: QQ-plot of the residuals for primer combination PC1

Chapter 4

Classifying the (normalised) fluorescence

The normalisation removes the effects of plate, replicate, capilar, marker,... for the fluorescence. We expect that the normalise fluoresence of polymorphic markers has a bimodal distribution: a baseline fluorescence when the marker is absent and a high fluorescence when the marker is present. On top of that there will be some measurement error changing the fluorescence. So we don't get two values but a range of values.

classify estimates the density distribution of the normalised fluorescence for each marker. These distributions are displayed on graphs when the user sets *output* to `"screen"` (default) or `"tex"`. Then the algorithm looks for different peaks in the distribution. A peak is a local maximum with a height which is at least *thresholdPeakRatio* of the largest peak (default = `0.03`). The border between two peaks is set at the local minimum between two sufficiently large local maxima. If this results in more than *maxBorder* border(s) (default = `1`), then only the *maxBorder* smallest border(s) are retained. Then there is only one sufficiently large local maximum, then the border is set at `Inf`. Such markers are considered monomorphic. The normalised fluorescence is binned with these borders and the result is stored in the `Score` variable in the `fluorescence data.frame` of the `AFLP` object.

```
> Tilia <- output$data
> Tilia <- classify(Tilia, output = "tex", device = "png")
```

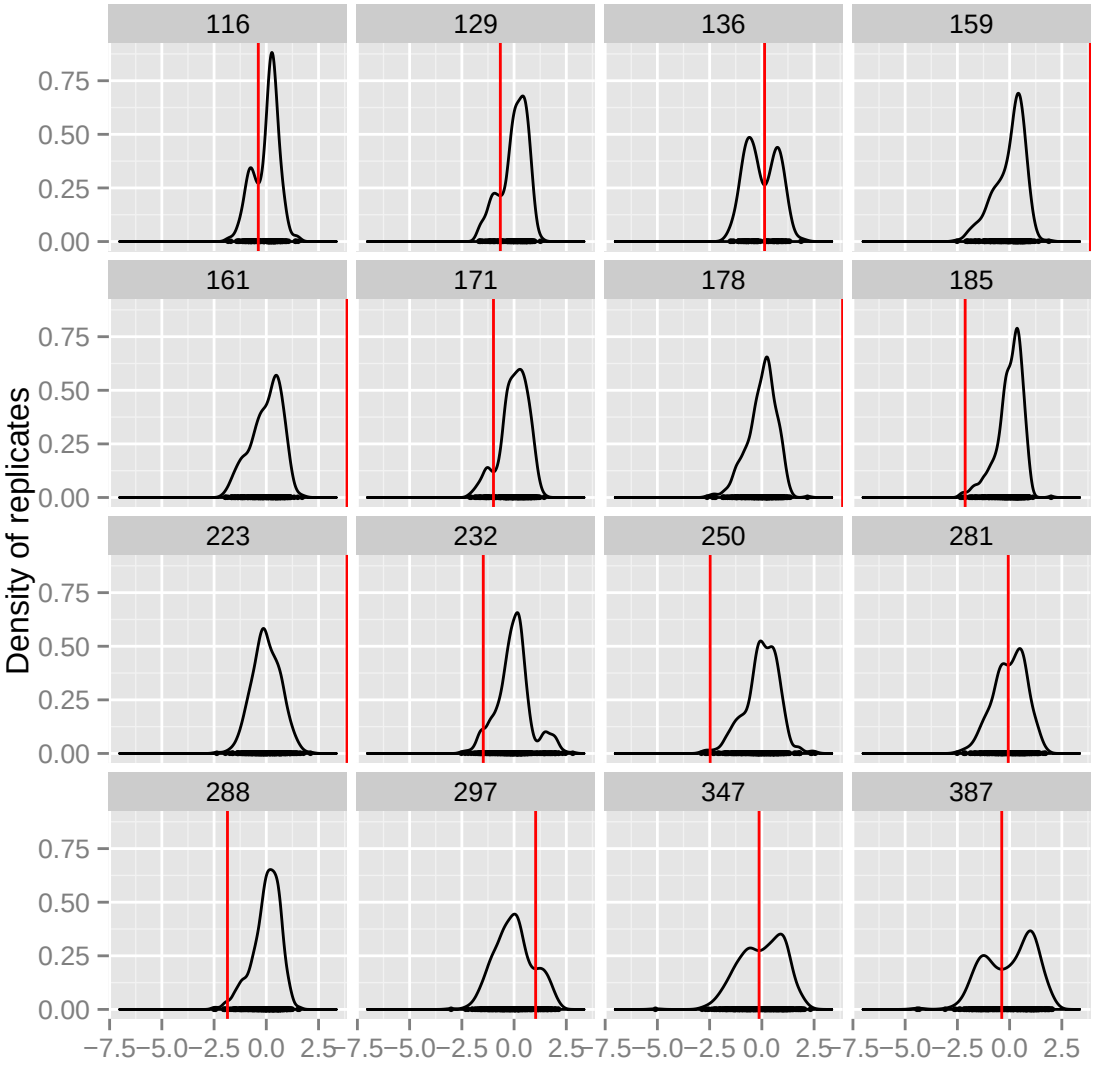

Figure 4.1: Density of normalised fluorescence and cut-off values per class for PC1

