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

Kaiqiong Zhao

Department of Epidemiology, Biostatistics and Occupational Health,

McGill University

Bioinformatics-Biostatistics Journal Club and Research Seminar Series

University of Manitoba

December 16, 2020

# Overview



**Please feel free to interrupt and ask questions at any time during the talk!**

- Background and motivation
- †New method 1: SOMNiBUS (SmOoth ModeliNg of BisUlfite Sequencing)
- ‡New method 2: dSOMNiBUS (dispersion-adjusted SmOoth ModeliNg of BisUlfite Sequencing)
- ⋆New method 3: sparseSOMNiBUS (SOMNiBUS with variable selection)

---

† Zhao, et.al (2020). A novel statistical method for modeling covariate effects in bisulfite sequencing derived measures of DNA methylation. *Biometrics*. Early-View

‡ Zhao, et.al (2020+). Detecting differentially methylated regions in bisulfite sequencing data using quasi-binomial mixed models with smooth covariate effect estimates.

⋆ Zhao, et.al (2020+). In preparation

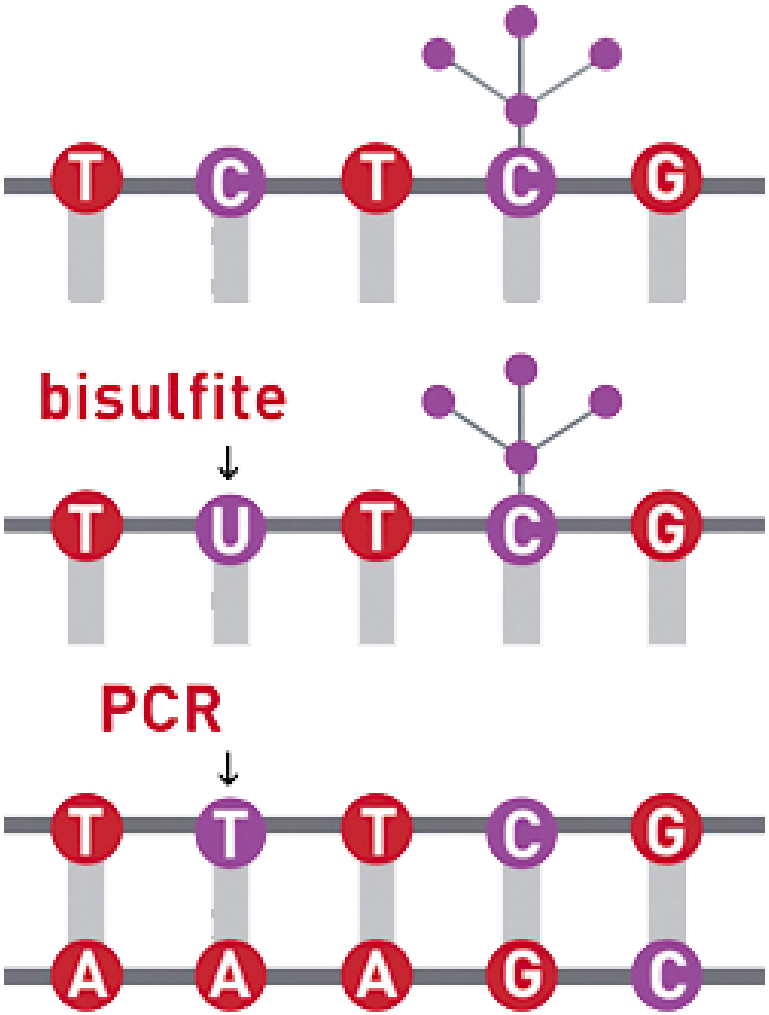
# Epigenetics and DNA Methylation



- change gene expression without changing DNA sequence
- can be altered by age, diet, stress and environmental exposures
- Localized abnormal methylation is a characteristic feature of many diseases

# Bisulfite Sequencing & Methylation





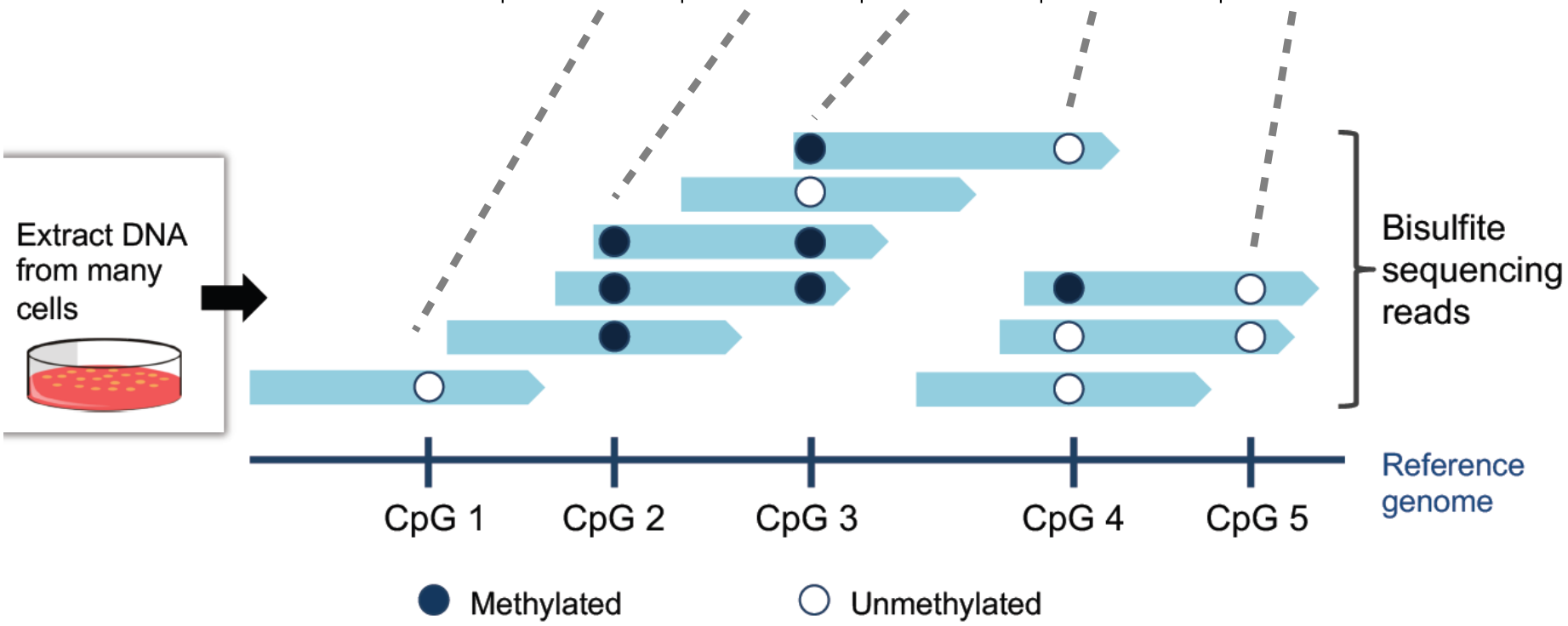

https://www.diagenode.com/en/applications/dna-bisulfite-conversion

Methylated cytosines are not converted by bisulfite treatment

# Sequencing-derived DNA methylation data



| | CpG 1 | CpG 2 | CpG 3 | CpG 4 | CpG 5 |
|---|---|---|---|---|---|
| Methylated Counts (Y) | 0 | 3 | 3 | 1 | 0 |
| Read depth (X) | 1 | 3 | 4 | 4 | 2 |
| Proportion (Y/X) | 0 | 1 | 0.75 | 0.25 | 0 |




http://kkorthauer.org/talks/korthauer_aisc_2018_static.pdf

# Motivating datasets

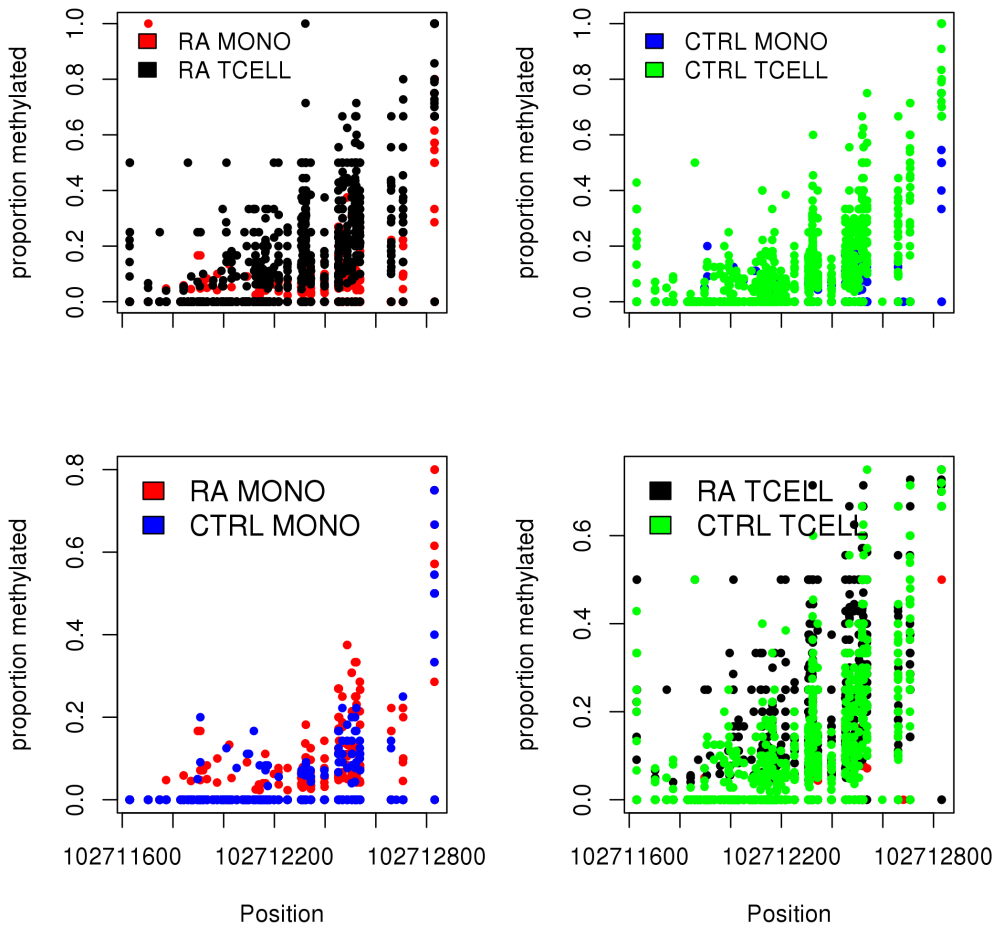
## Methylation profiles of Rheumatoid Arthritis (RA) patients and controls (from our collaborator Dr. Marie Hudson)



- Targeted Custom Capture Bisulfite Sequencing
  - predefined genomic regions
  - 5 million CpGs
- Cell-separated blood samples

| | Monocytes | T cells |
|---|---|---|
| RA | 10 | 12 |
| Controls | 8 | 13 |

- Small region on chromosome 4 near *BANK1*
- 123 CpGs



# Goal



**Find associations between**

- methylation patterns in each targeted region, and
- phenotypes or covariates

# Challenges / Opportunities



- Read depth at CpGs varies substantially
  - Need a model that can use all available data
- Cell-type mixture affects observed methylation levels
  - Adjust for this in model
- Sequencing errors, e.g. bisulfite conversion error
  - Build a model allowing for error
- Local correlations in methylation levels
  - Opportunity for imputing missing data or poorly measured signals
  - Opportunity for modelling smooth effects along the genome

# Existing methods appropriate for regions



| Method | regional | one-stage | count-based | read-depth variability | adjust for confounding | experimental errors |
|---|---|---|---|---|---|---|
| SOMNiBUS | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| BSmooth | ✓ | | | ✗ | | |
| SMSC | ✓ | | | ✗ | | ✓ |
| dmrseq | ✓ | | | ✓ | ✓ | |
| Biseq | ✓ | | | ✗ | ✓ | |
| GlobalTest | ✓ | ✓ | | | ✓ | |

