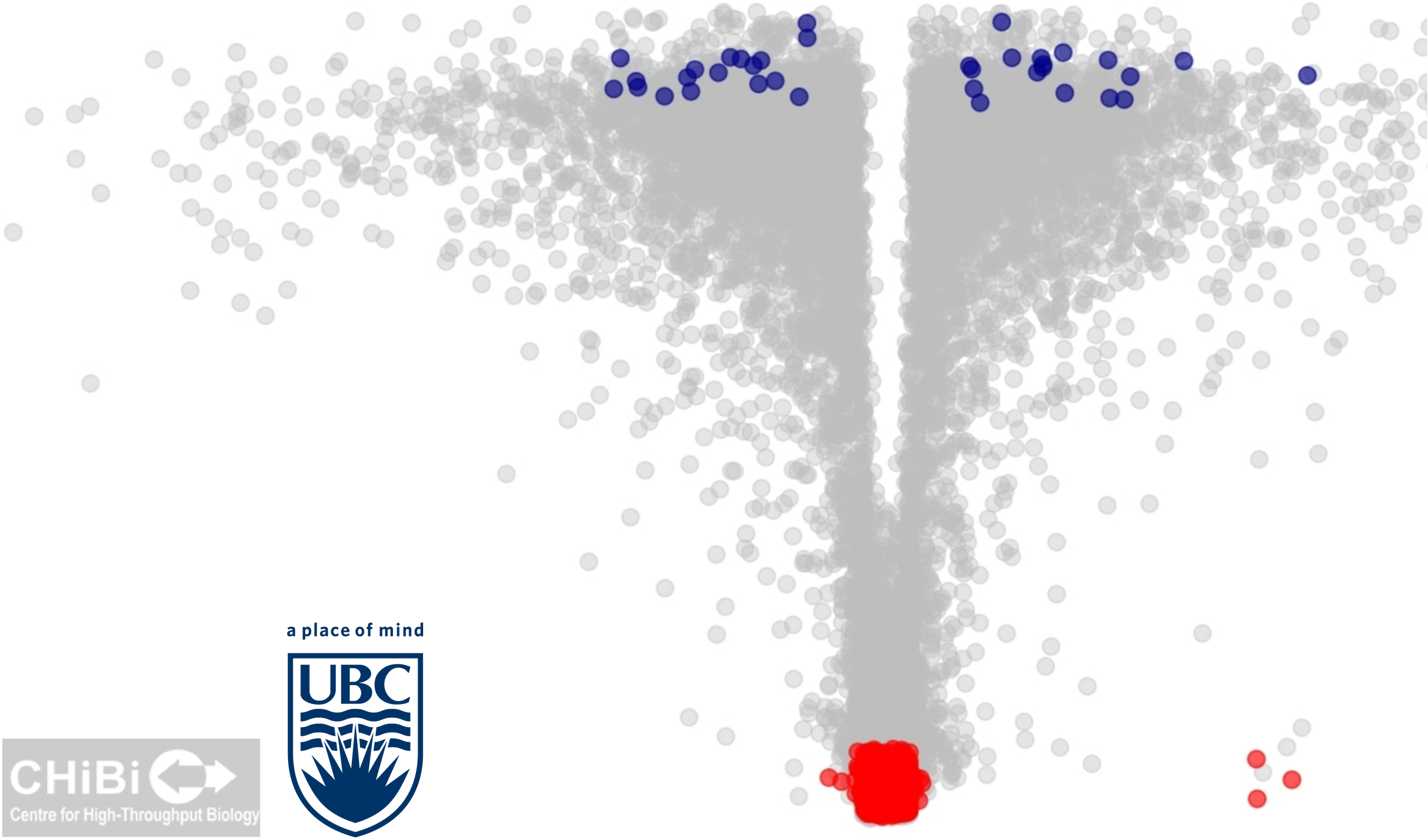
Meta-analysis of Human DNA Methylation Data



Rachel Edgar
GSAT M.Sc. Program, UBC
Bioinfo Journal Club
November 12, 2013

CHiBi
Centre for High-Throughput Biology

a place of mind
UBC

Pavlidis Lab

CHiBi Centre for High-Throughput Biology

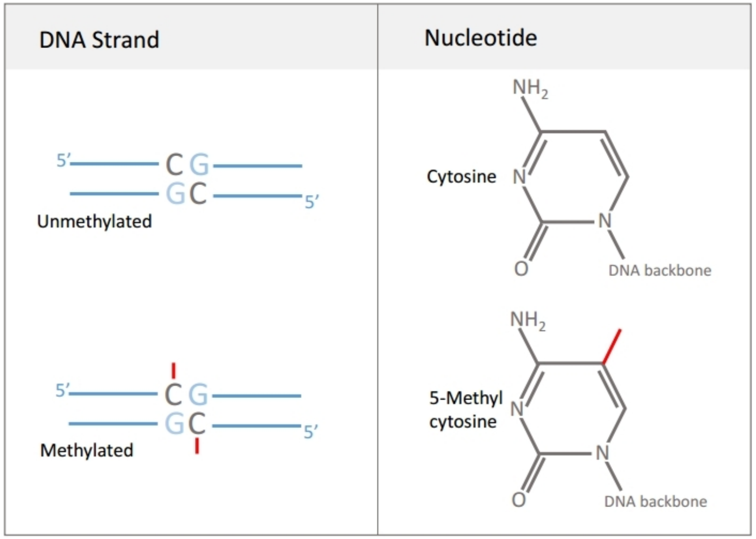
UBC DEPARTMENT of Psychiatry

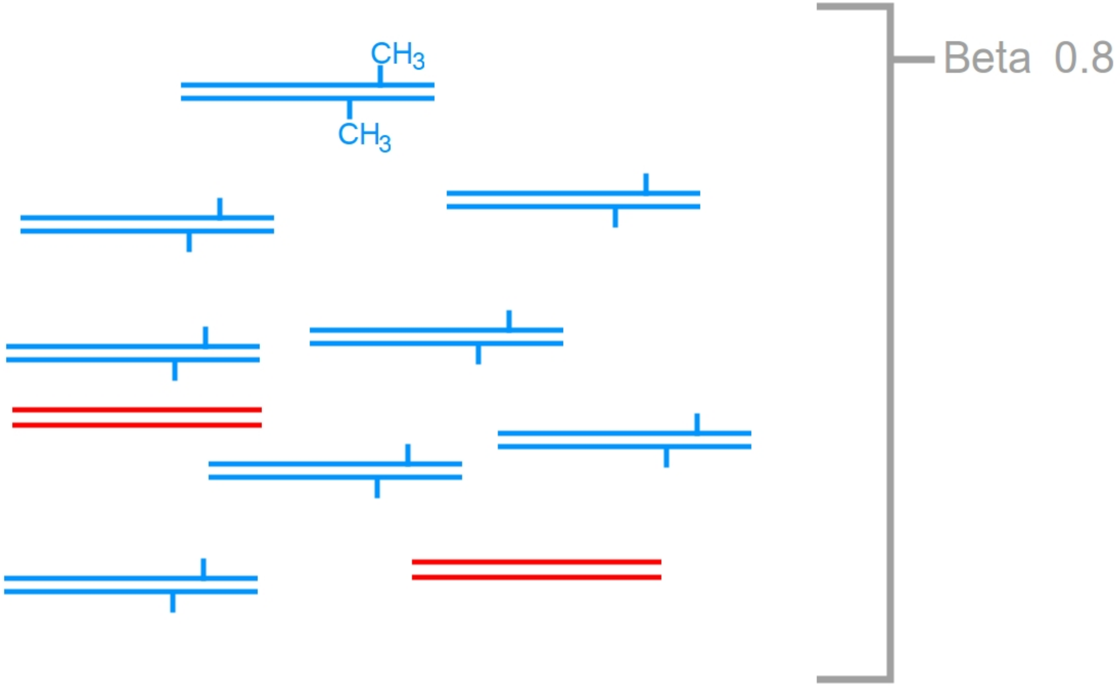
Most frequently used words in abstracts of the lab's papers from the last few years

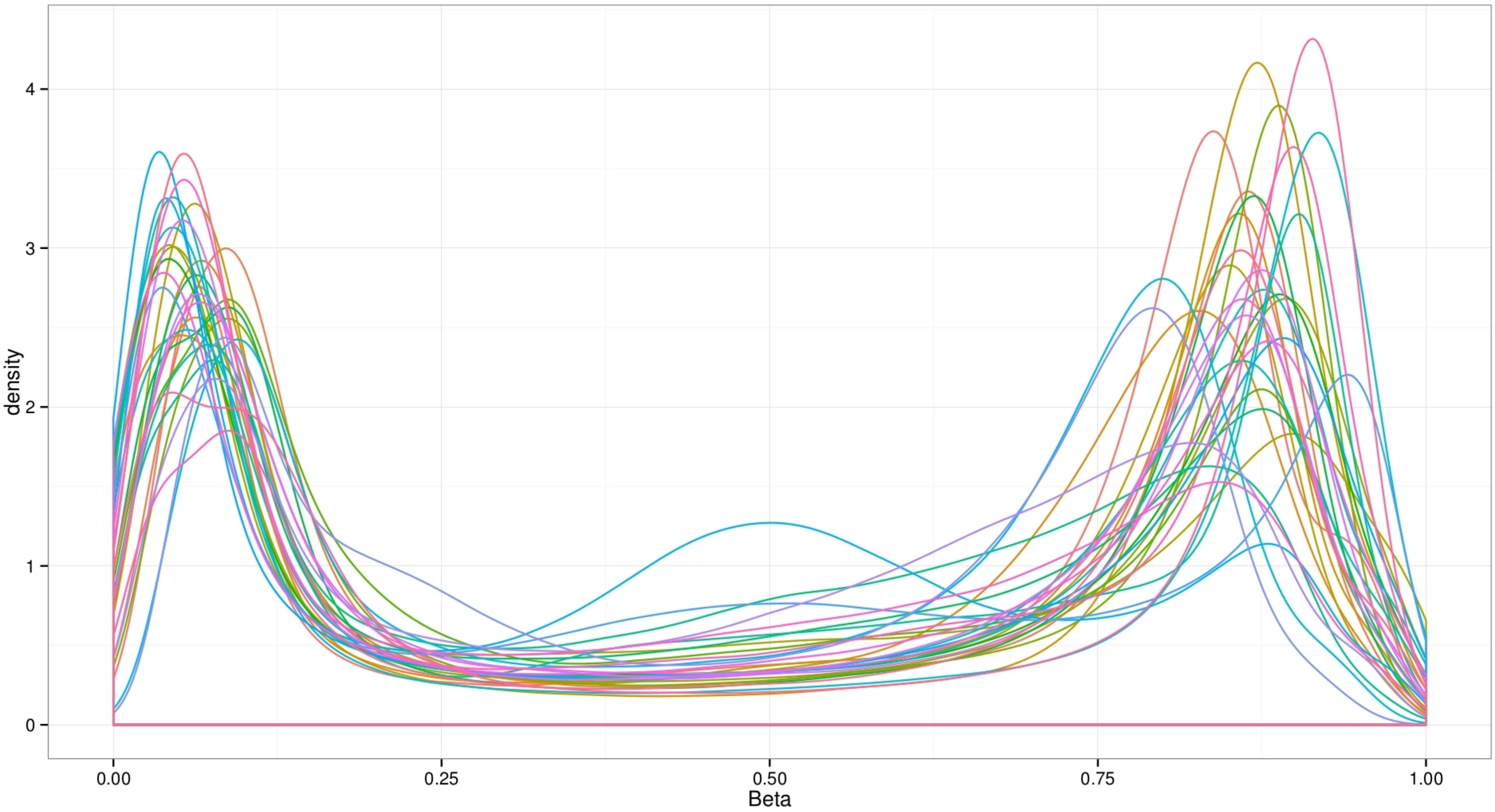
genome science + technology
GRADUATE PROGRAM

DNA Methylation

- Modification of cytosine base in a CG sequence (CpG) with a methyl
- Measured on the Illumina 450K as a beta value

