Smooth modeling of covariate effects in bisulfite sequencing-derived measures of DNA methylation

Kaiqiong Zhao

McGill University

Supervisors: Celia Greenwood & Karim Oualkacha

BIRS workshop

New Statistical Methods for Family-Based Sequencing Studies

Banff Centre – August 5 - 10, 2018

Epigenetics and DNA Methylation

- change gene expression without changing DNA sequence
- can be altered by age, diet, stress and environmental exposures

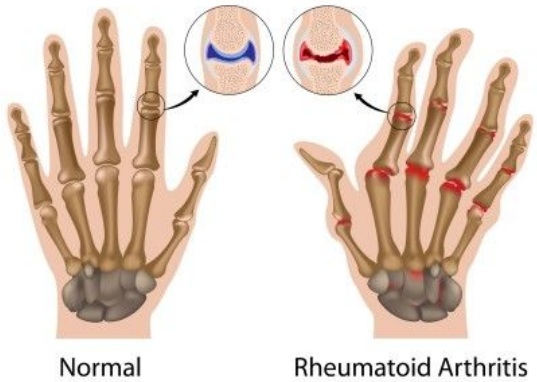
- interest: identify **genomic regions** where DNA methylation patterns demonstrate alterations associated with disease phenotypes (DMRs).

Motivating Dataset

Methylation profiles of Rheumatoid Arthritis (RA) patients and controls (provided courtesy of Dr. Marie Hudson, McGill)

- Samples: cell-separated blood samples of 22 RA patients and 21 healthy individuals from either T cells or monocytes.

	MONO	TCELL
RA	10	12
Control	8	13



- Methylation: targeted custom captured bisulfite sequencing †
  - Prior selection of predefined genomic regions of interest
  - $\sim 400,000$ regions in the genome
- This presentation focuses on one targeted region
  - **Chr4: 102,711,629 – 102,712,832** (near gene *BANK1*)
  - consists of 123 CpG sites

† Allum et al (2015) *Nature*.

Quantification – Bisulfite sequencing

Position	Unmeth counts	Meth counts	Total counts	Sample ID	Sample-level covariates
102711629	2	2	4	1	...
102711630	15	0	15	1	
102711649	15	0	15	1	
102711650	8	0	8	1	
102711850	15	0	15	2	
102711851	4	9	13	2	

Sample-level covariates: disease status, cell type composition, age, smoking status...

Challenges

- **Read-depth variability**
  - The total number of reads varies at different CpG sites. Modeling the proportion treats noisy measurements the same way as accurate ones $\left(\frac{5}{10} = \frac{50}{100}\right)$
  - missing values occur frequently

- **Variability in cell-type mixture proportions**

  $\rightarrow$ **adjusting for multiple covariates**
  - Methylation levels vary substantially across different cell types, which can confound the association of interest.

- **Experimental errors**
  - Sequencing errors: more mis-alignment of unmethylated reads after bisulfite sequencing
  - Bisulfite conversion errors: incomplete C-T conversion; or over-treatment with bisulfite leading to conversion of methylated C to T

Existing methods & Motivations

Challenges	BSmooth (Hansen et al., 2012)	BiSeq (Hebestreit et al., 2013)	SMSC (Lakhal-Chaieb et.al., 2017)
Variable read-depth	✓		✓
Experimental error			✓
Mixture of cell types		✓	
Multiple covariates		✓	

- Most of the existing methods are of two-stage nature

  (1) smooth the methylated proportions **for each sample**, and

  (2) fit model (t-test or beta regression) to the **smoothed methylation data**.

- Motivation: extend the work in Lakhal-Chaieb et.al., 2017 to enable an integrated analysis of **multiple samples** that allows for
  - **cell type mixtures**, and
  - multiple **covariates** in the model.