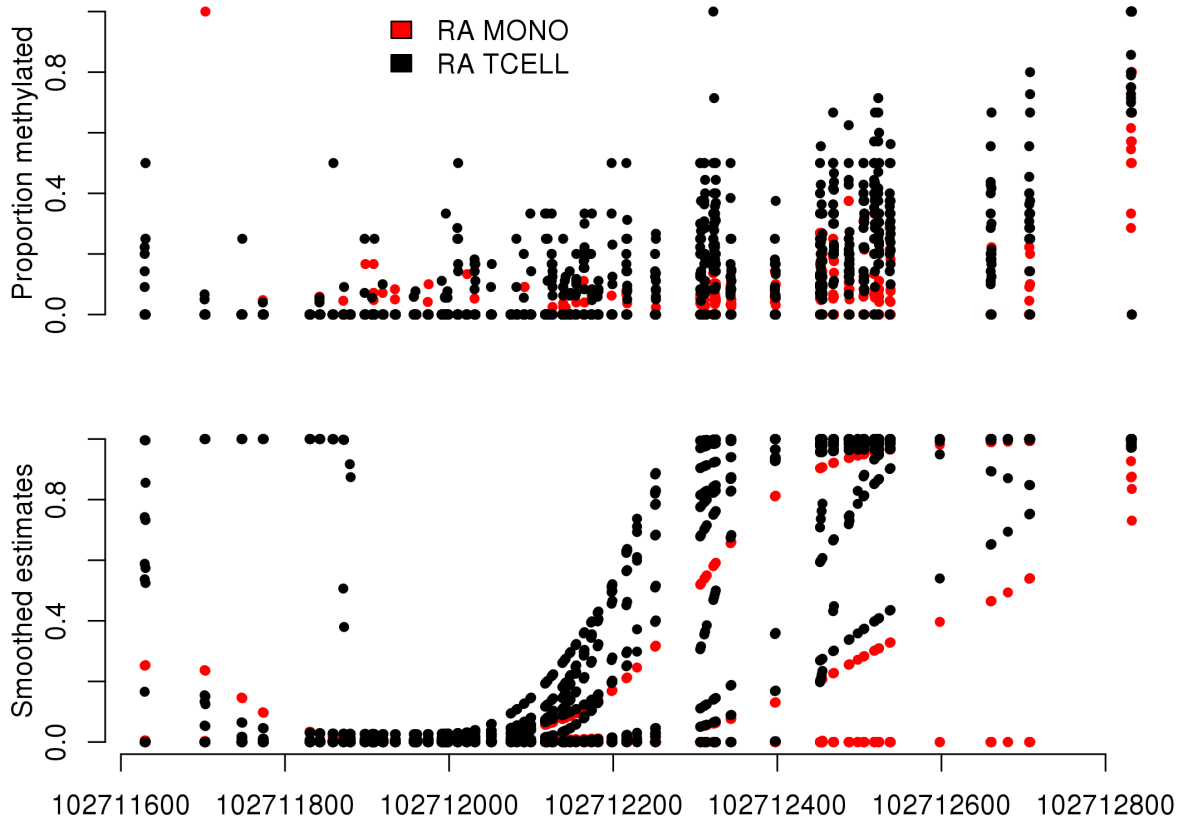
**BSmooth**: Hansen, 2012
**SMSC**: Lakhal-Chaieb, 2017
**dmrseq**: Korthauer, 2018
**BiSeq**: Hebestreit, 2013
**GlobalTest**: Goeman, 2006

# An example of two-stage method

## Raw data & per-sample smoothed estimates



Results from SMSC (Lakhal-Chaieb, 2017)




Genomic position (in bp)

# SOMNiBUS

SmOoth ModeliNg of BisUlfite Sequencing



| Method | regional | one-stage | count-based | read-depth variability | adjust for confounding | experimental errors |
|---|---|---|---|---|---|---|
| SOMNiBUS | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| BSmooth | ✓ | | | ✗ | | |
| SMSC | ✓ | | | ✗ | | ✓ |
| dmrseq | ✓ | | | ✓ | ✓ | |
| Biseq | ✓ | | | ✗ | ✓ | |
| GlobalTest | ✓ | ✓ | | | ✓ | |

Motivation: a novel one-stage method that

- collapses smoothing and testing steps into a single step
- allows for experimental errors, variable read depths and test samples with a mixture of cell types
- provides **rigorous uncertainty assessment** for differentially methylated regions

# Overview



- Background and motivation
- **[†]New method 1: SOMNiBUS (SmOoth ModeliNg of BisUlfite Sequencing)**
- [‡]New method 2: dSOMNiBUS (dispersion-adjusted SmOoth ModeliNg of BisUlfite Sequencing)

---

[†] Zhao, et.al (2020). A novel statistical method for modeling covariate effects in bisulfite sequencing derived measures of DNA methylation. *Biometrics*. Early-View

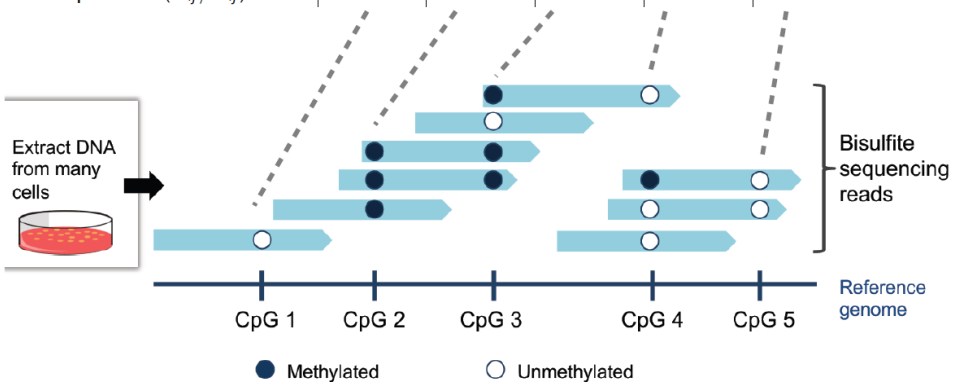
[‡] Zhao, et.al (2020+). Detecting differentially methylated regions in bisulfite sequencing data using quasi-binomial mixed models with smooth covariate effect estimates.

# Notations



- $X_{ij}$: total number of reads aligned to CpG $j$ from sample $i$
- $Y_{ij}$: observed methylated counts at CpG $j$ for sample $i$. $Y_{ij} = \sum_{k=1}^{X_{ij}} Y_{ijk}$
- $S_{ij}$: true methylated counts at CpG $j$ for sample $i$. $S_{ij} = \sum_{k=1}^{X_{ij}} S_{ijk}$

| | CpG 1 | CpG 2 | CpG 3 | CpG 4 | CpG 5 |
|---|---|---|---|---|---|
| Methylated Counts ($Y_{ij}$) | 0 | 3 | 3 | 1 | 0 |
| Read depth ($X_{ij}$) | 1 | 3 | 4 | 4 | 2 |
| Proportion ($Y_{ij}/X_{ij}$) | 0 | 1 | 0.75 | 0.25 | 0 |




- $t_{ij}$: the genome position (in bp) for sample $i$ at CpG $j$
- $Z_{1i}, Z_{2i}, \ldots Z_{Pi}$ are the $P$ covariates.
- $\pi_{ij}$: the methylation proportion parameter for sample $i$, CpG $j$

# SOMNiBUS[†]: Model



- Assume **known error parameters** $p_0$ and $p_1$,

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

- Specify the model

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

- Smooth curves along the genome for
  - Overall methylation
  - Covariate effects

[†] R package: `https://github.com/kaiqiong/SOMNiBUS`.

# Technical details 1: splines



- Use splines for smoothing

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

- Penalize roughness of effect curves $\beta_p(t_{ij})$.

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

$\{\lambda_0, \lambda_1, \ldots \lambda_P\}$ are the smoothing parameters.

  - Penalties go onto second derivatives
  - $P + 1$ penalization parameters for $P$ covariates

# Technical details 2: E-M algorithm



**E step:** Calculate $\eta_{ij}^{\star} = \mathbb{E}(S_{ij} \mid Y_{ijk}\ ; \boldsymbol{\alpha}^{\star})$

**M step:** ‡Maximize $Q(\boldsymbol{\alpha}, \boldsymbol{\lambda} \mid \boldsymbol{\alpha}^{\star}) = l(\boldsymbol{\eta}^{\star}; \boldsymbol{\alpha}) - \frac{1}{2}\boldsymbol{\alpha}^{T}\boldsymbol{A}_{\boldsymbol{\lambda}}\boldsymbol{\alpha} + \frac{1}{2}\log\{|\boldsymbol{A}_{\boldsymbol{\lambda}}|_{+}\}$

- Estimate $\boldsymbol{\alpha}$ given the value of $\boldsymbol{\lambda}$: P-IRLS

$$\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\lambda}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \left\{ l(\boldsymbol{\eta}^{\star}; \boldsymbol{\alpha}) - \frac{1}{2}\boldsymbol{\alpha}^{T}\boldsymbol{A}_{\boldsymbol{\lambda}}\boldsymbol{\alpha} \right\}$$

- Estimate $\boldsymbol{\lambda}$: maximize the Laplace-approximated restrictive (or marginal) likelihood

$$L^{M}(\boldsymbol{\lambda}) = \int \exp\left\{Q(\boldsymbol{\alpha}, \boldsymbol{\lambda} \mid \boldsymbol{\alpha}^{\star})\right\} d\boldsymbol{\alpha} \approx \text{Laplace}(\boldsymbol{\lambda}; \widehat{\boldsymbol{\alpha}}_{\boldsymbol{\lambda}}).$$

‡ Wood (2011), JRSSB; R package `mgcv`

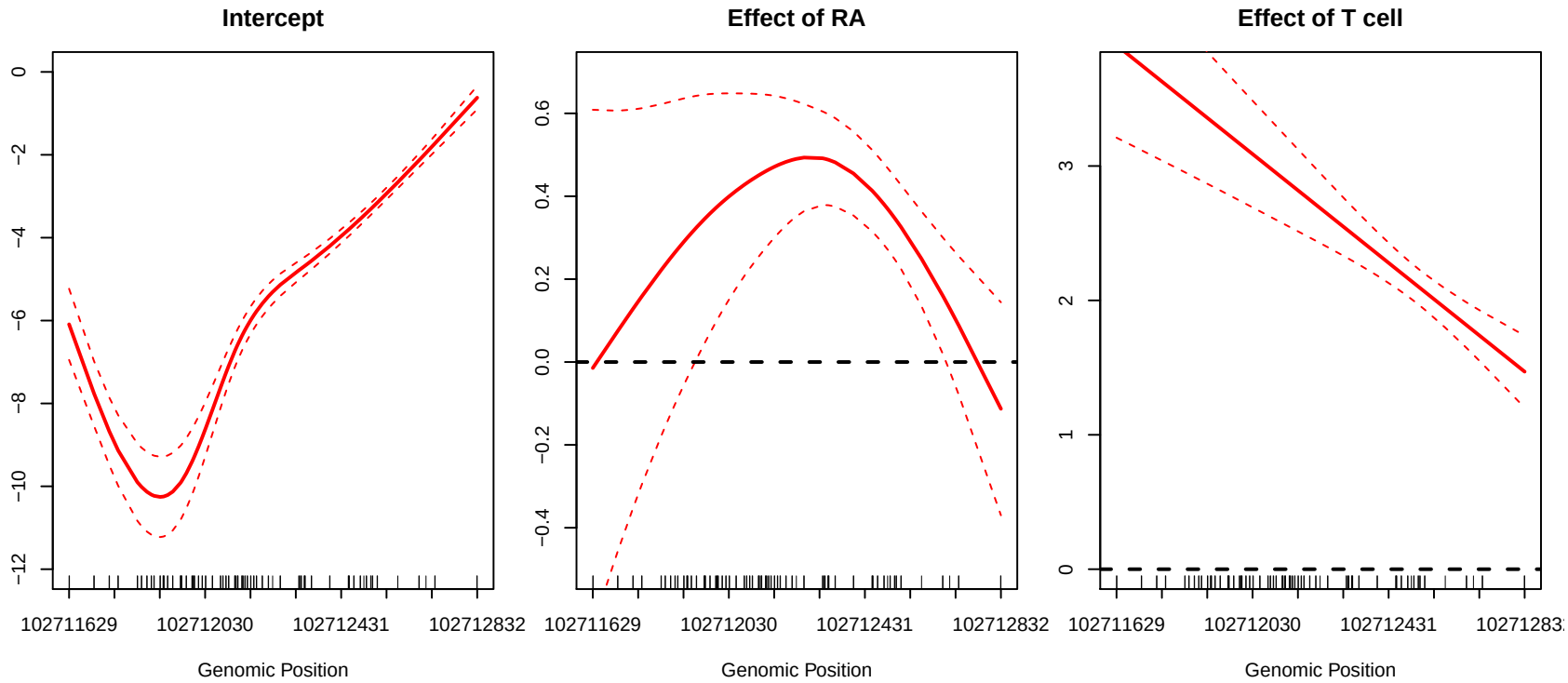
# Technical details 3: inference



- Pointwise confidence intervals
- Regional tests for non-zero covariate effects
  - for each covariate, or
  - for the combined effects of multiple covariates
- Penalization affects effective degree of freedom

# Results in *BANK1* region






$p = 1.11e-16$ $p = 6.37e-218$

- Error parameters $p_0 = 0.003$ and $1 - p_1 = 0.1$[‡]

