DNA Methylation Hierarchical Variance Model Paper Figures

Kevin Murgas

2020-08-13

Contents


```
## Load prepped data dir: data-stan_rg3ps3 (needs to be pre-processed)...
```

1 Figure 1: Model Overview, Priors, Example Site Fit

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

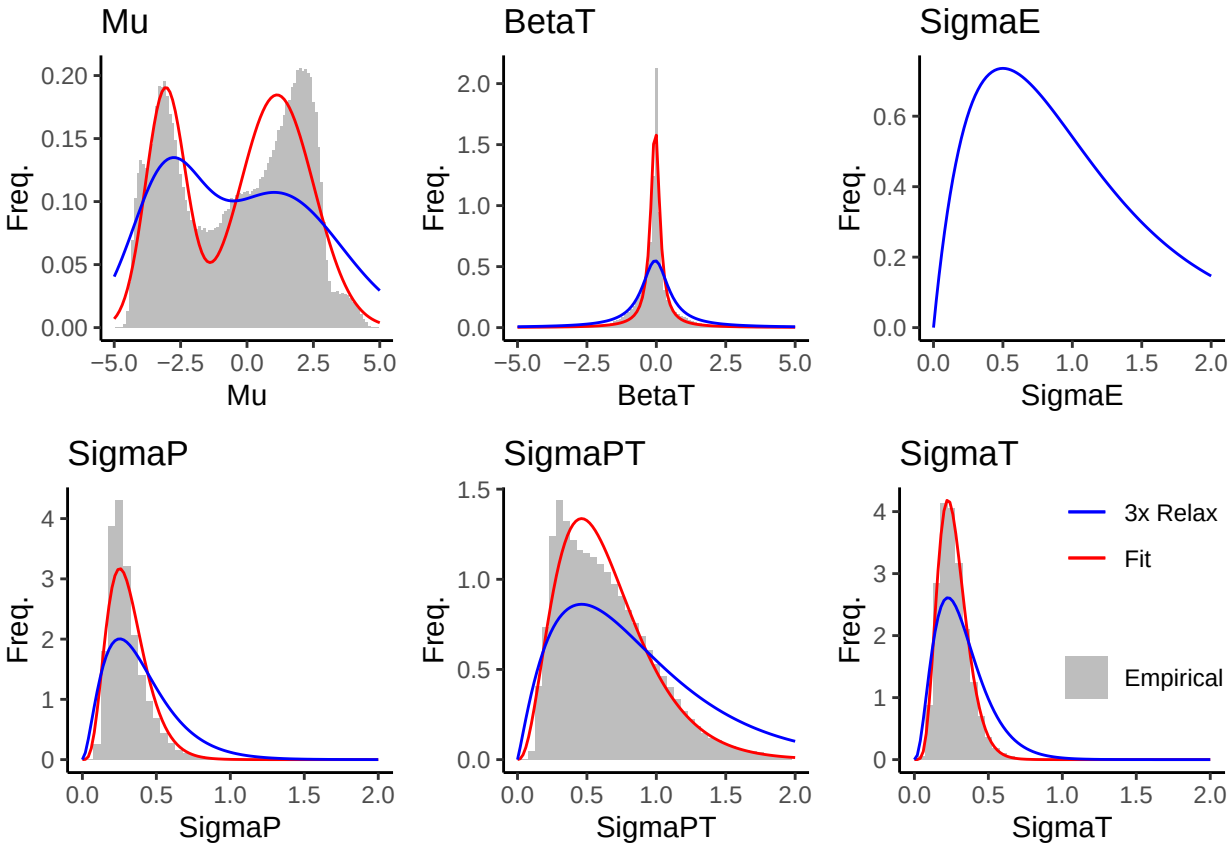
$$y_i = \mu + a_j + \text{tInd}(i) * (\beta_T + b_j + c_i)$$

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

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)