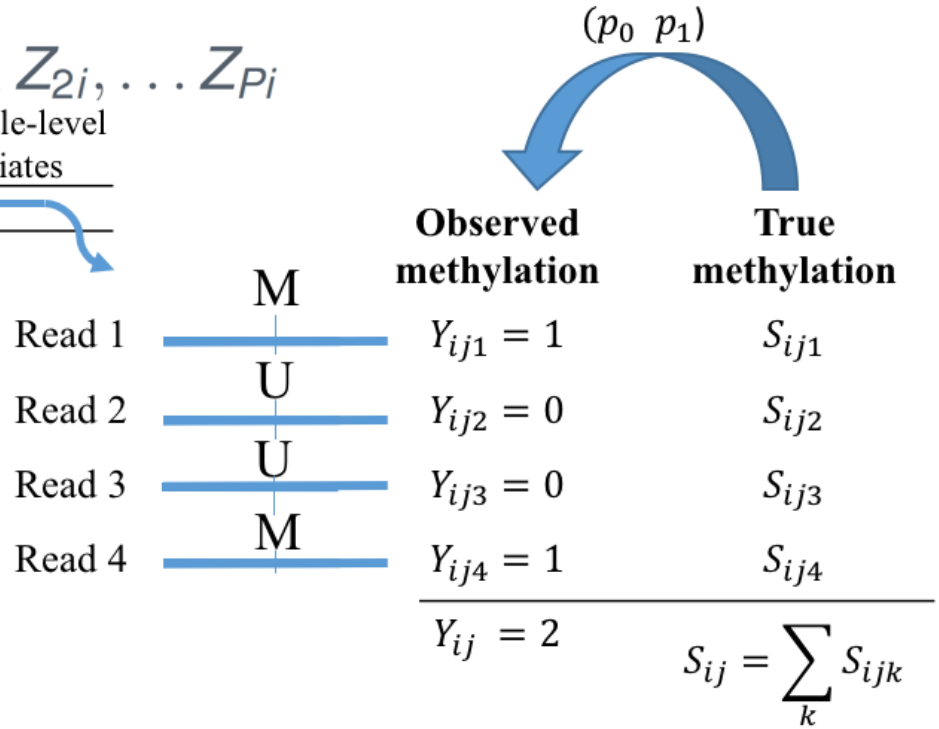
Notation & Model

Let (i, j, k) index sample, CpG sites and reads respectively.

$$i = 1, 2, \ldots I; \quad j = 1, 2, \ldots n_i; \quad k = 1, 2, \ldots X_{ij}, \quad N = \sum_{i=1}^{I} n_i.$$

Example: CpG site j for Individual i

(i, j)	t_{ij} Position	Y_{ij} Meth	X_{ij} Total	Sample ID	$Z_{1i}, Z_{2i}, \ldots Z_{Pi}$ Sample-level covariates
(1,1)	114354051	2	4	1	…
(1,2)	114354052	0	15	1	
…	114354053	0	15	1	
(1, n1)	114354054	0	14	1	
(2, 1)	114354056	0	15	2	
(2, 2)	114354057	9	13	2	



Model

- Assume **known error rates** p_0 and p_1,

$$\begin{aligned} p_0 &= \mathbb{P}(Y_{ijk} = 1 \mid S_{ijk} = 0) \\ p_1 &= \mathbb{P}(Y_{ijk} = 1 \mid S_{ijk} = 1). \end{aligned}$$

(Lakhal-Chaieb et.al., 2017)

- We specify model

$$\begin{aligned} S_{ij} \mid \boldsymbol{Z_i}, X_{ij} &\sim \text{Binomial}(X_{ij}, \pi_{ij}) \\ \theta_{ij} = \log\left\{\frac{\pi_{ij}}{1-\pi_{ij}}\right\} &= \beta_0(t_{ij}) + \beta_1(t_{ij})Z_{1i} + \beta_2(t_{ij})Z_{2i} + \ldots + \beta_P(t_{ij})Z_{Pi}. \end{aligned}$$

- Use splines to parametrize smooth covariate effects:

$$\beta_p(t_{ij}) = \sum_{l=1}^{L} \alpha_{pl} B_l(t_{ij}) \quad \text{for } p = 0, 1, \ldots P.$$

- Smoothing parameters $\{\lambda_0, \lambda_1, \ldots \lambda_P\}$ for controlling the smoothness of $\beta_p(\boldsymbol{t})$

$$\mathcal{L}^{\text{Penalization}} = \sum_{p=0}^{P} \lambda_p \int \left(\beta_p''(t)\right)^2 dt = \sum_{p=0}^{P} \lambda_p \boldsymbol{\alpha_p^T A \alpha_p}$$

Methods

- Penalized EM algorithm

Initialization: $\alpha^\star$ $\lambda^\star$;
repeat
1. E-step: calculate the conditional expected outcomes $\mathbb{E}(S_{ij} \mid Y_{ijk})$
2. M-step: $\left(\widehat{\alpha}, \widehat{\lambda}\right) = \arg\max Q(\alpha, \lambda \mid \alpha^\star)$. ($Q$ is the binomial likelihood replacing S_{ij} with its expectations)
   repeat
   2.1 P-IRLS iteration.
   2.2 Smoothing parameters estimated by REML[‡].
   until *estimates converge*;
   $\widehat{\alpha} \to \alpha^\star$ and $\widehat{\lambda} \to \lambda^\star$
until *estimates converge*;

$\Longrightarrow$ Estimates of the smooth functions of covariates effects $\widehat{\beta_1(\boldsymbol{t})}, \widehat{\beta_2(\boldsymbol{t})} \ldots \widehat{\beta_P(\boldsymbol{t})}$.