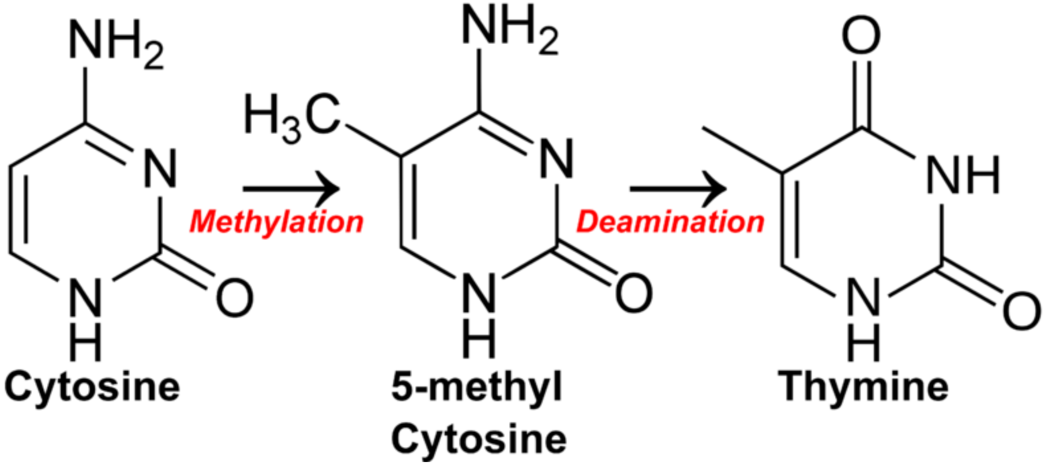





density
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
Beta

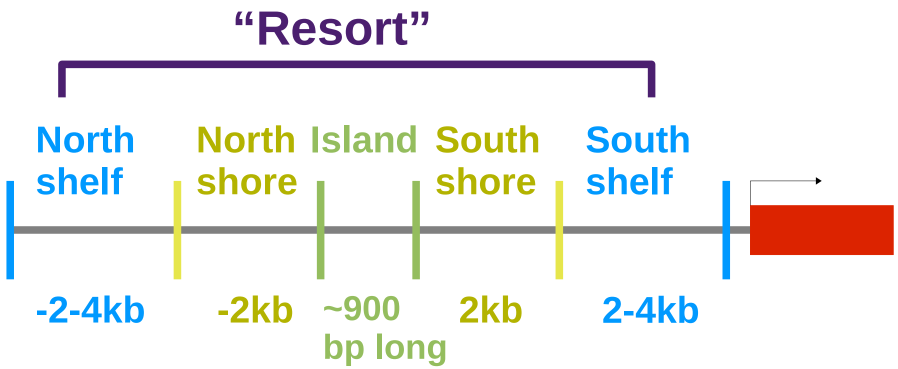
CpG distribution in the genome

- ~28 million CpGs in the human genome
- Expect 132 million CpGs (C 0.21 x G 0.21)
- mCpG -> TpG
- High CpG density does exist in 28,890 CpG islands (CpGI)



CpG Islands Often Promoter Associated

- What is a CpG island (CGI)?
  - GC content>50%
  - O/E CpG >0.6
  - >200 bp long

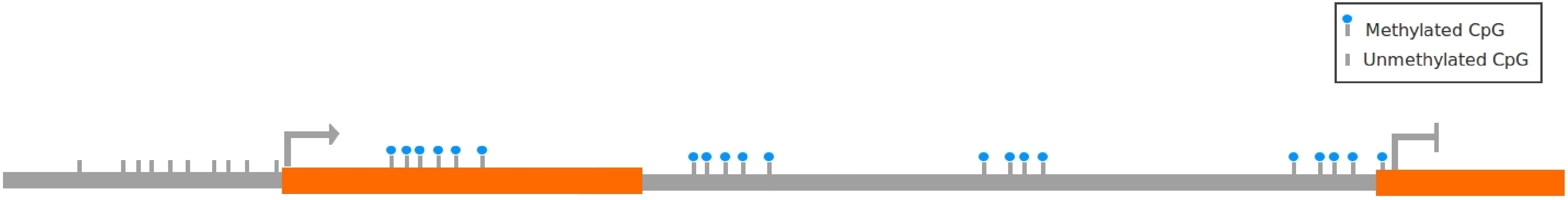


“Resort”

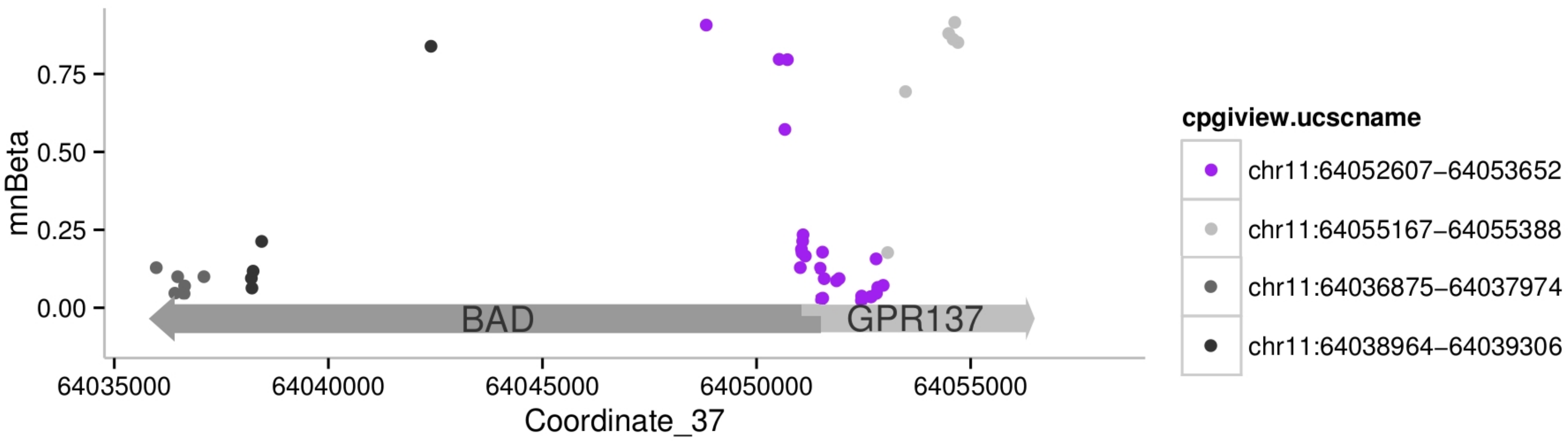
North shelf **North shore** **Island** **South shore** **South shelf**

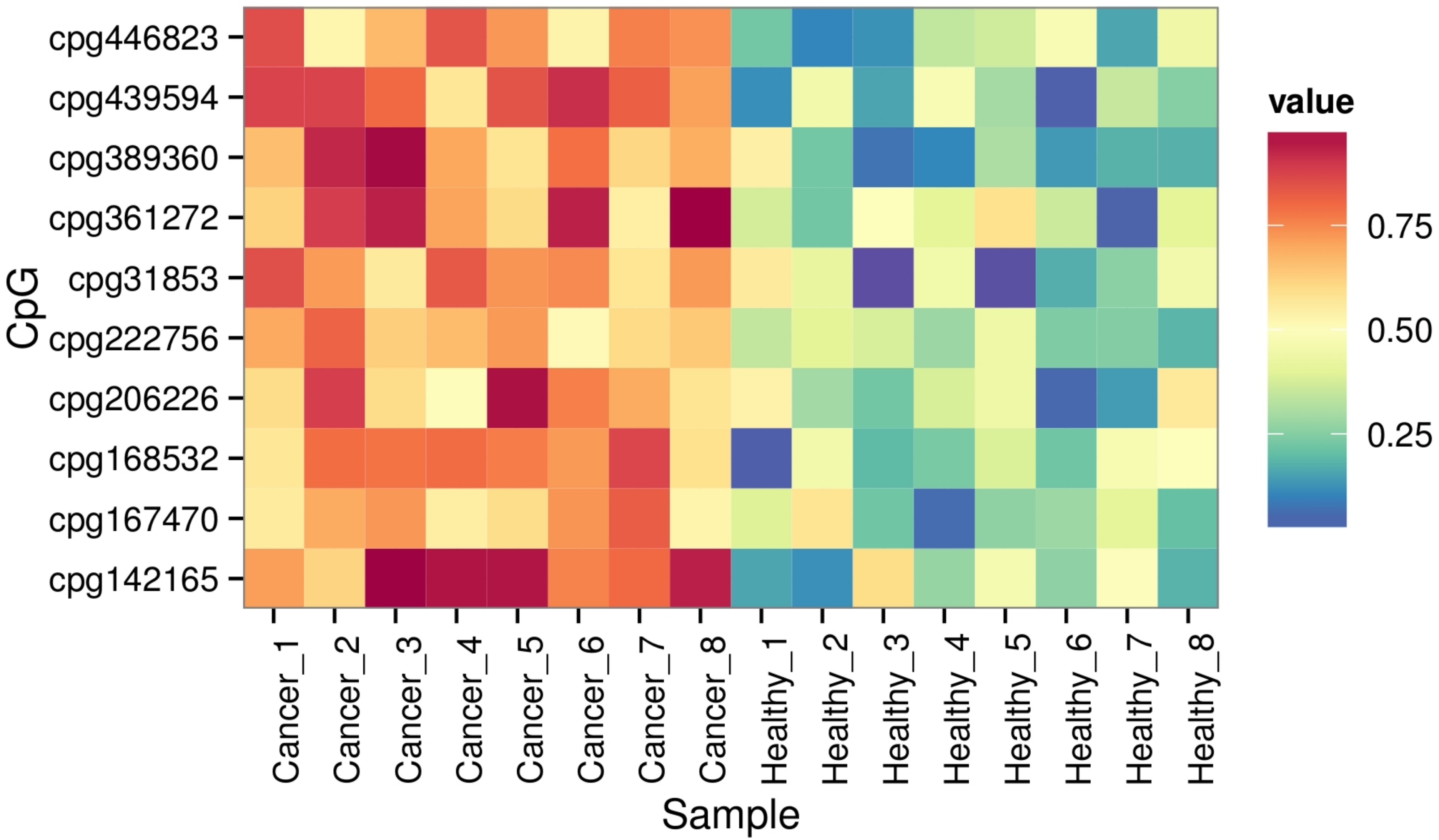
-2-4kb **-2kb** **~900 bp long** **2kb** **2-4kb**

- Classes
  - Promoter (48%)
  - Intragenic (34%)
  - Intergenic (18%)
  - 3' (5%)



● Methylated CpG
| Unmethylated CpG

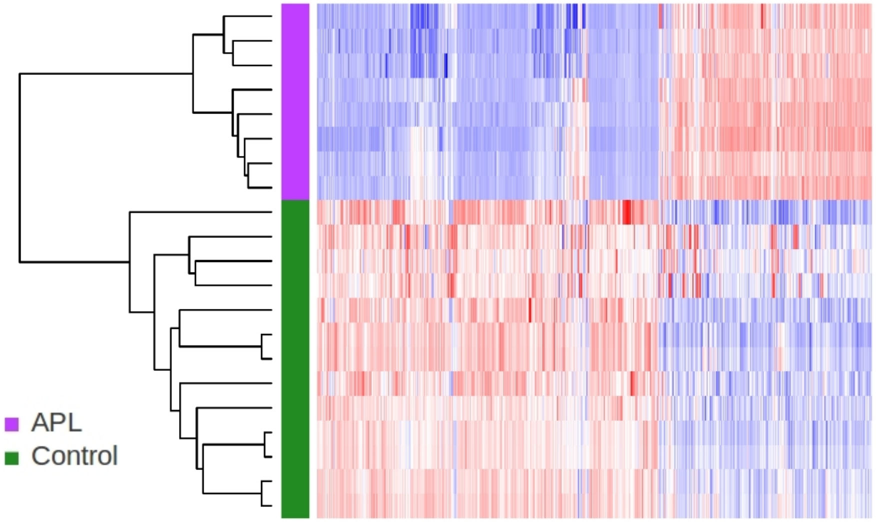

mnBeta
0.75
0.50
0.25
0.00
BAD
GPR137
64035000
64040000
64045000
64050000
64055000
Coordinate_37
cpgiview.ucscname
chr11:64052607−64053652
chr11:64055167−64055388
chr11:64036875−64037974
chr11:64038964−64039306

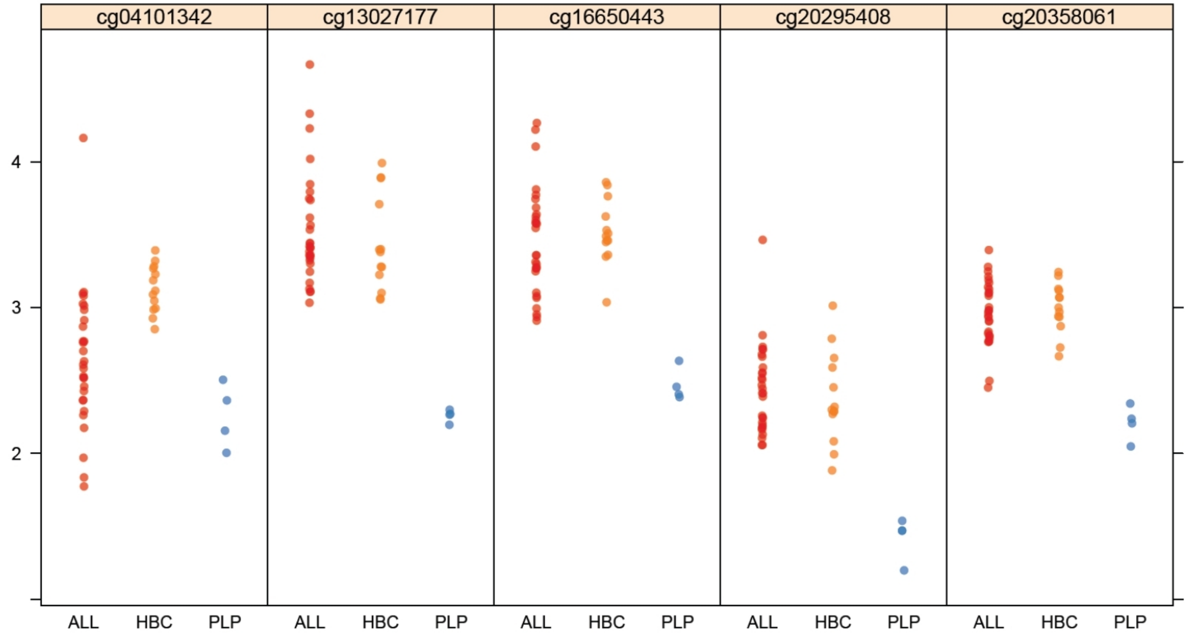

cpg446823
cpg439594
cpg389360
cpg361272
cpg31853
cpg222756
cpg206226
cpg168532
cpg167470
cpg142165
CpG
Cancer_1
Cancer_2
Cancer_3
Cancer_4
Cancer_5
Cancer_6
Cancer_7
Cancer_8
Healthy_1
Healthy_2
Healthy_3
Healthy_4
Healthy_5
Healthy_6
Healthy_7
Healthy_8
Sample
value
0.75
0.50
0.25

But what does Rachel do?