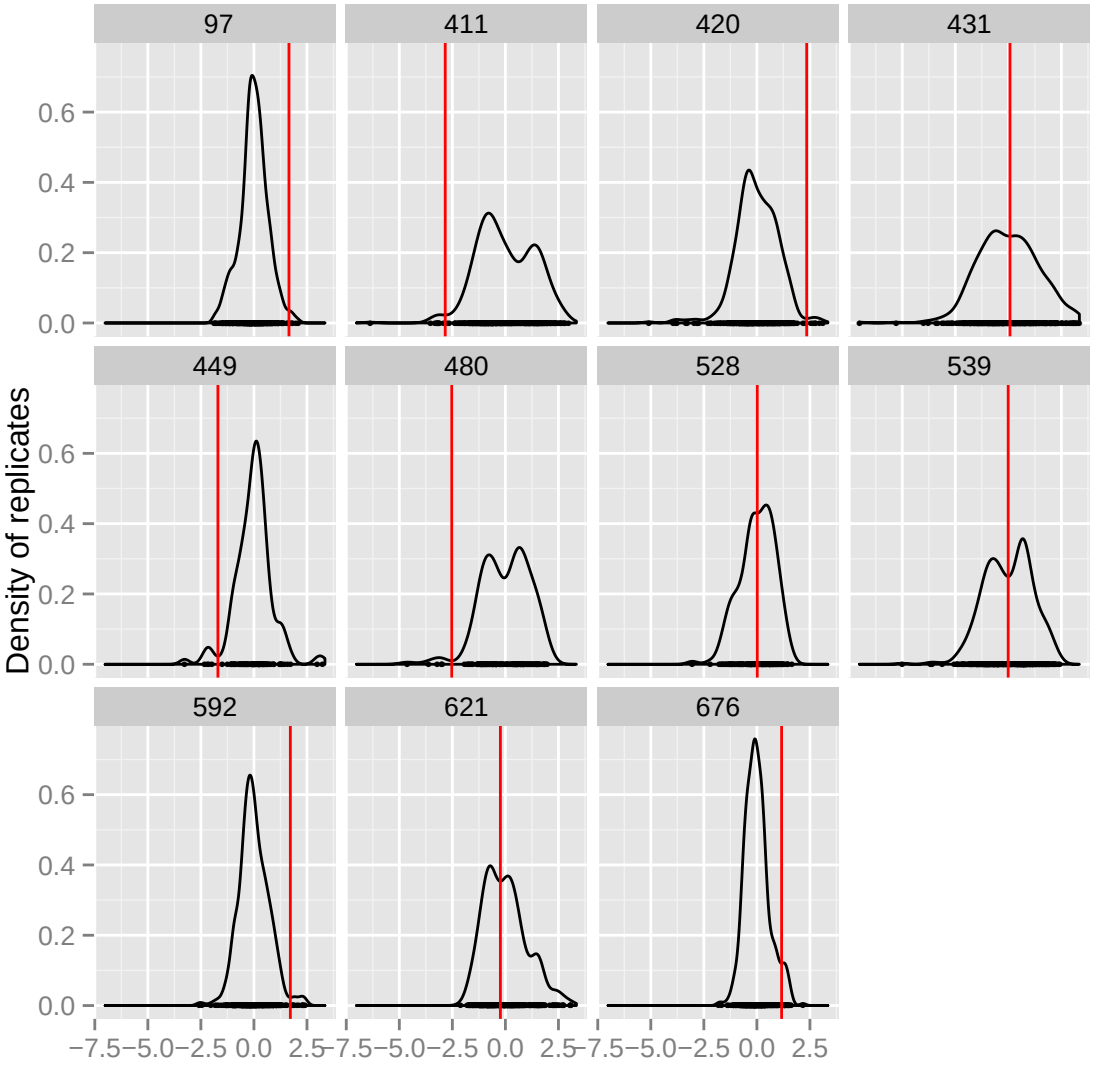

Figure 4.2: Density of normalised fluorescence and cut-off values per class for PC1

Chapter 5

Estimating repeatability

Any classification has a risk for misclassification. However, this risk is high variable. In order to work with thrustworthy data, we strongly recommend to assest the repeatability of the analysis. A second analysis of the same sample should, in theory, yield an identical DNA fingerprint. This is in practice hardly the case since the signal can be influenced by other factors than just the true underlying signal from the DNA.

5.1 Repeatability based on the fluorescence

A first way to investigate this, it to look at the variance in raw fluorescence within measurements for the same marker and the same specimen. The variance is zero when all measurements are equal and increases when the differences become larger. We calculate these variances for each combination of marker and specimen that have multiple measurement (e.g. specimens with more than one replicate per specimen). Averaging these variances over a given specimen gives an idea of the average repeatability of that specimen. A low score indicates very similar measurements within the specimen, whereas a high score is an indication of dissimilar measurements. E.g. when one of the replicates of a specimen has a problematic amplification, the fluorescence measurements will be much lower than the other replicates of that specimen. That would result in a high average variance. Likewise we can aggregate over a given marker, highlighting potential problems due to the marker and not the replicates. The interpretation is similar.

