Data Preparation and Modelling Report

Utrecht University

Mike Vink

May 1, 2021

Contents

1 Data selection

The data that we use in this work is based on the data used in the simon manuscript (Tomic et al., 2019). This subset of the fluprint database comprises data from 5 clinical studies, most importantly the longitudal study SLVP015. It only uses the first visits of donors, as the classification is the most complete in this dataset (Figure 1). We will use this dataset to model the high vaccine response with as predictors the in total 3285 features measured in assays done by the clinical studies. The data will be prepared in the same way as in the original work, using the mulset algorithm (Tomic et al., 2019).

In addition to repeating a similar procedure as in Tomic et al., 2019, we will compare the values of features selected by the models trained on the first visit data to those of second visit data. Initially the plan was to train new models on the second visit data, however the classes are extremely unbalanced in the data of repeat visits (Figure 1). For example in the first visit there are 65 high responders and 130 low responders, in the second visit there is only data available for 6 high responders and 44 low responders. Therefore we will only train model on first visit data, and use the knowledge gained to explore second visit data.

2 Clean data

The goal is to obtain one or more tables from the simon data suitable to train models, thus we are looking to change the data from the long format as in the database in a wide format where each

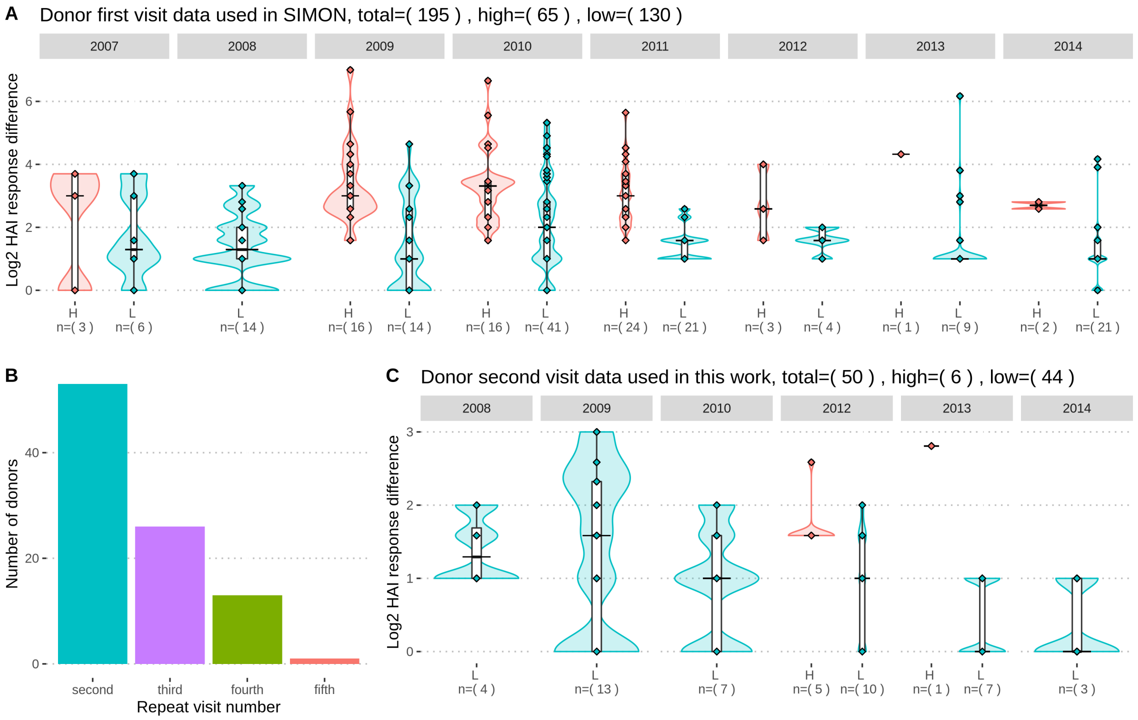

Figure 1: caption

donor_id	study	age	outcome	year	type	hai_response	name	data_name	assay	data	dup
285	18	9.47	0	2009	pre	1	CD4+ T cells	CD4_pos_T_cells	13	33.8	TRUE
285	18	9.47	0	2009	pre	1	CD4+ T cells	CD4_pos_T_cells	13	34.1	TRUE
285	18	9.47	0	2009	pre	1	CD4+ T cells	CD4_pos_T_cells	13	34.3	TRUE
285	18	9.47	0	2009	pre	1	CD4+ T cells	CD4_pos_T_cells	13	33.0	TRUE

Table 1

