DNA Methylation Hierarchical Variance Model Stan Results

Kevin Murgas

2020-08-12

Contents

1 Model Overview

1.1 Data

Logit-transformed beta values, from EPIC methylation array pre-processed with minfi Noob algorithm.

Sample counts (total): 48 samples from 21 patients.
By tissue: 6 normal samples, 42 paired bulk tumor samples.

1.2 Model equation

Let i=sample, j=patient

$$y_i = \mu + a_j + \text{tInd}(i) * (\beta_T + b_j + c_i)$$

$$a_j \sim N(0, \sigma_P), b_j \sim N(0, \sigma_{PT}), c_i \sim N(0, \sigma_T)$$

Model is fit at each individual CpG site using RStan's MCMC sampling algorithm.
Stan Parameters: 4 chains x 2000 iterations (200 warmup) = 7200 total draws. adapt_delta = 0.999

1.3 Prior Probabilities

Priors for each parameter (mu, betaT, sigmaP/PT/T) were defined using TCGA 450k array data from colorectal tumor patients with paired multiple tumor and/or normal+tumor samples.
Empirical values of mean and standard deviations were assesed for all 450k CpG sites.
The resulting distributions were fit with: bimodal gaussian (mu), Cauchy (betaT), or gamma distributions (sigmas).
SigmaE (error variance) prior was defined using a non-informative gamma distribution
Additionally: Prior relaxation was performed for the Cauchy and gamma distribtions by algebraically increasing the variance of the distributions while maintaining the same mode (peak)

1.4 Conservation score

We want to understand which sites of DNA methylation are fundamentally conserved within tumors, i.e. lower variation in methlyation within tumor (sigmaT) relative to the variation within normal healthy tissue (sigmaP). Therefore we choose the log-ratio of sigmaP/sigmaT to represent a conservation score, which is positive when tumor variation is lower than normal variation. We take the median of the posterior distribution for this parameter. Additionally, we calculate a regularization term based on the inverse of the L2 norm of sigmaP and sigmaT.

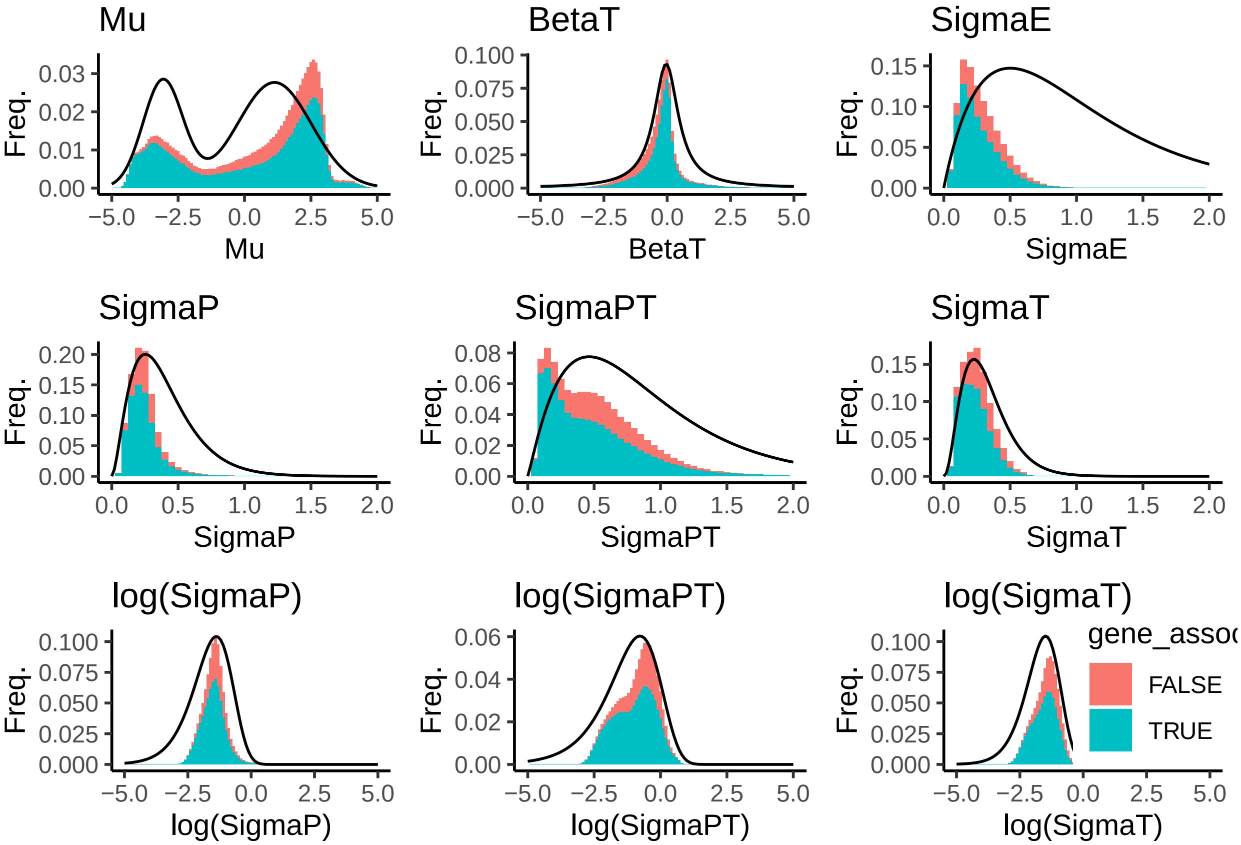
$$\text{score}_{\text{med}} = \text{Median}(\log \frac{\sigma_P}{\sigma_T})$$