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

# Simulation study

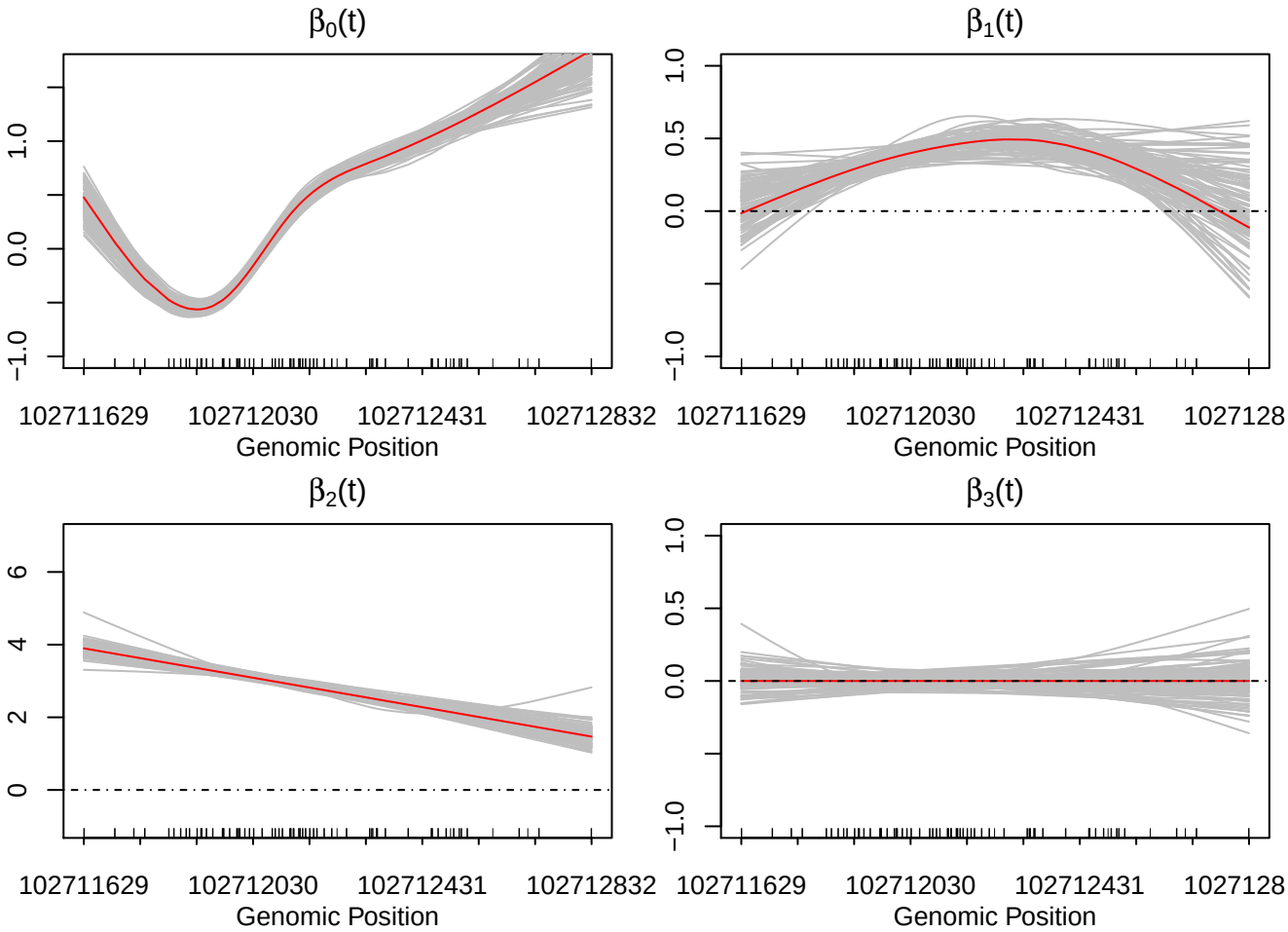


- Simulated dataset similar to the *BANK1* example
- One “null” covariate with no effect
- Two covariates with effects like those seen near *BANK1*
- Simulate the observed methylated counts $Y_{ij}$ from

$$Y_{ij} \mid S_{ij} \sim \text{Binomial}(S_{ij}, p_1) + \text{Binomial}(X_{ij} - S_{ij}, p_0).$$

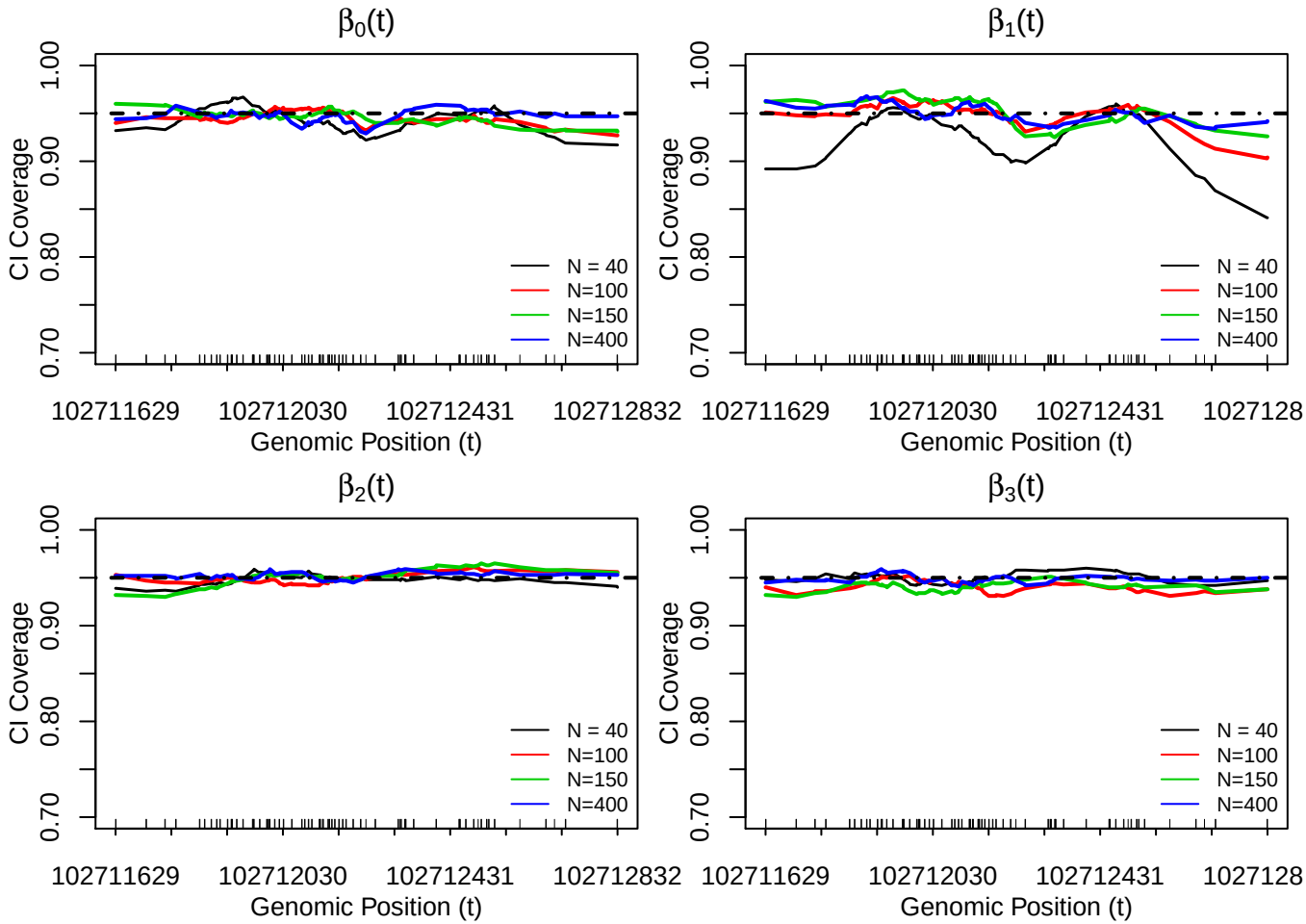
# Little bias in the curve estimates





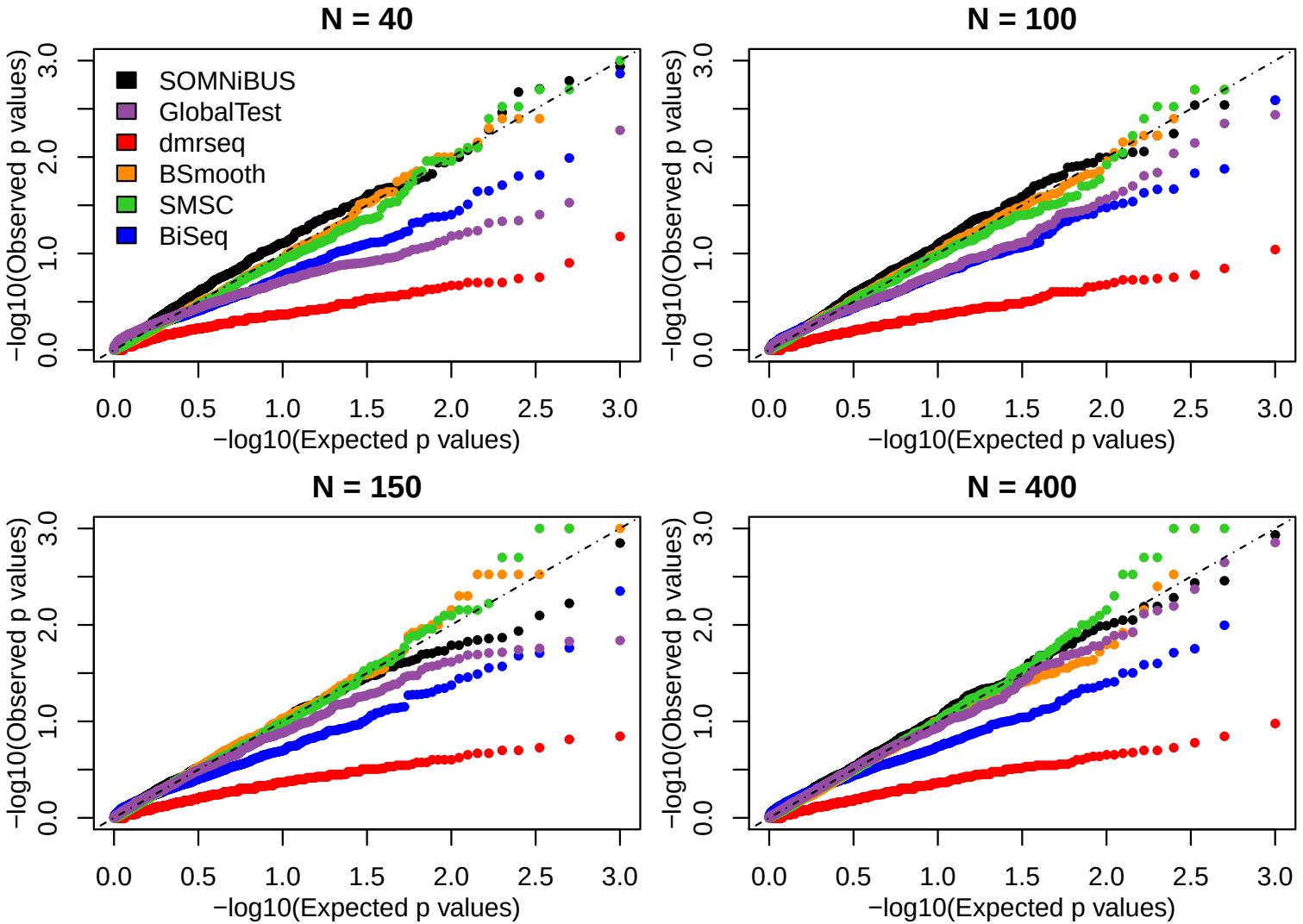
# Empirical confidence interval coverages