Normalising the fluorescence should remove some of the noise in the data due to batch effects. Therefore the repeatability score for the normalised fluorescence will on average be lower than the repeatability score of the raw fluorescence. The **repeatability** function create two scatterplots displaying the repeatability score for the raw versus the normalised fluorescence score (provided `output = "screen"` or `output = "tex"`). The line indicates equal repeatability scores. Thus most of the points should be on the lower righthand side of this line (indicating more variability in the raw fluorescence than in the normalised

fluorescence).

5.2 Repeatability based on scores

When the (normalised) fluorescence is classified into presence-absence data, then we expect that all replicates of a given specimen get the same state for a given marker. E.g. the marker is either present or absent in all replicates. If not, then we clearly have misclassifications. Unfortunalty, we don't know the truth: is the marker is reality present or absent in a given specimen? We assume that the majority is correctly classified. Assume we have five replicates of a given specimen and for a given marker we get 4 absences and 1 presence. The majority is absent, thus we assume 1 misclassification. Due to our assumption, the maximum number of misclassifications is not equal to the number of replicates but the half that number rounded down to an integer. Thus we have maximum 2 misclassifications when we have 5 replicates (see table˜5.2). Based on the observed scores on a given marker for all replicate of a given specimen and our assumption, we can calculate the observed number of errors and the maximum number of errors.

Absent	Present	Assumption	Misclassifications
0	5	Present	0
1	4	Present	1
2	3	Present	2
3	2	Absent	2
4	1	Absent	1
5	0	Absent	0

Table 5.1: Number of absent and present scores of a given marker for the replicates of a given specimen and the resulting assumption and number of misclassifications.

When we aggregate these numbers we get meaningfull information on markers or specimens. Our repeatability scores are based on the sums of the observed errors and the sum of the theoretically maximum number of errors. (5.1) defines the repeatability R_i for a marker i and (5.2) the reapeatability R_j for specimen j . t_{ij} is the theoretical maximum number of errors for marker i and specimen j, o_{ij} is the observed maximum number of errors for marker i and specimen j, s is the number of specimens and m the number of markers. R_i and R_j are restricted to the interval $[0, 1]$, with 1 indicating no observed errors (perfect repeatability) and 0 indicating that the number of observed errors is always equal to the theoretical maximum (not repeatable at all).

$$R_i = \frac{\sum_{j=1}^{s} t_{ij} - \sum_{j=1}^{s} o_{ij}}{\sum_{j=1}^{s} t_{ij}} \tag{5.1}$$

$$R_j = \frac{\sum_{i=1}^{m} t_{ij} - \sum_{j=1}^{m} o_{ij}}{\sum_{j=1}^{m} t_{ij}} \tag{5.2}$$

```
> output <- repeatability(Tilia, output = "tex", device = "png")
```

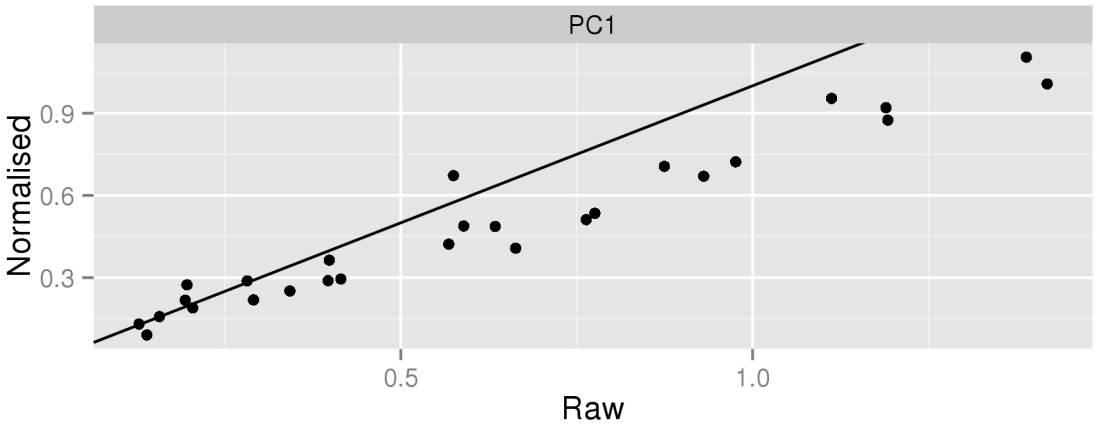

Figure 5.1: Repeatability for markers based on fluorescence

PC	Marker	Raw	Normalised
PC1	97	0.398	0.364
PC1	116	0.157	0.158
PC1	129	0.128	0.131
PC1	136	0.575	0.672
PC1	159	0.204	0.190
PC1	161	0.282	0.288
PC1	171	0.194	0.218
PC1	178	0.415	0.295
PC1	185	0.342	0.251
PC1	223	0.196	0.274
PC1	232	0.776	0.535
PC1	250	0.763	0.512
PC1	281	0.589	0.488
PC1	288	0.397	0.289
PC1	297	0.663	0.408
PC1	347	0.976	0.723
PC1	387	1.189	0.920
PC1	411	0.875	0.706
PC1	420	1.192	0.875
PC1	431	1.389	1.104
PC1	449	1.419	1.007
PC1	480	0.930	0.670
PC1	528	0.139	0.091

PC1	539	1.112	0.954
PC1	592	0.568	0.422
PC1	621	0.634	0.487
PC1	676	0.291	0.219

Table 5.2: Repeatability for markers based on fluorescence. Smaller is better.