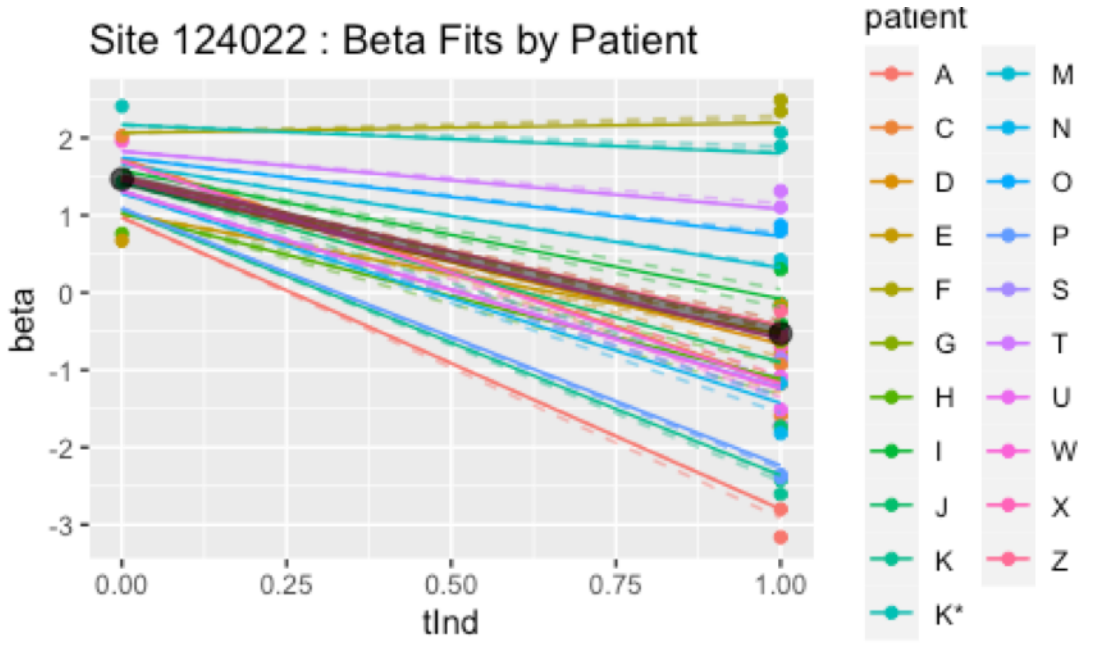


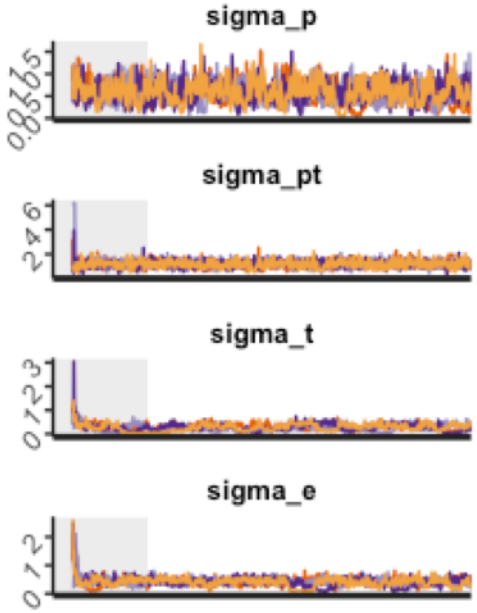
1.4 CpG Conservation score

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.

