Team Meeting

22.04. 2022 / 2:15 PM / Zoom

Attendees

Ate, Xavier, Ni, Stefan, Zexi, Jiadong

Agenda

Group 1

Ate, Zexi
Summary table:

	Dataset 1	Dataset2	Dataset3
name	NTR(Netherland twin Register) Whole blood Illumina 450 K array data	BSGS from Australia (blood, 450K)	NTR buccal methylation data set from children (EPIC array)
Data size	1989 - 2155 for training 1100-934 for testing	606	1237
Purpose wrt classification	Split randomly for training and testing (70%), the testing data (30%)	Independent testing	Independent testing

Experimental Set up: 2 input sets × 2 phenotypes = 4 classifiers
DNA methylation predictor of MZ twinning We compared models based on two input sets (genome-wide methylation sites versus metaanalysis DMPs), and trained on two phenotypes (MZ versus DZ twins, and MZ twins versus everyone else (including DZ twins and family

members of twins). Regressions returned predictors based on 232-1867 methylation sites (Supplementary Data 12).
Question

1. What are genome-wide methylation sites? - from memory, I think it's a machine
2. What are metaanalysis DMPs?
3. What is the difference between these two?

Pre-processing constraints:

1. In training data, zygosity was regressed on all methylation sites (N = 381,376) that 1) were present both on the Illumina 450 K and EPIC array 2) survived quality control in the training set(NTR-blood) and in the test data sets (NTR-buccai, Australia-blood).
2. We tested training on the subset of epigenome-wide significant CpGs from the meta-analysis (833 CpGs that were also present on the EPIC array).

Missing-value handling:

1. In the NTR data sets, missing values for probes (probes with missing values in more than 5% of the sample had been removed) were imputed (for penalized regression models only) with the function imputePCA from the package missMDA as implemented in the pipeline for DNA methylation-array analysis developed by the Biobank-based Integrative Omics Study (BIOS) consortium (https://molepi.github.io/DNAmArray_workflow/).

Supplementary Note 7

Result:
In NTR test data from blood (which were left out of the training dataset), the area under the curve (AUC) ranged from 0.69 to 0.77, with up to 84% of MZ twins correctly classified, up to 57% of DZ twins correctly classified, and up to 63% of family members correctly classified as non-MZ.

We tested prediction in two independent datasets (Table 1): BSGS (blood from MZ twins, DZ twins, and family members, 450k array) and NTR children (buccal from MZ and DZ twins, EPIC array). AUCs ranged from 0.67 to 0.80 in BSGS, and from 0.63 to 0.76 in buccal data from NTR children. The predictors performed best when trained on genome-wide significant CpGs from the meta-analysis (rather than genome-wide methylation data). Weights of these scores are provided in Supplementary Data 13 and Supplementary Data 14. In the group of NTR children with buccal methylation data and information on chorionicity available, we compared the performance of the predictor for MZ twins with different chorionicities. The performance was similar across chorionicities. For the predictor that performed best on data from buccal (trained to distinguish MZ versus DZ twins, on genomewide significant CpGs), the percentage of correctly predicted MZ twins were: 76% for monochorionic monoamniotic twins, 72% for monochorionic diamniotic twins, and 75% for dichorionic twins.

Next step: Writing the part of the final report

Group 2

Xavier, Ni, Stefan

Normalized table “GSE105018_NormalisedData.csv”: 450k * 1658 [E-risk dataset]

It has been transposed to 1658 * 450k, so that we’ve got 1658 samples and 450k features.
The transposed table has dimension of 1658*450k

	A	B	C	D	E	F
1		0	1	2	3	4
2	Unnamed: 0	cg00000029	cg00000108	cg00000109	cg00000165	cg00000236
3	2621	0.53971412	0.8836497	0.84234682	0.20386056	0.76007919
4	2622	0.51425345	0.85103299	0.80248834	0.15259106	0.76779622
5	2691	0.4551858	0.90991701	0.74850277	0.15960055	0.67251187
6	2692	0.44264955	0.88844601	0.75662747	0.14796141	0.73094653
7	3391	0.42130992	0.89781832	0.795476	0.18977541	0.73248031
8	3392	0.47463141	0.88258471	0.79887455	0.15440726	0.70809676
9	2791	0.47084655	0.89979905	0.81155519	0.18954872	0.73536882
10	2792	0.465576	0.87785056	0.79908371	0.1752903	0.69025831

Summarised table “GSE105018_series_matrix.txt.gz”: 1658 * 40

Under “title” column of this table we can extract the sample name, so that this table can be merged with the normalised table.

```
36 ```{r}
37 train = gse$GSE105018_series_matrix.txt.gz
38 gse.df <- as(train,"data.frame")
39 gse.df
40 ```
```

	title <fctr>
GSM2814050	whole blood taken from sample 2621
GSM2814051	whole blood taken from sample 2622
GSM2814052	whole blood taken from sample 2691
GSM2814053	whole blood taken from sample 2692
GSM2814054	whole blood taken from sample 3391
GSM2814055	whole blood taken from sample 3392
GSM2814056	whole blood taken from sample 2791
GSM2814057	whole blood taken from sample 2792
GSM2814058	whole blood taken from sample 3161
GSM2814059	whole blood taken from sample 3162

1–10 of 1,658 rows | 1–3 of 40 columns Previous

Next step:

1. Merge the transposed normalised data with the summarised table
2. Test the classifier of the nature paper

Note:

1- The datasets mentioned in the nature paper are discussed.

What is the meaning of meta analysis? Do they deploy different machines (What's the difference between 450k dataset and 80k dataset? Different machine? Same machine but different techniques?)?

Which part of the dataset is used as validation (developing data? The 30% of the NTR dataset?)

2- It is suggested to start writing a literature review by reviewing the relevant papers with focus on the classification. What are the drawbacks and advantages of the papers that are reviewed compared to the nature paper. For methodological papers, we should criticize the method.

3- Use strategies / existing packages to deal with big data in R ,
data.table, RHadoop, parallel algorithms / computing

4- If we can write a code that can be run on a PC instead of a cloud, that should be our motivation.

Steps:

1. Set up meeting for the 29th (including the minutes) - Zexi
2. Chase up Note 1
3. Ask Michael for access to the cloud server - Ate
3. Big datasets -team 2
4. Meeting Agenda (client)