Specimen	Raw	Normalised
7109	0.487	0.446
7203	0.309	0.251
7220	0.259	0.173
C/07/3323	0.346	0.254
C/07/3327	0.487	0.416
C/07/3331	0.287	0.270
C/07/3336	1.323	0.859
C/07/3337	0.665	0.545
C/07/3348	0.438	0.353
C/07/3353	0.099	0.099
C/07/3361	0.295	0.288
C/07/3363	0.652	0.563
C/07/3387	0.570	0.498
C/07/3392	0.623	0.310
C/07/3410	0.100	0.077
C/07/3571	0.083	0.081
C/07/3589	0.288	0.220
C/07/3600	0.773	0.751
C/07/3629	0.303	0.250
C/07/3633	0.912	0.830
C/08/1121	2.724	2.195
C/08/1134	1.324	0.948
C/08/1139	0.597	0.332
C/08/1145	0.229	0.202
C/08/1149	0.170	0.138
C/08/1159	3.202	3.093
C/08/1173	0.892	0.456
C/08/1220	0.910	0.658
C/08/1223	0.337	0.166
C/08/1731	0.290	0.153
C/08/1751	0.916	0.353
C/08/1786	0.546	0.456
C/08/1787	0.434	0.247
T*eur 130	0.508	0.497
T*eur 137	0.192	0.174
T. cor 69	0.179	0.144
T.eur 132	0.243	0.216
T.pla 157	0.226	0.168
T.pla 161	0.527	0.420
T.pla Savel 08	0.271	0.173
TC SI22BDO	0.869	0.732
TV BAVEVZA	1.164	0.950
QC method	0.588	0.387

Table 5.3: Repeatability for specimens based on fluorescence. Smaller is better.

PC	Marker	Polymorph	Score	Errors	MaxErrors	nBin
PC1	136	TRUE	0.400	6	10	32
PC1	539	TRUE	0.413	27	46	131
PC1	621	TRUE	0.522	22	46	131
PC1	347	TRUE	0.565	20	46	131
PC1	431	TRUE	0.587	19	46	130
PC1	387	TRUE	0.658	13	38	113
PC1	281	TRUE	0.667	11	33	97
PC1	528	TRUE	0.824	3	17	50
PC1	297	TRUE	0.842	6	38	113
PC1	480	TRUE	0.857	2	14	46
PC1	116	TRUE	0.857	3	21	63
PC1	232	TRUE	0.870	6	46	131
PC1	129	TRUE	0.882	2	17	49
PC1	676	TRUE	0.891	5	46	131
PC1	449	TRUE	0.909	1	11	32
PC1	171	TRUE	0.914	3	35	99
PC1	592	TRUE	0.935	3	46	131
PC1	288	TRUE	0.957	2	46	131
PC1	411	TRUE	0.957	2	46	131
PC1	420	TRUE	0.957	2	46	130
PC1	97	TRUE	0.978	1	46	131
PC1	185	TRUE	0.978	1	46	131
PC1	159	FALSE	1.000	0	27	81
PC1	161	FALSE	1.000	0	30	84
PC1	178	FALSE	1.000	0	38	113
PC1	223	FALSE	1.000	0	42	117
PC1	250	TRUE	1.000	0	46	131