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

1.5 Example site fit



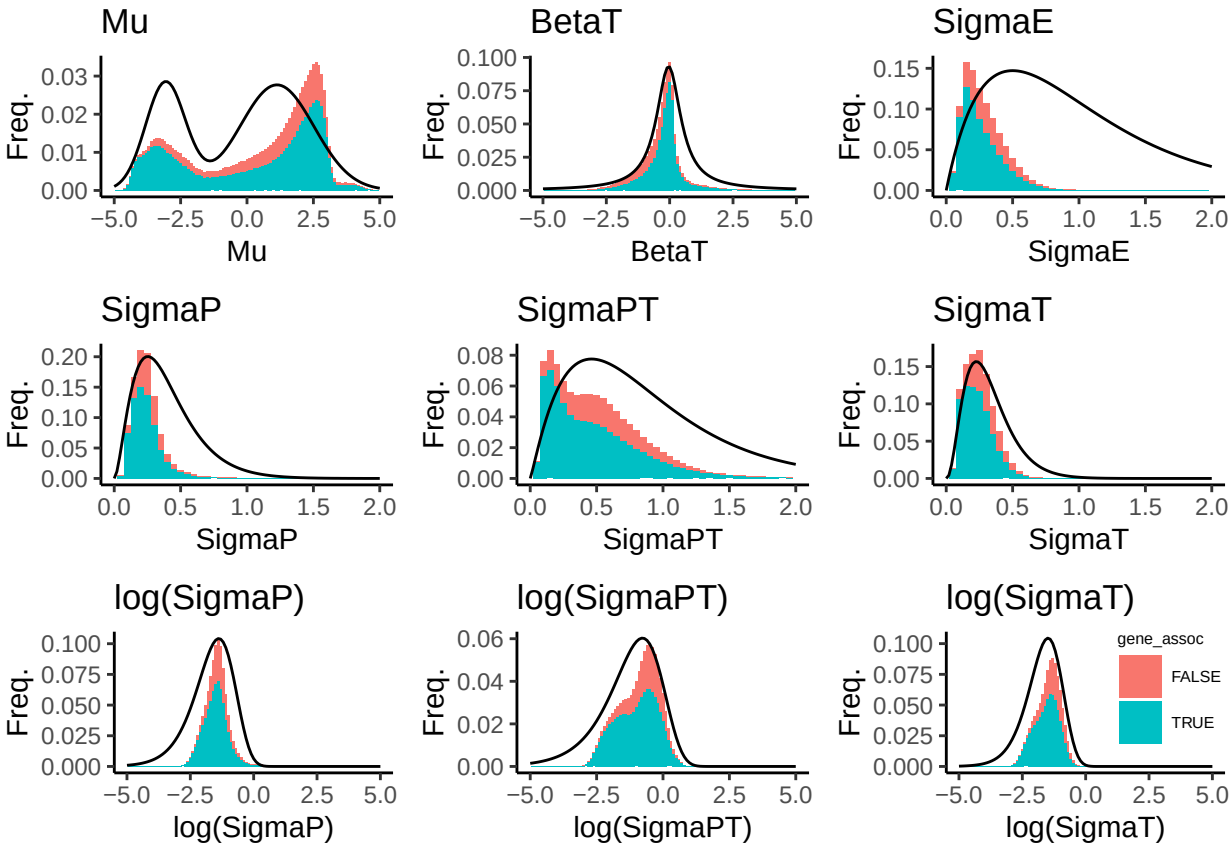


2 Figure 2: Model Fit Results

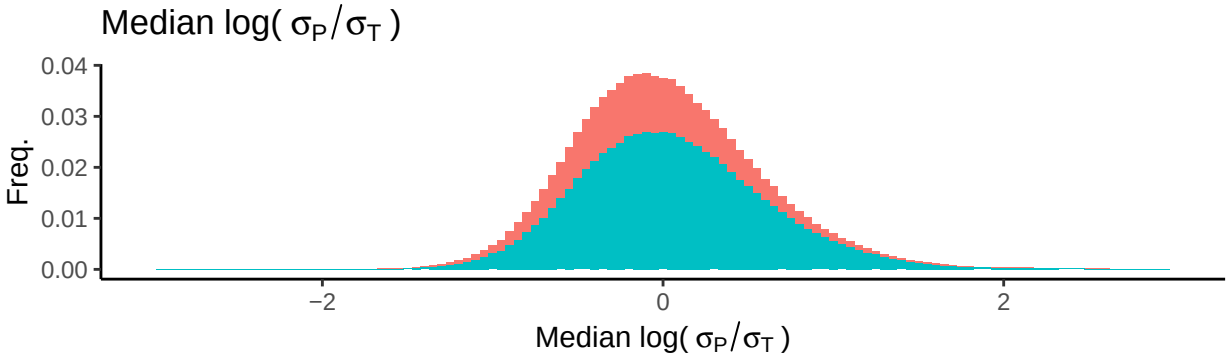
2.1 Parameter Distributions

Histograms are median values of posterior distribution

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



2.2 Conservation score



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.