column is a features measured in an assay. When attempting to do this it was discovered that some assay data contained duplicate readouts (Table 1). Since the values were all similar it was decided to aggregate the values to unique features using the mean value.

The obtained first visit wide data had dimensions of 195x3284 (donors by measured features), with 596736 missing value cells (93% sparsity). The second visit data had dimensions of 50x3251, with a lower sparsity (58%) since the donor population is smaller and they are from the same clinical study.

2.1 Mulset alogrithm

Following the procedure in Tomic et al., 2019 we deal with this data sparsity by applying the mulset algorithm created by Tomic et al., 2019. This is necessary since data is missing in every column and the lack of prior knowledge doesn't allow for imputation of missing values, precluding conventional measures of missing data cleaning.

The mulset algorithm uses the intersection of features sets of donors to calculate pairwise shared feature sets. For every shared feature set it then retrieves all donors that have values for these features (Figure 2).

Listing 1: Applying the mulset algorithm

```
% Step 1: generate re-sampled intersection datasets suitable for analysis
for {each subject in data} do:
        Calculate intersection between subject and all other subjects using mulset algorithm
        Skip sets that have less than 5 features and less than 15 donors in common
end for;
# Save all shared intersections to corresponding datasets
```

Applying the algorithm resulted in 47 different datasets without missing values which contained a subset of donors and features, and the vaccine response classification. Further, the same criteria as in the original work were applied, the number of datasets was filtered down to 36 by excluding datasets that had less than 15 donors or less than 5 features (Listing 1).

To prepare the datasets for modelling they were partitioned into trianing and test sets consisting of 75% and 25% of the data respectively. To ensure that out of sample point estimates were not based on nonsense, datasets with a test set containing less than 10 donors/rows were discarded.