Most CpGs Stable in Individual DNA Methylation Studies

- Differential methylation analysis common method to study DNA methylation
    - i.e which CpG sites have different methylation level between cancer and healthy
- Typically 0.5-20% of CpGs significantly differentially methylated

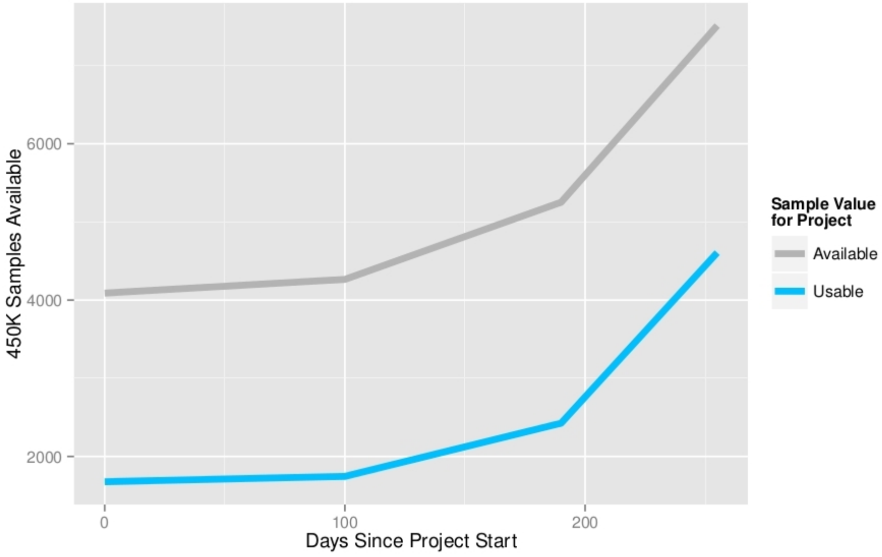




Are there CpGs with a consistent methylation state across many human methylones? If so, what are the particular biological properties of the genomic regions and genes associated with the sites?

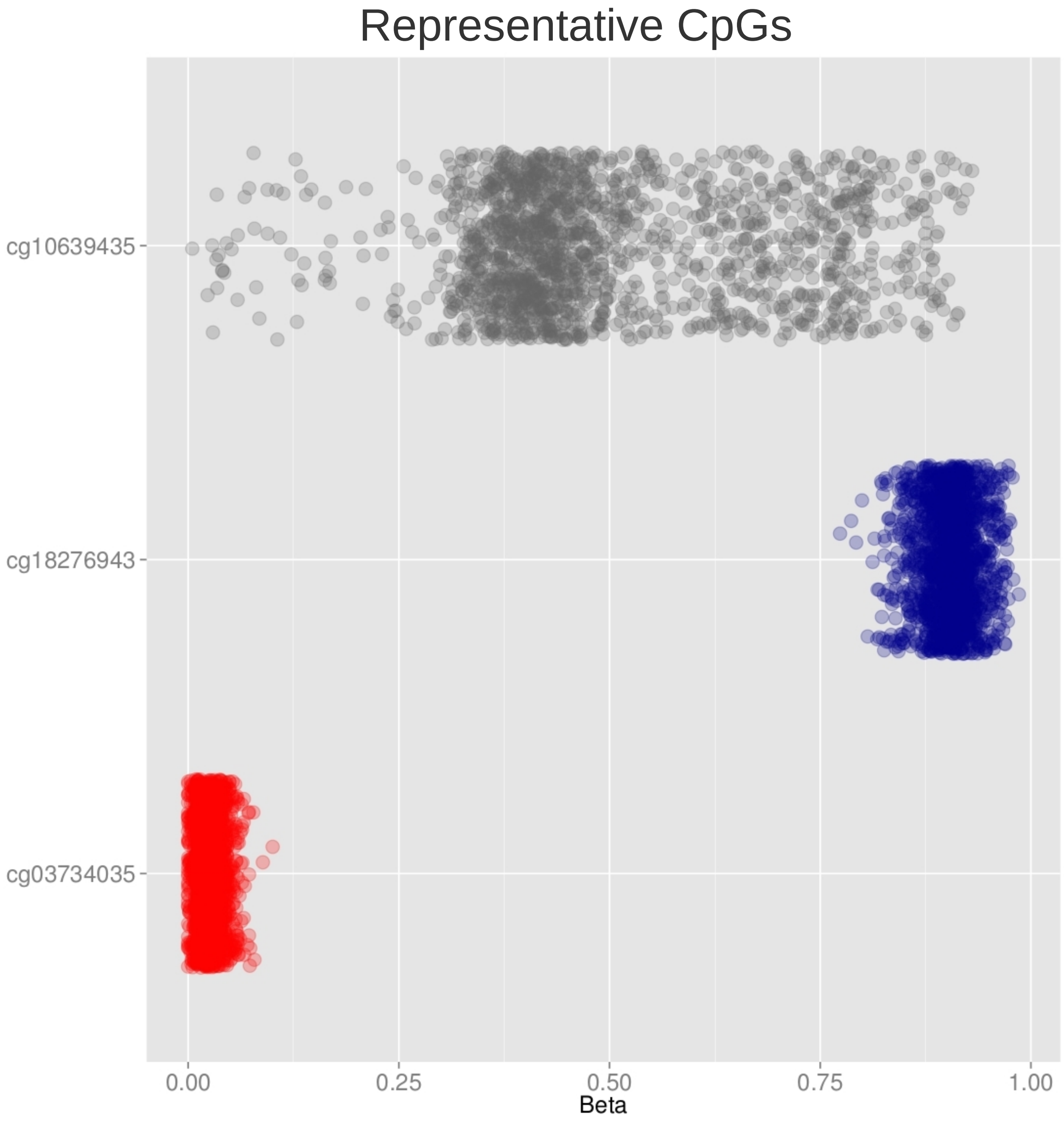
Large Amount Of Data Available

- Large pool of GEO data allows for exploring general trends in methylation data
- 485577 CpGs measured on 450K
- Tissues available and diseases
    - Blood, skin, bone, prostate, breast, lung, brain, muscle, colon, placenta, germ cells, stem cells, and cell line
    - Cancer, arthritis, crohns, Rett, Werner, schizophrenia, Fragile X



Consistent CpG Sites Exist

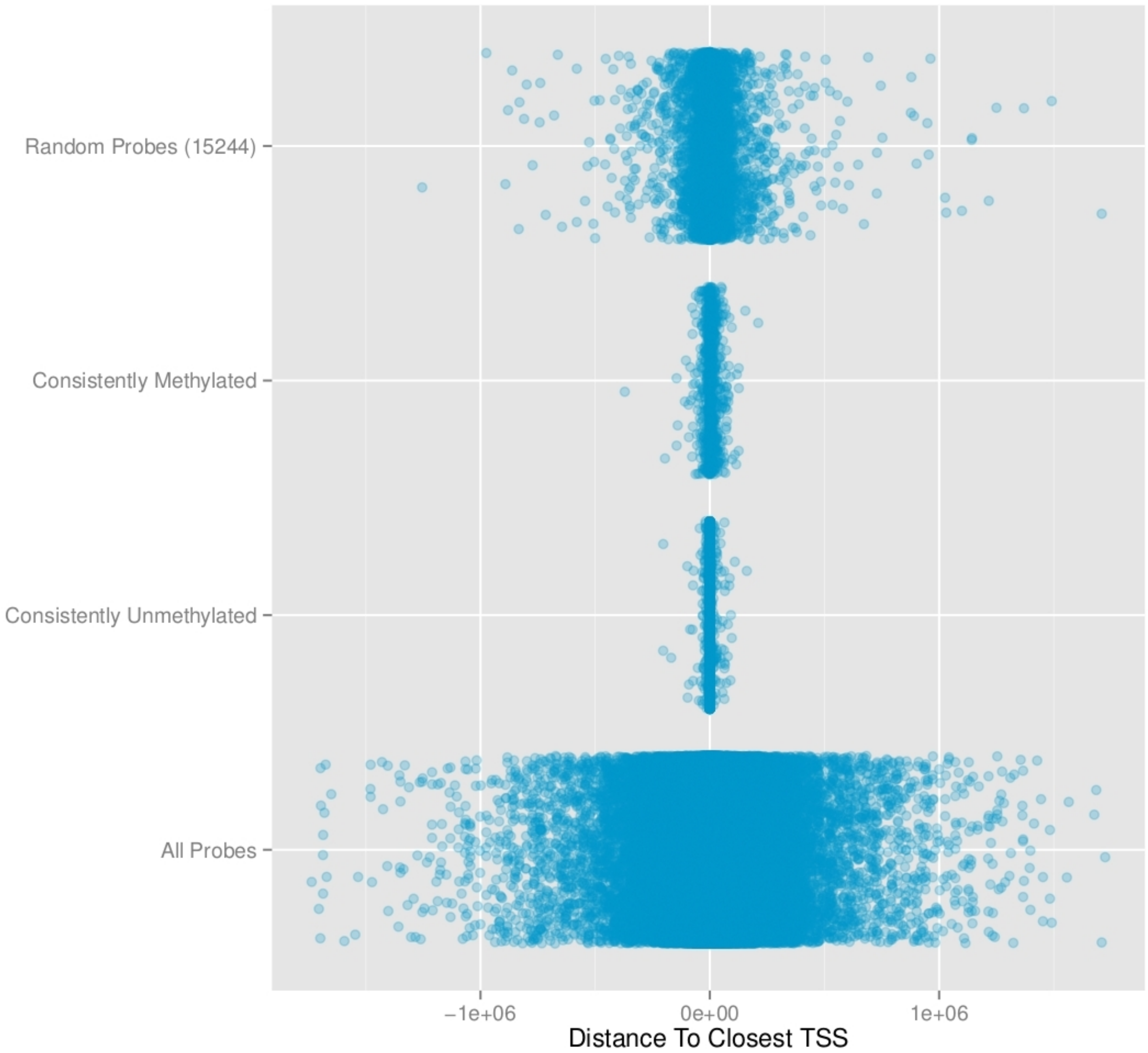
- 14250 consistently unmethylated probes
- 974 consistently methylated probes
  - Confirmed with ENCODE RRBS data and supported by methyltransferase studies