2.4 Intra-gene conservation

We examine intra gene conservation 2 ways: a table looking at gene regulatory regions, and a CpG-distance analysis

2.5 Regulatory regions

Table comparing different groups of CpG sites:

	Hierarchical Variances			
Island relation	sigmaP	sigmaPT	sigmaT	CpG Conservation Score
Island	0.257	0.474	0.191	0.284
North Shore	0.252	0.495	0.236	0.049
South Shore	0.251	0.488	0.234	0.052
North Shelf	0.253	0.489	0.250	0.000
South Shelf	0.253	0.492	0.252	−0.005
Sea	0.258	0.531	0.273	−0.057

region	cons_in	cons_out	nsites	pval
Gene-Associated	0.064	−0.057	610691	0.000
TSS1500	0.086	0.020	125463	0.000
TSS200	0.281	0.004	79745	0.000
5'UTR	0.140	0.013	110073	0.000
1stExon	0.264	0.016	46906	0.000
Body	0.013	0.041	357172	0.000
3'UTR	−0.000	0.030	18339	0.000

2.6 Figure 1: CpG Distance Analysis

Corrected for gene direction, based on Ensembl gene database

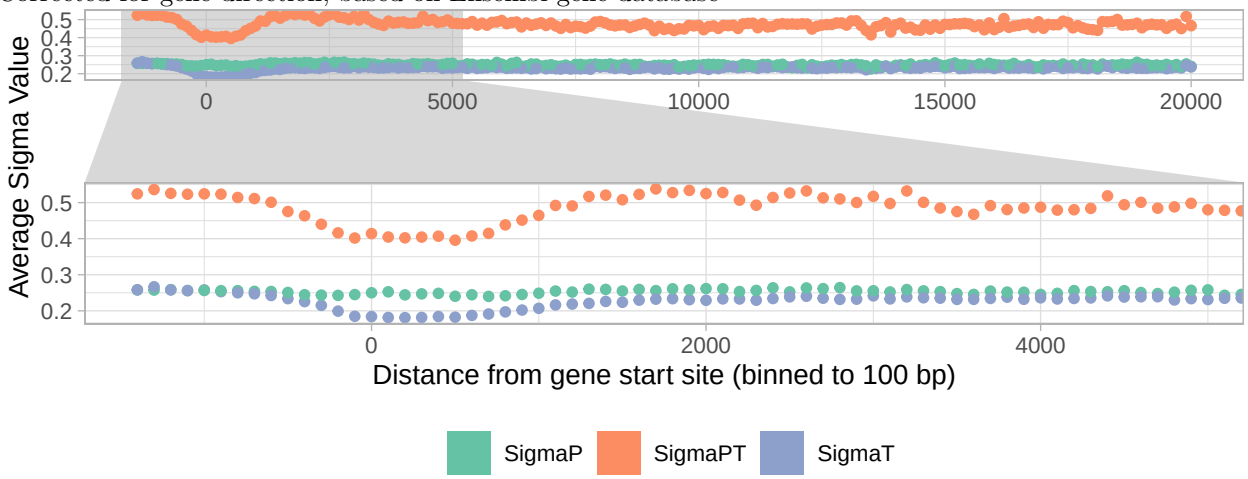


Figure 1: Average Sigma values of single CpGs as a function of position relative to gene.

