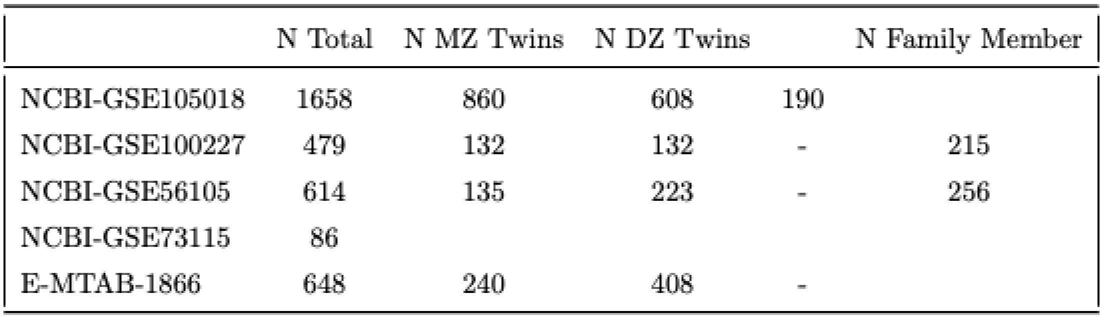
Team Meeting

13 May 2022 / 2:15 PM / ZOOM

Attendees

Ate, Stefan, Xavier, Jiadong, Zexi, Ni

Agenda



	N Total	N MZ Twins	N DZ Twins		N Family Member
NCBI-GSE105018	1658	860	608	190	
NCBI-GSE100227	479	132	132	-	215
NCBI-GSE56105	614	135	223	-	256
NCBI-GSE73115	86				
E-MTAB-1866	648	240	408	-	

Results - Ni

Southern Denmark - GSE73115

- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73115
- For each sample, whole blood was drawn in year 1997 and in year 2007 (86 * 2 != 180?)
- Couldn't find the response value (ie. no column classifying MZ/DZ)
- (Show data2.csv)
- May need to send a email to Qihua Tan qtan@health.sdu.dk

BSGS - GSE56105

- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56105
- 135 MZ, 223 DZ, 256 Family Member and 614 samples in total

  (There is a small difference between our dataset the BSGS dataset the nature paper is using)

- Supposed to be a multiclass classification. However, since our coefficients only have 1 column, only binary classification can be done.
- If family members are removed the auc would be 1
- ‘0’ (representing the family) in true label has been converted to ‘DZ’ when calculating the auc - Show R code
- The results are different from the nature paper - Show the excel file
- We have also considered that ‘are we supposed to use coefficients of 13 and 14 together when testing for datasets containing family members’. So we may have two columns for our predictors and would be able to do this 3-class classification. However, some of the predictors in 14 are not presented in 13, so we rejected this hypothesis.

Results - Xavier

- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100227

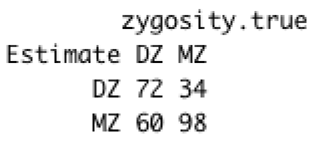
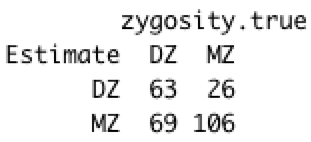
Dataset Summary(GSE100227):

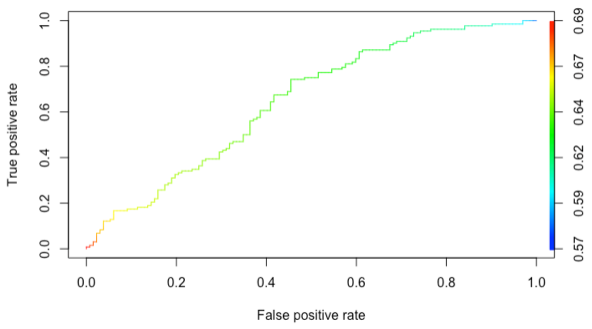
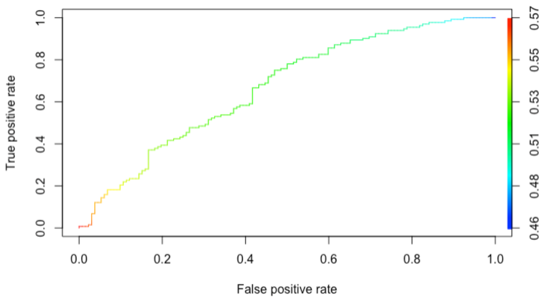
- Total sample: 479
- MZ:132
- DZ:132
- Sister:215