Table 5.4: Repeatability for markers based on score

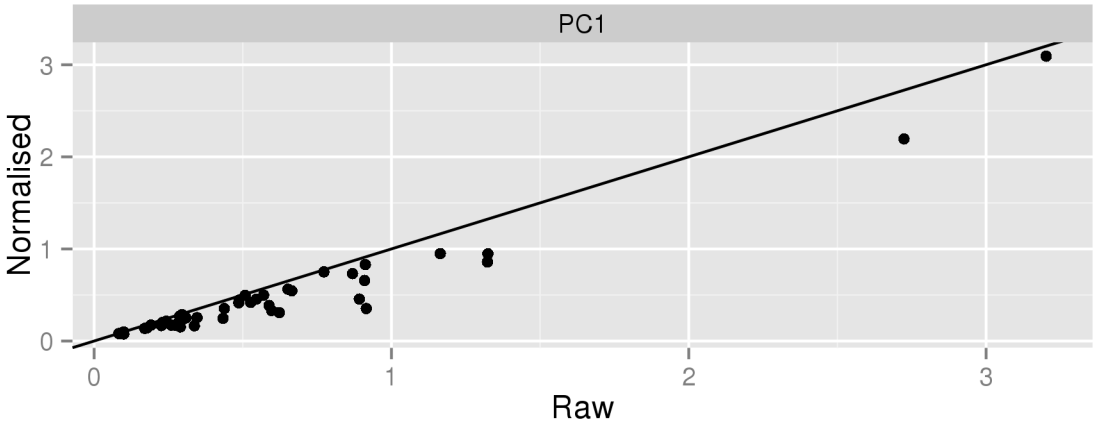

Figure 5.2: Repeatability for specimens based on fluorescence

PC	Specimen	Score	Errors	MaxErrors	nBin	MaxErrorsAll	nBinAll
PC1	C/08/1159	0.600	8	17	52	20	62
PC1	TV BAVEVZA	0.609	9	19	60	23	72
PC1	C/07/3633	0.652	8	20	62	23	72
PC1	C/08/1121	0.654	9	22	63	26	75
PC1	C/07/3363	0.696	7	19	60	23	72
PC1	C/07/3600	0.731	7	22	63	26	75
PC1	TC SI22BDO	0.739	6	19	56	23	68
PC1	C/08/1134	0.739	6	20	58	23	65
PC1	C/07/3336	0.750	5	17	52	20	61
PC1	C/08/1731	0.762	5	18	53	21	62
PC1	C/07/3387	0.769	6	22	62	26	73
PC1	C/08/1139	0.783	5	20	58	23	65
PC1	C/08/1787	0.789	4	18	58	19	64
PC1	C/08/1173	0.800	4	17	52	20	59
PC1	C/08/1220	0.800	4	17	54	20	63
PC1	T*eur 130	0.808	5	22	63	26	75
PC1	7109	0.818	4	18	54	22	65
PC1	C/07/3337	0.818	4	18	36	22	44
PC1	C/07/3392	0.818	4	18	55	22	67
PC1	C/08/1751	0.818	4	18	54	22	65
PC1	C/08/1786	0.824	3	15	52	17	60
PC1	C/07/3327	0.826	4	19	56	23	65
PC1	7203	0.857	3	18	54	21	64
PC1	C/07/3331	0.870	3	19	55	23	66
PC1	T.pla 161	0.870	3	20	58	23	68
PC1	QC method	0.877	10	71	147	81	171
PC1	C/07/3348	0.882	2	15	48	17	56
PC1	C/07/3629	0.882	2	15	48	17	53
PC1	T.pla 157	0.900	2	17	53	20	63
PC1	C/08/1145	0.905	2	18	51	21	59
PC1	T. cor 69	0.905	2	18	56	21	65
PC1	C/07/3323	0.909	2	18	51	22	61
PC1	C/08/1223	0.913	2	20	58	23	68
PC1	C/07/3361	0.941	1	15	30	17	34
PC1	7220	0.944	1	15	36	18	43
PC1	C/08/1149	0.947	1	18	53	19	58
PC1	T.eur 132	0.947	1	18	51	19	55
PC1	T.pla Savel 08	0.950	1	17	54	20	63
PC1	C/07/3589	0.957	1	20	57	23	66
PC1	C/07/3353	1.000	0	15	48	17	53
PC1	C/07/3410	1.000	0	15	48	17	53
PC1	C/07/3571	1.000	0	15	48	17	55
PC1	T*eur 137	1.000	0	20	58	23	67

Table 5.5: Repeatability for specimens based on score

PC	Specimen	ReplicateA	ReplicateB	Score	Errors	MaxErrors
PC1	7109	C.09.2067	C.09.2088	0.926	2	27
PC1	7109	C.09.2067	C.09.2246	0.889	3	27
PC1	7109	C.09.2088	C.09.2246	0.889	3	27
PC1	7203	C.09.2085	C.09.2202	0.889	3	27
PC1	7203	C.09.2085	C.09.2209	0.889	3	27
PC1	7203	C.09.2202	C.09.2209	1.000	0	27
PC1	7220	C.09.1874	C.09.2002	0.926	2	27
PC1	C/07/3323	C.09.2083	C.09.2103	0.963	1	27
PC1	C/07/3323	C.09.2083	C.09.2138	0.852	4	27
PC1	C/07/3323	C.09.2103	C.09.2138	0.889	3	27
PC1	C/07/3327	C.09.1903	C.09.1914	0.963	1	27
PC1	C/07/3327	C.09.1903	C.09.1983	0.815	5	27
PC1	C/07/3327	C.09.1914	C.09.1983	0.778	6	27
PC1	C/07/3331	C.09.1878	C.09.1884	0.926	2	27
PC1	C/07/3331	C.09.1878	C.09.2008	0.889	3	27
PC1	C/07/3331	C.09.1884	C.09.2008	0.815	5	27
PC1	C/07/3336	C.09.2001	C.09.2049	0.852	4	27
PC1	C/07/3336	C.09.2001	C.09.2242	0.852	4	27
PC1	C/07/3336	C.09.2049	C.09.2242	0.852	4	27