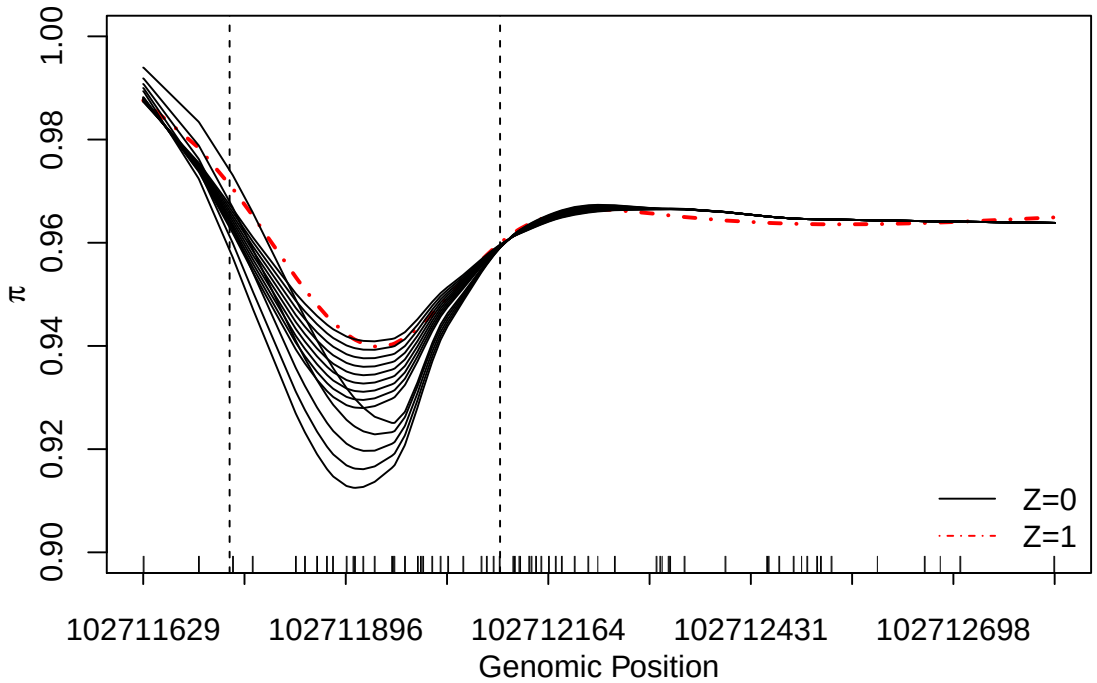
# Accurate type I error rates





# Simulation to evaluate power



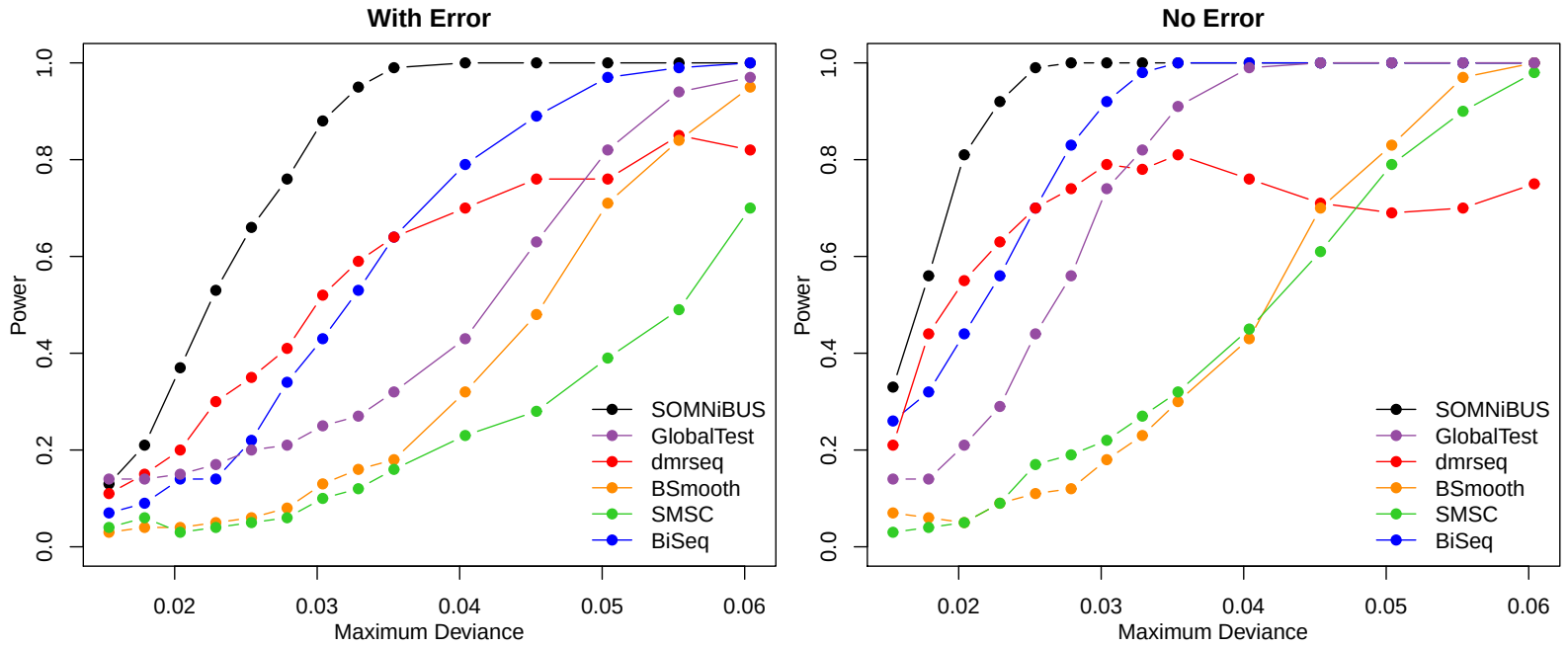



$Z = 1$ curve in red (fixed)

$Z = 0$ curve varied to give various sizes of differences

# Increased power to detect DMRs






Maximum difference between curves

- With Error: $p_0 = 0.003, p_1 = 0.9$
- No Error: $p_0 = 0, p_1 = 1$

# SOMNiBUS: Summary



**Advantages**

- Able to use data from many more CpGs where univariate analysis fails / power gain
- One-stage nature
- Explicitly allows for experimental errors
- Inference!

# SOMNiBUS: Summary


**Advantages**

- Able to use data from many more CpGs where univariate analysis fails / power gain
- One-stage nature
- Explicitly allows for experimental errors
- Inference!

**Room for improvements**

- Its underlying binomial assumption may be overly restrictive
- It is only applicable for data with negligible (within-group) variability (such as data from inbred animal or cell line experiments)

# Overview



- Background and motivation
- [†]New method 1: SOMNiBUS (SmOoth ModeliNg of BisUlfite Sequencing)
- **[‡]New method 2: dSOMNiBUS (dispersion-adjusted SmOoth ModeliNg of BisUlfite Sequencing)**

---

† Zhao, et.al (2020). A novel statistical method for modeling covariate effects in bisulfite sequencing derived measures of DNA methylation. *Biometrics*. Early-View

‡ Zhao, et.al (2020+). Detecting differentially methylated regions in bisulfite sequencing data using quasi-binomial mixed models with smooth covariate effect estimates.

# Motivating datasets
(from our collaborator Dr. Sasha Bernatsky)


- CARTaGENE is an ongoing population-based cohort, including $\sim$43,000 participants aged 40 to 69 years in Quebec
- The level of anti-citrullinated protein antibodies (ACPA) is a marker of rheumatoid arthritis (RA) risk that often presents prior to any clinical manifestations
- Aim: detect differentially methylated regions (DMRs) associated with ACPA

# Motivating datasets
(from our collaborator Dr. Sasha Bernatsky)



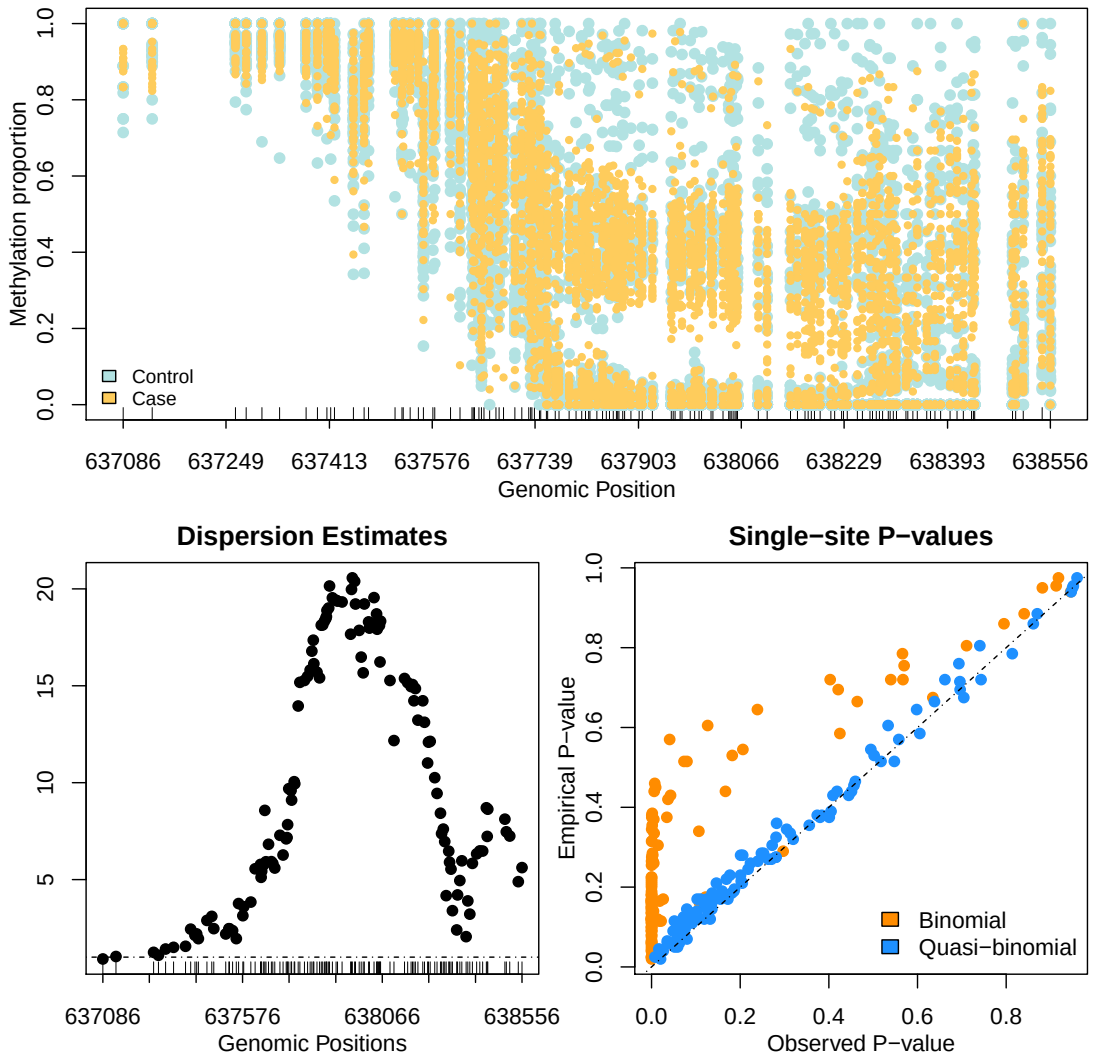
- blood samples of ACPA positive and ACPA negative subjects
  - **covariate of primary interest**: ACPA status
  - **adjusting variables**: age, sex, smoking status and cell type composition(captured by the top 4 PCs)

- two batches of data, referred to as data 1 and data 2, were collected in 2017 and 2019, respectively.

| | data 1 (N =116 ) | data 2 (N = 102) |
|---|---|---|
| ACPA Positives | 55 | 48 |
| ACPA Negatives | 61 | 54 |
| Number of targeted regions (with at least 50 CpGs) | 10,759 | 12,985 |

# Observed dispersion in a targeted region





# New method 2: dSOMNiBUS
(dispersion-adjusted SmOoth ModeliNg of BisUlfite Sequencing)



- The same error model

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

- A quasi-binomial mixed model with the **combination** of
  - a *multiplicative* dispersion, $\phi$
  - an *additive* dispersion, $\boldsymbol{u}$, (i.e. a subject-specific RE)

$$\begin{aligned} \log \frac{\pi_{ij}}{1-\pi_{ij}} &= \beta_0(t_{ij}) + \beta_1(t_{ij})Z_{1i} + \beta_2(t_{ij})Z_{2i} + \ldots + \beta_P(t_{ij})Z_{Pi} + u_i, \\ u_i &\overset{iid}{\sim} N(0, \sigma_0^2) \\ \mathbb{V}\text{ar}(S_{ij} \mid u_i) &= \phi X_{ij}\pi_{ij}(1-\pi_{ij}) \end{aligned}$$