Ways to Define Biological Relevance of Consistent CpGs

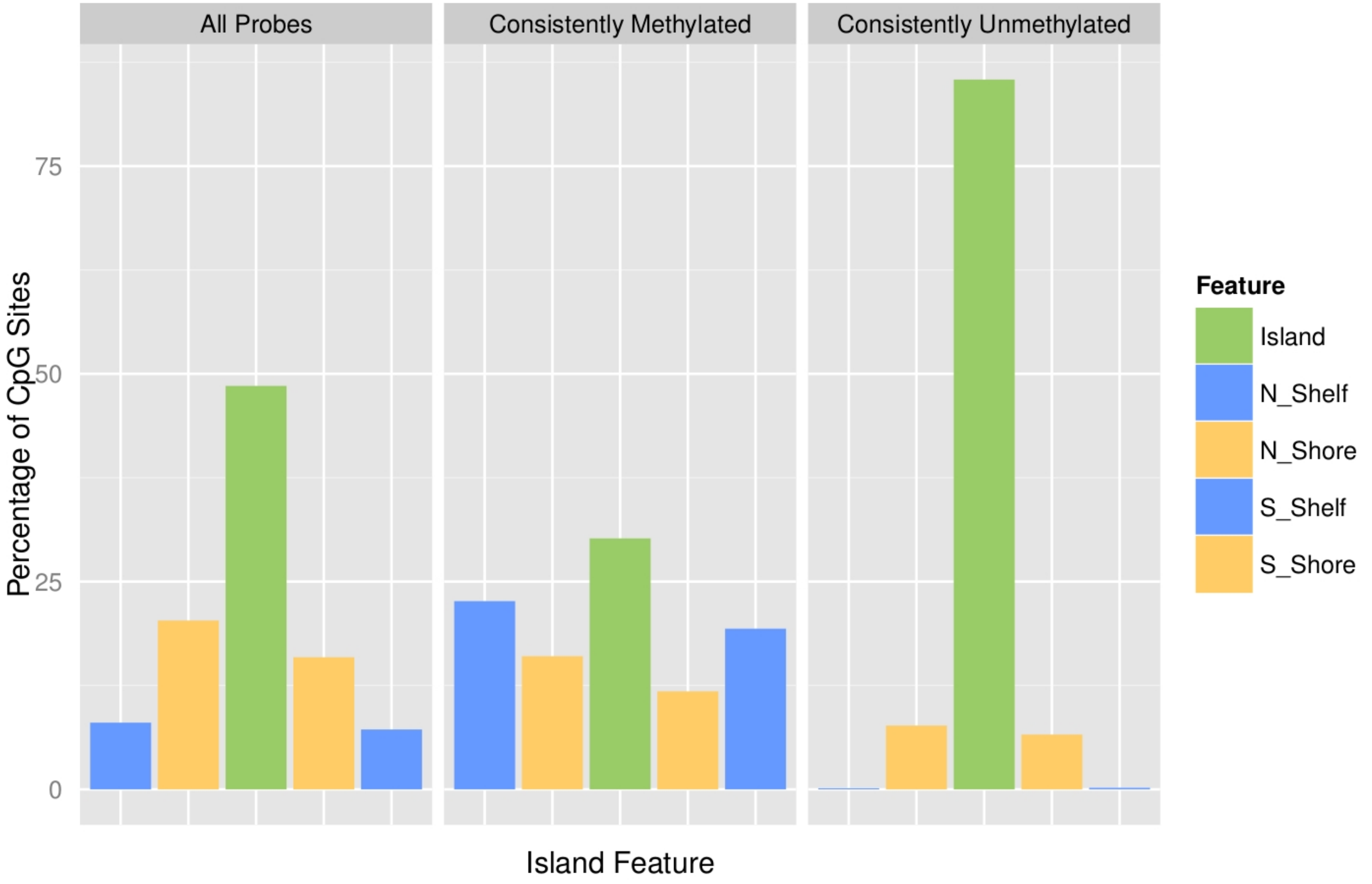
- Consistent CpG distribution in genomic features
    - Islands
    - Genes
- Pattern in CpG islands
    - Associated gene expression
    - Regions DNAse sensitivity
    - Gene set enrichment

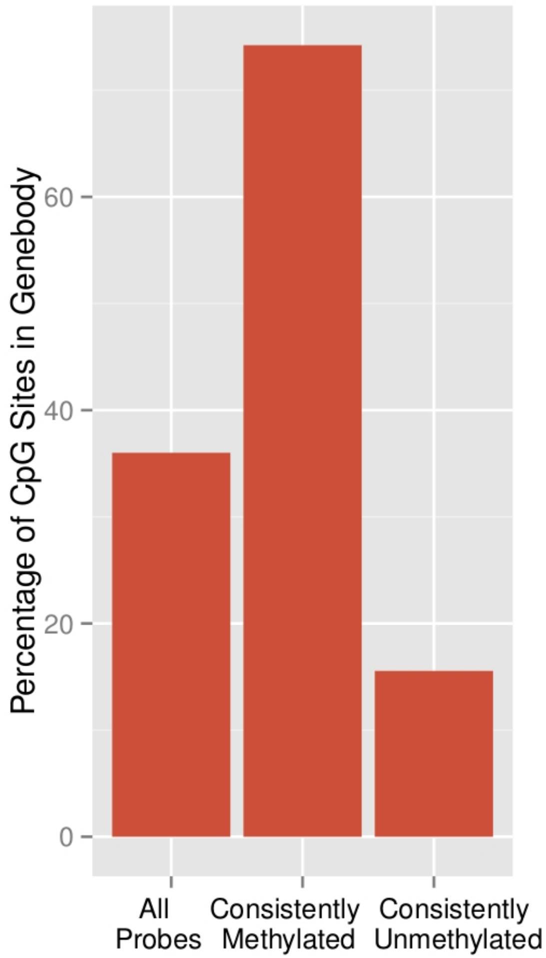
Consistent Probes Closer to Transcription Start Sites

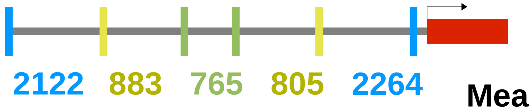


t test between consistent and random probes TSS distance p<0.001

Consistently Unmethylated in Islands
Consistently Methylated in Shores, Shelves and Genes



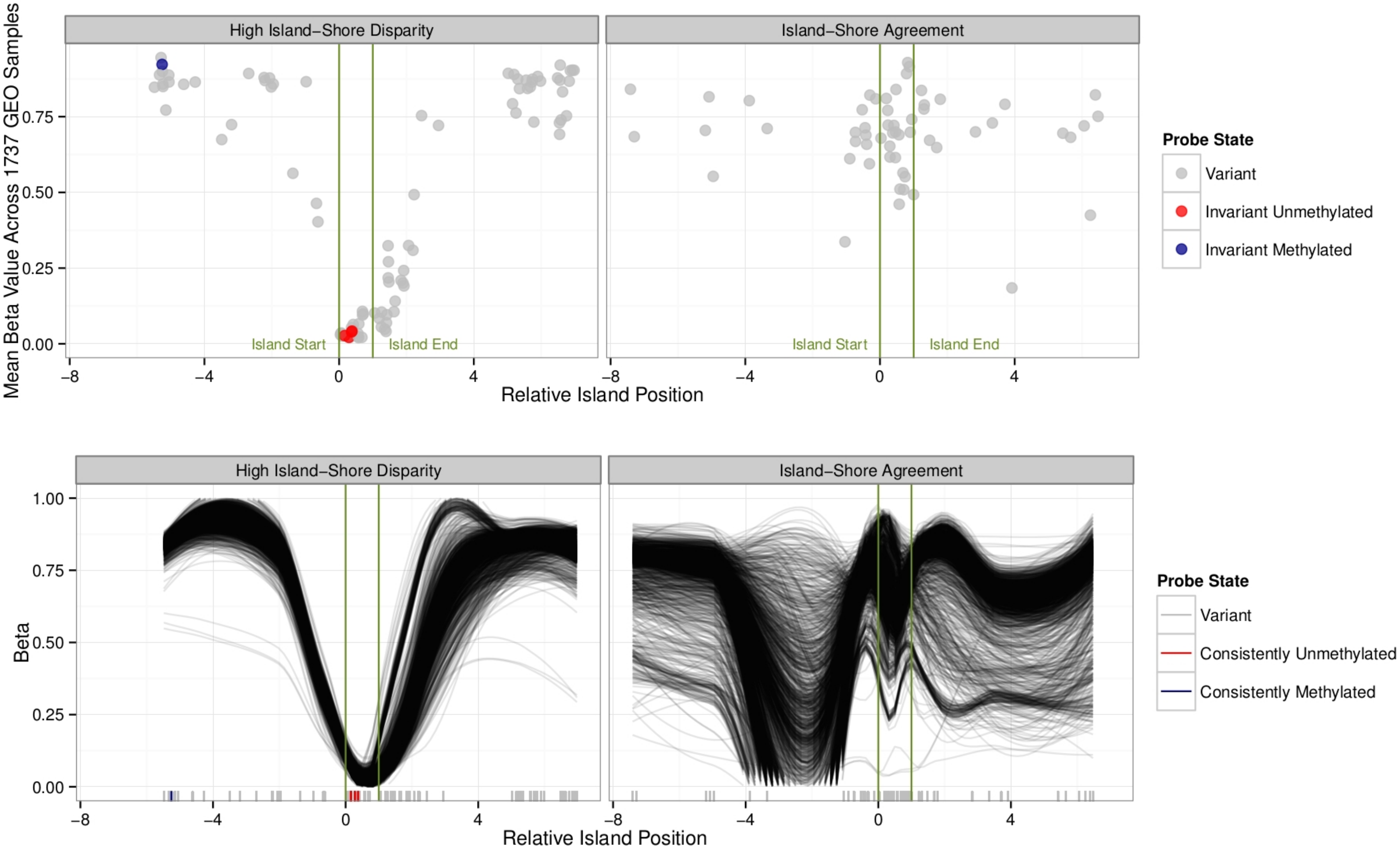




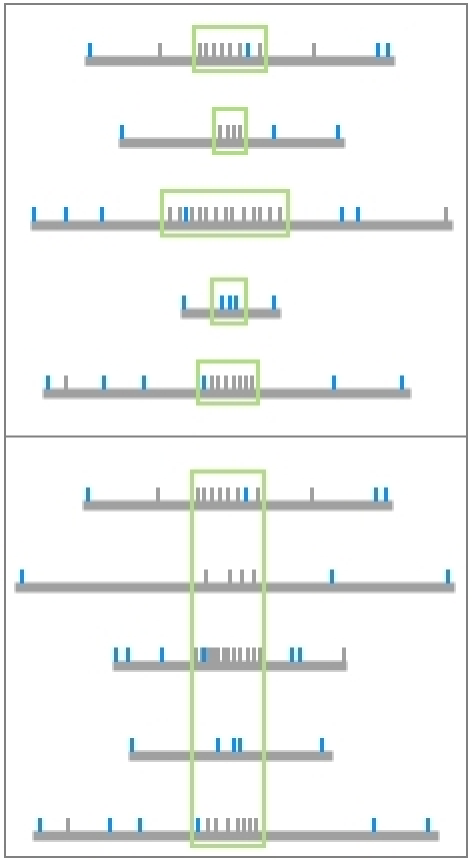
Mean length (bp)