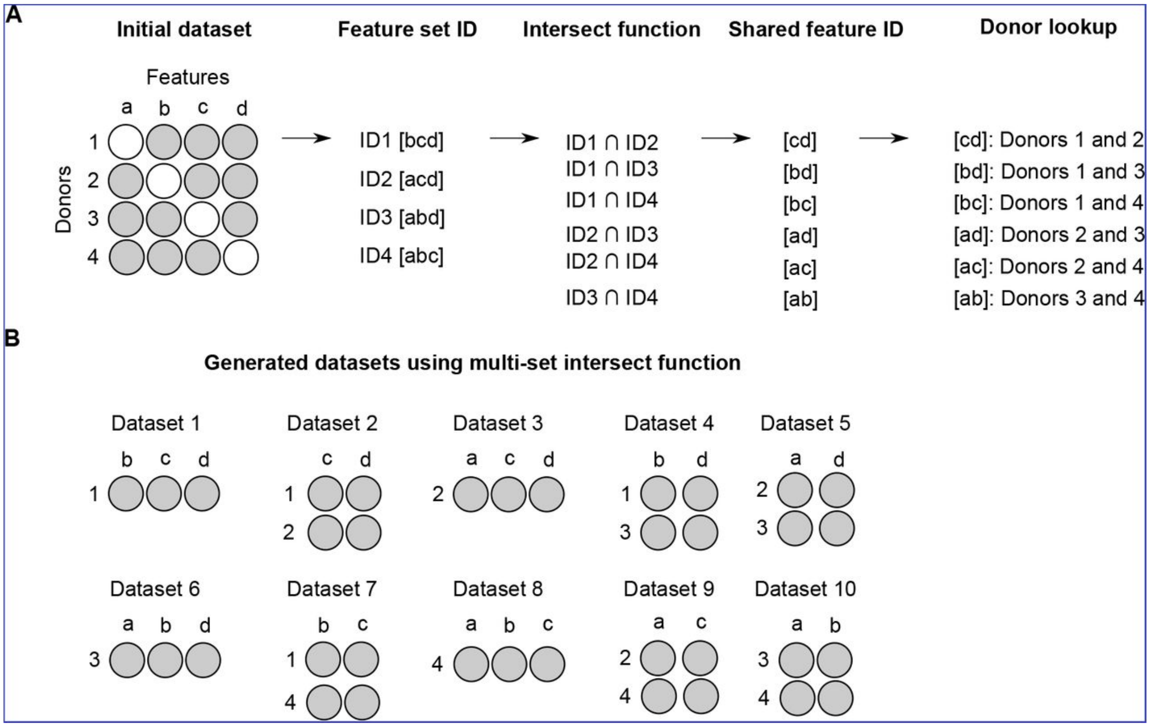

Figure 2: **taken from original work**

The resulting number of cleaned datasets for modelling purposes was 20 (Table 2). A significant number of datasets contained more predictors than samples, however we consider this as an inevitible phenomenon and not an absolute obstacle since the purpose of the models is not to discriminate vaccine responders with the highest accuracy, but to select features from that correlate with a vaccine response from the great number of features.

3 Modelling

The modelling techniques of choice were to be resistent to the ”too many features” problem and suitable for selecting features in an embedded based approach (Hira and Gillies, 2015). Technically, the approach used here is a wrapper approach since we are using the mulset algorithm to generate different subsets of features and training machine learning models on those features. However, in this work we train three models that have an embedded mechanism for obtaining the most important predictors of vaccine response. This is done for every feature set, and manually we chose the best and most interesting trained models and their obtained features. The end goal was to identify important features and investigating the change in those features for the second visit/influenza season of donors.

wieojfwoiejf

dataset	Rows x Cols	total (low / high (low %))	train (low / high)	test (low / high)
1	61 x 78	43 / 18 (0.7)	33 / 14	10 / 4
2	105 x 101	62 / 43 (0.59)	47 / 33	15 / 10
3	140 x 50	94 / 46 (0.67)	71 / 35	23 / 11
4	63 x 269	38 / 25 (0.6)	29 / 19	9 / 6
5	62 x 293	38 / 24 (0.61)	29 / 18	9 / 6
6	68 x 237	42 / 26 (0.62)	32 / 20	10 / 6
7	67 x 44	47 / 20 (0.7)	36 / 15	11 / 5
8	111 x 93	66 / 45 (0.59)	50 / 34	16 / 11
9	73 x 54	58 / 15 (0.79)	44 / 12	14 / 3
10	40 x 105	28 / 12 (0.7)	21 / 9	7 / 3
11	46 x 97	32 / 14 (0.7)	24 / 11	8 / 3
12	137 x 53	78 / 59 (0.57)	59 / 45	19 / 14
13	48 x 42	35 / 13 (0.73)	27 / 10	8 / 3
14	91 x 38	62 / 29 (0.68)	47 / 22	15 / 7
15	42 x 37	36 / 6 (0.86)	27 / 5	9 / 1
16	92 x 26	62 / 30 (0.67)	47 / 23	15 / 7
17	88 x 6	68 / 20 (0.77)	51 / 15	17 / 5
18	82 x 87	56 / 26 (0.68)	42 / 20	14 / 6
19	151 x 51	92 / 59 (0.61)	69 / 45	23 / 14
20	83 x 75	56 / 27 (0.67)	42 / 21	14 / 6

