New Software for Cell-type-specific EWAS & DE analysis; Evaluation in Real Data

Fumihiko TAKEUCHI

fumihiko@takeuchi.name

National Center for Global Health and Medicine (NCGM), Japan

2019.11.14 @CSHL

R package omicwas available from

https://github.com/fumi-github/omicwas

This poster can be downloaded from

http://103.253.147.127/PUBLICATIONS/191114cshl.pdf

Background: EWAS & DE analysis

- Epigenome-wide association study (EWAS)
  - Measure CpG methylation genome-wide using 450K or EPIC arrays
  - Test association with disease
  - Mostly done in tissue (= mixture of cell types), mostly blood
- Differential gene expression (DE) analysis
  - Measure transcriptome using array or RNA-seq
  - Test association with disease, exposure
  - Human/animal samples are mostly tissue

Background: Estimation of cell type proportion

- In a tissue sample, the cell type proportion can be estimated statistically
  - Established for methylome of blood
    - RefFreeEWAS [Houseman 2012], EpiDISH [Teschendorff 2017], GLINT [Rahmani 2017]
  - Possible for transcriptome of any tissue, using single-cell RNA-seq as reference
    - Bisque [Jew 2019], MuSiC [Wang 2019], CPM [Frishber 2019], DWLS [Tsoucas 2019], ADAPTS [Dnziger 2019], MOMF [Sun 2019]
- *Not discussed in this poster*

Aim & Problem

- Predict markers (i.e., CpG sites, genes) that are differentially methylated/expressed in each cell type, by measuring bulk tissue samples
- Previous cell-type specific EWAS methods
  - CellDMC [Zheng 2018], TCA [Rahmani 2019], HIRE [Luo 2019]
  - Not systematically evaluated in real data
- Multicollinearity of predictors hinders linear regression (shown later)

➔ Use ridge regression

Statistical models (1/2)

- Indexes
  - Cell type $h = 1, \ldots, H$
  - Sample $i = 1, \ldots, I$
  - Marker (CpG site, gene) $j = 1, \ldots, J$
  - Traits (disease, age, sex)
    - have cell-type specific effect $k = 1, \ldots, K$
    - have bulk tissue effect $l = 1, \ldots, L$
- Input data
  - Cell type proportion $W_{h\,i}$
  - Trait value $X_{i\,k}$ and $C_{i\,l}$
  - Marker level $Y_{i\,j}$
- Hidden variable
  - Cell-type-specific marker level $Z_{h\,i\,j}$

Parameters to estimate:

- Effect of traits
  - cell-type specific $\beta_{h\,j\,k}$
  - bulk tissue $\gamma_{j\,l}$
- Basal marker level $\mu_{h\,j}$

Statistical models (2/2)

- Cell-type-specific marker level

$$Z_{h,i,j} \sim N(\mu_{h,j} + \sum_k \beta_{h,j,k} X_{i,k}, \sigma^2_{h,j})$$

- Observable, bulk marker level

$$Y_{i,j} \sim N(\sum_h W_{h,i} \left\{ \mu_{h,j} + \sum_k \beta_{h,j,k} X_{i,k} \right\} + \sum_l \gamma_{j,l} C_{i,l}, \tau^2_j + \sum_h {W_{h,i}}^2 \sigma^2_{h,j})$$

- Full regression model; df=(H+1)*K+L

$$Y_{i,j} = \sum_h \mu_{h,j} W_{h,i} + \sum_{h,k} \beta_{h,j,k} W_{h,i} X_{i,k} + \sum_l \gamma_{j,l} C_{i,l} + \text{error}$$

- "Marginal" model, for cell-type h; df=H+K+L

$$Y_{i,j} = \sum_{h'} \mu_{h',j} W_{h',i} + \sum_k \beta_{h,j,k} W_{h,i} X_{i,k} + \sum_l \gamma_{j,l} C_{i,l} + \text{error}$$