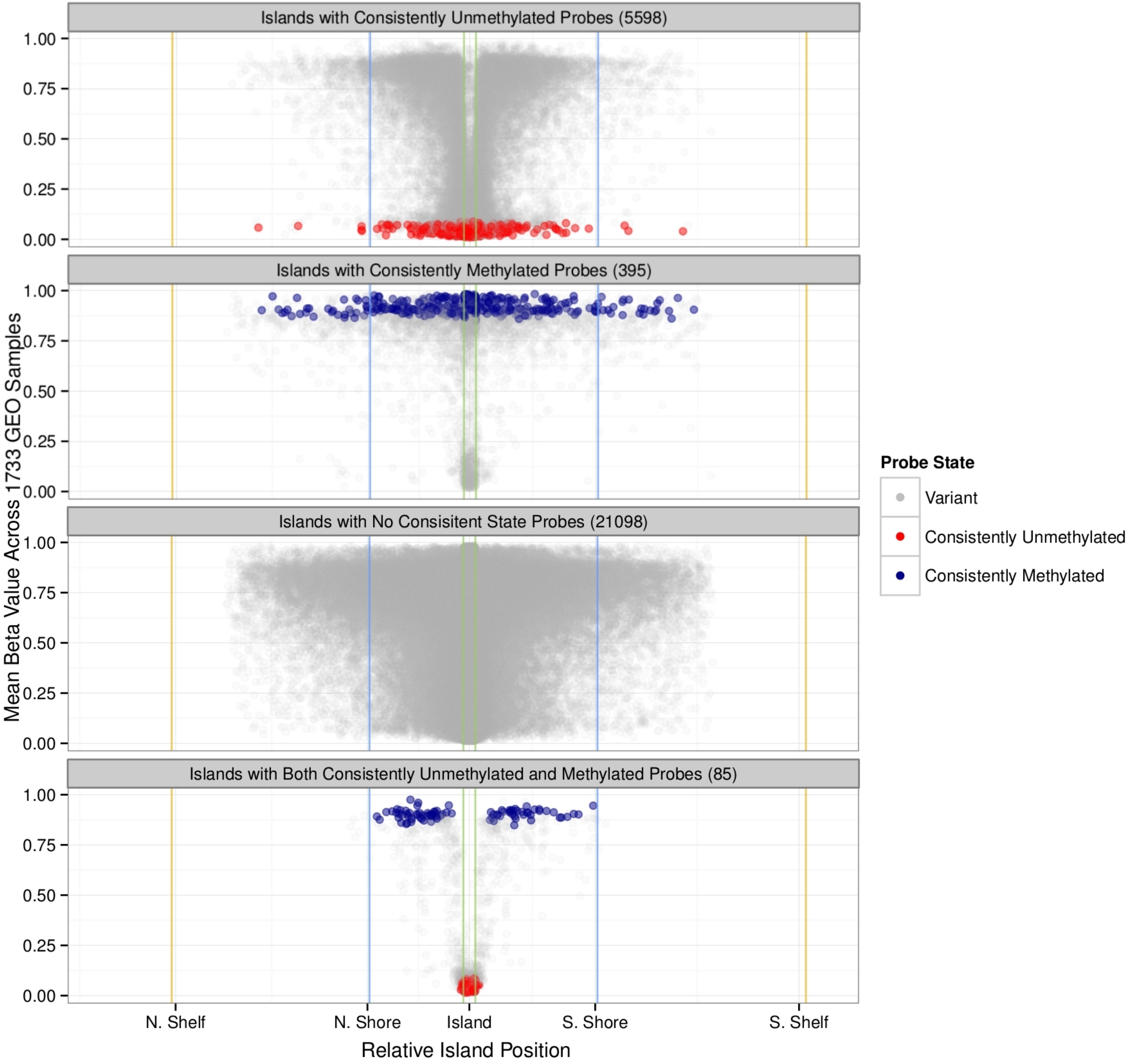
Genomic CpG density

Representative Islands

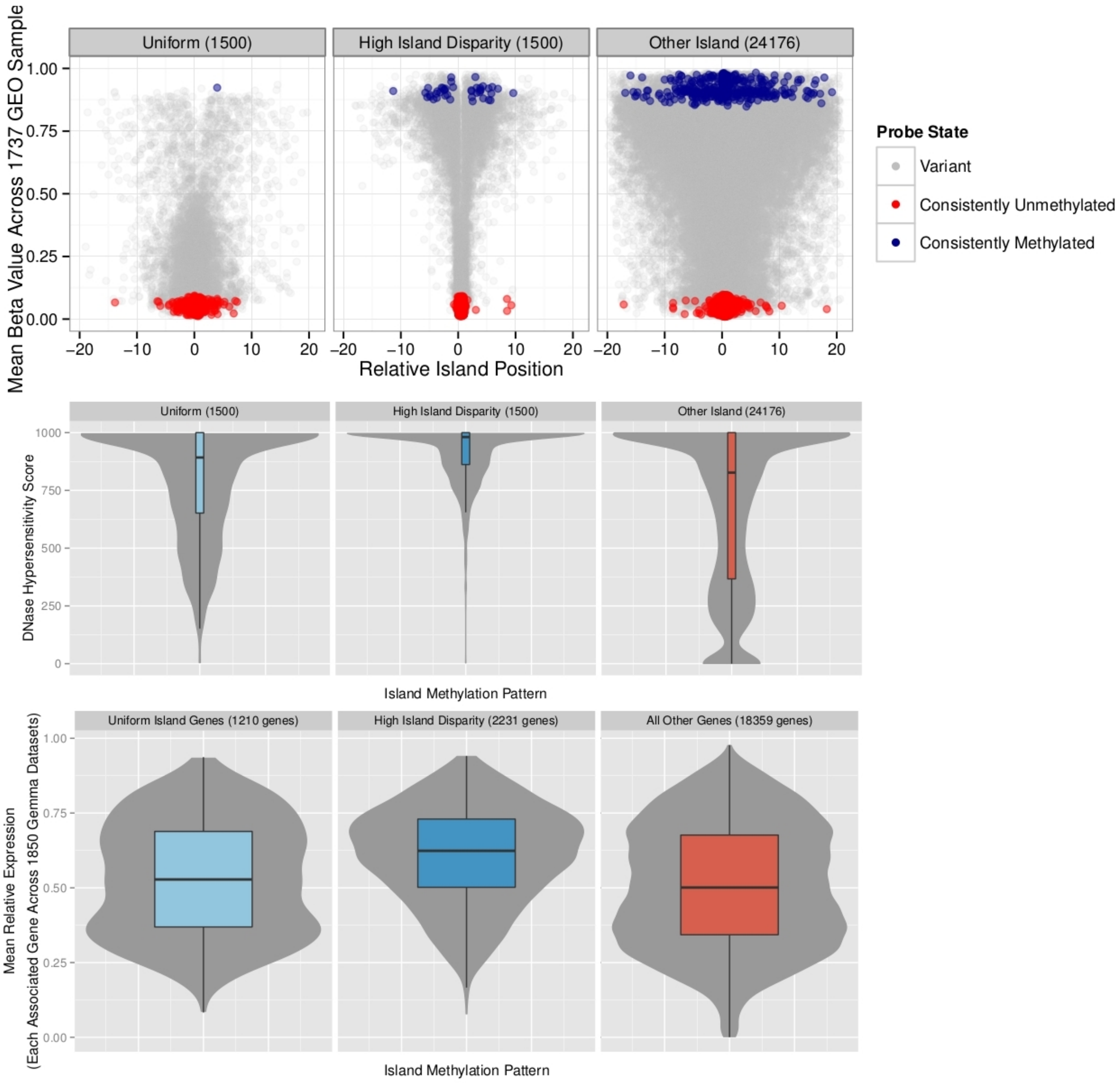


Pattern Seen In Islands With Consistently Unmethylated CpG



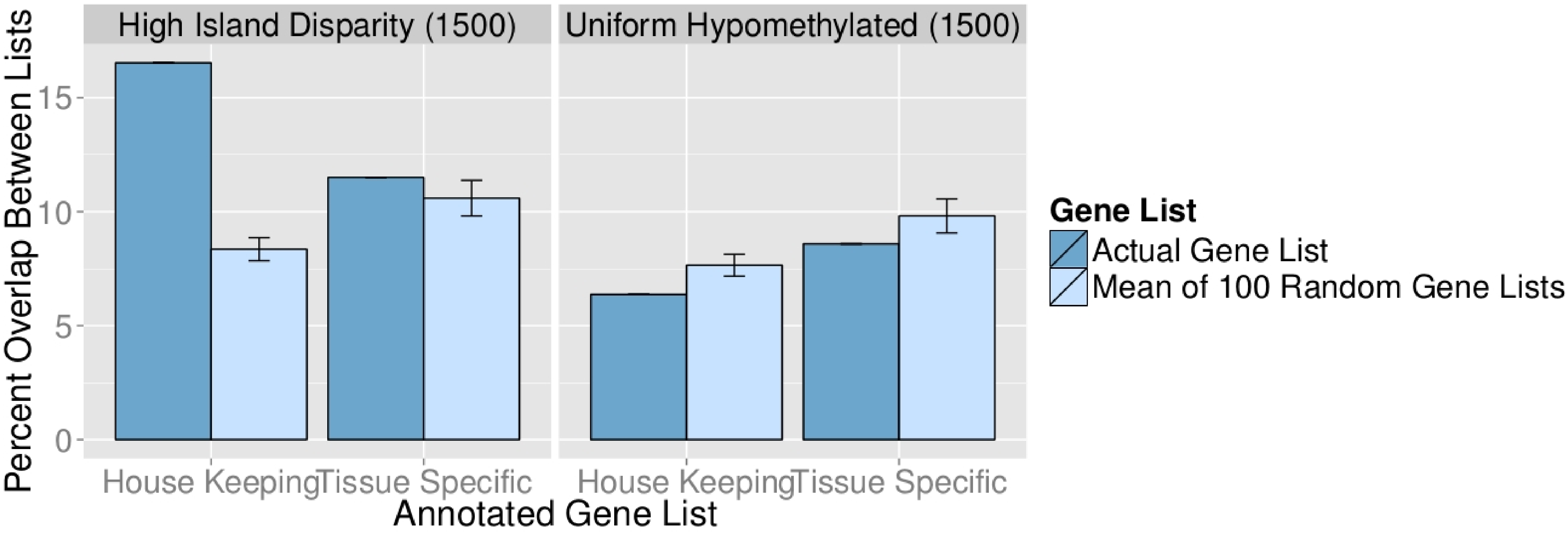


Dipped Islands Are More Highly Expressed



KS test between all island classes DNAse scores and gene expression p<0.001

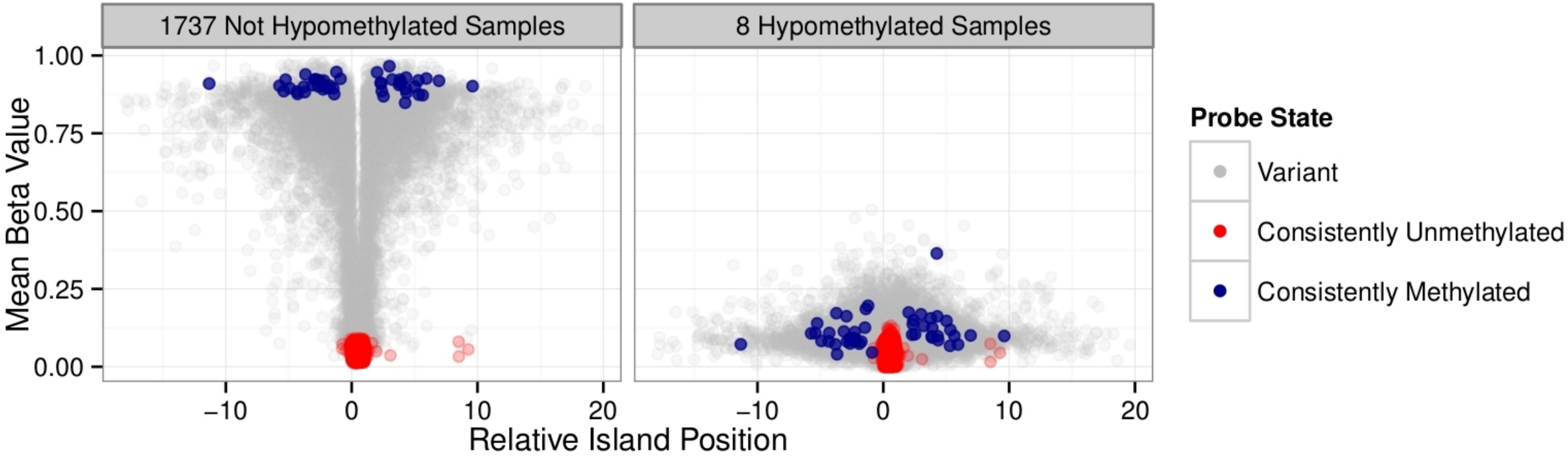
Dipped Island Associated Genes Enriched for Housekeeper Genes



(Cheng *et al.*, *PloS ONE* 2011)

Dipped Pattern Disrupted in Hypomethylated Samples

- Gene expression in hypomethylated samples could help define importance of dipped pattern in CpG islands



Consistent Sites Improve Overall Understanding of Human Methylome

- Existence of consistent CpG sites highlights a level of stability in the human methylome
- Dipped pattern may define a subset of islands associated with universally active genes in the human genome.