Table 2: caption

dataset	model	SENS	SPEC	MCC	PREC	NPV	FPR	F1	TP	FP	TN	FN	train AUC	test AUC
14	rrlda	0.091	0.915	0.010	0.333	0.683	0.085	0.143	2	4	43	20	0.50	0.62
	nb	0.636	0.702	0.321	0.500	0.805	0.298	0.560	14	14	33	8	0.67	0.59
	rf	0.364	0.851	0.243	0.533	0.741	0.149	0.432	8	7	40	14	0.65	0.61
	reglog	0.227	0.766	-0.007	0.312	0.679	0.234	0.263	5	11	36	17	0.49	0.48
16	rrlda	0.000	1.000	NaN	NaN	0.671	0.000	0.000	0	0	47	23	0.48	0.61
	nb	0.652	0.617	0.253	0.455	0.784	0.383	0.536	15	18	29	8	0.68	0.55
	rf	0.261	0.851	0.135	0.462	0.702	0.149	0.333	6	7	40	17	0.65	0.69
	reglog	0.391	0.723	0.116	0.409	0.708	0.277	0.400	9	13	34	14	0.64	0.47
19	rrlda	0.533	0.391	-0.075	0.364	0.562	0.609	0.432	24	42	27	21	0.47	0.41
	nb	0.489	0.565	0.053	0.423	0.629	0.435	0.454	22	30	39	23	0.54	0.48
	rf	0.244	0.739	-0.018	0.379	0.600	0.261	0.297	11	18	51	34	0.54	0.52
	reglog	0.267	0.754	0.023	0.414	0.612	0.246	0.324	12	17	52	33	0.51	0.32