Multicollinearity of predictor variables

- Rheumatoid arthritis cases 336, controls 322 (GSE42861)

	Nue	CD4+	CD8+	NK	mono	Bcells	Eos	
Mean	0.59	0.10	0.08	0.08	0.07	0.07	0.01	
SD	0.11	0.06	0.05	0.04	0.02	0.03	0.02	
r	Nue	CD4+	CD8+	NK	mono	Bcells	Eos	Disease
Nue	1	-0.68	-0.60	-0.46	-0.06	-0.49	-0.48	0.44
CD4+	-0.68	1	0.14	0.05	-0.17	0.38	0.26	-0.33
CD8+	-0.60	0.14	1	0.08	-0.05	0.19	0.13	-0.27
NK	-0.46	0.05	0.08	1	-0.04	0.01	0.11	-0.27
mono	-0.06	-0.17	-0.05	-0.04	1	-0.17	0.05	0.10
Bcells	-0.49	0.38	0.19	0.01	-0.17	1	0.11	-0.22
Eos	-0.48	0.26	0.13	0.11	0.05	0.11	1	-0.10

- Cell type proportions W_h are moderately correlated

- Proportion-disease interaction term $W_h * X$ is strongly correlated

➔ hinder linear regression

r	Nue*X	CD4+*X	CD8+*X	NK*X	mono*X	Bcells*X	Eos*X
Nue*X	1	0.83	0.80	0.85	0.93	0.90	0.27
CD4+*X	0.83	1	0.78	0.78	0.83	0.88	0.42
CD8+*X	0.80	0.78	1	0.77	0.82	0.83	0.35
NK*X	0.85	0.78	0.77	1	0.85	0.83	0.35
mono*X	0.93	0.83	0.82	0.85	1	0.88	0.35
Bcells*X	0.90	0.88	0.83	0.83	0.88	1	0.36
Eos*X	0.27	0.42	0.35	0.35	0.35	0.36	1

Statistical methods

- Several methods proposed to cope with statistical difficulties:
  - large number, especially H*K, of parameters, and
  - multicollinearity of independent variables.

Method	#Parameters	Idea	Reference
Full	(H+1)*K+L		
Marginal	H+K+L		
Ridge	(H+1)*K+L	Ridge regression (R package ridge), to estimate $\beta_{h\,j\,k}$	*New!*
TCA	(H+1)*(K-1)+L, K	Two steps, to limit df	[Rahmani 2019]
cellDMC	(H+1)*K+L	Ranking postprocess, to decrease false positive	[Zheng 2018]

Evaluation of statistical methods

- Cell-type-specific association of a target trait
  - Predict in whole blood sample using each method
  - “True” association is determined in sorted blood cells
    - Nominal $P < 0.05$
    - True set of markers that are positively (or negatively) associated with the trait
- Evaluation of prediction
  - AUC of ROC curve
  - Robustness evaluated by comparing with the best method in each scenario
    - $(AUC - 0.5) / (AUC_{Best} - 0.5)$

Rheumatoid arthritis associated methylome

- Prediction in whole blood samples
  - GSE42861 (Liu et al. 2013, PMID: 23334450)
  - 336 cases, 322 controls
  - Whole blood methylation measured with Illumina 450K
- ”True” data from sorted blood cells
  - GSE131989 (Rhead et al. 2017, PMID: 27723282)
  - GSE87095 (Julia et al. 2017, PMID: 28475762)
  - CD14+ monocytes: 63 cases, 31 controls
  - CD19+ B cells: 108 cases, 95 controls