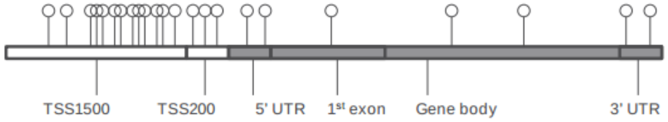
NSERC
CRSNG

NeuroDevNet

NATIONAL INSTITUTES OF HEALTH

Supplemental Figures

Figure 2: HumanMethylation450 BeadChip Provides Coverage Throughout Gene Regions

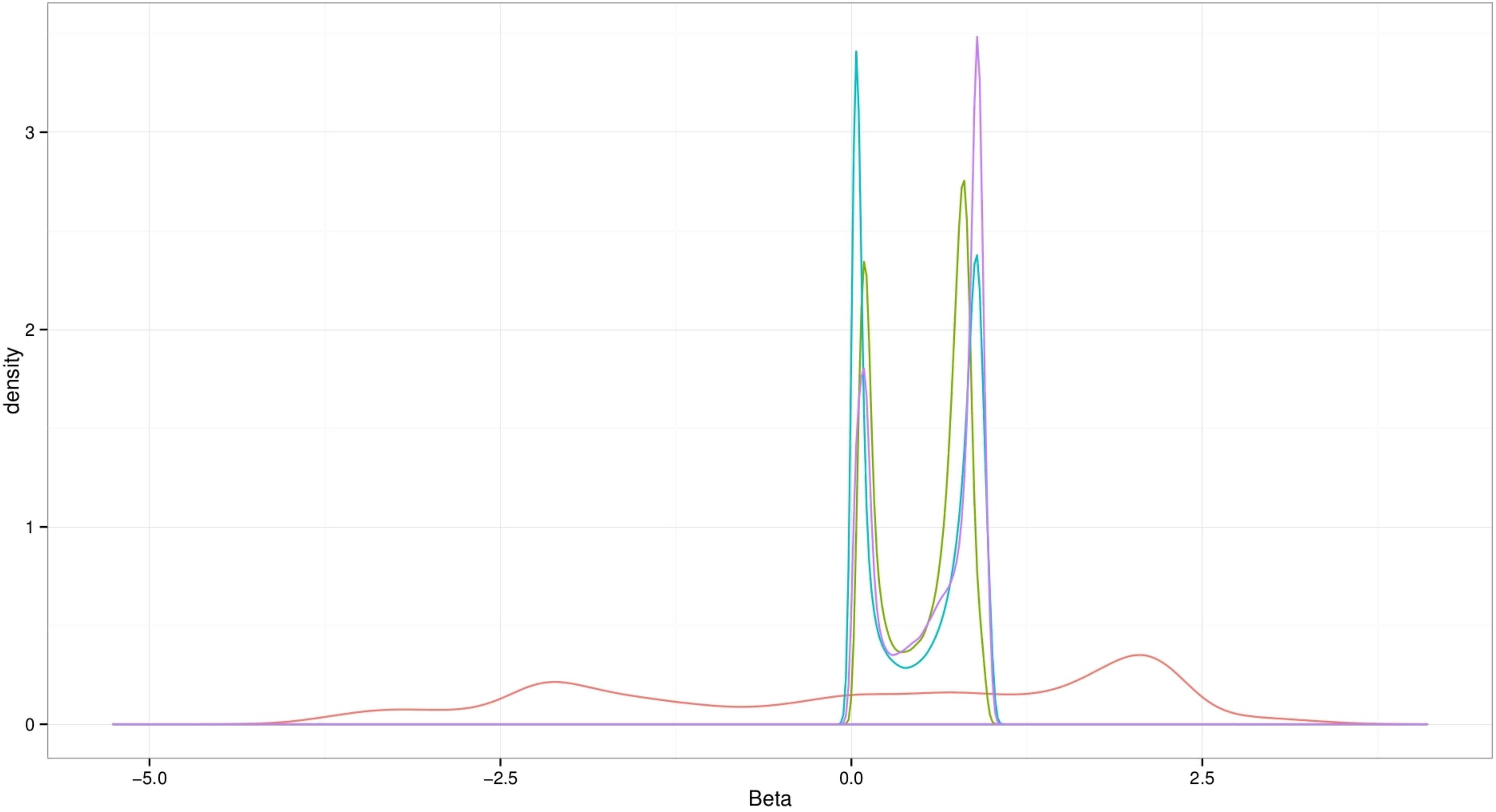


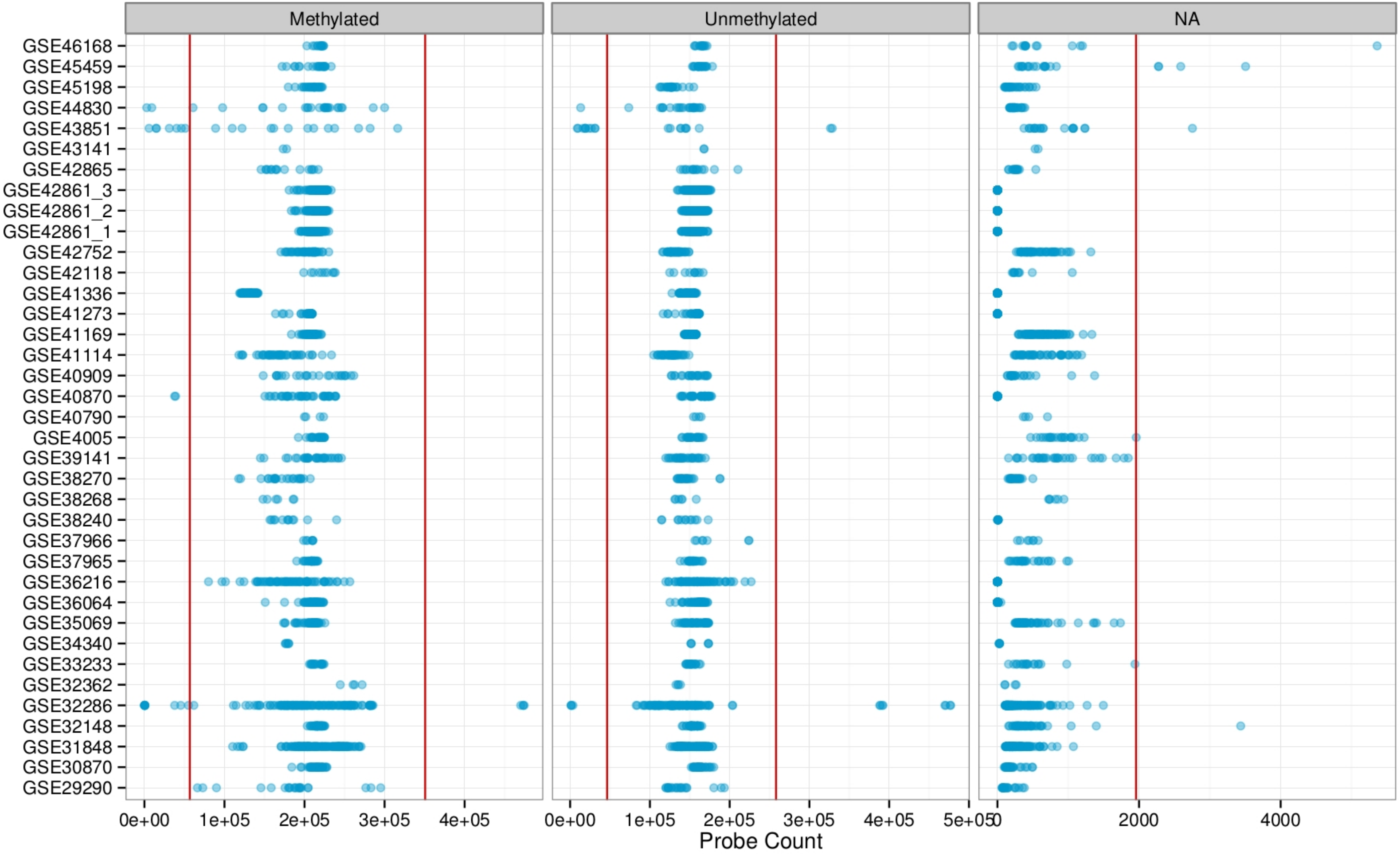
Feature Type	Genes Mapped	Percent Genes Covered	Number of Loci on Array
NM_TSS200	14895	0.79	2.56
NM_TS1500	17820	0.94	3.41
NM_5'UTR	13865	0.78	3.34
NM_1stExon	15127	0.80	1.62
NM_3'UTR	13042	0.72	1.02
NM_GeneBody	17071	0.97	8.97
NR_TSS200	1967	0.65	1.84
NR_TSS1500	2672	0.88	2.92
NR_GeneBody	2345	0.77	5.34

Feature Type	Islands Mapped	Percent Islands Covered	Average Number of Loci on Array
Island	26153	0.94	5.08
N_Shore	25770	0.93	2.74
S_Shore	25614	0.92	2.66
N_Shelf	23896	0.86	1.97
S_Shelf	23968	0.86	1.94

Quality Controls

1. Methyltransferase
2. Filtered Probes
3. Not Betas
4. Detection Pvalues
5. Beta Distribution
6. NA Counts
7. Methylated and Unmethylated Totals

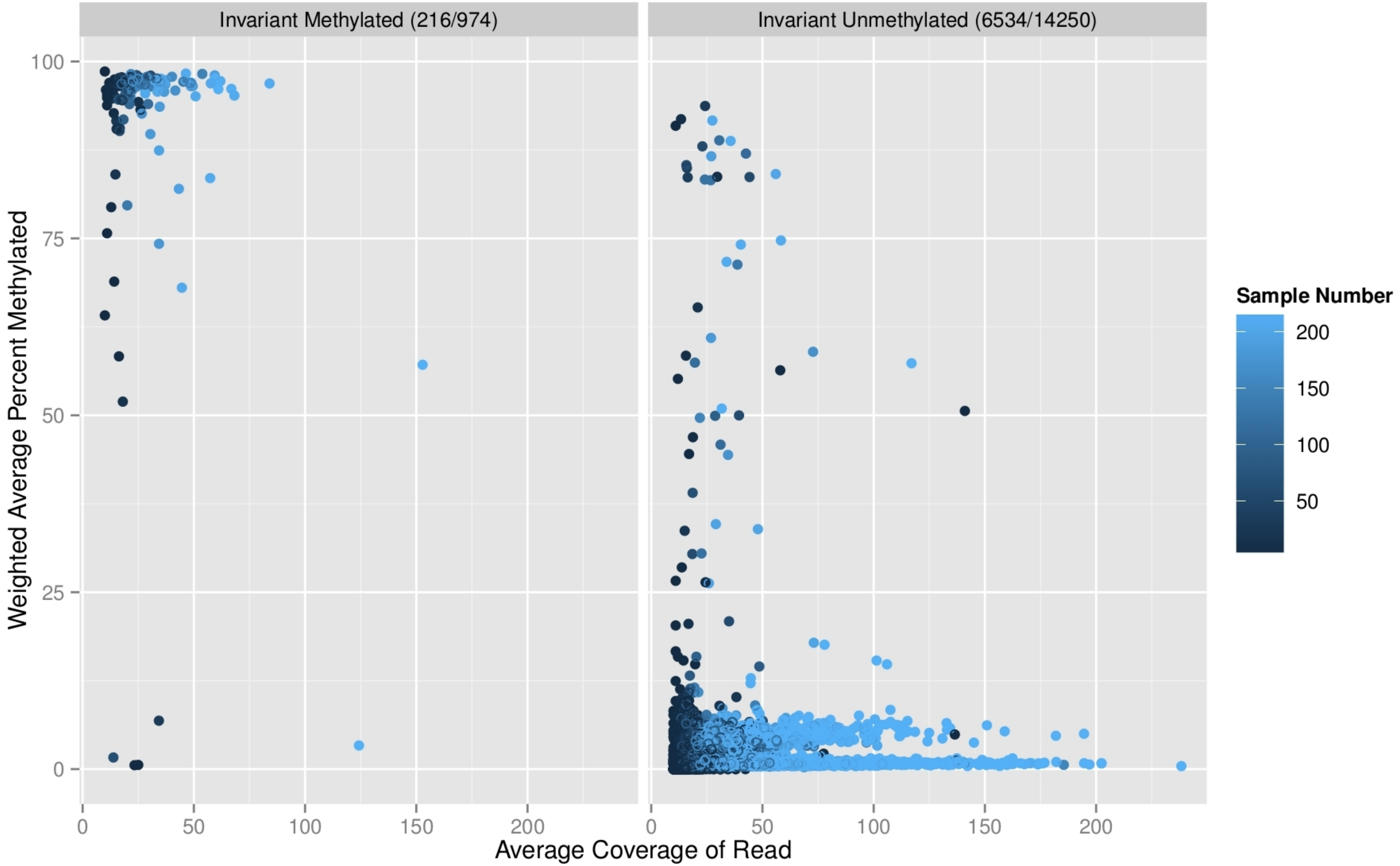

density
3
2
1
0
−5.0
−2.5
0.0
2.5
Beta



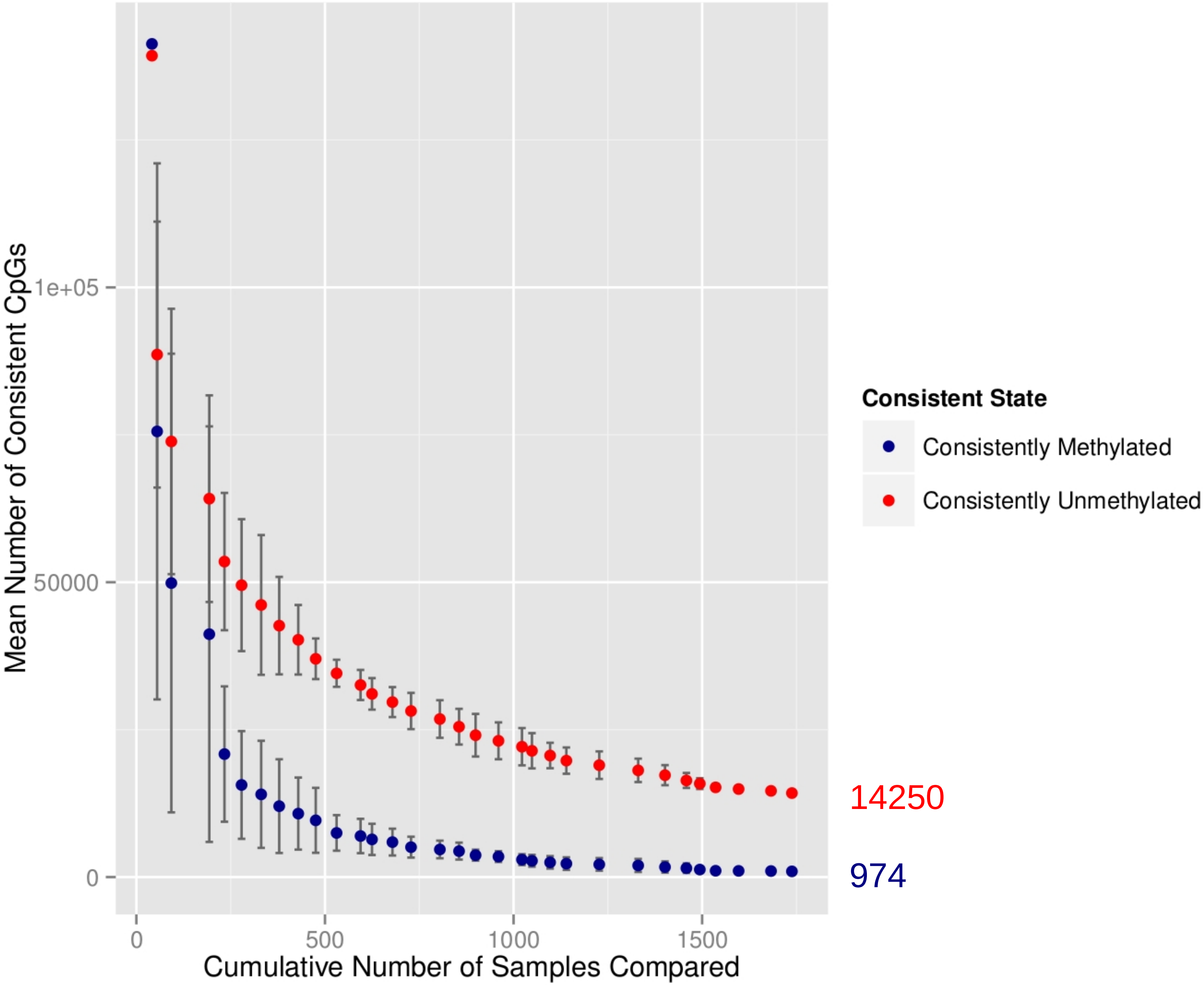
- 72 series to **32 series**
- 4219 samples to **1737 samples**