PC1	C/07/3571	C.09.2125	C.09.2130	1.000	0	27
PC1	C/07/3571	C.09.2125	C.09.2195	0.963	1	27
PC1	C/07/3571	C.09.2130	C.09.2195	0.963	1	27
PC1	C/07/3589	C.09.1841	C.09.1866	1.000	0	27
PC1	C/07/3589	C.09.1841	C.09.2019	0.889	3	27
PC1	C/07/3589	C.09.1866	C.09.2019	0.889	3	27
PC1	C/07/3600	C.09.1756	C.09.1786	0.963	1	27
PC1	C/07/3600	C.09.1756	C.09.1918	0.630	10	27
PC1	C/07/3600	C.09.1786	C.09.1918	0.667	9	27
PC1	C/07/3629	C.09.1990	C.09.2126	0.889	3	27
PC1	C/07/3629	C.09.1990	C.09.2162	0.852	4	27
PC1	C/07/3629	C.09.2126	C.09.2162	0.963	1	27
PC1	C/07/3633	C.09.1773	C.09.1833	0.815	5	27
PC1	C/07/3633	C.09.1773	C.09.1845	0.815	5	27
PC1	C/07/3633	C.09.1833	C.09.1845	0.704	8	27
PC1	C/08/1121	C.09.1774	C.09.1799	0.963	1	27
PC1	C/08/1121	C.09.1774	C.09.1911	0.593	11	27
PC1	C/08/1121	C.09.1799	C.09.1911	0.630	10	27
PC1	C/08/1134	C.09.1865	C.09.1867	1.000	0	27
PC1	C/08/1134	C.09.1865	C.09.1953	0.667	9	27
PC1	C/08/1134	C.09.1867	C.09.1953	0.667	9	27
PC1	C/08/1139	C.09.1853	C.09.1868	1.000	0	27
PC1	C/08/1139	C.09.1853	C.09.1936	0.704	8	27
PC1	C/08/1139	C.09.1868	C.09.1936	0.704	8	27
PC1	C/08/1145	C.09.2123	C.09.2201	0.963	1	27
PC1	C/08/1145	C.09.2123	C.09.2244	0.926	2	27
PC1	C/08/1145	C.09.2201	C.09.2244	0.963	1	27
PC1	C/08/1149	C.09.1934	C.09.1985	0.963	1	27
PC1	C/08/1149	C.09.1934	C.09.2021	0.926	2	27
PC1	C/08/1149	C.09.1985	C.09.2021	0.889	3	27
PC1	C/08/1159	C.09.1780	C.09.1997	0.593	11	27
PC1	C/08/1159	C.09.1780	C.09.2042	0.556	12	27
PC1	C/08/1159	C.09.1997	C.09.2042	0.815	5	27
PC1	C/08/1173	C.09.1937	C.09.1996	0.741	7	27
PC1	C/08/1173	C.09.1937	C.09.2050	0.889	3	27
PC1	C/08/1173	C.09.1996	C.09.2050	0.852	4	27
PC1	C/08/1220	C.09.1819	C.09.2055	0.889	3	27
PC1	C/08/1220	C.09.1819	C.09.2058	0.778	6	27
PC1	C/08/1220	C.09.2055	C.09.2058	0.889	3	27
PC1	C/08/1223	C.09.1810	C.09.1860	0.963	1	27
PC1	C/08/1223	C.09.1810	C.09.2066	0.852	4	27
PC1	C/08/1223	C.09.1860	C.09.2066	0.889	3	27
PC1	C/08/1731	C.09.1998	C.09.2207	0.889	3	27
PC1	C/08/1731	C.09.1998	C.09.2226	0.852	4	27
PC1	C/08/1731	C.09.2207	C.09.2226	0.889	3	27
PC1	C/08/1751	C.09.2063	C.09.2064	1.000	0	27
PC1	C/08/1751	C.09.2063	C.09.2227	0.815	5	27
PC1	C/08/1751	C.09.2064	C.09.2227	0.815	5	27
PC1	C/08/1786	C.09.1794	C.09.2155	0.815	5	27
PC1	C/08/1786	C.09.1794	C.09.2174	0.852	4	27
PC1	C/08/1786	C.09.2155	C.09.2174	0.963	1	27
PC1	C/08/1787	C.09.1754	C.09.1979	0.778	6	27
PC1	C/08/1787	C.09.1754	C.09.1995	0.778	6	27
PC1	C/08/1787	C.09.1979	C.09.1995	1.000	0	27
PC1	T*eur 130	C.09.1767	C.09.1784	0.963	1	27
PC1	T*eur 130	C.09.1767	C.09.1890	0.815	5	27
PC1	T*eur 130	C.09.1784	C.09.1890	0.778	6	27
PC1	T*eur 137	C.09.1840	C.09.1857	1.000	0	27
PC1	T*eur 137	C.09.1840	C.09.2192	1.000	0	27
PC1	T*eur 137	C.09.1857	C.09.2192	1.000	0	27
PC1	T. cor 69	C.09.1808	C.09.2189	0.926	2	27
PC1	T. cor 69	C.09.1808	C.09.2198	0.963	1	27
PC1	T. cor 69	C.09.2189	C.09.2198	0.963	1	27
PC1	T.eur 132	C.09.1951	C.09.1969	0.963	1	27
PC1	T.eur 132	C.09.1951	C.09.2166	0.926	2	27
PC1	T.eur 132	C.09.1969	C.09.2166	0.963	1	27
PC1	T.pla 157	C.09.1877	C.09.2010	0.926	2	27
PC1	T.pla 157	C.09.1877	C.09.2022	0.889	3	27
PC1	T.pla 157	C.09.2010	C.09.2022	0.963	1	27
PC1	T.pla 161	C.09.1826	C.09.1848	1.000	0	27
PC1	T.pla 161	C.09.1826	C.09.2099	0.815	5	27
PC1	T.pla 161	C.09.1848	C.09.2099	0.815	5	27
PC1	T.pla Savel 08	C.09.1839	C.09.2012	0.926	2	27
PC1	T.pla Savel 08	C.09.1839	C.09.2039	0.926	2	27
PC1	T.pla Savel 08	C.09.2012	C.09.2039	1.000	0	27
PC1	TC SI22BDO	C.09.1893	C.09.1909	0.815	5	27
PC1	TC SI22BDO	C.09.1893	C.09.2101	0.852	4	27
PC1	TC SI22BDO	C.09.1909	C.09.2101	0.815	5	27
PC1	TV BAVEVZA	C.09.1764	C.09.1923	0.741	7	27
PC1	TV BAVEVZA	C.09.1764	C.09.1926	0.778	6	27
PC1	TV BAVEVZA	C.09.1923	C.09.1926	0.741	7	27
PC1	QC method	qc_method_a1	qc_method_a2	0.889	3	27
PC1	QC method	qc_method_a1	qc_method_a3	0.852	4	27
PC1	QC method	qc_method_a1	qc_method_a4	0.852	4	27
PC1	QC method	qc_method_a1	qc_method_a5	0.889	3	27
PC1	QC method	qc_method_a1	qc_method_a6	0.926	2	27
PC1	QC method	qc_method_a1	qc_method_a7	0.889	3	27
PC1	QC method	qc_method_a1	qc_method_a8	0.926	2	27
PC1	QC method	qc_method_a2	qc_method_a3	0.963	1	27
PC1	QC method	qc_method_a2	qc_method_a4	0.963	1	27
PC1	QC method	qc_method_a2	qc_method_a5	0.852	4	27