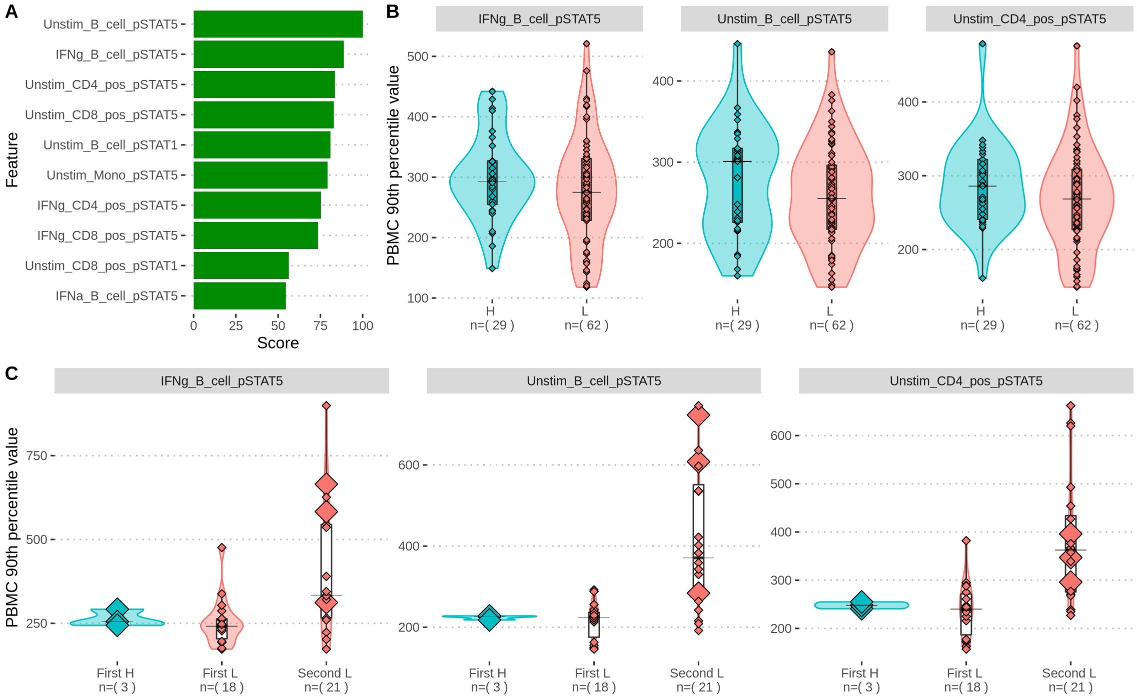

Figure 3: dataset1-nb-feature-exploration

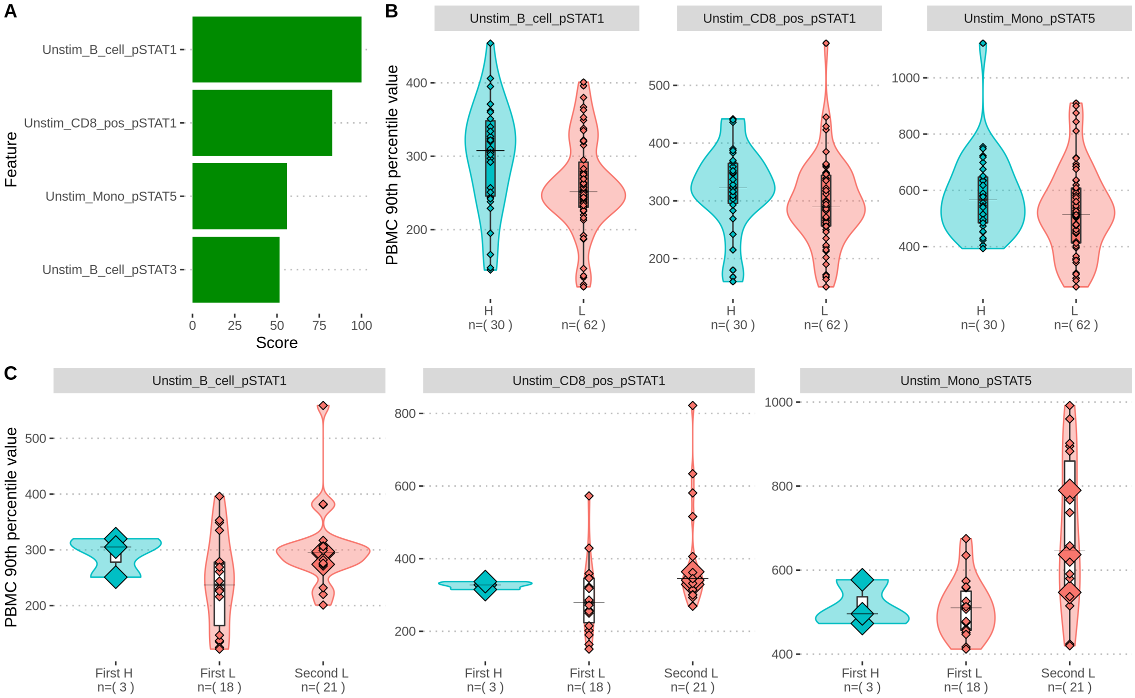