General Agreement Between 450K and RRBS Consistent CpGs

98% and 80% of unmethylated and methylated 450K consistent calls, respectively, are confirmed in 90% of samples of a given CpG



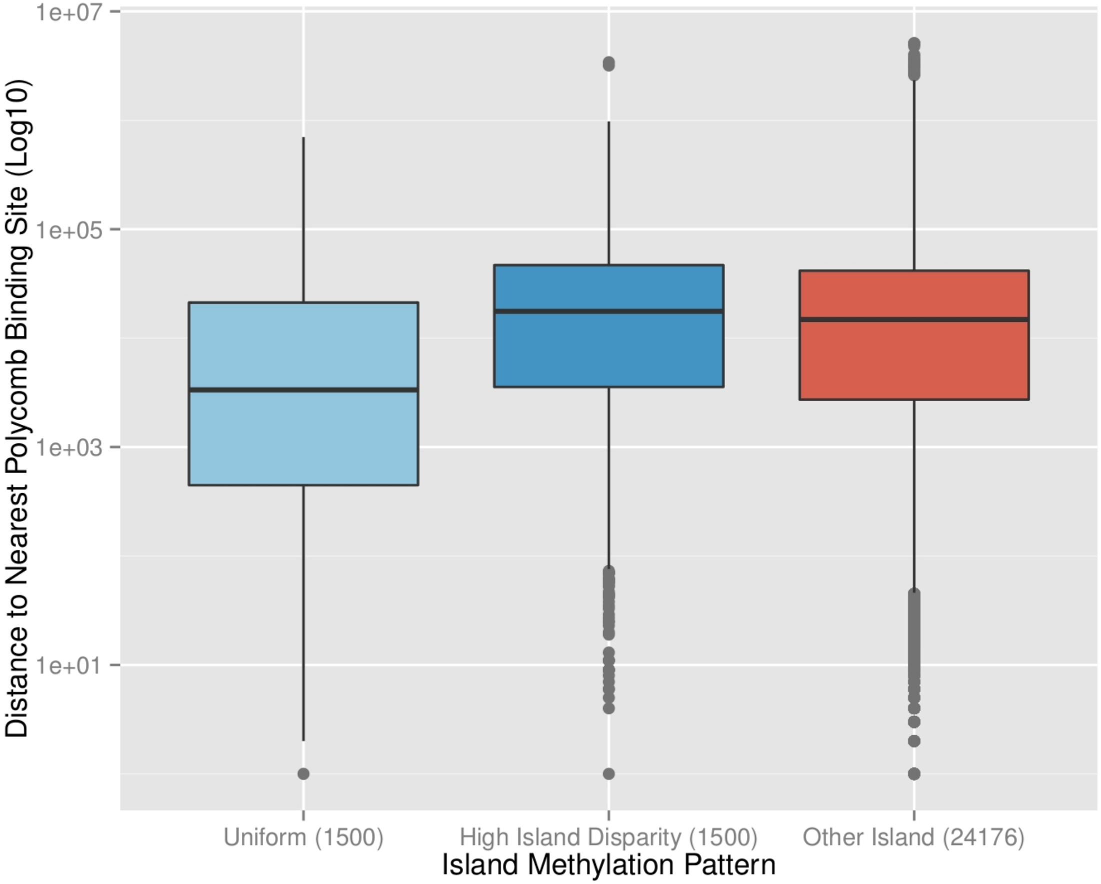
Consistent Probe Plateau

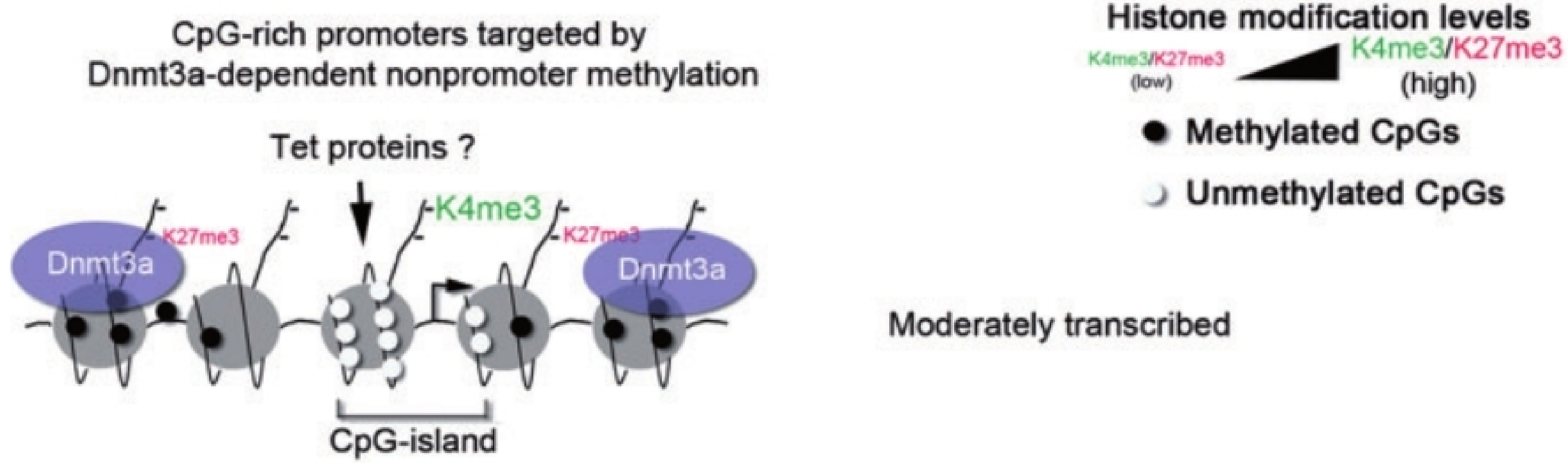


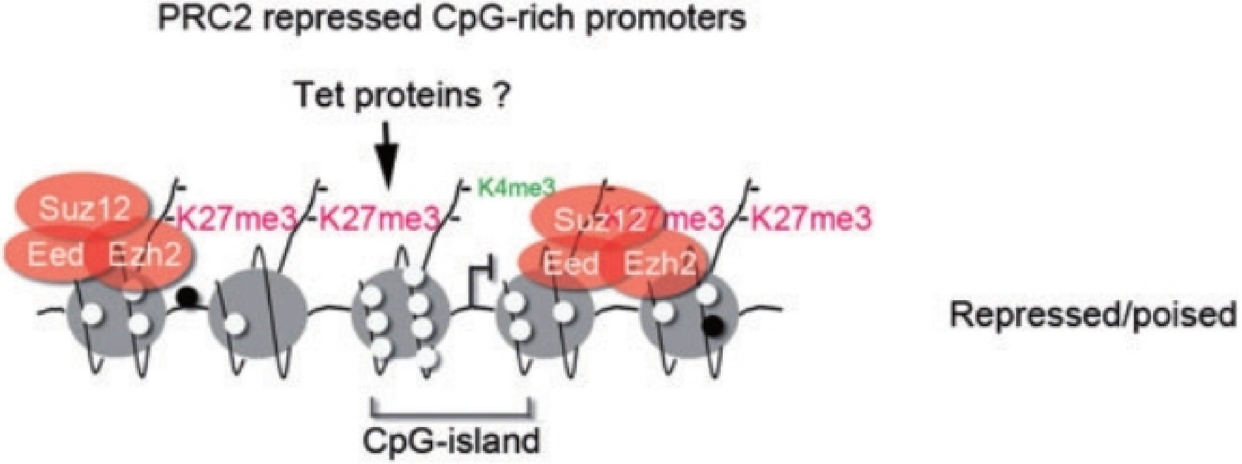
14250

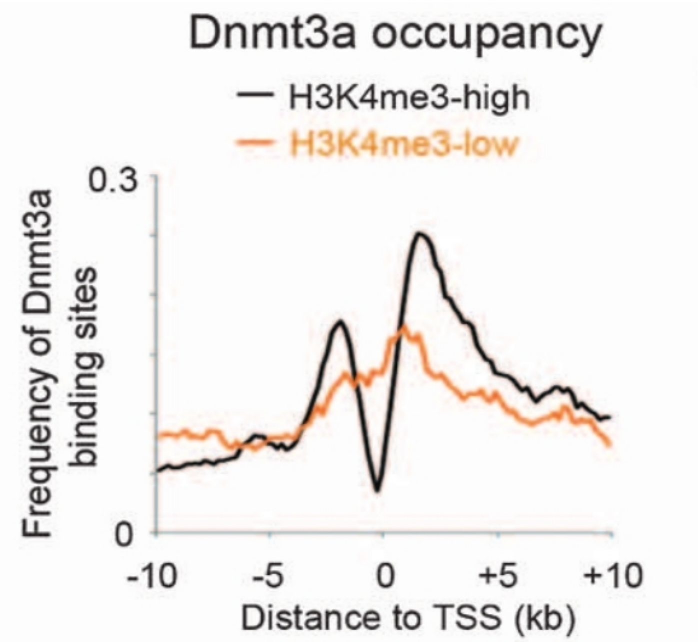
974

- Association of high disparity islands with 3710 polycomb binding sites

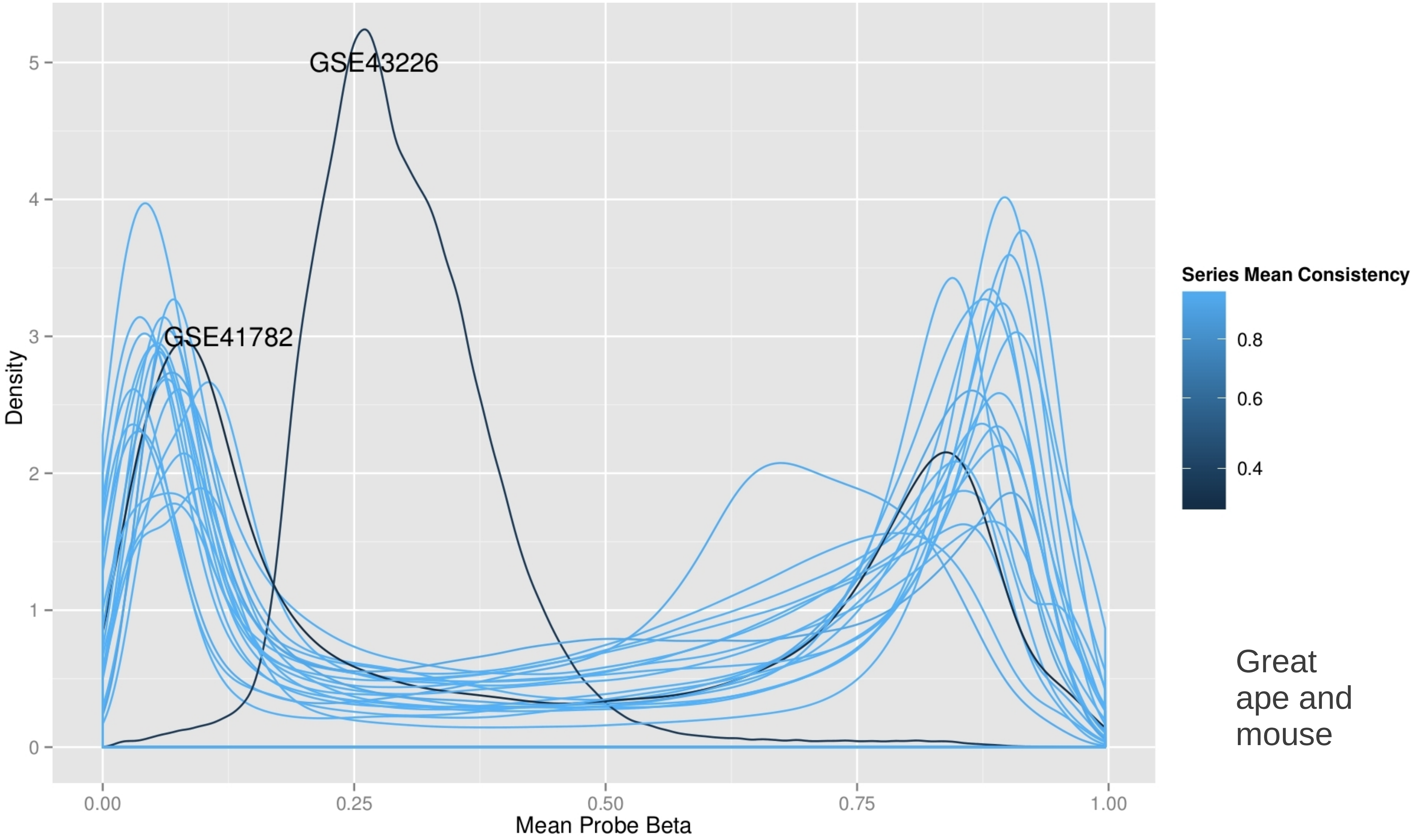



CpG-rich promoters targeted by Dnmt3a-dependent nonpromoter methylation
Tet proteins ?
K27me3
K4me3
K27me3
Dnmt3a
Dnmt3a
CpG-island
Moderately transcribed
Histone modification levels
K4me3/K27me3 (low)
K4me3/K27me3 (high)
● Methylated CpGs
○ Unmethylated CpGs