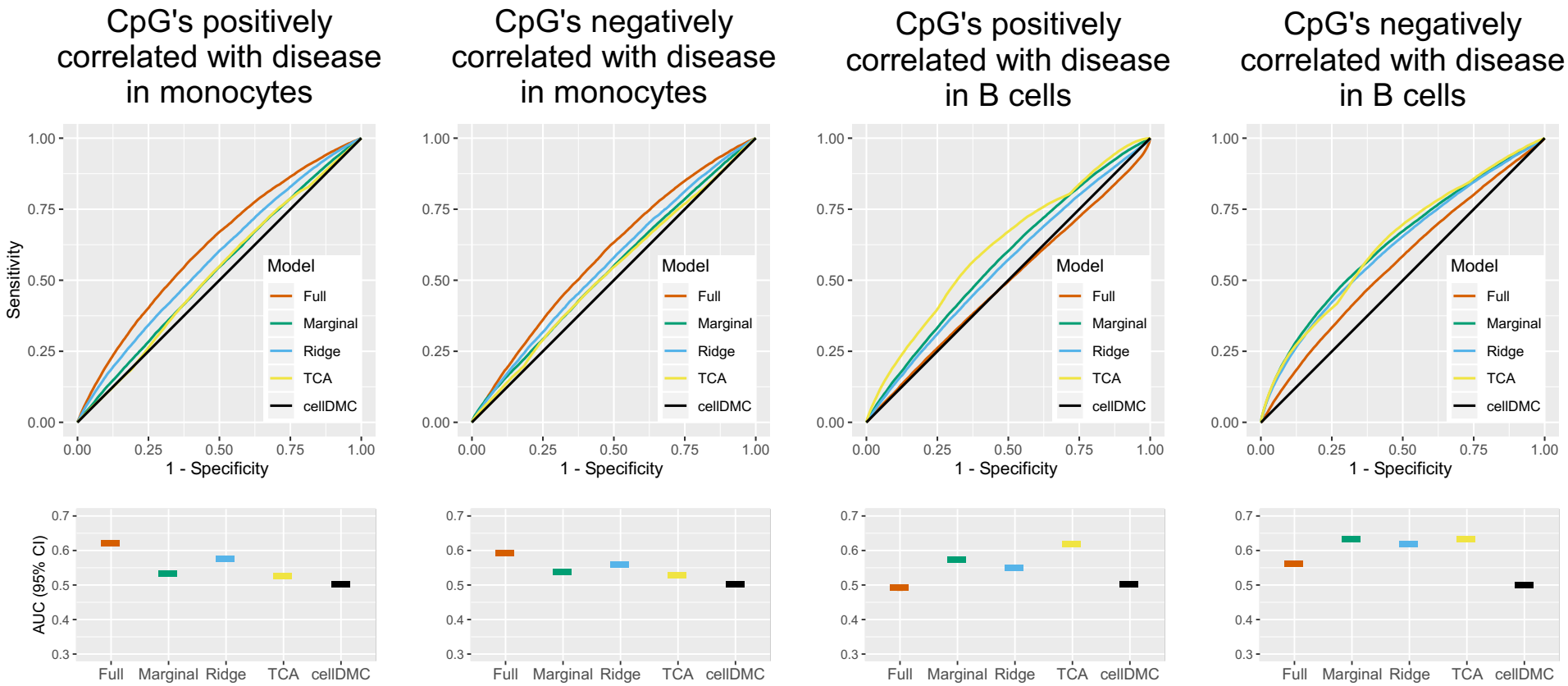
#CpG's		Disease correlation of methylation in CD19+ B cells			
		Positive	No	Negative	
Disease correlation of methylation in CD14+ monocytes	Positive	2293	17743	509	20545
	No	25468	363077	23853	412398
	Negative	455	12202	2085	14742
		28216	393022	26447	

Age-associated transcriptome

- Prediction in whole blood samples
  - GTEx v7 (PMID: 29022597)
  - 389 samples
  - Whole blood transcriptome measured with RNA-seq
- ”True” data from sorted blood cells
  - GSE56047 (Reynolds et al. 2014, PMID: 25404168)
  - CD14+ monocytes: 1202 samples
  - CD4+ T cells: 214 samples