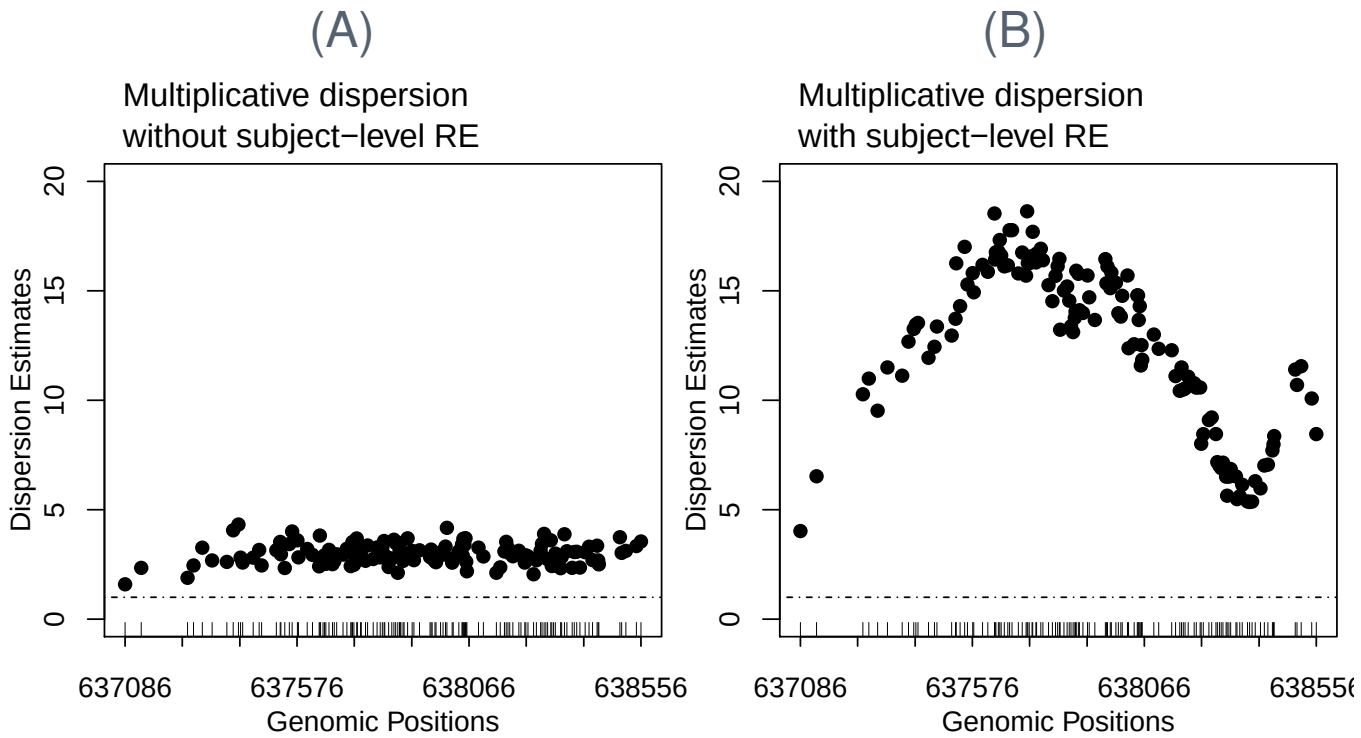
- Smoothness parameters to penalize the roughness of effect curves.

R package: `https://github.com/kaiqiong/SOMNiBUS`

# RE term enables flexible dispersion patterns in a region



A byproduct of introducing a subject-level RE to a model with smooth covariate effects is a regional dispersion pattern of varying degree.




$$\mathbb{V}\text{ar}(S_{ij}) \approx X_{ij}\pi^{\star}_{ij}(1-\pi^{\star}_{ij})\left\{\phi + \sigma_0^2\left(X_{ij}-\phi\right)\pi^{\star}_{ij}(1-\pi^{\star}_{ij})\right\}$$

# Technical details 1: difficulties



- Three sets of unknown parameters
  - conditional mean parameters (REs): $\boldsymbol{\mathcal{B}} = (\boldsymbol{\alpha}, \boldsymbol{u}) \in \mathbb{R}^{N+\sum_0^P L_p}$
  - variance component parameters: $\boldsymbol{\Theta} = (\boldsymbol{\lambda}, \sigma_0^2) \in \mathbb{R}^{P+2}$
  - **multiplicative dispersion parameter**: $\phi$
- The conditional 'distribution' of $\boldsymbol{S} \mid \boldsymbol{\mathcal{B}}$ is not available
- Joint estimation of $\phi$ and $\boldsymbol{\Theta}$ is required, as $\text{Laplace}(\phi, \boldsymbol{\Theta}; \widehat{\boldsymbol{\mathcal{B}}}) \neq f(\phi)g(\boldsymbol{\Theta})$
- In the presence of data errors, one cannot easily estimate $\phi$ using the EM algorithm
  - The estimating equation for $\phi$ is not linear in the unknown methylated counts $\boldsymbol{S}$

# Technical details 2: Estimation & Inference



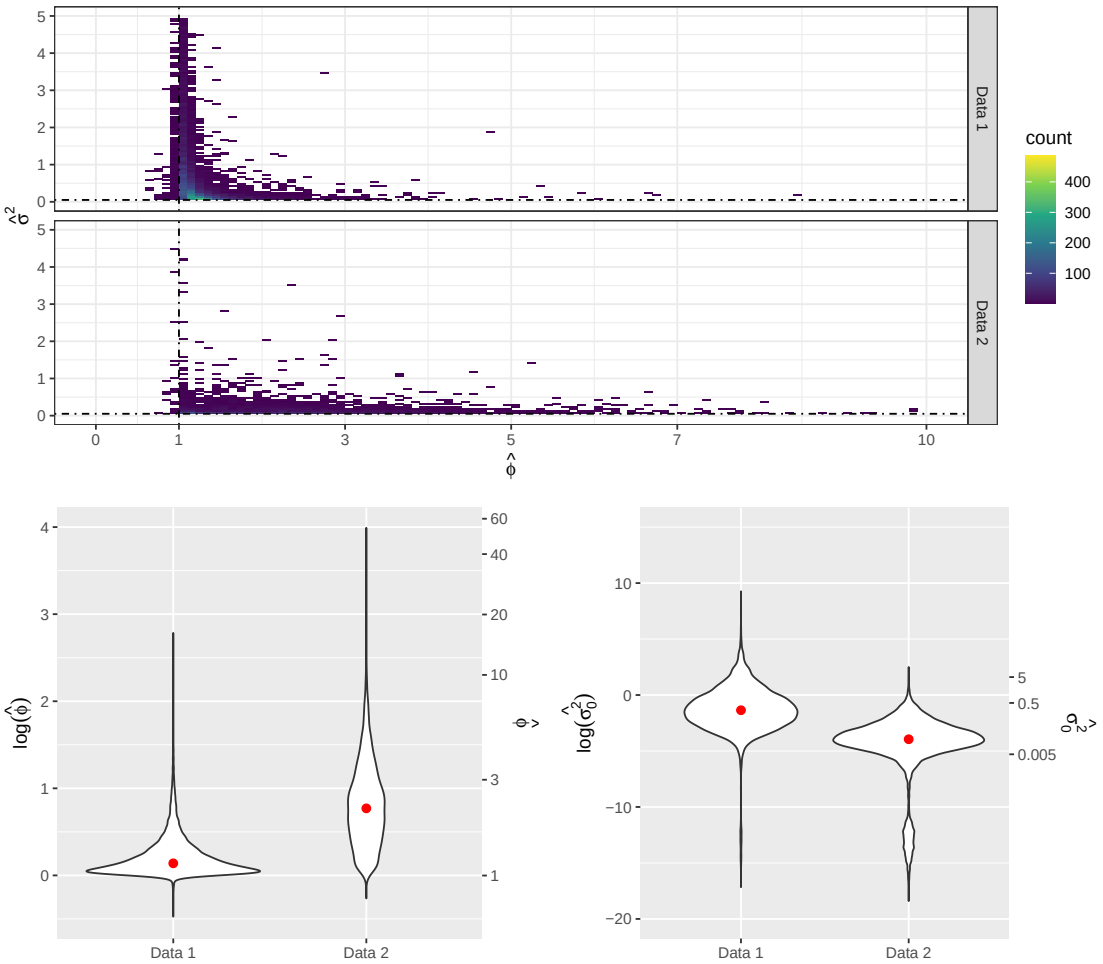
- Use the notation of extended quasi-likelihood to write the conditional quasi-likelihood function
- Calculate Laplace-approximated marginal quais-likelihood function and its derivatives
- A hybrid ES algorithm
  - A plug-in estimator for $\phi$ by exploting its relationship with the dispersion for the contaminated outcome $\boldsymbol{Y}$
  - Estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ using ES iterations[‡] assuming $\phi$ is fixed and known
- Inference using the observed quasi-Fisher information

† Efron (1986), Jorgensen (1987), McCullagh and Nelder (1989)

‡ Elashoff and Ryan (2004)

# Both additive and multiplicative dispersion is present in the data

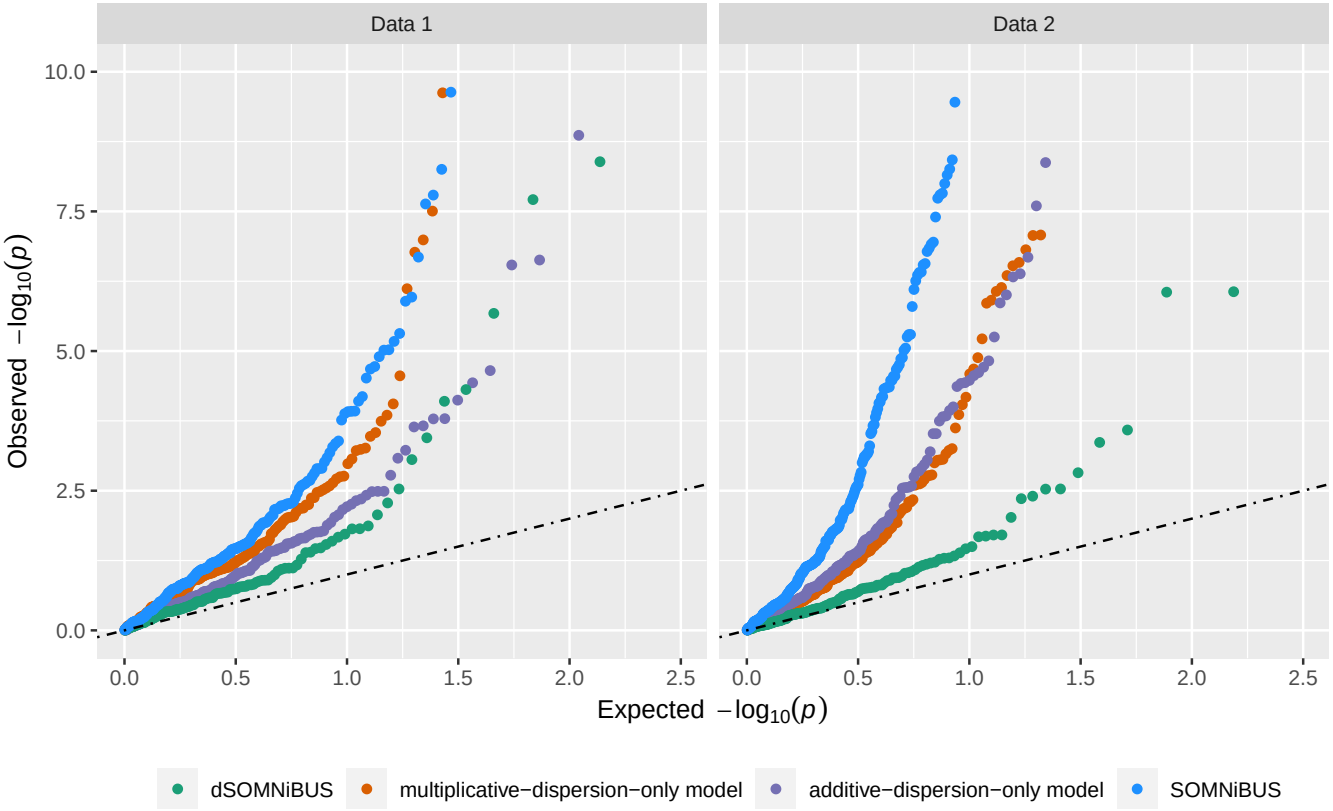





The distribution of estimated $\phi$ and $\sigma_0^2$ for the 10,759 and 12,985 regions in dataset 1 and 2, respectively.