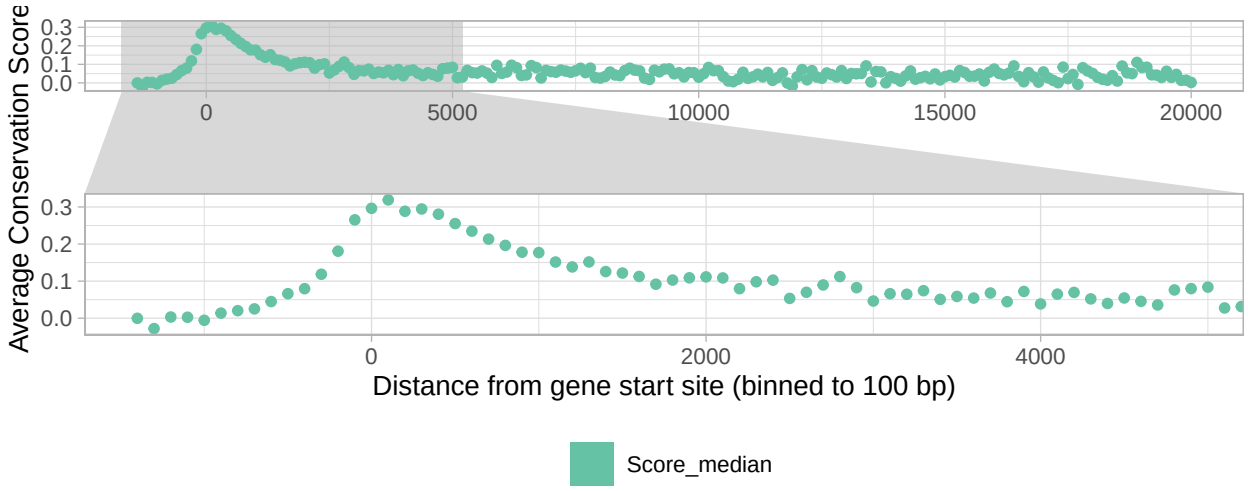


Figure 1: Average conservation scores of single CpGs as a function of