Figure 4: dataset2-nb-feature-exploration

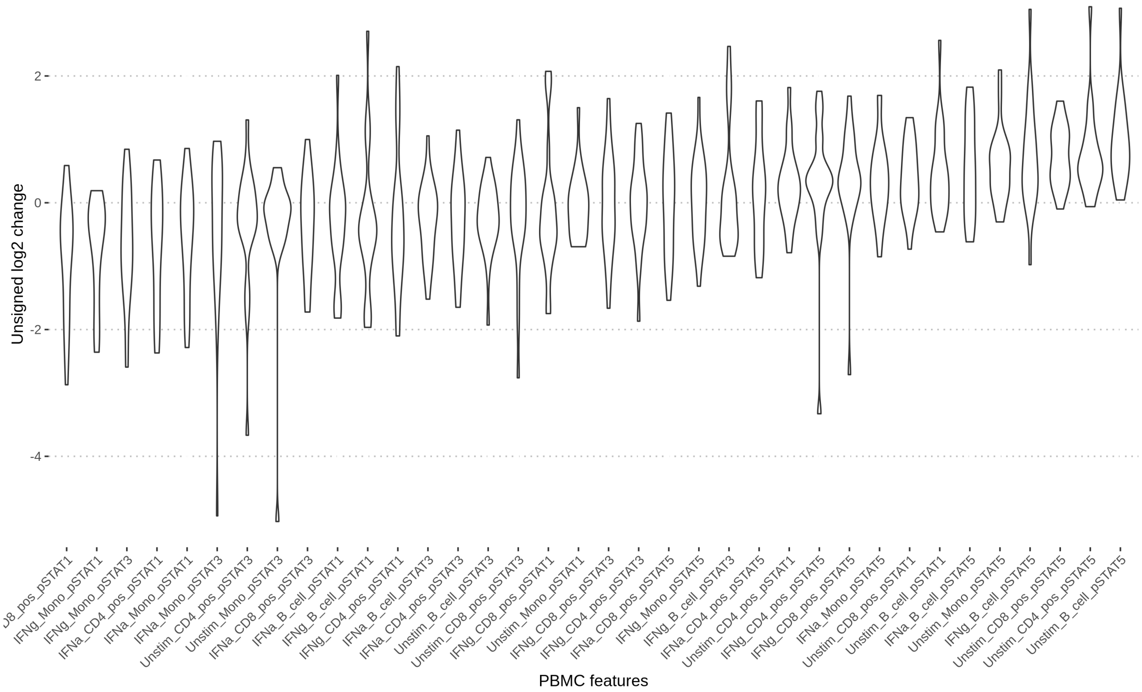