# Ignoring either type of dispersion leads to inflated type I errors



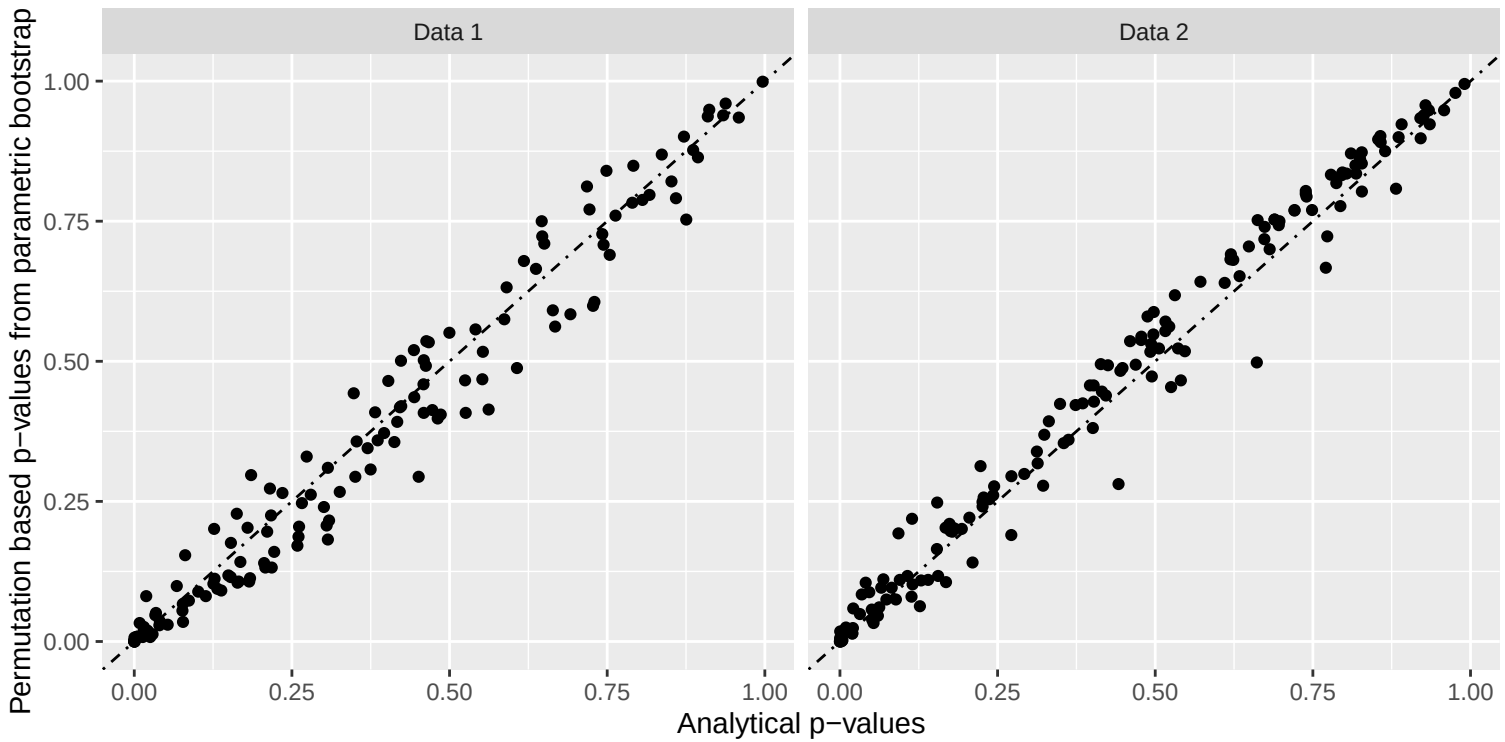



dSOMNiBUS: $\phi > 0$, $\sigma_0^2 > 0$; multiplicative-dispersion-only model: $\phi > 0$, $\sigma_0^2 = 0$

SOMNiBUS: $\phi = 1$, $\sigma_0^2 = 0$; additive-dispersion-only model: $\phi = 1$, $\sigma_0^2 > 0$

# Analytical v.s. bootstrap based p-values






Our inference procedure provides well-calibrated regional p-values.

# Simulation



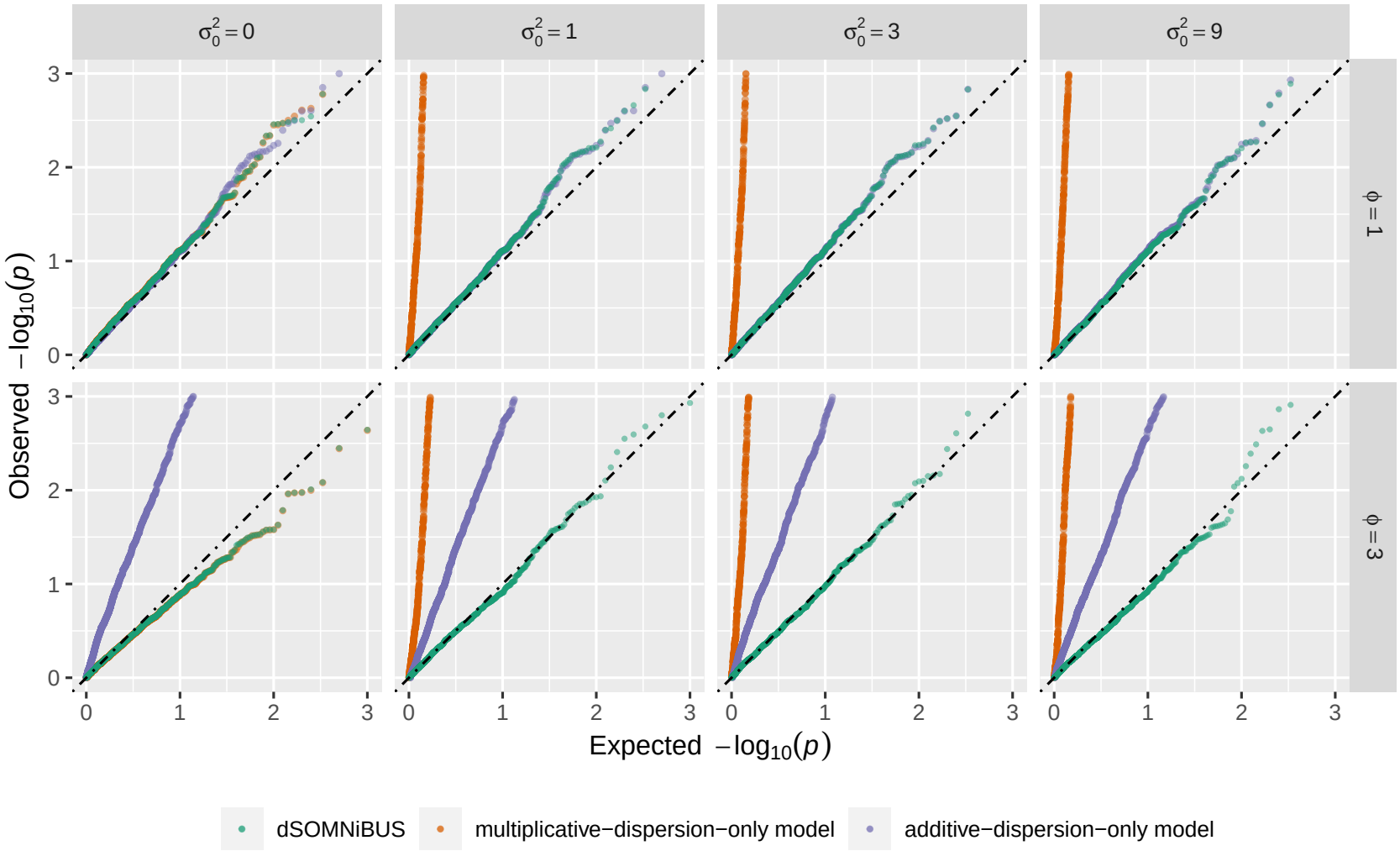
- Specify the same $\beta_p(t)$ and $Z_p$ as paper 1.
- $S_{ij} \sim$ Beta-binomial $\left(\mu_{ij} = \pi_{ij}, \rho_{ij} = \dfrac{\phi - 1}{X_{ij} - 1}, size = X_{ij}\right)$
- In this way, we can always guarantee $\dfrac{\mathbb{V}\text{ar}(S_{ij})}{X_{ij}\pi_{ij}(1 - \pi_{ij})} \equiv \phi$.
- Recall: If $S \sim$ Beta-binomial $(\mu, \rho, size = X)$,

$$\mathbb{V}\text{ar}(S) = \underbrace{[1 + (X - 1)\rho]}_{\text{dispersion}} \underbrace{X\mu(1 - \mu)}_{V(\mathbb{E}(Y))}.$$

