Applied Machine Learning - Exercise 8 (15.06.2017)

Ben Wulf, Lie Hong, Amnon Bleich

Task 1)

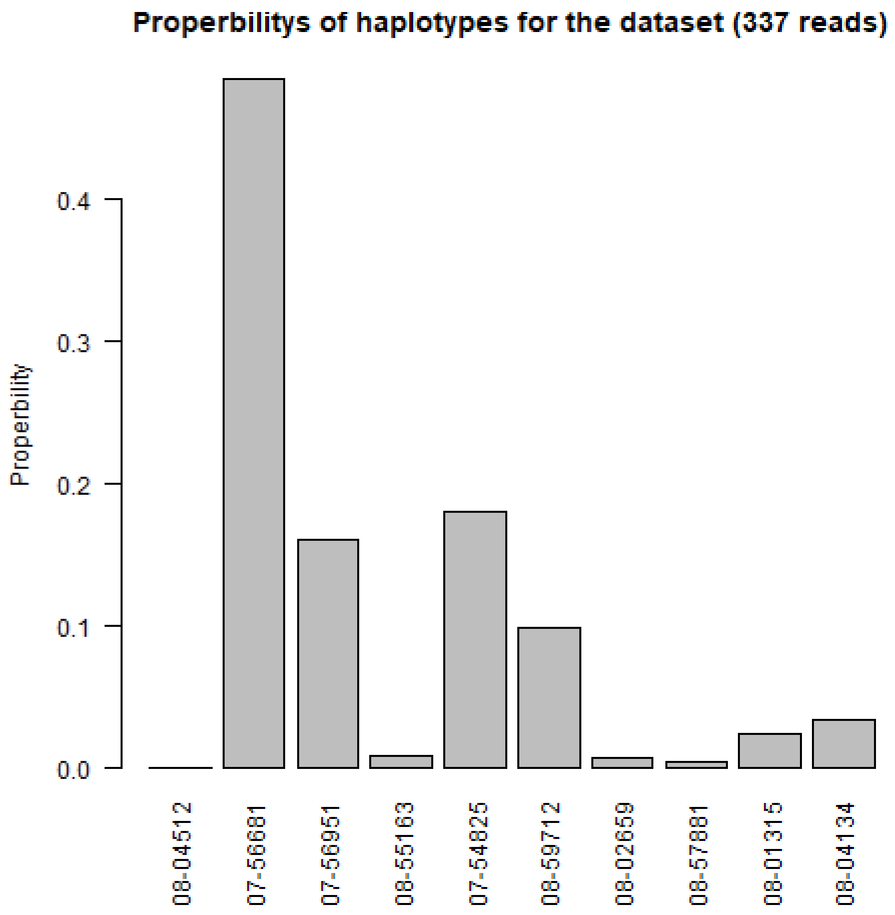
We should use the ‘clone mix’ dataset. Therefor, we used reads which has the prefix ‘acgagtgcgt’. We discharge any read which will not exactly map to this prefix. We count the occurance of every sequence identical copy of the reads. Later, these counts are used to get the U in the EM.

Task 2)

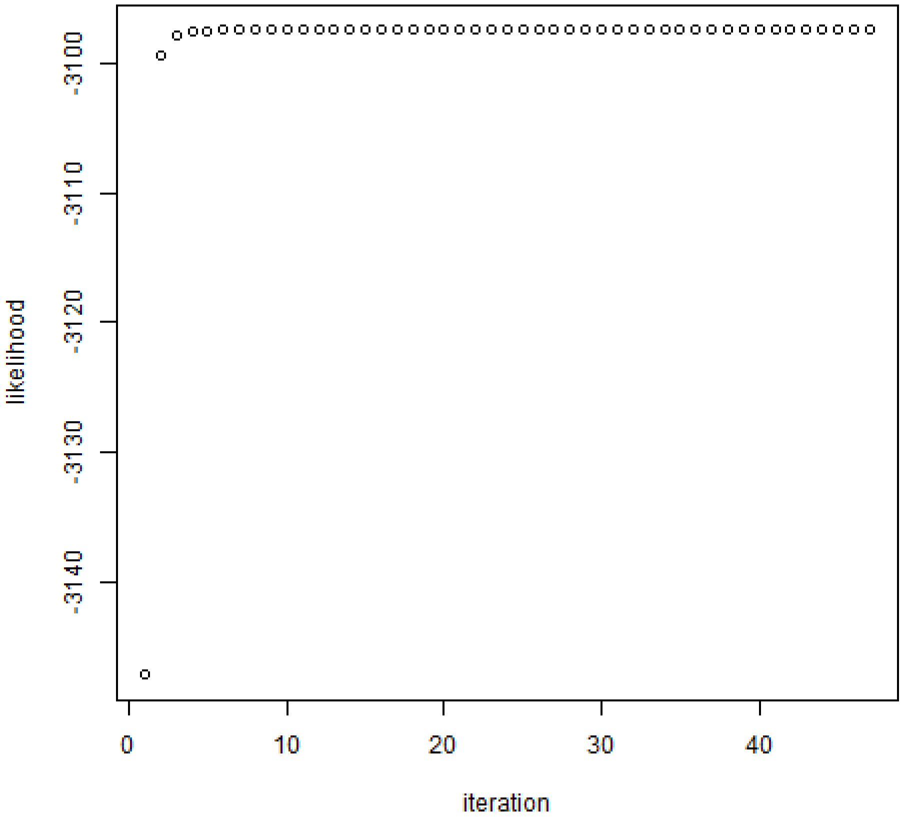
We create a boolean matrix , where every (unique sequence) read (r) is a row and every haplotype (h) is a column. An entry in this matrix is true if read Rr is in sequence identical containt in the the haplotype Hh, otherwise the entry is false. We create a subset of this matrix which will only contain sequences (reads) which will occure in at least one haplotype h. We also subset the counting vetor of the read sequences to sequences that occure in at least one haplotype h. 337 different read sequences will remain.