Figure 5: second-visit-change1

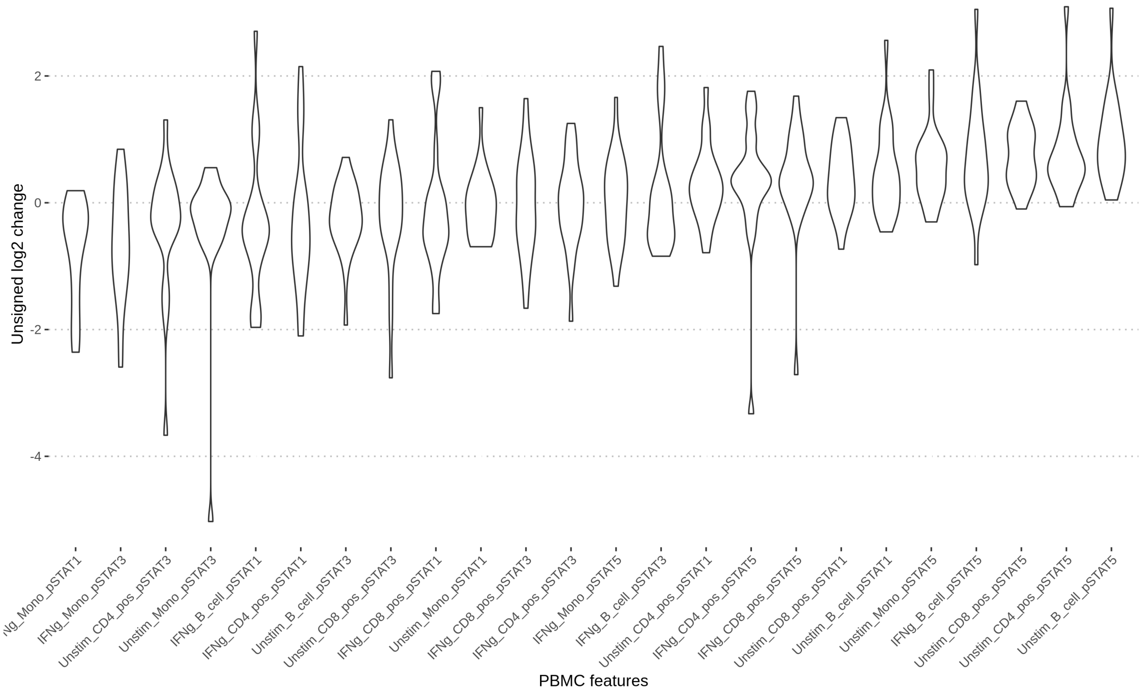

Figure 6: second-visit-change1