- Inference of smooth covariate effects taking account of the uncertainty in both E step and M step

$$\mathcal{H}(\boldsymbol{\alpha}) = \left\{ \frac{\partial^2 Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^\star)}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\alpha}^T} + \frac{\partial^2 Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^\star)}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\alpha}^{\star T}} \right\} \Bigg|_{\boldsymbol{\alpha}^\star = \boldsymbol{\alpha}}^{\ddagger}$$

$\Longrightarrow$ Variance of the smooth estimates $\widehat{\mathbb{V}\mathrm{ar}(\beta_p(\boldsymbol{t}))}$

$\Longrightarrow$ A region-wise test statistic (p-value) for $H_0 : \beta_p(\boldsymbol{t}) = \boldsymbol{0}$

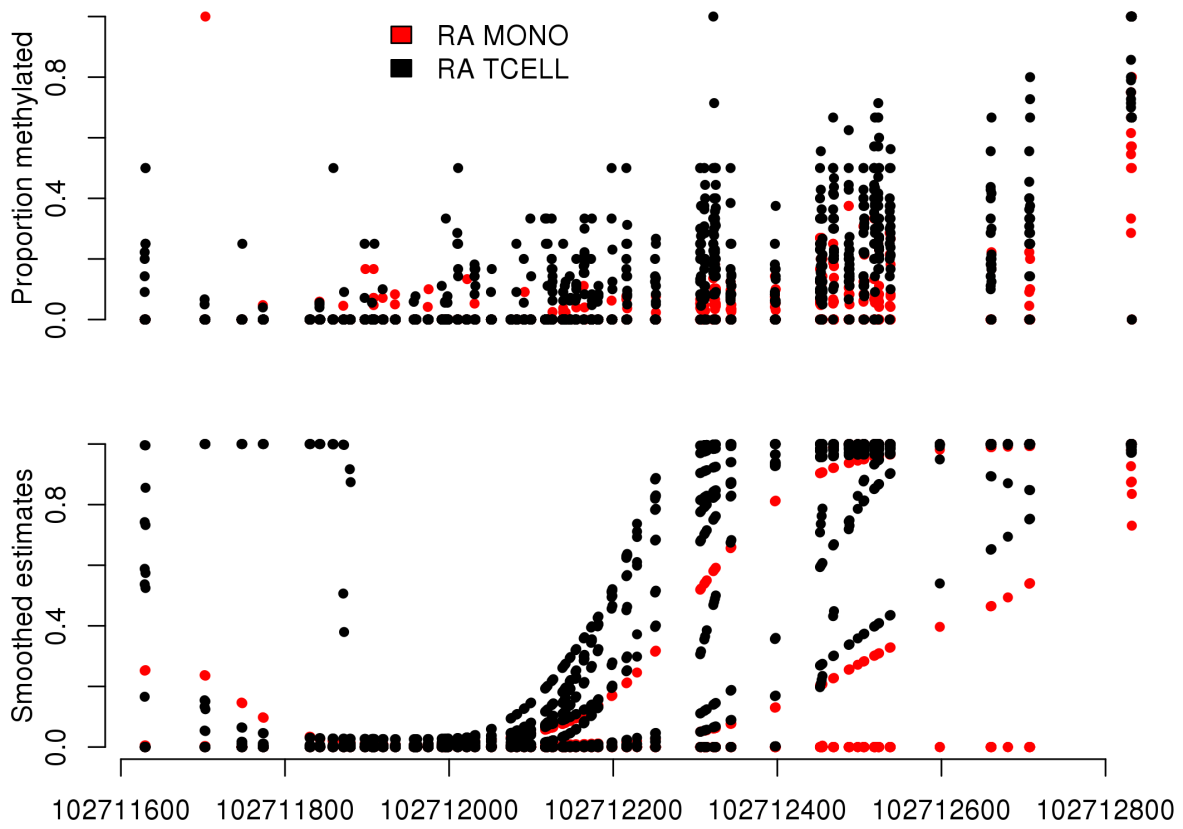
$\widehat{\boldsymbol{\alpha}_p} \left\{ \widehat{\mathbb{V}\mathrm{ar}}(\widehat{\boldsymbol{\alpha}_p}) \right\}^{-1} \widehat{\boldsymbol{\alpha}_p}^T \sim \chi^2_{edf}$ where