PC1	QC method	qc_method_a2	qc_method_a6	0.963	1	27
PC1	QC method	qc_method_a2	qc_method_a7	0.852	4	27
PC1	QC method	qc_method_a2	qc_method_a8	0.963	1	27
PC1	QC method	qc_method_a3	qc_method_a4	0.926	2	27
PC1	QC method	qc_method_a3	qc_method_a5	0.815	5	27
PC1	QC method	qc_method_a3	qc_method_a6	0.926	2	27
PC1	QC method	qc_method_a3	qc_method_a7	0.815	5	27
PC1	QC method	qc_method_a3	qc_method_a8	0.926	2	27
PC1	QC method	qc_method_a4	qc_method_a5	0.815	5	27
PC1	QC method	qc_method_a4	qc_method_a6	0.926	2	27
PC1	QC method	qc_method_a4	qc_method_a7	0.815	5	27
PC1	QC method	qc_method_a4	qc_method_a8	0.926	2	27
PC1	QC method	qc_method_a5	qc_method_a6	0.889	3	27
PC1	QC method	qc_method_a5	qc_method_a7	1.000	0	27
PC1	QC method	qc_method_a5	qc_method_a8	0.889	3	27
PC1	QC method	qc_method_a6	qc_method_a7	0.889	3	27
PC1	QC method	qc_method_a6	qc_method_a8	1.000	0	27
PC1	QC method	qc_method_a7	qc_method_a8	0.889	3	27

Table 5.6: Repeatability for replicates based on score

PC	PlateA	PlateB	Score	Errors	MaxErrors
PC1	1	1	0.970	12	405
PC1	1	2	0.815	10	54
PC1	2	2	0.963	18	486
PC1	1	3	0.753	40	162
PC1	2	3	0.963	1	27
PC1	3	3	0.922	36	459
PC1	1	4	0.889	6	54
PC1	2	4	0.864	11	81
PC1	3	4	0.926	4	54
PC1	4	4	0.989	4	378
PC1	1	5	0.833	9	54
PC1	2	5	0.917	9	108
PC1	3	5	0.917	9	108
PC1	4	5	0.901	8	81
PC1	5	5	0.947	26	486
PC1	1	6	0.963	1	27
PC1	2	6	0.926	6	81
PC1	3	6	0.864	11	81
PC1	4	6	0.963	1	27
PC1	5	6	0.926	2	27
PC1	6	6	0.963	16	432
PC1	1	7	0.926	4	54
PC1	2	7	0.889	3	27
PC1	3	7	0.889	3	27
PC1	4	7	0.956	6	135
PC1	5	7	1.000	0	27
PC1	6	7	0.926	6	81
PC1	7	7	0.988	6	486
PC1	1	8	0.870	7	54
PC1	2	8	0.963	3	81
PC1	3	8	0.926	2	27
PC1	4	8	0.963	1	27
PC1	5	8	0.852	12	81
PC1	6	8	0.898	11	108
PC1	7	8	0.951	4	81
PC1	8	8	0.975	10	405

Table 5.7: Repeatability for plates based on score

PC	PlateA	PlateB	Score	Errors	MaxErrors
PC1	1	1	0.970	12	405
PC1	1	2	0.815	10	54
PC1	2	2	0.963	18	486
PC1	1	3	0.753	40	162
PC1	2	3	0.963	1	27
PC1	3	3	0.922	36	459
PC1	1	4	0.889	6	54
PC1	2	4	0.864	11	81
PC1	3	4	0.926	4	54
PC1	4	4	0.989	4	378
PC1	1	5	0.833	9	54
PC1	2	5	0.917	9	108
PC1	3	5	0.917	9	108
PC1	4	5	0.901	8	81
PC1	5	5	0.947	26	486
PC1	1	6	0.963	1	27
PC1	2	6	0.926	6	81
PC1	3	6	0.864	11	81
PC1	4	6	0.963	1	27
PC1	5	6	0.926	2	27
PC1	6	6	0.963	16	432
PC1	1	7	0.926	4	54