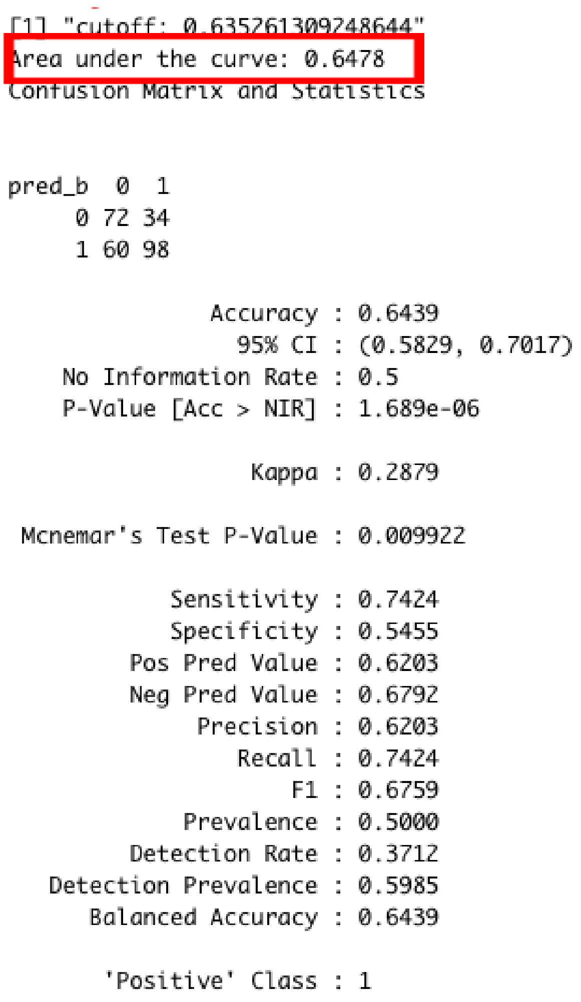
Method 1:

Remove all ‘Sister’ from the true label. (264 samples left)