$edf = \mathrm{trace}(2\boldsymbol{H} - \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})$

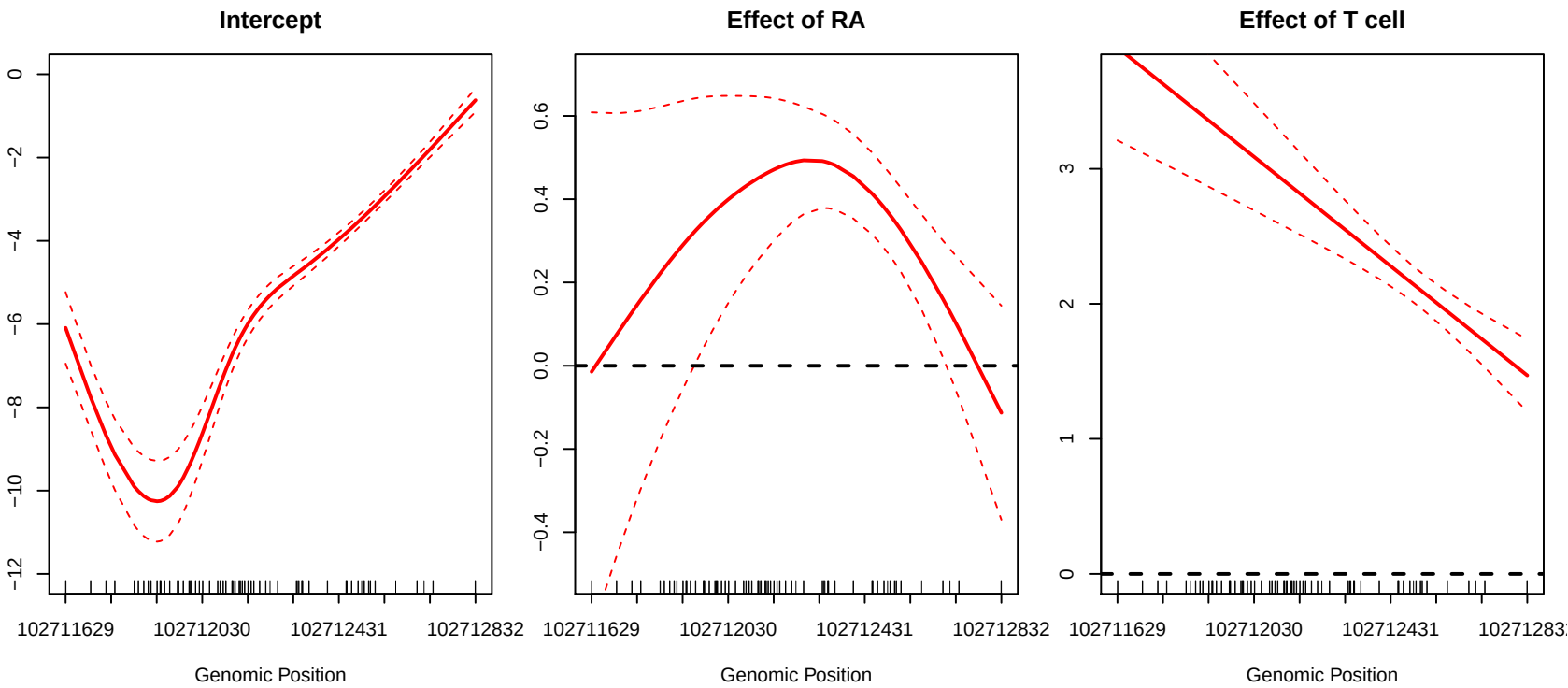
‡ Oakes, D. (1999) Direct calculation of the information matrix via the EM. JRSSB