position relative to gene.

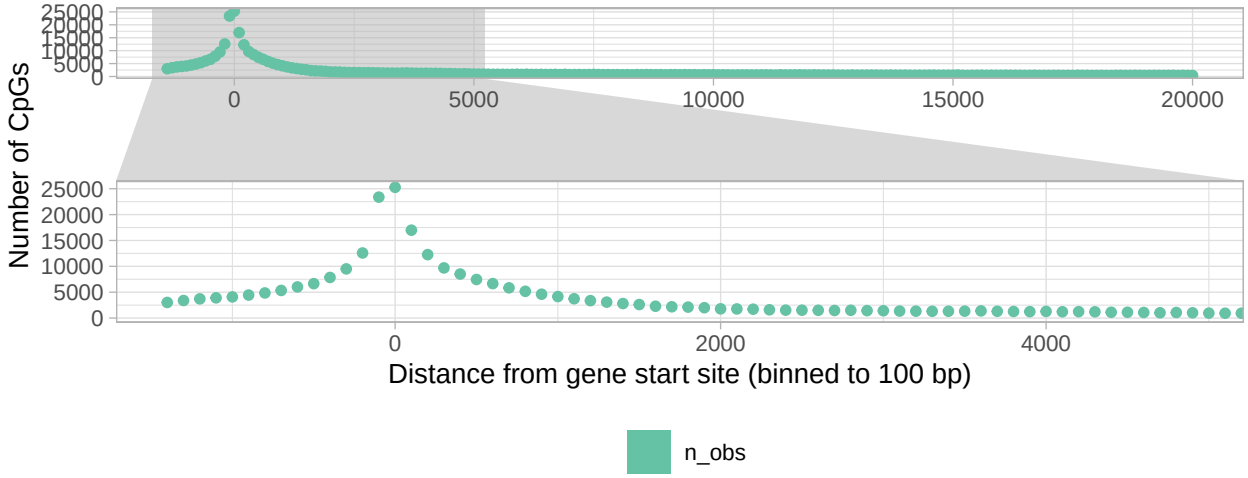
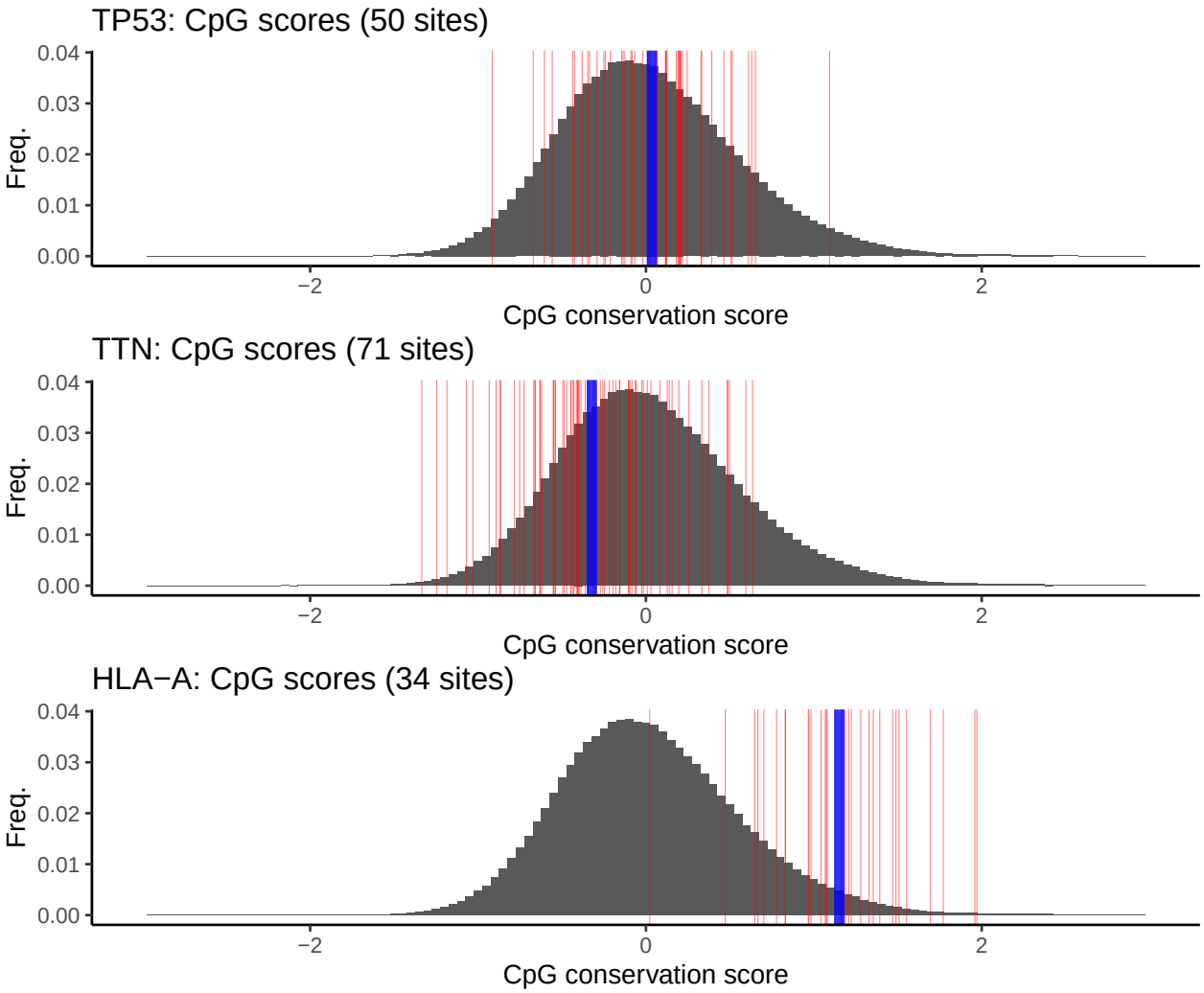