2 Fit Results

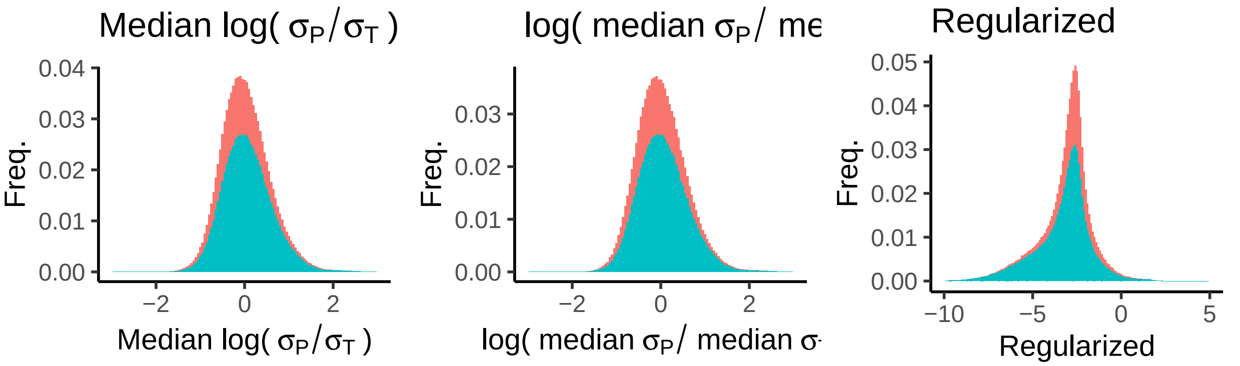
2.1 Parameter Distributions

Histograms are median values of posterior distribution

```
N sites: total=866091, gene-associated=616747, non-gene=249344
```



2.2 Conservation scores



2.3 Summary of neff and Rhat for each parameter

Parameter	Mean n_eff	Min n_eff	# <500	% <500	# <100	% <100
mu	3303.43	2.36	9370	1.08	1476	0.17
betaT	3168.31	2.95	6470	0.75	1457	0.17
sigmaP	726.65	2.45	293358	33.87	3022	0.35
sigmaPT	1568.36	2.42	108245	12.50	2142	0.25
sigmaT	241.96	2.50	859118	99.19	35044	4.05
sigmaE	356.47	2.37	707670	81.71	60477	6.98
lp	207.29	2.08	864363	99.80	89816	10.37

Parameter	Mean Rhat	Max Rhat	# >1.1	% >1.1	# >1.01	% >1.01
mu	1.00	2.44	373	0.04	7891	0.91
betaT	1.00	1.66	373	0.04	5684	0.66
sigmaP	1.01	2.53	594	0.07	176300	20.36
sigmaPT	1.00	2.40	574	0.07	76270	8.81
sigmaT	1.02	2.15	1810	0.21	642831	74.22
sigmaE	1.02	2.43	5411	0.62	537758	62.09
lp	1.03	4.82	8803	1.02	738015	85.21

For the remainder of these results, a filter is applied to remove all sites with Rhat > 1.1 in lp (log-posterior likelihood of entire site fit). Another filter is later applied to remove all genes with <5 CpG sites.

Prep data

```
Param:prep=TRUE -> prepare data...

Prep data with annotation, filter out sites Rhat>1.1 (and save)

6.73 0.827 8.018 0 0

Separate by gene

17.732 1.043 19.118 0 0

Summarize by gene, taking mean value of all sites in gene, filter out sites n<5

2.891 0.065 2.963 0 0

Permutate data with bootstrapping

1814.079 19.683 1839.227 0 0

Calculate mean methylation conservation in promoters

Warning: Missing column names filled in: 'X48' [48]

Parsed with column specification:
cols(
  .default = col_character(),
  Genome_Build = col_double(),
  MAPINFO = col_double(),
  `450k_Enhancer` = col_logical(),
  DNase_Hypersensitivity_Evidence_Count = col_double(),
  OpenChromatin_Evidence_Count = col_double(),
  TFBS_Evidence_Count = col_double(),
  Methyl27_Loci = col_logical(),
  Methyl450_Loci = col_logical(),
  Coordinate_36 = col_double(),
  Random_Loci = col_logical()
)

See spec(...) for full column specifications.

Warning: 866669 parsing failures.
row col   expected     actual
  1  -- 48 columns 47 columns '/Users/kevinmurgas/Documents/Data+ project/EPIC data/data-raw/Methylation
  2  -- 48 columns 47 columns '/Users/kevinmurgas/Documents/Data+ project/EPIC data/data-raw/Methylation
  3  -- 48 columns 47 columns '/Users/kevinmurgas/Documents/Data+ project/EPIC data/data-raw/Methylation
  4  -- 48 columns 47 columns '/Users/kevinmurgas/Documents/Data+ project/EPIC data/data-raw/Methylation
  5  -- 48 columns 47 columns '/Users/kevinmurgas/Documents/Data+ project/EPIC data/data-raw/Methylation
... ... ......... ......... ..........................................................................
See problems(...) for more details.

`summarise()` ungrouping output (override with `.groups` argument)

18.629 1.372 22.316 0 0

Done with data preparation (BOTH)
```