Data example

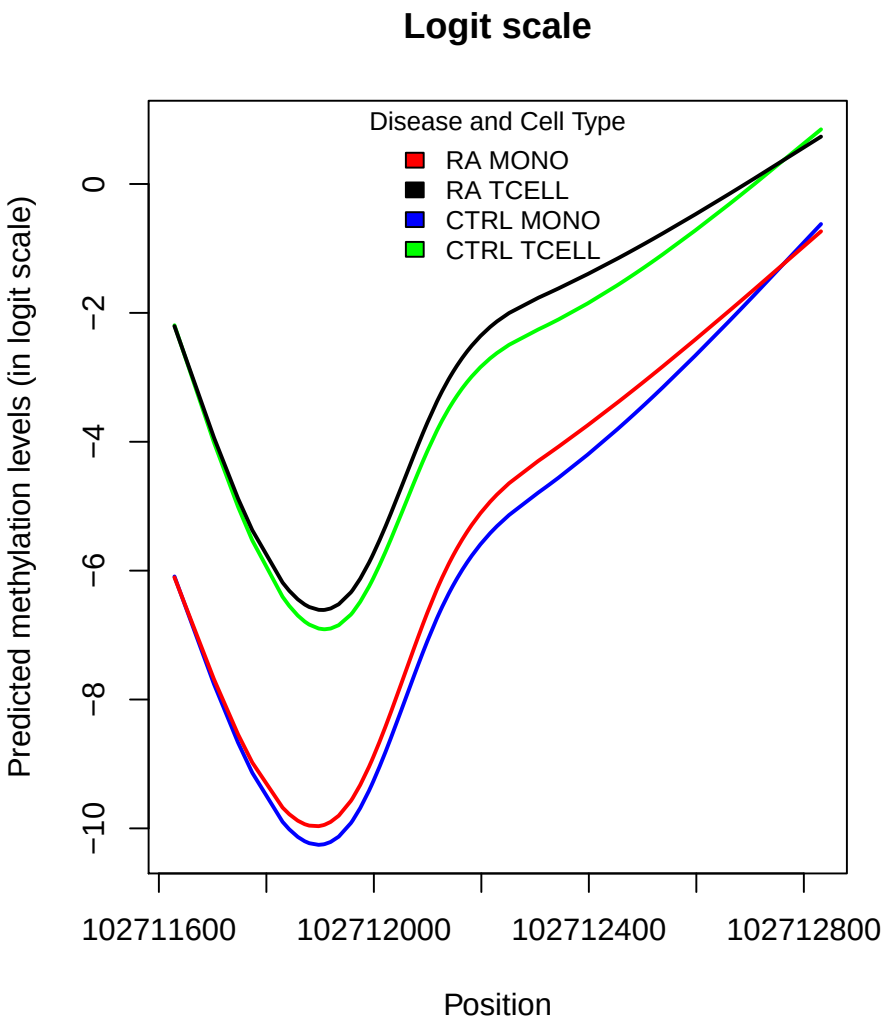
Raw data & per-sample smoothed estimates