Figure 1: Frequency of single CpGs as a function of position relative to gene start site.

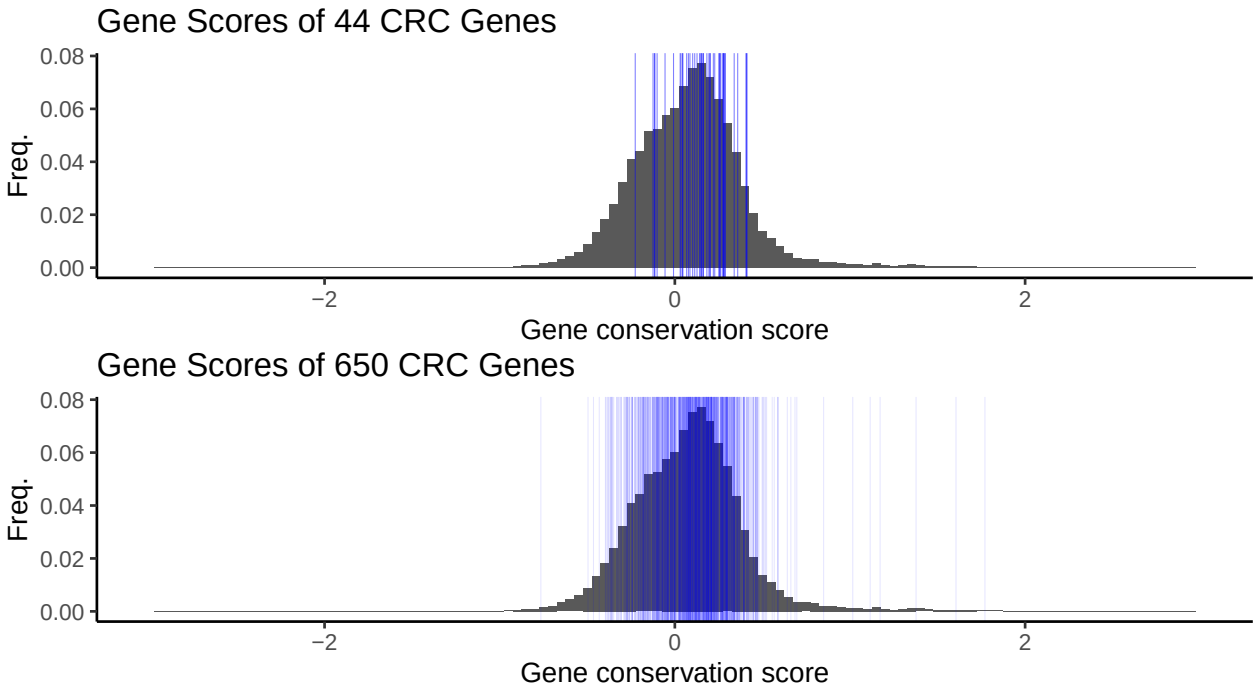
3 Gene Conservation Scores

Genes are scored by average of all CpG sites within gene

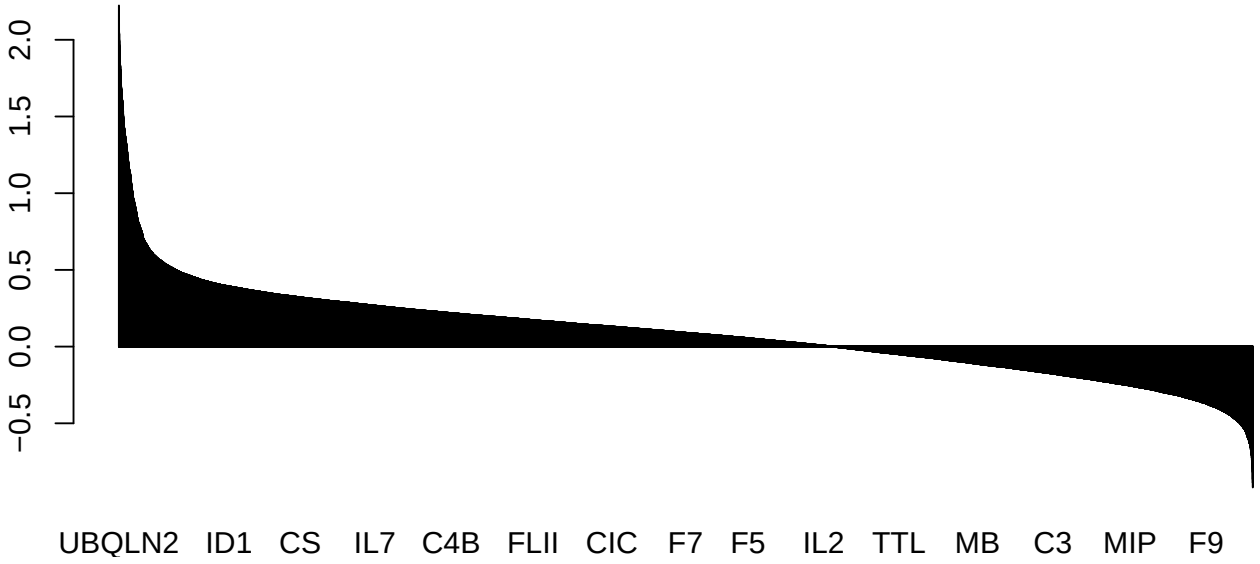
3.1 Example Genes: APC, TP53, TTN, HLA-A, GAPDH