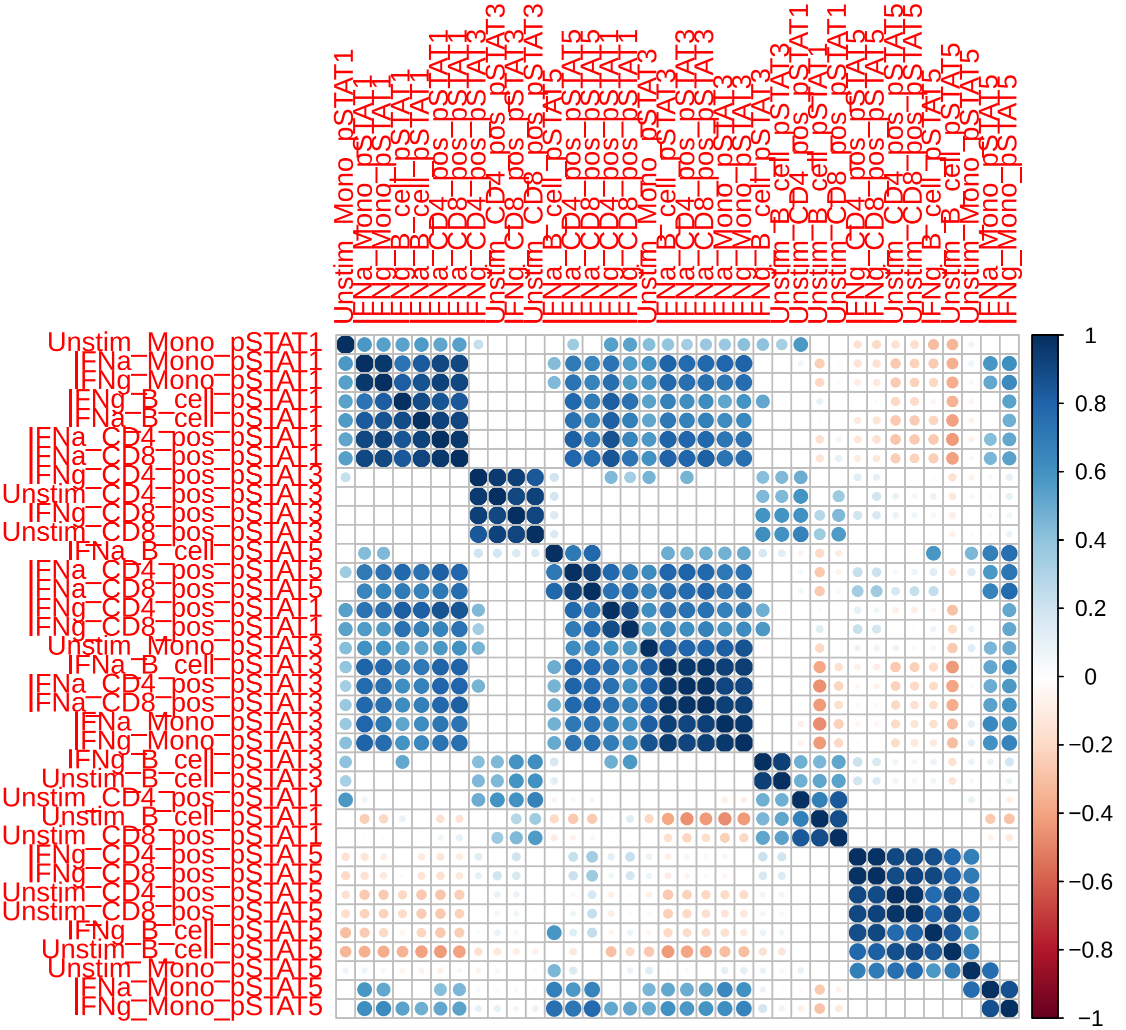

Figure 7: cor-dataset1

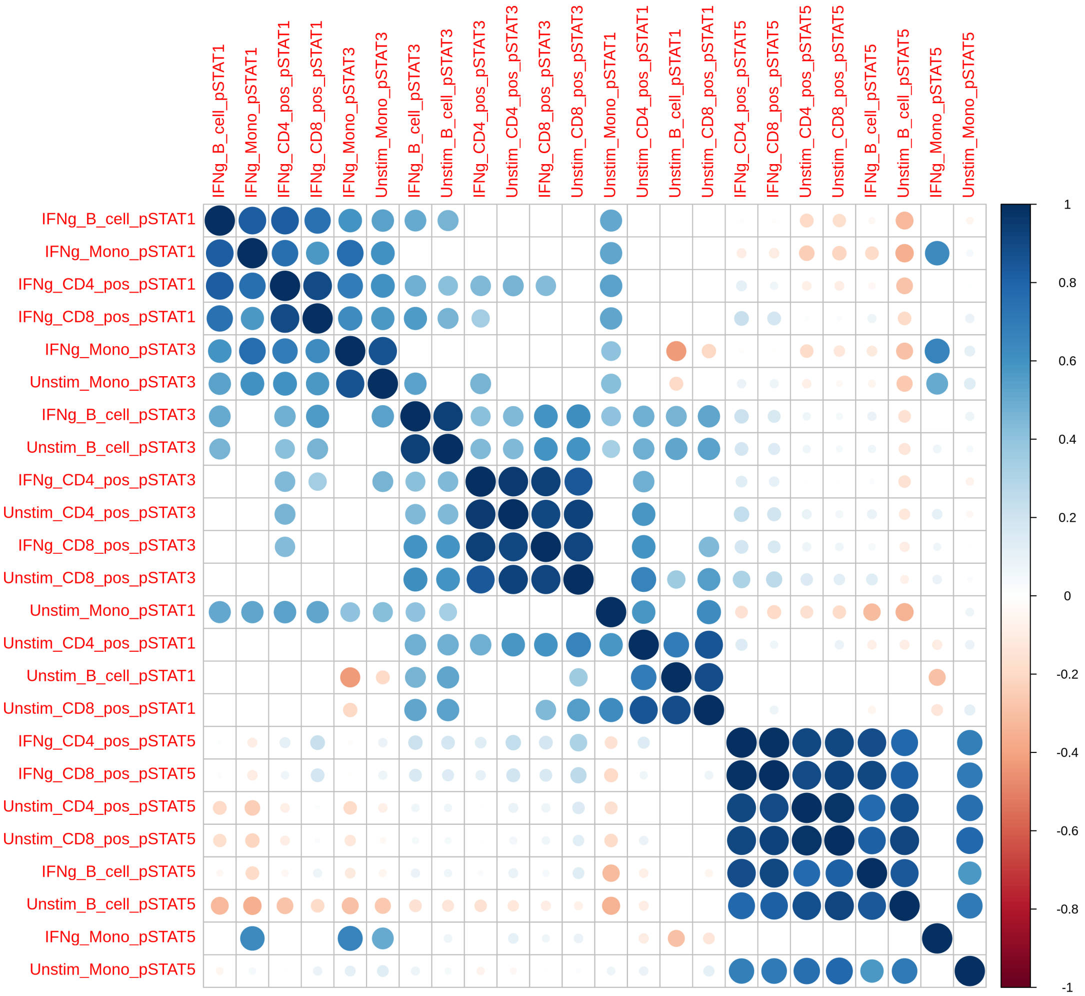

Figure 8: cor-dataset2