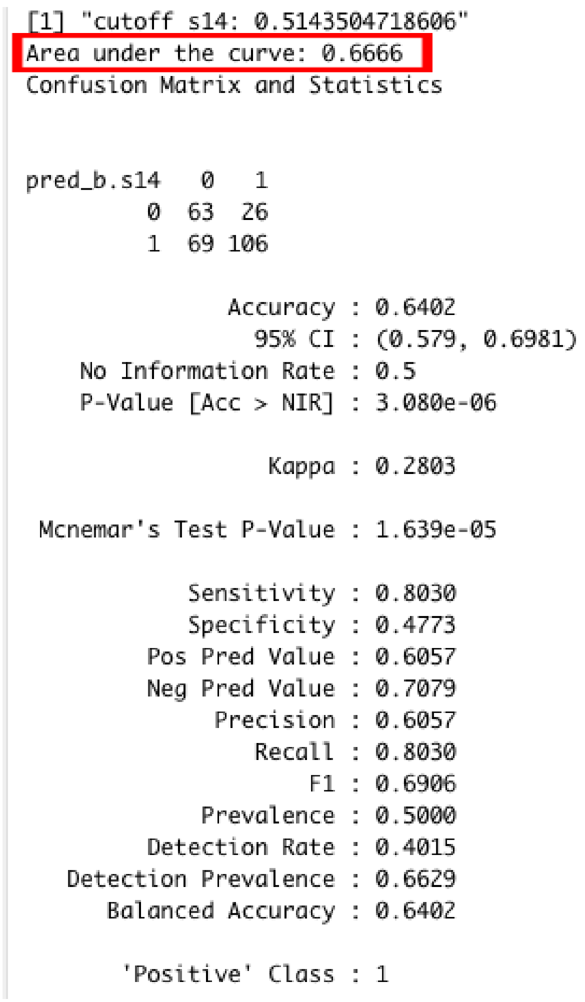
Supplementary13	**Supplementary14**


`          zygosity.true`
`Estimate DZ MZ`
`      DZ 72 34`
`      MZ 60 98`
	
	`          zygosity.true`
	`Estimate  DZ  MZ`
	`      DZ  63  26`
	`      MZ  69 106`



```
[1] "cutoff: 0.635261309248644"
Area under the curve: 0.6478
Confusion Matrix and Statistics


pred_b  0  1
     0 72 34
     1 60 98

               Accuracy : 0.6439
                 95% CI : (0.5829, 0.7017)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 1.689e-06

                  Kappa : 0.2879

 Mcnemar's Test P-Value : 0.009922

            Sensitivity : 0.7424
            Specificity : 0.5455
         Pos Pred Value : 0.6203
         Neg Pred Value : 0.6792
              Precision : 0.6203
                 Recall : 0.7424
                     F1 : 0.6759
             Prevalence : 0.5000
         Detection Rate : 0.3712
   Detection Prevalence : 0.5985
      Balanced Accuracy : 0.6439

       'Positive' Class : 1
```

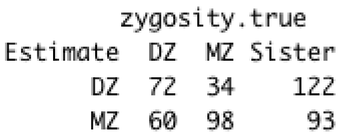


```
[1] "cutoff s14: 0.5143504718606"
Area under the curve: 0.6666
Confusion Matrix and Statistics


pred_b.s14   0   1
         0  63  26
         1  69 106

               Accuracy : 0.6402
                 95% CI : (0.579, 0.6981)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 3.080e-06

                  Kappa : 0.2803

 Mcnemar's Test P-Value : 1.639e-05

            Sensitivity : 0.8030
            Specificity : 0.4773
         Pos Pred Value : 0.6057
         Neg Pred Value : 0.7079
              Precision : 0.6057
                 Recall : 0.8030
                     F1 : 0.6906
             Prevalence : 0.5000
         Detection Rate : 0.4015
   Detection Prevalence : 0.6629
      Balanced Accuracy : 0.6402

       'Positive' Class : 1
```

Method 2:

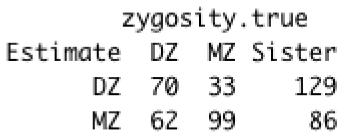
Replace all ‘Sister’ to ‘dz’ from the true label.(MZ:132, DZ:347)

Supplementary13

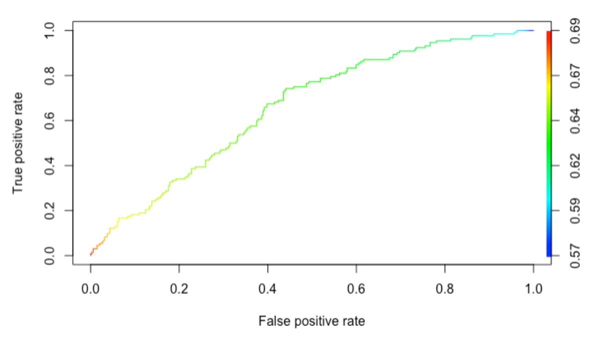


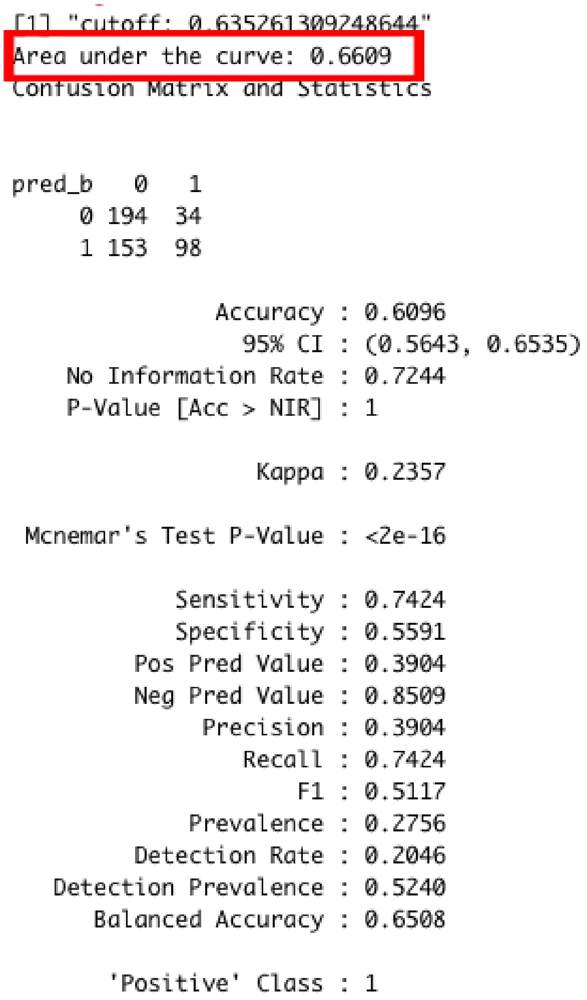
```
        zygosity.true
Estimate  DZ  MZ Sister
      DZ  72  34    122
      MZ  60  98     93
```