3.2 COSMIC Cancer Genes



4 Gene Set Conservation Analysis




```
183.398 5.61 189.129 0 0
```

```
Processed 23343 genes, 2232 pathways. Top 50 paths:
```

score	num_gene	GeneSet
0.6561522	10	BIOCARTA_CPSF_PATHWAY
0.6533607	11	REACTOME_ENDOSOMAL_VACUOLAR_PATHWAY
0.6482742	25	REACTOME_CHOLESTEROL_BIOSYNTHESIS
0.6419905	10	REACTOME_FOLDING_OF_ACTIN_BY_CCT_TRIC
0.6380963	25	REACTOME_ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND_PEPTIDE_LO
0.6278815	11	BIOCARTA_RANBP2_PATHWAY
0.5967514	11	REACTOME_ESTABLISHMENT_OF_SISTER_CHROMATID_COHESION
0.5896483	14	BIOCARTA_SM_PATHWAY
0.5854369	10	BIOCARTA_RANMS_PATHWAY
0.5842369	11	PID_RANBP2_PATHWAY
0.5809904	15	KEGG_TERPENOID_BACKBONE_BIOSYNTHESIS
0.5626221	17	REACTOME_PROTEIN_METHYLATION
0.5587931	16	BIOCARTA_EIF_PATHWAY
0.5546991	13	REACTOME_MITOTIC_TELOPHASE_CYTOKINESIS
0.5531247	94	REACTOME_EUKARYOTIC_TRANSLATION_ELONGATION
0.5515276	88	KEGG_RIBOSOME
0.5498442	14	REACTOME_GLUTAMATE_AND_GLUTAMINE_METABOLISM
0.5485958	11	REACTOME_SLBP_DEPENDENT_PROCESSING_OF_REPLICATION_DEPENDENT_H
0.5469242	13	BIOCARTA_IL5_PATHWAY
0.5455714	14	BIOCARTA_MHC_PATHWAY
0.5437614	10	BIOCARTA_DNAFRAGMENT_PATHWAY
0.5410644	28	REACTOME_PROCESSING_OF_CAPPED_INTRONLESS_PRE_MRNA
0.5398819	12	BIOCARTA_ACTINY_PATHWAY
0.5346642	19	REACTOME_PROCESSING_OF_INTRONLESS_PRE_MRNAS
0.5339227	17	KEGG_STEROID_BIOSYNTHESIS
0.5334301	102	REACTOME_RESPONSE_OF_EIF2AK4_GCN2_TO_AMINO_ACID_DEFICIENCY
0.5314487	20	BIOCARTA_ARENRF2_PATHWAY
0.5311252	15	REACTOME_PENTOSE_PHOSPHATE_PATHWAY
0.5272805	116	REACTOME_NONSENSE_MEDIATED_DECAY_NMD

0.5254578	113	REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_
0.5239131	10	REACTOME_ATF6_ATF6_ALPHA_ACTIVATES_CHAPERONE_GENES
0.5237324	10	REACTOME_PROCESSING_AND_ACTIVATION_OF_SUMO
0.5226514	15	REACTOME_RESPONSE_OF_EIF2AK1_HRI_TO_HEME_DEFICIENCY
0.5213085	12	REACTOME_ATF6_ATF6_ALPHA_ACTIVATES_CHAPERONES
0.5193469	23	REACTOME_MTORC1_MEDIATED_SIGNALLING
0.5155710	73	REACTOME_DEPOSITION_OF_NEW_CENPA_CONTAINING_NUCLEOSOMES_AT_T
0.5147243	15	REACTOME_GLYCOGEN_BREAKDOWN_GLYCOGENOLYSIS
0.5144657	73	REACTOME_PRC2_METHYLATES_HISTONES_AND_DNA
0.5080633	120	REACTOME_EUKARYOTIC_TRANSLATION_INITIATION
0.5058242	29	REACTOME_G1_S_SPECIFIC_TRANSCRIPTION
0.5032537	11	BIOCARTA_NPC_PATHWAY
0.4994374	54	REACTOME_SNRNP_ASSEMBLY
0.4982944	118	REACTOME_SELENOAMINO_ACID_METABOLISM
0.4975480	33	KEGG_PROPANOATE_METABOLISM
0.4973464	39	REACTOME_FANCONI_ANEMIA_PATHWAY
0.4957187	60	REACTOME_ACTIVATION_OF_THE_MRNA_UPON_BINDING_OF_THE_CAP_BIND
0.4949670	12	REACTOME_PURINE_RIBONUCLEOSIDE_MONOPHOSPHATE_BIOSYNTHESIS
0.4941383	15	BIOCARTA_BCELLSURVIVAL_PATHWAY
0.4930410	13	REACTOME_LOSS_OF_FUNCTION_OF_MECP2_IN_RETT_SYNDROME
0.4926574	13	SA_REG_CASCADE_OF_CYCLIN_EXPR