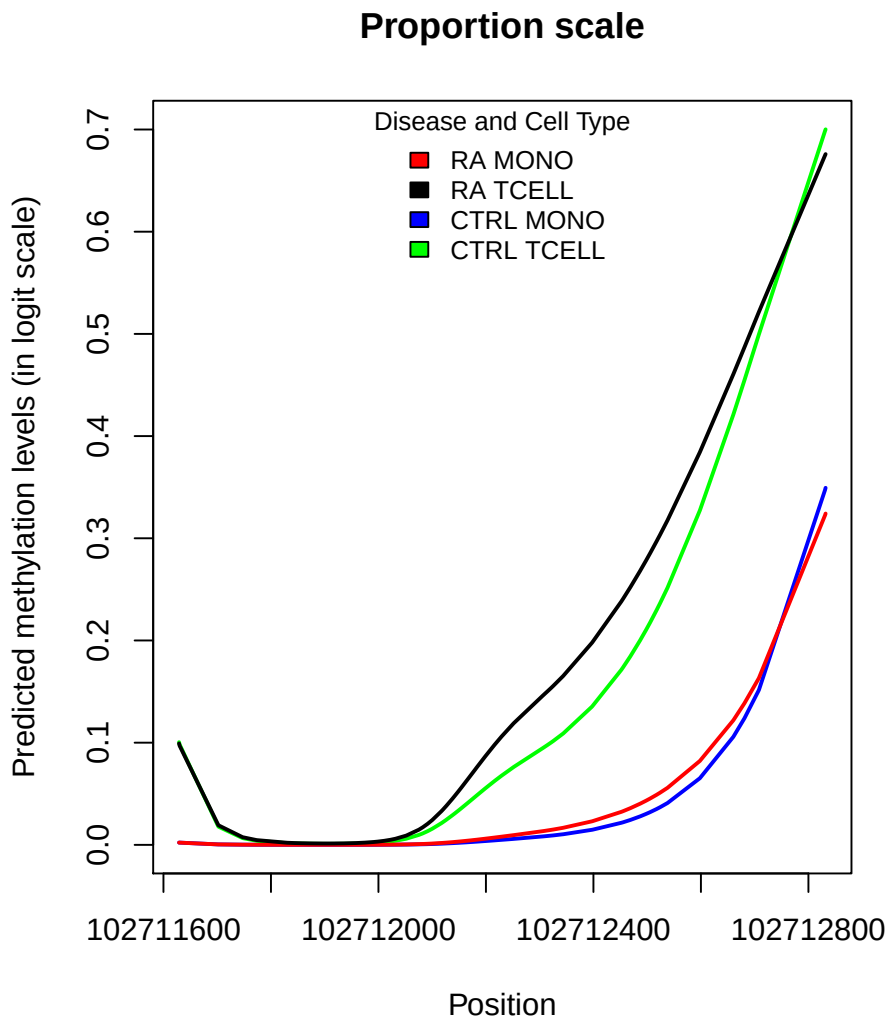


Inference of the smooth covariate effects



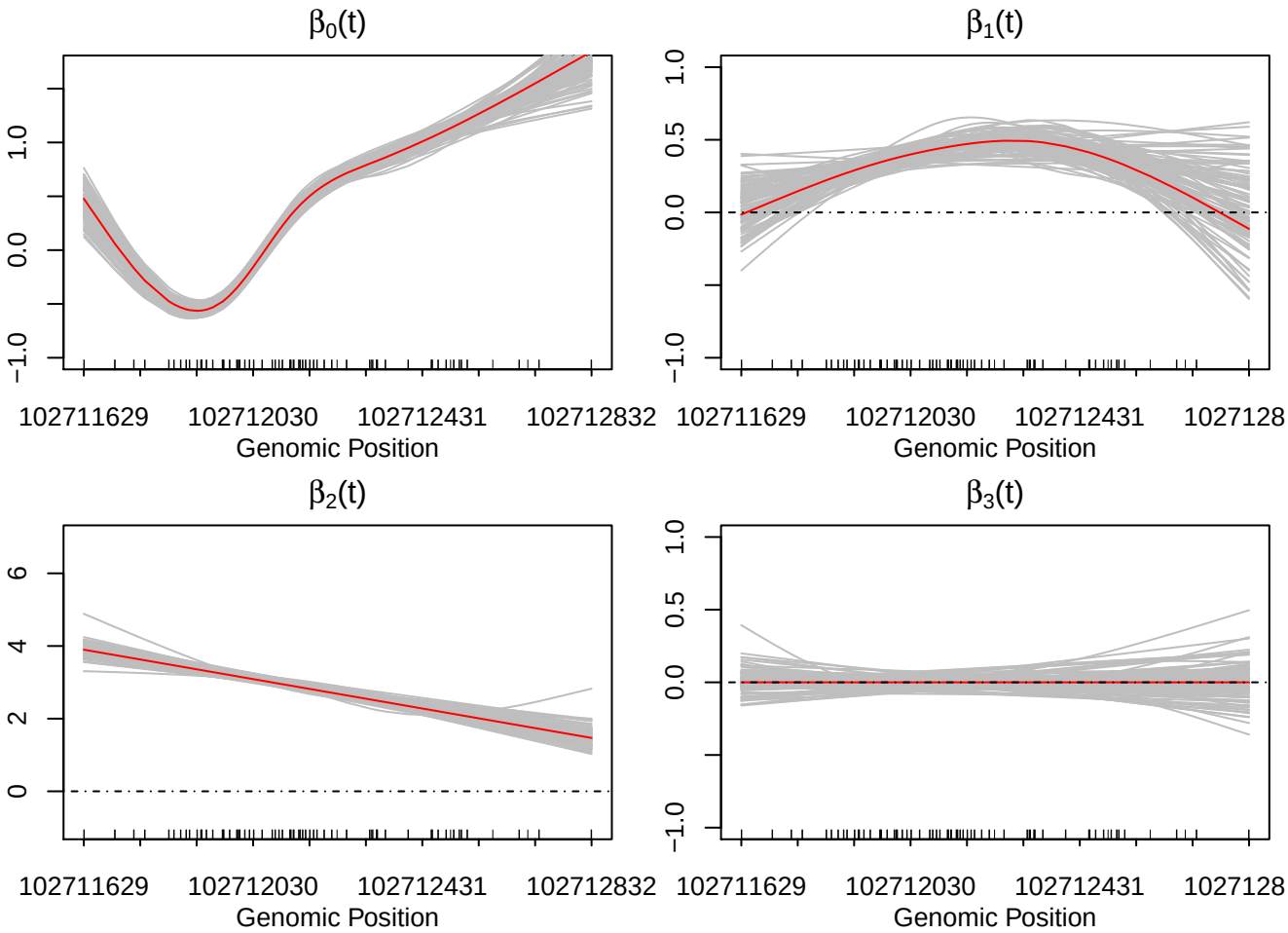
Predicted methylation levels