Supplementary14

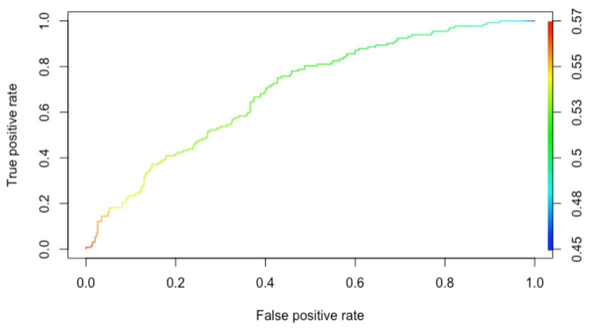


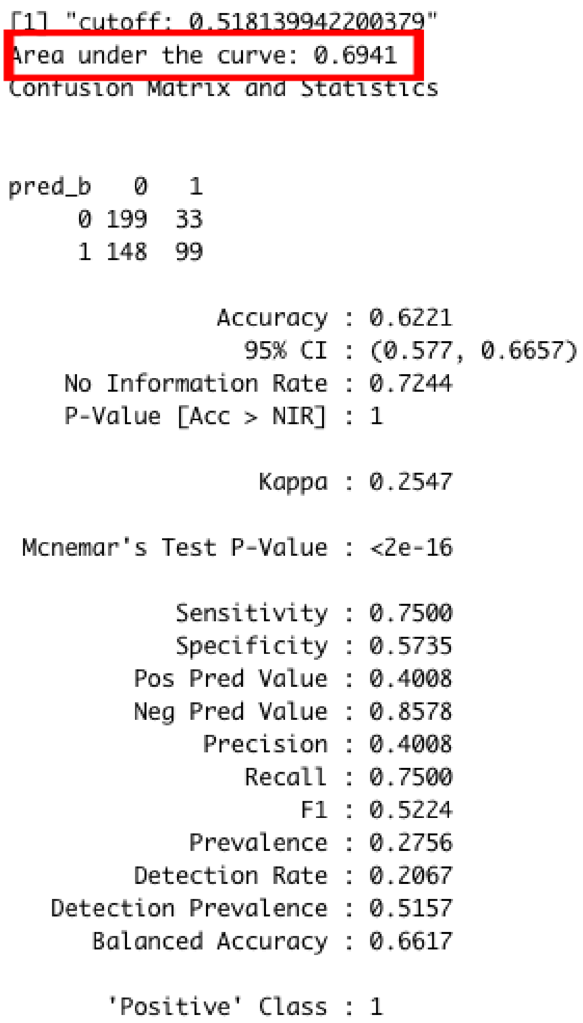
```
        zygosity.true
Estimate  DZ  MZ Sister
      DZ  70  33    129
      MZ  62  99     86
```





```
[1] "cutoff: 0.635261309248644"
Area under the curve: 0.6609
Confusion Matrix and Statistics


pred_b   0   1
     0 194  34
     1 153  98

               Accuracy : 0.6096
                 95% CI : (0.5643, 0.6535)
    No Information Rate : 0.7244
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2357

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7424
            Specificity : 0.5591
         Pos Pred Value : 0.3904
         Neg Pred Value : 0.8509
              Precision : 0.3904
                 Recall : 0.7424
                     F1 : 0.5117
             Prevalence : 0.2756
         Detection Rate : 0.2046
   Detection Prevalence : 0.5240
      Balanced Accuracy : 0.6508

       'Positive' Class : 1
```





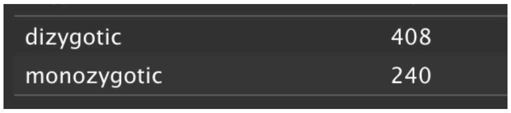
```
[1] "cutoff: 0.518139942200379"
Area under the curve: 0.6941
Confusion Matrix and Statistics


pred_b   0   1
     0 199  33
     1 148  99

               Accuracy : 0.6221
                 95% CI : (0.577, 0.6657)
    No Information Rate : 0.7244
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2547

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7500
            Specificity : 0.5735
         Pos Pred Value : 0.4008
         Neg Pred Value : 0.8578
              Precision : 0.4008
                 Recall : 0.7500
                     F1 : 0.5224
             Prevalence : 0.2756
         Detection Rate : 0.2067
   Detection Prevalence : 0.5157
      Balanced Accuracy : 0.6617

       'Positive' Class : 1
```