#Genes		Age correlation of gene expression in CD4+ T cells			
		Positive	No	Negative	
Age correlation of gene expression in CD14+ monocytes	Positive	178	1762	141	2081
	No	458	8029	535	9022
	Negative	57	1761	166	1984
		693	11552	842	

Results: Rheumatoid arthritis associated methylome



- AUC for monocytes: Full > Ridge > Marginal > TCA
- AUC for B cells: TCA > Marginal > Ridge > Full
- Marginal and Ridge perform robustly

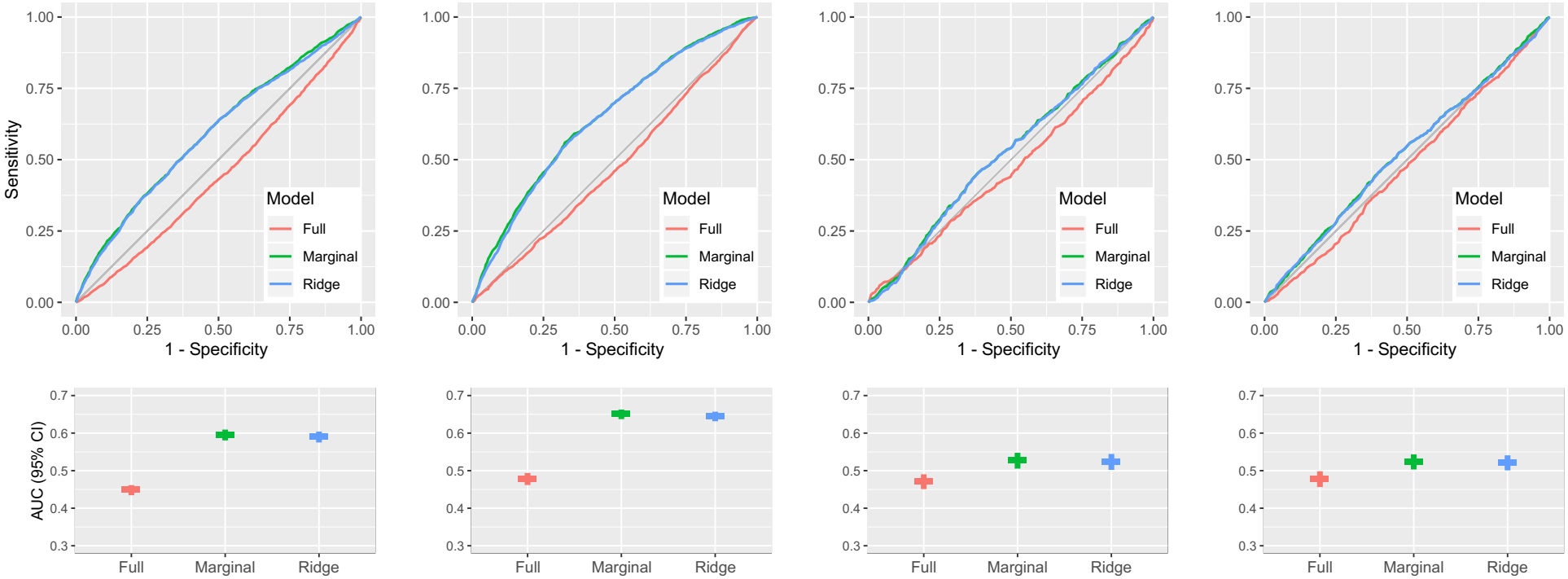
Results: Age-associated transcriptome

Genes positively correlated with age in monocytes

Genes negatively correlated with age in monocytes

Genes positively correlated with age in CD4+ T cells

Genes negatively correlated with age in CD4+ T cells



- AUC for either cell types: Marginal > Ridge > Full

Summary

- Ridge regression was the most robust, performing 41% to 94% relative to the best method.
- Marginal model performed 27% to 100% relative to the best method.
- The naive full model needs caution. It could be severely biased, with AUC < 0.5

- The ridge regression in R package omicwas enhances cell-type-specific association study of methylome and transcriptome data.