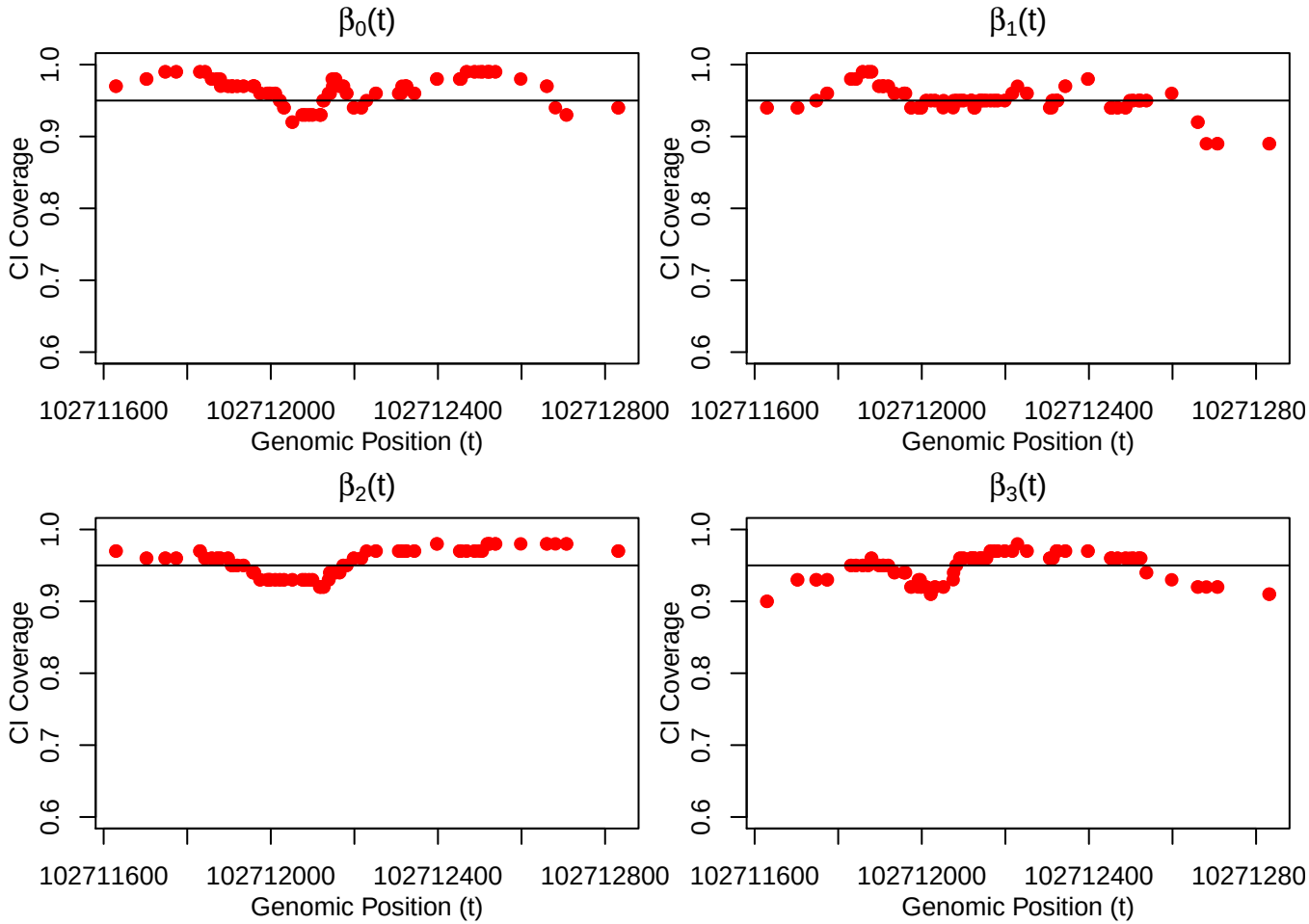
Simulation

- $\beta_3(\boldsymbol{t}) = 0$ and error parameters $p_0 = 0.003$ and $1 - p_1 = 0.1$[‡]