PC1	2	7	0.889	3	27
PC1	3	7	0.889	3	27
PC1	4	7	0.956	6	135
PC1	5	7	1.000	0	27
PC1	6	7	0.926	6	81
PC1	7	7	0.988	6	486
PC1	1	8	0.870	7	54
PC1	2	8	0.963	3	81
PC1	3	8	0.926	2	27
PC1	4	8	0.963	1	27
PC1	5	8	0.852	12	81
PC1	6	8	0.898	11	108
PC1	7	8	0.951	4	81
PC1	8	8	0.975	10	405

Table 5.8: Repeatability for plates based on score

5.3 Generating a design with replication

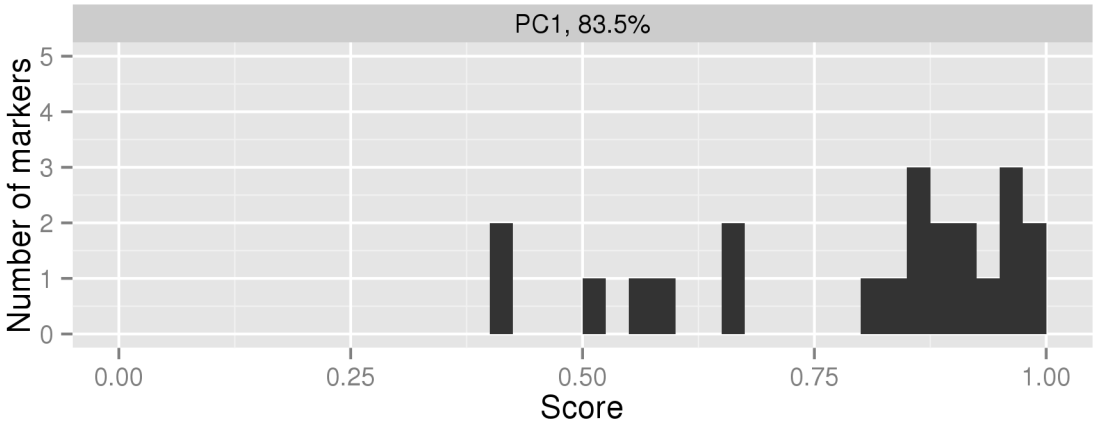

Figure 5.3: Repeatability for markers based on score

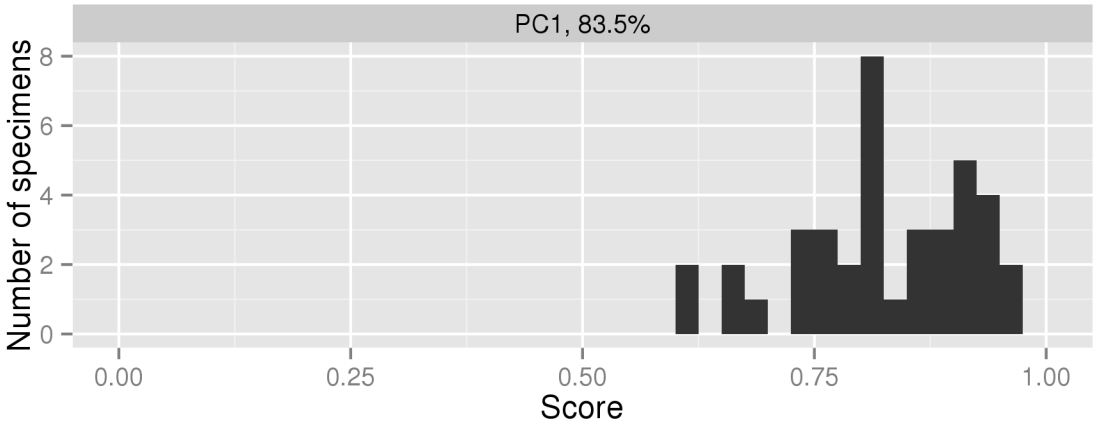

Figure 5.4: Repeatability for specimens based on score

Chapter 6

Further analysis

Appendix A

Tilia dataset

The Tilia dataset is a collection of leaf samples from 489 specimens of lime trees collected in Belgium. Field specialists determined the species of each tree and classified the specimens as belonging to *Tilia cordata*, *Tilia platyphylos* or their hybrid *Tilia europea (x)*. Some specimens were not classified in the field (table~**??**). The research goal was to see whether the field determination matches with the genetical information.

Group	Specimen
Tilia cordata	87
Tilia platyphyllos	122
Tilia europea (x)	132
Unknown	82

Table A.1: Number of specimen per group.

The AFLP was run on a LiCor Sequencer system with primer combination from tableA.

The LiCor images were imported in the Saga MX software. The images were sized using a size standard **(which?)** placed on lane 1. Desmiling was done using about xxx monomorphic markers in the 50 - 400 bp range. Then potential markers were selected. The fluoresence and size of of all selected bands were measured by the software and exported as tab-delimited text files. These files are available in the `extdata` folder of the package.

Code	Primer combination	Label
PC1	E-ACT M-CTG	700
PC2	E-ACG M-CAA	800
PC3	E-AGC M-CAT	700
PC4	E-ACA M-CTG	PC4

Table A.2: Used primer combination and labels

The design of the experiment is available as the *Tiliadesign* `data.frame`.

We would like to thank Kristien Vander Mijnsbrugge and An Vanden Broeck for their permission to use the dataset in the AFLP package.