# The impact of dispersion

$p_0 = 0.003, p_1 = 0.9$

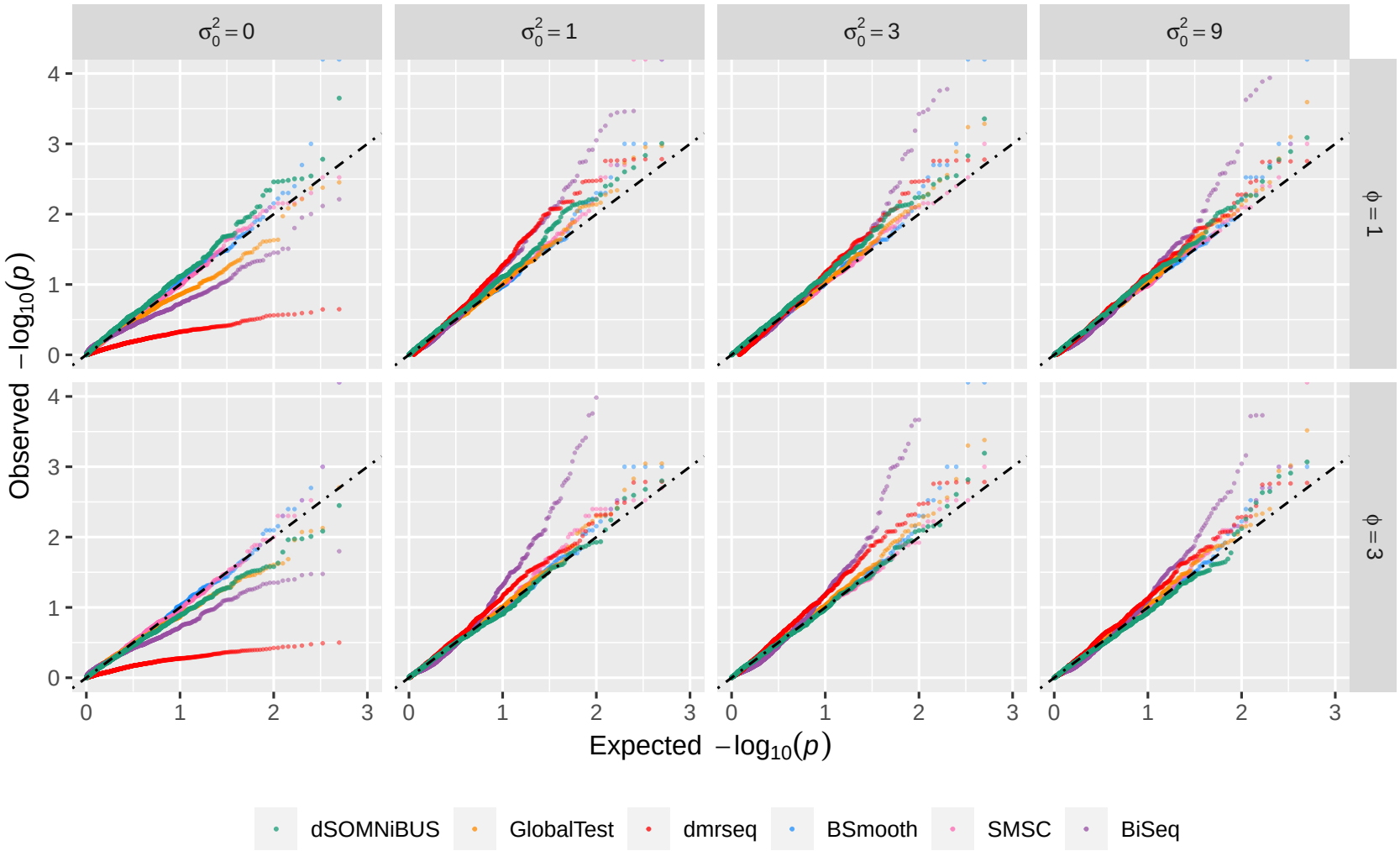




# Type I Error

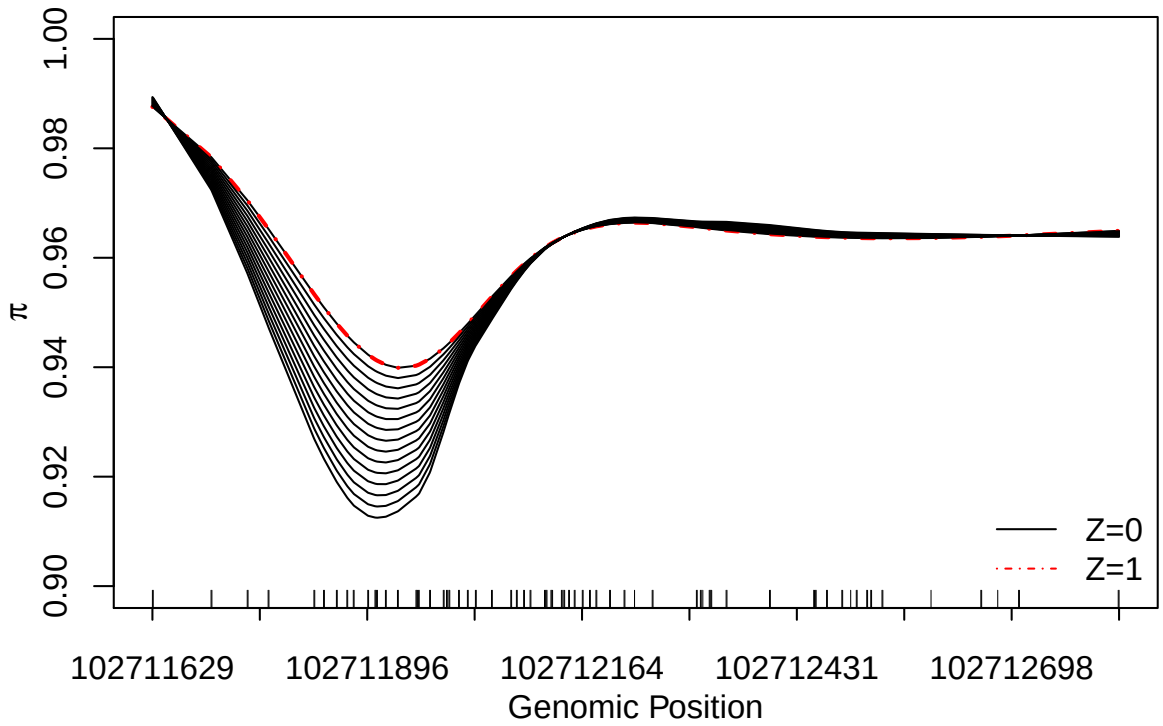
$p_0 = 0.003, p_1 = 0.9$





# Simulation to evaluate power



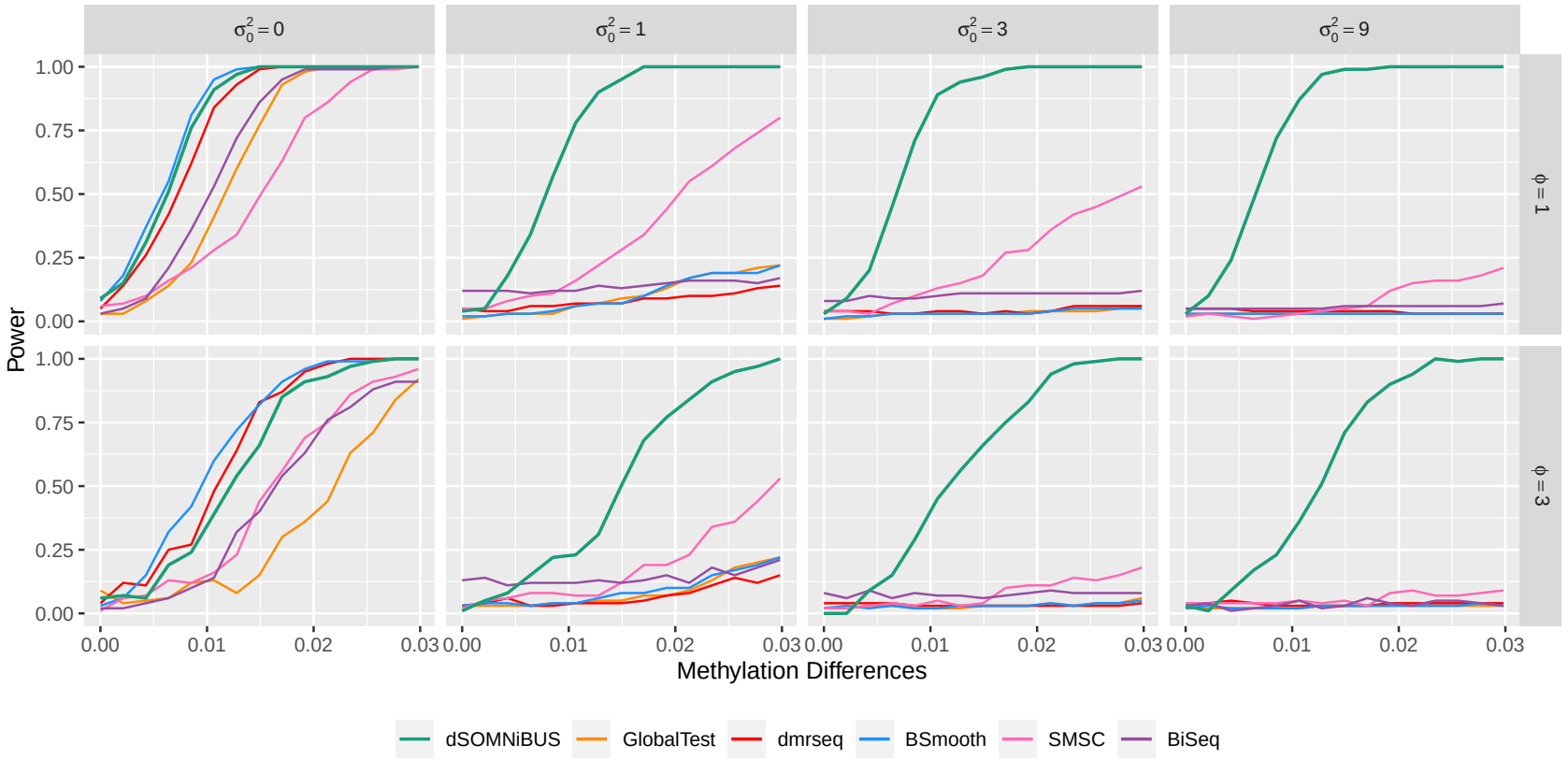



$Z = 1$ curve in red (fixed)

$Z = 0$ curve varied to give various sizes of differences

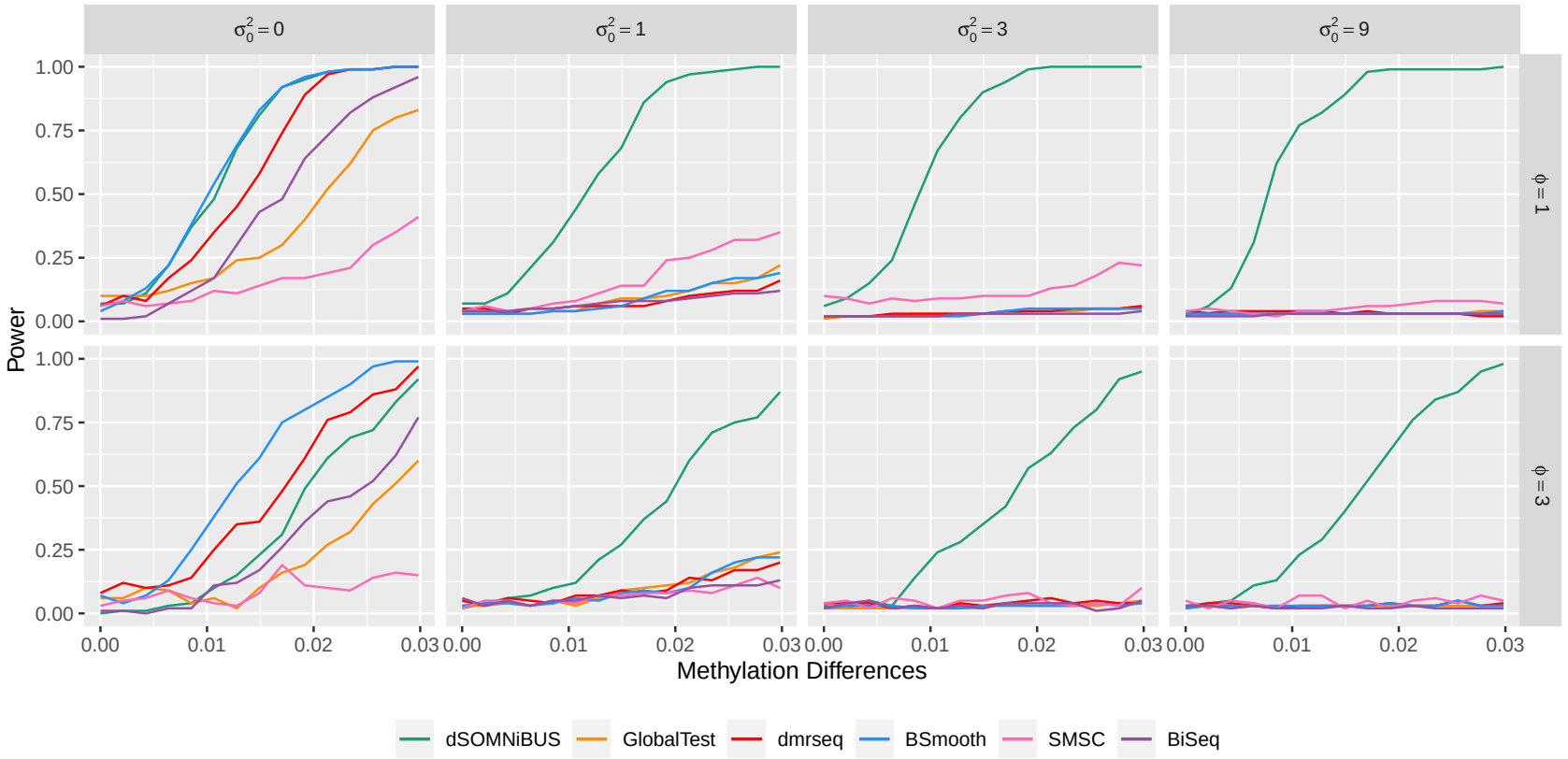
# Power without errors: $p_0 = 0, p_1 = 1$





# Power with errors: $p_0 = 0.003, p_1 = 0.9$





# Summary



- an adequate representation of realistic dispersion trends in regional methylation data
- well-founded theoretical properties accounting for all (known) sources of data variability and possible experimental errors
- increased power; correct control of the type I error rate

- methodologies can be generally applied to other types of count data
  - allele-specific gene expression (ASE) measured from RNA-seq data
  - any type of count data for a more comprehensive representation of dispersion
  - varying-coefficients models in other context, e.g. temporal trend

# Next step plans



- **integrate SNP information** (automatic variable selection)
- covariates (eg. disease status) may influence the variability/dispersion of DNA methylation (model $\phi(Z)$)
- correlated samples (additional set of random effects)

## methylation QTL mapping

Given: a set of CpGs & a set of nearby SNPs ($P >> N$)
Output: genetic variants associated with methylation levels in the test region

### sparseSOMNiBUS

- a sparsity-smoothness penalty on each functional component $\beta_p(t)$

$$J(\beta_p) = \lambda\sqrt{(1-\alpha)J_1(\beta_p) + \alpha J_2(\beta_p)}$$

where

$$J_1(\beta_p) = \int (\beta_p(t))^2\, dt$$

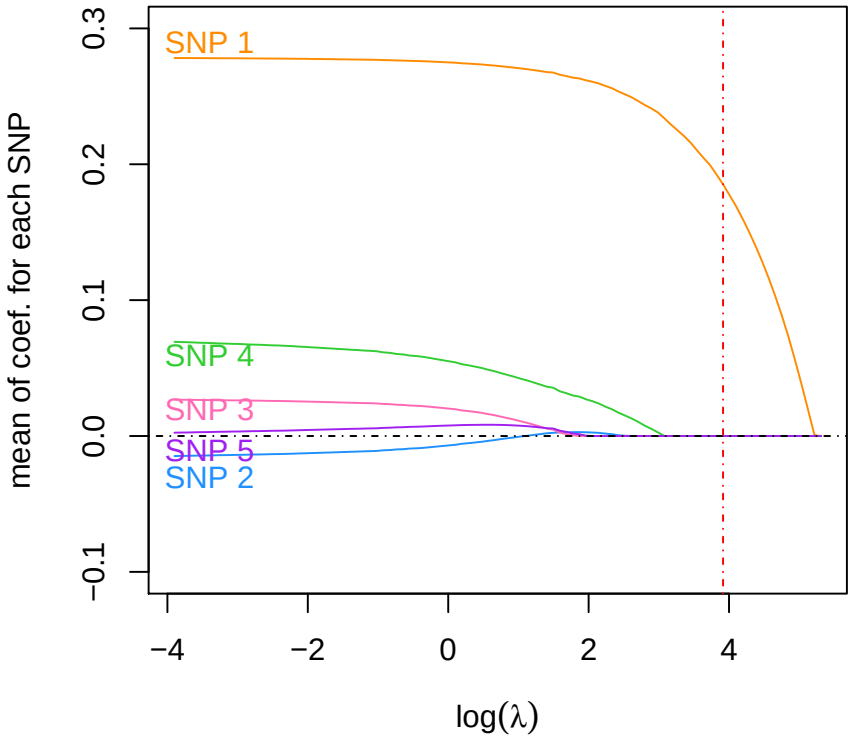
$$J_2(\beta_p) = \int (\beta_p''(t))^2\, dt$$