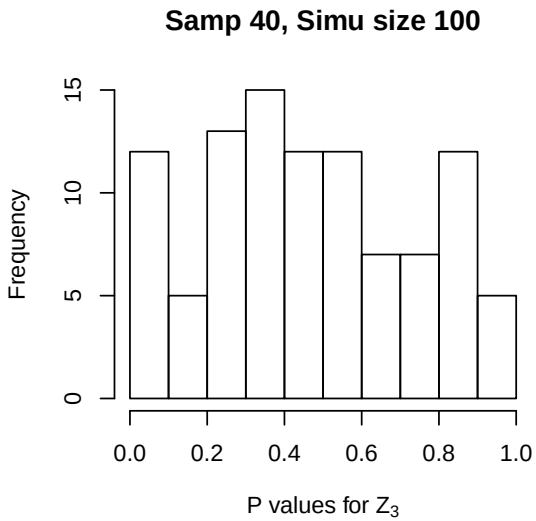


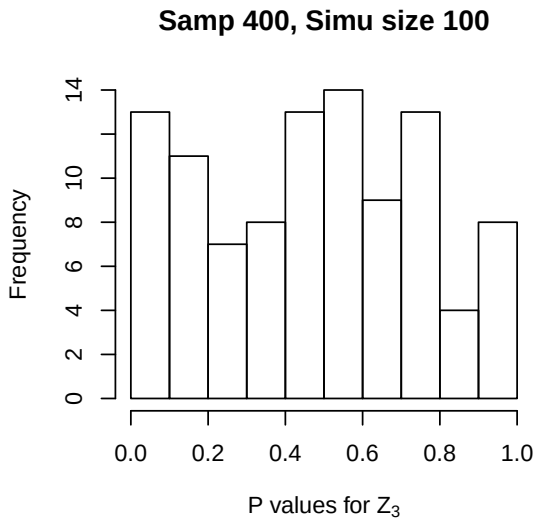
‡ Prochenka.et al. (2015) *Bioinformatics*.

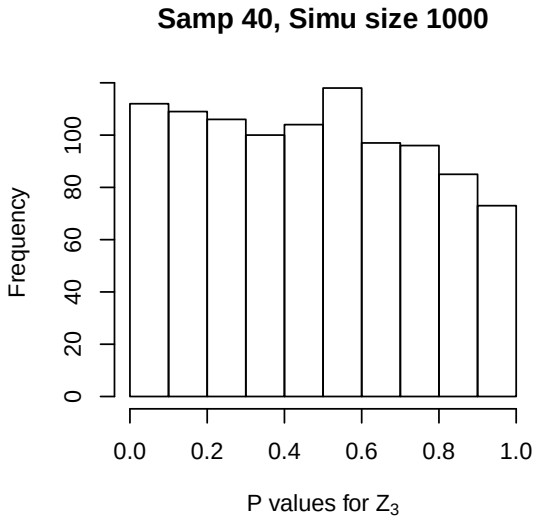
Coverage probability of confidence intervals

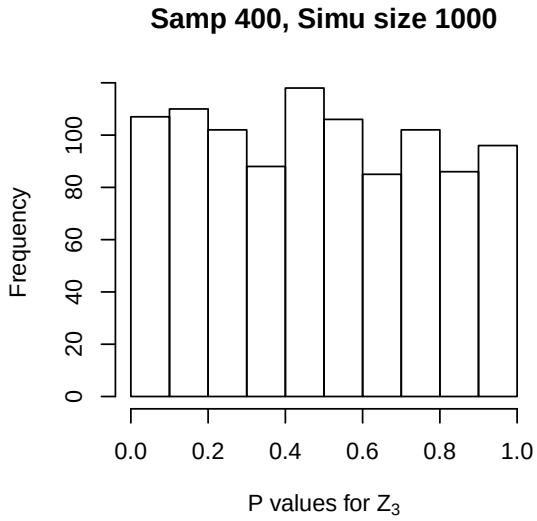


Region-based p-values under null









Summary & Discussion

- a **model** characterizing bisulfite sequencing data from multiple samples, which copes naturally with variable read depth, experimental errors and test samples with a mixture of cell types.
- a **smoothed EM method** to **make inference about smooth exposure/covariate effects**.
- the method is shown to be capable in **capturing the major underlying patterns** in the data.
- **Next step plans**
  - estimating error rate p_0 by calculating C-T conversion rate at non-CpG Cs
  - adding subject-specific random effects to account for the heterogeneity in methylation profiles that cannot be explained by covariates Z_is in the model
  - correlated samples

Acknowledgement

Dr. Celia Greenwood (McGill)
Dr. Karim Oualkacha (UQAM)
Dr. Lajmi Lakhal-Chaieb (Laval)
Dr. Kathleen Klein (JGH)
Dr. Marie Hudson (McGill)

Thanks

Questions & Comments