CELL-TYPE CHARACTERIZATION

Enumeration of cell subsets from blood methylation profiles

Introduction

- Changes in cell composition
  - Biologically meaningful
  - Common biotechnological methods are error prone
  - Computational methods performs well

Novelty

- Previous studies *(Newman, 2015, Nature)*:
  - Predicting fractions of multiple cell types in gene expression profiles (GEPs) of admixtures.
  - Model an mRNA mixture m by a system of linear regressions, corresponding to a weighted sum of cell type-specific GEPs.

- Here:
  - In methylation profiles
  - Model based on a novel application of nu-support vector regression that outperformed other approaches in benchmarking experiment *(Newman, 2015, Nature)*.

Data

- cell_sorted_data
  - Matrix of cell type-specific methylation data

    for B cells, Cord Blood Mononuclear Cells (CBMC), CD4T cells, CD8T cells, Granulocytes cells, Monocytes cells, NK cells, nucleated Red Blood Cells (nRBC) and Whole Blood, in 7 Samples for 130381 probes.

- mixed_blood_data
  - Matrix of methylation data for the whole blood
    - in 24 samples for 128606 probes.

- cell_counts
  - Matrix of relative fractions of diverse cell types for the above 24 samples.

Method

- Step1- Extract differentially methylated probes for each cell type

(from cell_sorted_data)

- Step2- Extract the whole blood methylation value (from mixed_blood_data) for the selected probes from previous step.
- Step3- run SVM nu-regression over the counterparts data from 2 previous steps to resolve the model.
- Step4- Use the result model on each sample in mixed_blood_data and compare the coefficients for each with their available cell type fractions in cell_counts (validation).

1- Extracting Significant Probes

- Data from cell_sorted_data
- for each cell type methylation values are compared with a the values for all the other cell types (two-sided unequal variance t-test).
- Probes with adjusted-bonferroni-Pvalue < 5.47844942e-8* in all 7 samples are extracted as the signature probes.

*5.47844942e-8 = 0.05/(7 * 130381), 7 samples, 130381 probes

Result Step1

- 1039 probes extracted as significant
  - 512 probes for cell type B, 17 for CBMC, 30 for CD4T, 37 for CD8T, 67 for G, 374 for Mo and 7 for NK.
- Only 6 of these probes are not exclusive for just one cell type

2- Constructing Signature Matrix

- the selected probes as significant mapped to mixed-blood data to build the Signature Matrix, B.
- SignatureMatrix B includes the methylation data for each cell type (from sample TS190 of *cell_sorted_data*) and the methylation data for whole blood (from *mixed_blood_data*, 24 sample)

Result Step2

- Signature matrix, B, dimension: 7 X 1027
  - 7 cell types-methylation values, 1027 probes
- 24 different mix-blood vector, m(m1:m24), each with dimension:1 x1027
  - 1 mixed-blood-methValue for the related sample, 1027 probes

- The model f (for the next step) would be
  m = f X B.

3- SVM nu-regression modeling

```
svr_model <-
      svm(m ~ v1 + v2 +v3 + v4 +v5 +v6 + v7 , data ,
       scale = TRUE, type = "nu-regression", nu = k)
```

m = m1:m24

data = data1:data24

m1	v1	v2	v3	v4	v5	v6	v7
0.814596724	0.50054996	0.858630979	0.836566605	0.887863227	0.861524046	0.868909385	0.878994429
0.778129733	0.942892844	0.774541317	0.94461769	0.957234351	0.540580653	0.085896055	0.896737912
0.87107906	0.951561624	0.783082477	0.941140199	0.959776498	0.888082287	0.160978377	0.953539977
0.899081155	0.103015804	0.870718764	0.91169051	0.912399436	0.963428129	0.936061865	0.935034866
0.839683241	0.42698429	0.854760225	0.897173642	0.898623099	0.917245249	0.902917289	0.897056114
0.876815577	0.316651908	0.872844684	0.94688105	0.950694595	0.931345868	0.948558812	0.909093958
0.553489373	0.770327919	0.569528675	0.715324306	0.76843584	0.491768786	0.135391752	0.670146509

Figure1- view of data1 matrix

Result Step3

- coef <- t(svr_model$coefs) %*% svr_model$SV

	B cells	RBC	CD4+ T	CD8+ T cells	Granulocytes	Monocytes	NK cells
TS222	8.948585	41.90394	20.12339	18.89167	30.36748	15.80051	34.44191
TS223	6.148418	39.03582	19.30757	18.54371	30.13637	15.27804	30.49882
TS224	4.783611	36.63695	17.65853	16.74137	35.61677	16.91282	32.51285
TS225	4.994721	37.56088	17.87967	16.52425	32.76946	16.65471	30.05376
TS226	5.32897	37.61555	17.66444	18.35042	33.20551	17.63865	32.80373
TS227	5.71860567	34.6158023	15.096439	14.1584571	38.0901894	17.9569454	32.6322974
TS228	6.501933	37.75186	19.32156	16.34389	31.62959	15.3397	31.32743
TS229							
TS230	5.668659	37.84035	20.89385	17.0872	31.89174	15.84016	31.71086
TS231	5.018291	38.44982	21.4076	19.5915	29.16452	12.3829	31.22587
TS232	6.368722	39.50675	18.16497	18.43254	32.2401	15.51551	32.17512
TS233	5.597027	34.59524	14.76108	13.80439	36.15251	17.26693	30.0111
TS234	4.341966	35.47518	16.2106	16.25549	34.04067	17.10636	30.28684
TS235	6.35418	36.51175	14.18407	15.2456	34.819	19.23605	32.77584
TS236	6.412261	36.80342	18.97075	18.1116	29.93219	15.58252	35.78581
TS237	6.05675	38.62614	19.31181	18.1674	29.46724	15.42951	31.76806
TS239	6.411933	38.12245	18.28073	16.74256	32.45468	16.03179	31.90393
TS240	6.639489	37.58026	21.2262	17.66985	30.77553	14.32196	33.12017
TS241	6.043703	39.25947	21.77882	20.9588	26.33933	13.11052	30.61423
TS242	5.648718	37.25825	17.59338	17.12071	33.51944	16.22146	30.20444
TS243	6.899902	38.3979	24.83413	21.1656	25.80583	10.31104	31.25041
TS244	7.255153	37.33957	16.41062	14.57156	36.03776	18.35367	32.34884
TS245	7.525724	39.42849	21.28104	18.97878	27.5205	13.44905	32.09462
TS246	4.743352	38.30643	20.24508	19.2791	27.6968	12.22309	27.69817
TS247	7.01175	39.04799	17.71864	18.39964	30.28076	17.31934	31.50815

Figure 2- view of the coefficients values for the svr model

4- Validation

- Result from step 3, last slide, should match the relative fraction cell type of cell_counts data.