Results - Stefan

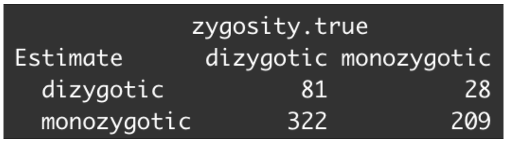
- E-MTAB-1866

https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1866/
DNA methylation was quantified in subcutaneous fat and skin derived from 648 TwinsUK participants using the Infinium HumanMethylation450 BeadChips.

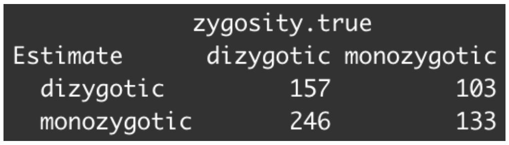
- Large dizygotic samples size
- Low AUC for both classifiers
- Need improvement



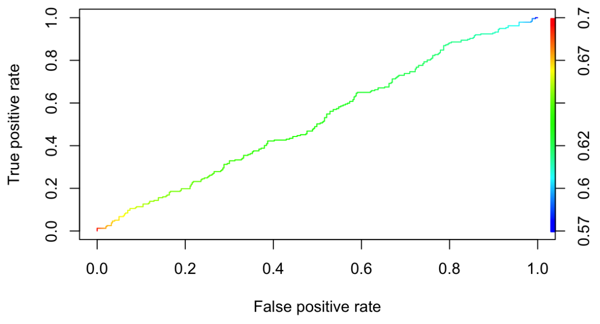
```
dizygotic                        408
monozygotic                      240
```

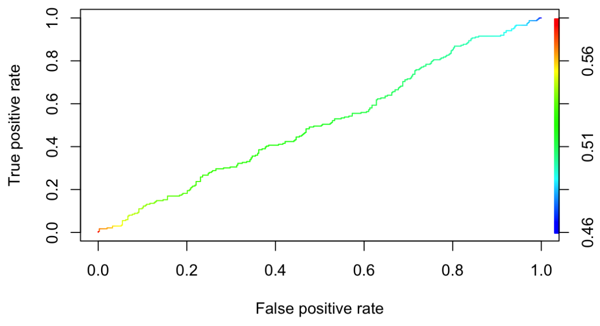


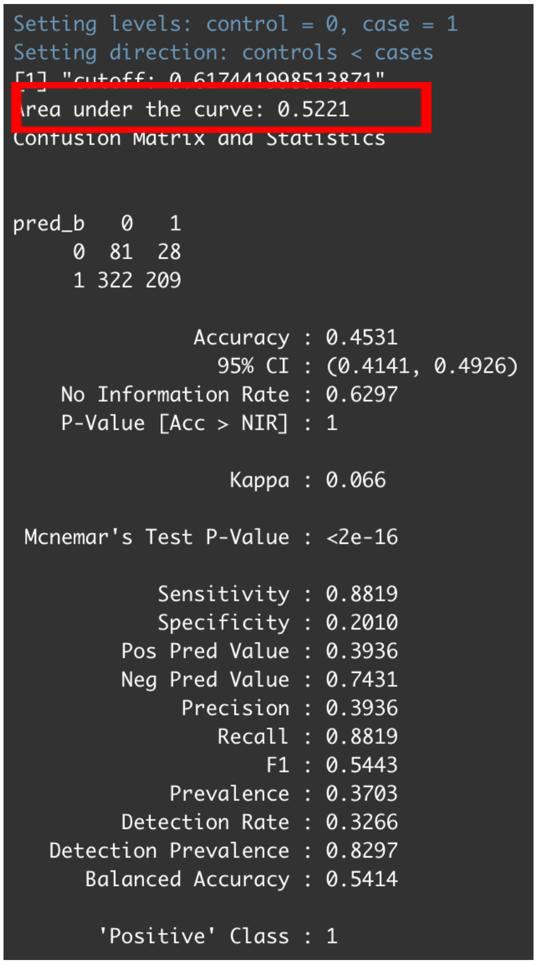
```
               zygosity.true
Estimate       dizygotic monozygotic
  dizygotic           81          28
  monozygotic        322         209
```