Task 3)

We were not sure what the Ur really means. We assume that Ur is the count of times a read sequence occurs in the dataset.
We decide to break the iterations for the EM if the changes of every p is smaller than a given epsilon. In the try with the full dataset we are using an epsilon of 10^{-10}. It seems that with a smaller epsilon the EM will never break.



We can see that haplotype 07-56681 is most likely for the given dataset and 08-04512 is extreme unlikely.



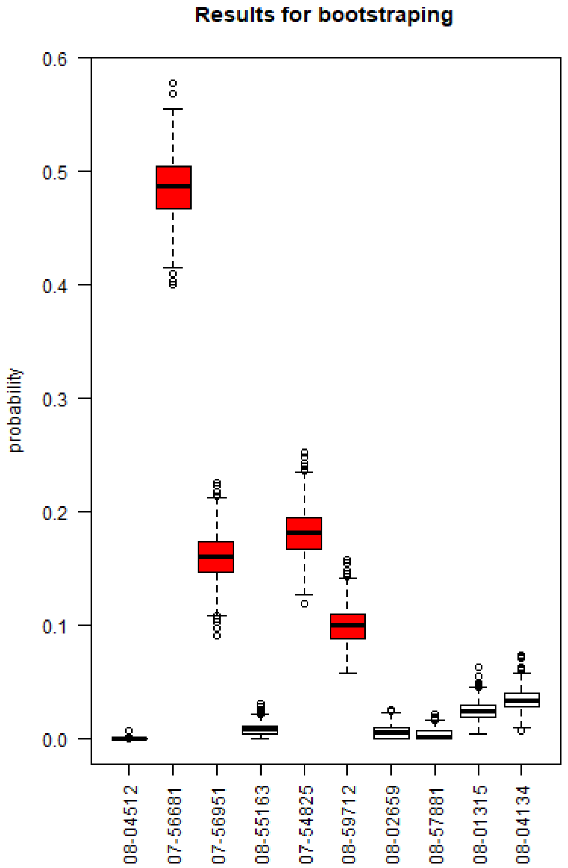
The likelihoods converging rapidly against the maximum and after the 5th iteration they are quite similar.

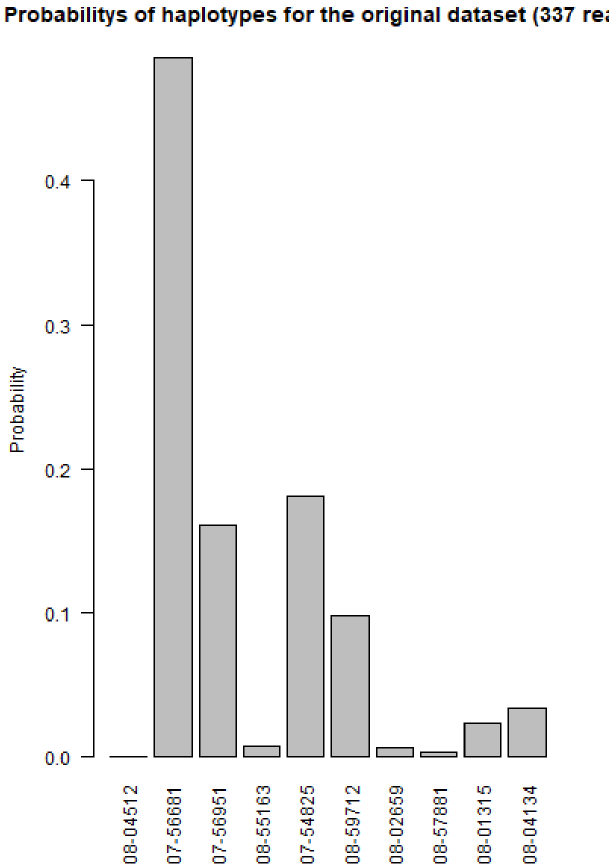
Task 4)

In each bootstrap iteration, we use again only unique read sequences for the matrix of read occurrences in the different haplotypes. The U vector is resampled for every iteration according to the sequences which are chose more than once (weights). The sequences are more often chosen if they are more frequently in the raw data. For time reasons, we decide to use a bigger epsilon of 10^{-5} in the bootstrap to break the EM iterations.

```
          value
haplotypes          mean conf_intervall_low_95 conf_intervall_up_95
  08-04512  7.070228e-06           0.004506558          0.021233792
  07-56681 -7.062025e-06           0.024039395          0.003745523
  07-56951  9.037720e-09           0.180221848          0.002826499
  08-55163  4.861100e-01           0.155959452          0.016173270
  07-54825  1.407876e-01           0.238793950          0.023622862
  08-59712  2.473457e-01           0.099389040          0.020630499
  08-02659  1.595668e-01           0.161083793          0.052335715
  08-57881  2.228195e-01           0.225566051          0.033650407
  08-01315  3.008985e-01           0.006084773          0.051964298
  08-04134  7.601707e-03           0.003926379          0.089706124
```

Task 5)





Which haplotypes are most dominant?

The most dominant haplotype is the second (07-56681), but also the 3rd, the 5th and the 6th haplotypes are common. The rest less common.

How reliable are results based on the bootstrapping?

The overlap between the boxes are not big. That means, that rank of the haplotypes driven by the bootstrap results should lead to a stable chance.

How common is a change in the order of importance of the haplotypes?

The change in the top 4 haplotypes is uncommon, because they differ allot in their distribution, but it is possible that the 5th haplotype is more likely than the 3rd. In the rest of the haplotypes are changes extremely common, because their distributions do not really differ.