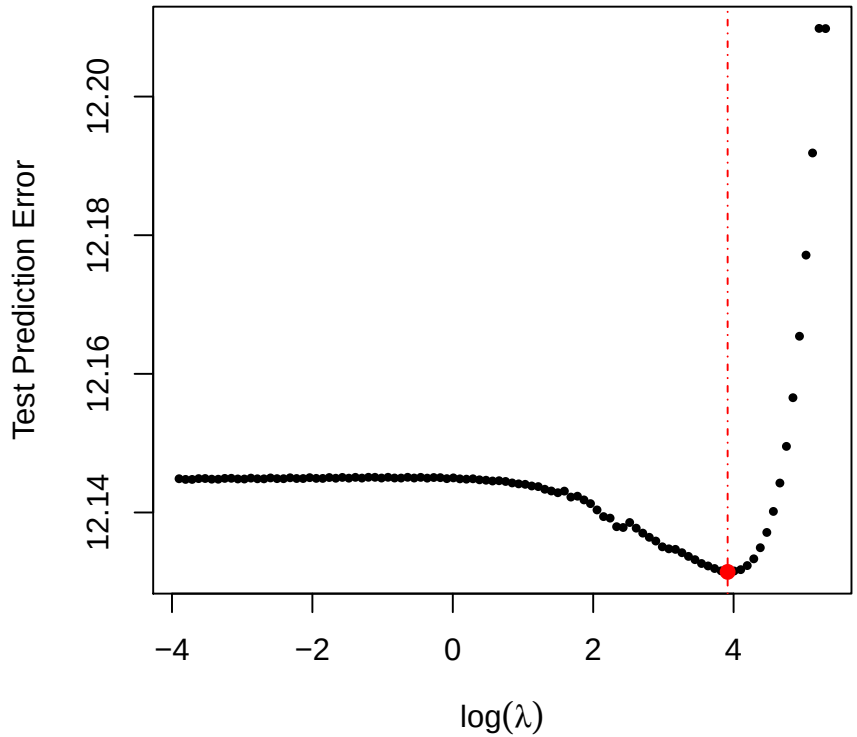
- proximal gradient descent + backtracking line search (Rcpp)
- tunning parameters $\lambda$ and $\alpha$, selected by cross-validation

# A simple illustration of sparseSOMNiBUS



- a methylation region with 123 CpG sites
- 5 SNPs: 1 mQTL and 4 negative controls




under the best chosen $\alpha = 0.55$

# Acknowledgement




- Dr. Celia Greenwood and Dr. Karim Oualkacha
- Dr. Lajmi Lakhal-Chaieb, Dr. Aurélie Labbe
- Dr. Yi Yang
- Dr. Kathleen Klein
- Dr. Sasha Bernatsky, Dr. Marie Hudson, Dr. Inés Colmegna
- the CARTaGENE study investigators
- the participants in the CARTaGENE study




# Thanks

Questions & Comments

# SOMNiBUS†: Model



- Assume **known error parameters** $p_0$ and $p_1$,

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

- Specify the model

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

- Consider basis expansion: $\beta_p(t_{ij}) = \sum_{l=1}^{L_p} \alpha_{pl} B_l(t_{ij})$ for $p = 0, 1, \ldots P$.
- ‡Smoothness parameters to penalize the roughness of effect curves

$$\mathcal{L}^{\text{Smooth}} = \sum_{p=0}^{P} \lambda_p \int \left(\beta_p''(t)\right)^2 dt = \sum_{p=0}^{P} \lambda_p \boldsymbol{\alpha_p^T A_p \alpha_p} = \boldsymbol{\alpha}^T \boldsymbol{A_\lambda \alpha},$$

† R package: `https://github.com/kaiqiong/SOMNiBUS`. ‡ Wahba (1980), Parker and Rice (1985)

# Technical detail 1: E-M algorithm



**Complete joint likelihood**

- [†]Random-effect view of the smoothness penalty: $\alpha \sim MVN(\mathbf{0}, \boldsymbol{A}_\lambda^{-})$
- $l^{\text{complete}}(\boldsymbol{S}; \alpha, \lambda) = l(\boldsymbol{S}; \alpha) - \frac{1}{2}\alpha^T \boldsymbol{A}_\lambda \alpha + \frac{1}{2}\log\{|\boldsymbol{A}_\lambda|_+\}$

**E step:** Calculate $\eta_{ij}^\star = \mathbb{E}(S_{ij} \mid Y_{ijk}\ ; \alpha^\star)$

**M step:** [‡]Maximize $Q(\alpha, \lambda \mid \alpha^\star) = l(\eta^\star; \alpha) - \frac{1}{2}\alpha^T \boldsymbol{A}_\lambda \alpha + \frac{1}{2}\log\{|\boldsymbol{A}_\lambda|_+\}$

- Estimate $\alpha$ given the value of $\lambda$: P-IRLS

$$\widehat{\alpha}_\lambda = \underset{\alpha}{\text{argmax}} \left\{ l(\eta^\star; \alpha) - \frac{1}{2}\alpha^T \boldsymbol{A}_\lambda \alpha \right\}$$

- Estimate $\lambda$: maximize the Laplace-approximated restrictive likelihood

$$L^M(\lambda) = \int \exp\left\{Q(\alpha, \lambda \mid \alpha^\star)\right\} d\alpha \approx \text{Laplace}(\lambda; \widehat{\alpha}_\lambda).$$

[†] Wahba (1983), JRSSB; Silverman (1985), JRSSB. [‡] Wood (2011), JRSSB; R package `mgcv`

# Technical detail 2: Inference



- Conditional on the values of smoothing parameter $\boldsymbol{\lambda}$
- Estimate the variance of EM estimator $\widehat{\alpha}$, $\boldsymbol{V}$, using the observed Fisher information[†]
- Hypothesis testing for a regional zero effect $H_0 : \beta_p(t) = 0$.
  - Wald-type statistic

$$T_p = \widehat{\boldsymbol{\alpha}_p}^T \{\boldsymbol{V_p}\}^{-1} \widehat{\boldsymbol{\alpha}_p} \sim \chi^2_{\tau_p}$$

  - Penalization affects effective degree of freedom[‡]; $\tau_p < L_p = \dim(\boldsymbol{\alpha_p})$

$$\tau_p = \sum_{l=a_p}^{b_p} (2\boldsymbol{F} - \boldsymbol{FF})_{(l,l)}\,, \text{ for } p = 0, 1, \ldots P,$$

  - $\boldsymbol{F}$ is the 'hat' matrix and has the form $\boldsymbol{F} = (\mathbb{X}^T \widehat{\boldsymbol{W}} \mathbb{X} + \boldsymbol{A}_{\widehat{\boldsymbol{\lambda}}})^{-1} \mathbb{X}^T \widehat{\boldsymbol{W}} \mathbb{X}$

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

[‡] Wood, S.N. (2013) On p-values for smooth components of an extended generalized additive model. Biometrika

# dSOMNiBUS: Estimation



- Random-effect view of the smoothness penalty: $\boldsymbol{\alpha} \sim MVN(\mathbf{0}, \boldsymbol{A}_{\boldsymbol{\lambda}}{}^{-})$
- conditional mean parameters (REs): $\boldsymbol{\mathcal{B}} = (\boldsymbol{\alpha}, \boldsymbol{u}) \in \mathbb{R}^{N+\sum_0^P L_p}$
- variance component parameters: $\boldsymbol{\Theta} = (\boldsymbol{\lambda}, \sigma_0^2) \in \mathbb{R}^{P+2}$
- **multiplicative dispersion parameter**: $\phi$

**Complete joint log-quasi-likelihood function**

$$q\ell^{(\boldsymbol{S},\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}}, \phi, \boldsymbol{\Theta}) = ql^{(\boldsymbol{S}|\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}}, \phi) \underbrace{-\frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{A}_{\boldsymbol{\lambda}} \boldsymbol{\alpha} - \frac{1}{2\sigma_0^2}\boldsymbol{u}^T \boldsymbol{u}}_{-\frac{1}{2\phi}\boldsymbol{\mathcal{B}}^T \boldsymbol{\Sigma}_{\boldsymbol{\Theta}} \boldsymbol{\mathcal{B}}} + \underbrace{\frac{1}{2}\log\{|\boldsymbol{A}_{\boldsymbol{\lambda}}|_+\} + \frac{N}{2}\log\left(1/\sigma_0^2\right)}_{1/2\log\{|\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}/\phi|_+\}}$$

# dSOMNiBUS: Estimation



## Conditional quasi-likelihood function

$$qL^{(S|\mathcal{B})}(\mathcal{B}, \phi) \propto \exp\left\{ -\frac{1}{2\phi} \sum_{i,j} d_{ij}\left(S_{ij}, \pi_{ij}\right) - \frac{M}{2} \log \phi \right\},$$

- $d_{ij}(S_{ij}, \pi_{ij}) = -2 \int_{S_{ij}/X_{ij}}^{\pi_{ij}} \frac{S_{ij} - X_{ij}\pi_{ij}}{\pi_{ij}(1-\pi_{ij})} d\pi_{ij}$ is the quasi-deviance function
- This is the extended quasi-likelihood for the joint parameter $(\mathcal{B}, \phi)$
- It exhibits the properties of log-likelihood, with respect to both $\mathcal{B}$ (exact) and $\phi$ (approximate)
- †The assumptions required are that $\phi$ be small and that $\kappa_r = O(\phi^{r-1})$

† Efron (1986), Jorgensen (1987), McCullagh and Nelder (1989)

# dSOMNiBUS: Estimation



- **Marginal quasi-likelihood function**

$$qL^M(\phi, \Theta) = \int \exp\left\{q\ell^{(S,\mathcal{B})}(\mathcal{B}, \phi, \Theta)\right\} d\mathcal{B} \approx \text{Laplace}(\phi, \Theta; \widehat{\mathcal{B}}) \neq f(\phi)g(\Theta).$$

- **A similar E-M algorithm**

Initialize $\Theta^{(0)}, \phi^{(0)}, \mathcal{B}^{(0)}$ (estimates ignoring errors); Choose $\varepsilon = 10^{-6}$; Set $\ell = 0$;
**repeat**
- E step: $\eta_{ij}^{(\ell)} = \mathbb{E}(S_{ij} \mid Y_{ij}; \mathcal{B}^{(\ell)})$;
- M step: $(\mathcal{B}^{(\ell)}, \phi^{(\ell)}, \Theta^{(\ell)}) = \text{argmax}_{\mathcal{B},\phi,\Theta}\, \ell^{\text{Joint}}\left(\mathcal{B}, \phi, \Theta; \eta_{ij}^{(\ell)}\right)$. Specifically
  **repeat**
  - Solve $\boldsymbol{U}(\mathcal{B}; \Theta^{(s)}) = \mathbf{0}$ to obtain $\mathcal{B}^{(s)}$ using data $\eta_{ij}^{(\ell)}$;
  - Newton's update for the Laplace approximated marginal likelihood evaluated at data $\eta_{ij}^{(\ell)}$ :

$$(\phi, \Theta)^{(s+1)} = (\phi, \Theta)^{(s)} - \left[\nabla^2 \text{Laplace}(\mathcal{B}^{(s)})\right]^{-1} \nabla \text{Laplace}(\mathcal{B}^{(s)});$$

  $s \leftarrow s + 1$;
  **until** $\|\mathcal{B}^{(s)} - \mathcal{B}^{(s-1)}\|_2 < \varepsilon$;
  $\ell \leftarrow \ell + 1$;
**until** $\|\mathcal{B}^{(\ell)} - \mathcal{B}^{(\ell-1)}\|_2 < \varepsilon$;
Return $\Theta^{(\ell)}, \mathcal{B}^{(\ell)}, \phi^{(\ell)}$ ;

- **Estimating $\phi$**
  - Likelihood-based estimator
  - Moment-based estimator (better)

# Inference for smooth covariate effects



- Estimate the variance of EM estimator $\widehat{\alpha}$, $\boldsymbol{V}$, using the observed (quasi-)Fisher information[†]
- Hypothesis testing for a regional zero effect $H_0 : \beta_p(t) = 0$.
  - Regional statistic

$$T_p = \frac{\widehat{\alpha}_p^T \left\{ \widehat{\boldsymbol{V}}_{\boldsymbol{p}} \right\}^{-1} \widehat{\alpha}_p}{\tau_p} \sim F_{\tau_p, M-\tau}$$

  - $\tau_p$ : EDF for smooth term $\beta_p(t)$. $\tau$: total EDF of the model
  - This F null distribution relies on the assumption that $(M - \tau)\widehat{\phi}/\phi \sim \chi^2_{M-\tau}$, which is approximately true for moment-based dispersion estimator

[†] Elashoff and Ryan (2004) An EM algorithm for estimating equations. Journal of Computational and Graphical Statistics