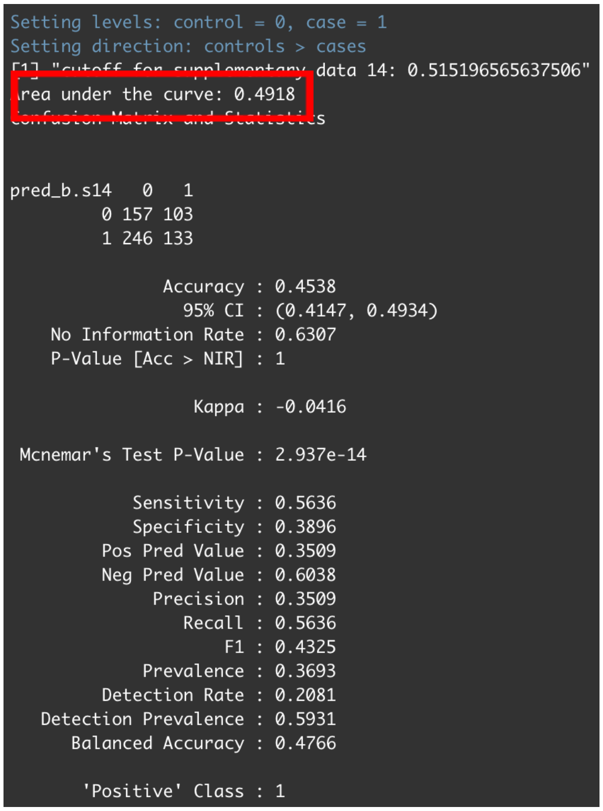
```
               zygosity.true
Estimate       dizygotic monozygotic
  dizygotic          157         103
  monozygotic        246         133
```







```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
[1] "cutoff: 0.617441998513871"
Area under the curve: 0.5221
Confusion Matrix and Statistics


pred_b   0   1
     0  81  28
     1 322 209

               Accuracy : 0.4531
                 95% CI : (0.4141, 0.4926)
    No Information Rate : 0.6297
    P-Value [Acc > NIR] : 1

                  Kappa : 0.066

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8819
            Specificity : 0.2010
         Pos Pred Value : 0.3936
         Neg Pred Value : 0.7431
              Precision : 0.3936
                 Recall : 0.8819
                     F1 : 0.5443
             Prevalence : 0.3703
         Detection Rate : 0.3266
   Detection Prevalence : 0.8297
      Balanced Accuracy : 0.5414

       'Positive' Class : 1
```



```
Setting levels: control = 0, case = 1
Setting direction: controls > cases
[1] "cutoff for supplementary data 14: 0.515196565637506"
Area under the curve: 0.4918
Confusion Matrix and Statistics


pred_b.s14   0   1
         0 157 103
         1 246 133

               Accuracy : 0.4538
                 95% CI : (0.4147, 0.4934)
    No Information Rate : 0.6307
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0416

 Mcnemar's Test P-Value : 2.937e-14

            Sensitivity : 0.5636
            Specificity : 0.3896
         Pos Pred Value : 0.3509
         Neg Pred Value : 0.6038
              Precision : 0.3509
                 Recall : 0.5636
                     F1 : 0.4325
             Prevalence : 0.3693
         Detection Rate : 0.2081
   Detection Prevalence : 0.5931
      Balanced Accuracy : 0.4766

       'Positive' Class : 1
```

Notes

- In our classifier, include 3
- Identify MZ, it seems they want to classify MZ from everything else,
- Consider paris of methylation
- Unbalanced number of sample sizes in training, result in strange results in the last datasets.

Action Items

1. Find algorithms that works good with imbalanced datasets
    - Oversampling (resampling)
    -
2. Email Shuai, 1. Clarify what they mean by pair data, 2. Which one do we use for training our own classifier? 3. Ask if the Brisbane datasets are exactly the same as in the paper,

3. Double check codes again, to makes sure have the correct AUC
4. Choices of datasets for aim 2
    - Large one, with family members
5. Choices between 3 multinomial,
    - Use pca/pls to visualize the distribution of data

Next Meeting Agenda