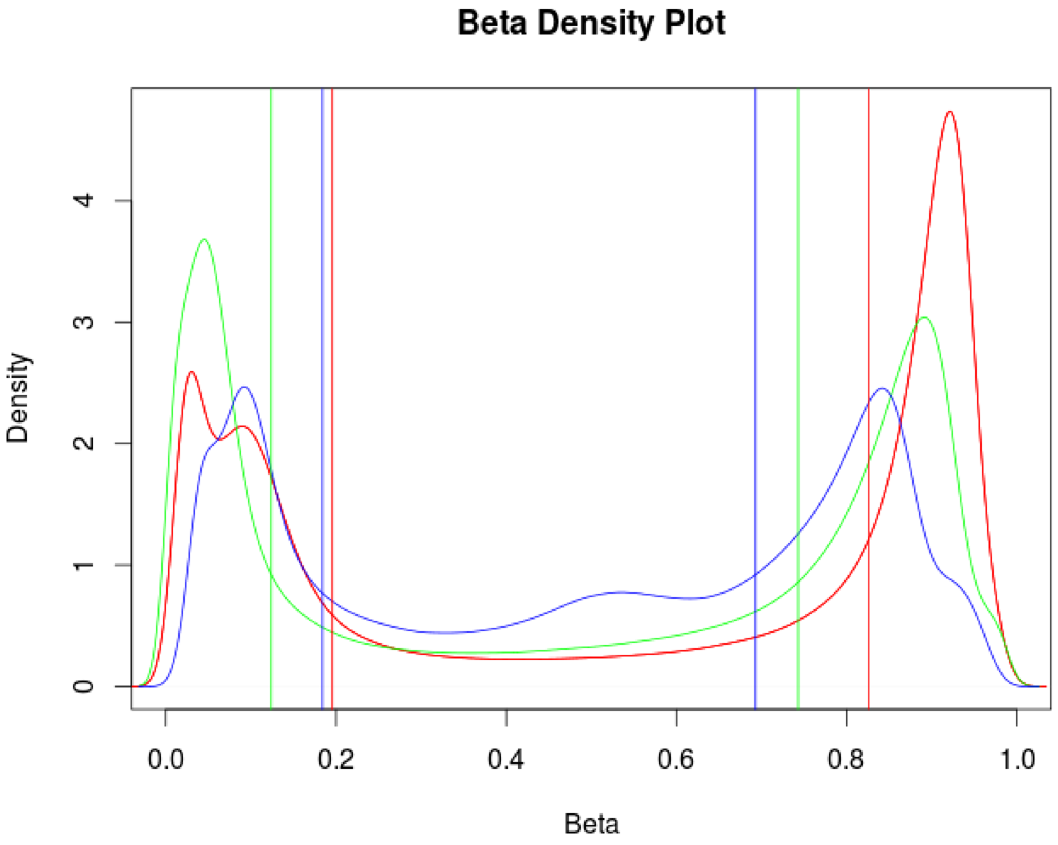

PRC2 repressed CpG-rich promoters
Tet proteins ?
Suz12
Eed
Ezh2
K27me3
K27me3
K4me3
Suz12
Eed
Ezh2
K27me3
K27me3
CpG-island
Repressed/poised

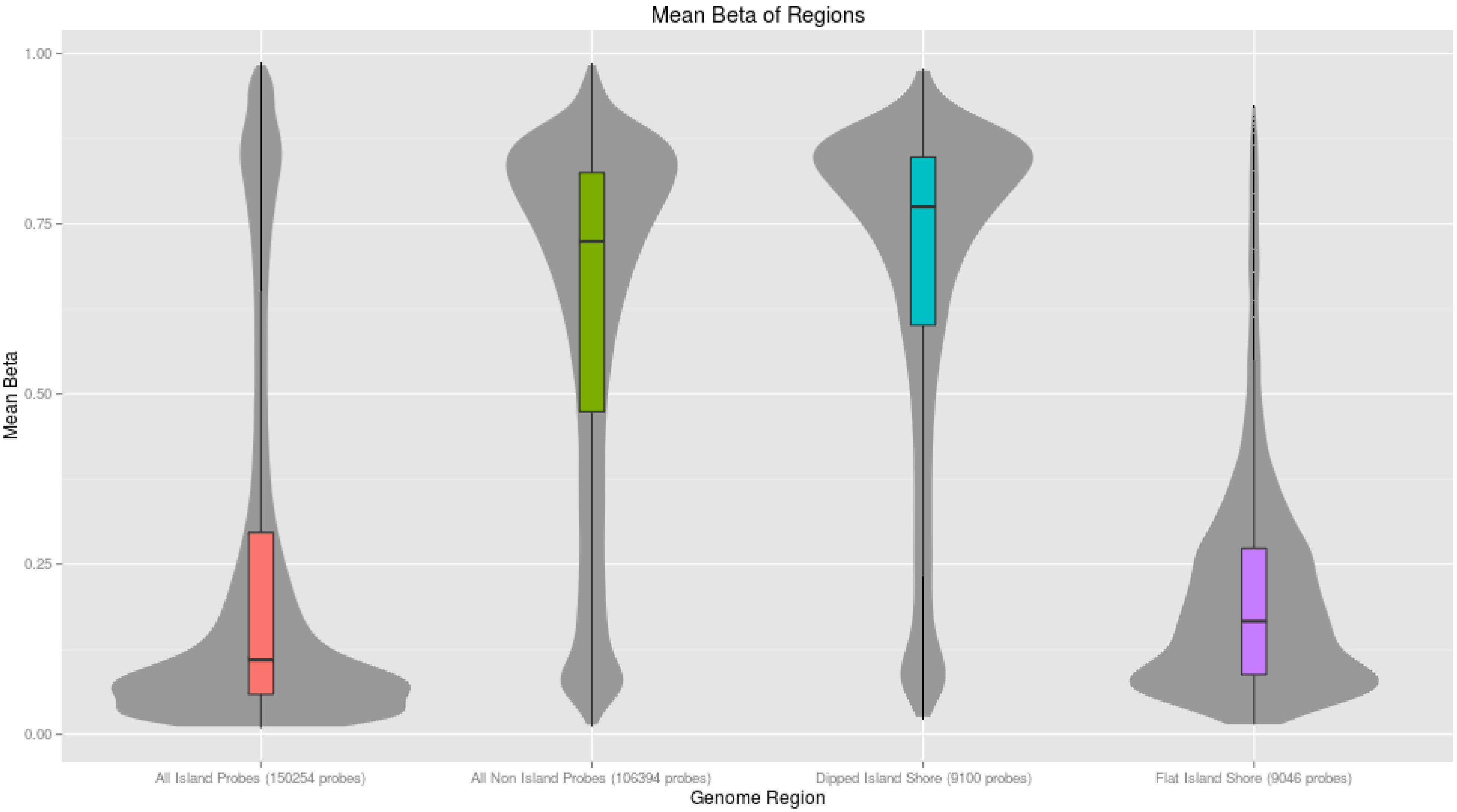

Dnmt3a occupancy
H3K4me3-high
H3K4me3-low
Frequency of Dnmt3a binding sites
0.3
0
-10
-5
0
+5
+10
Distance to TSS (kb)

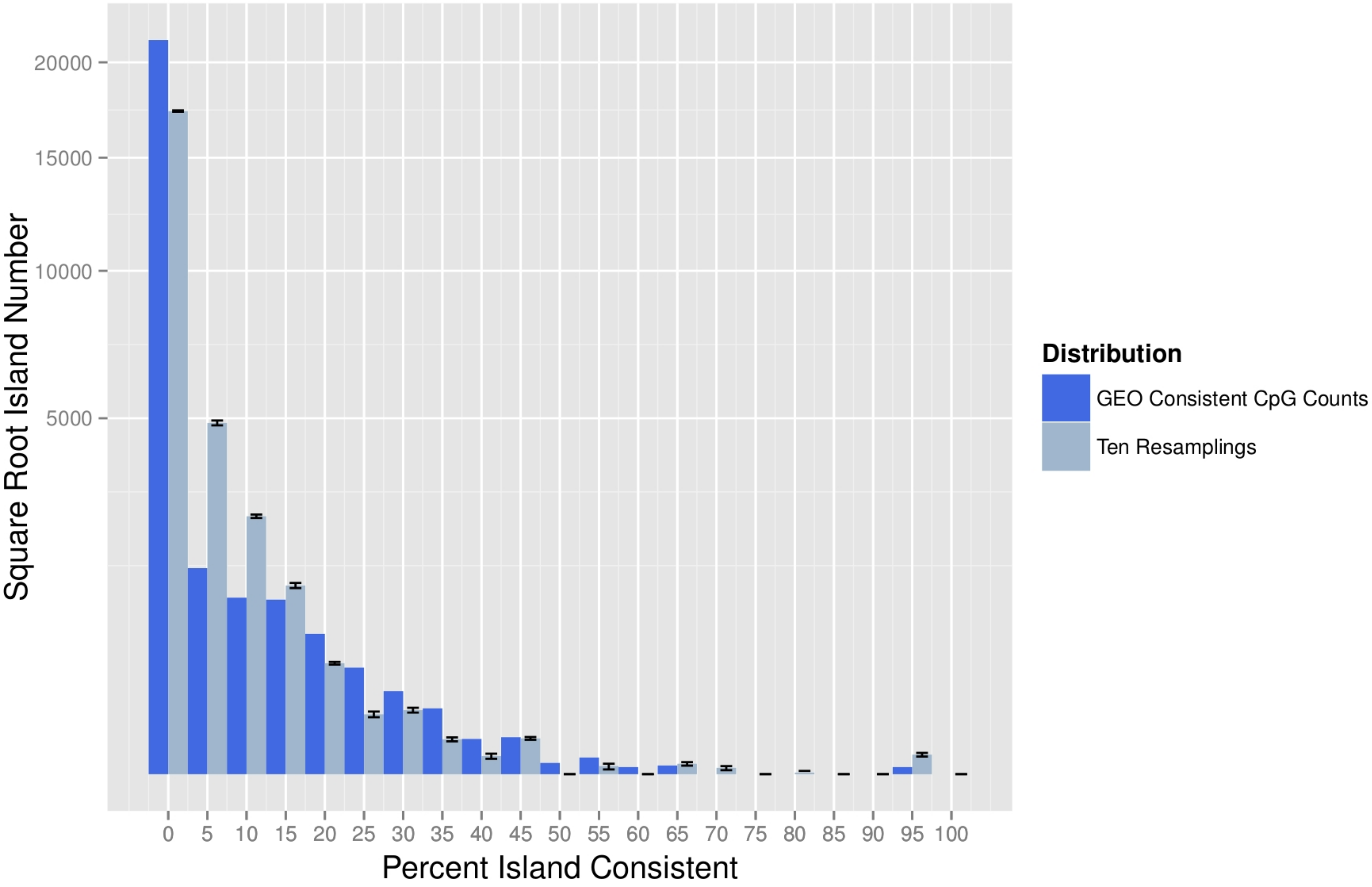
GEO Series Consistency as QC Measure



Great ape and mouse

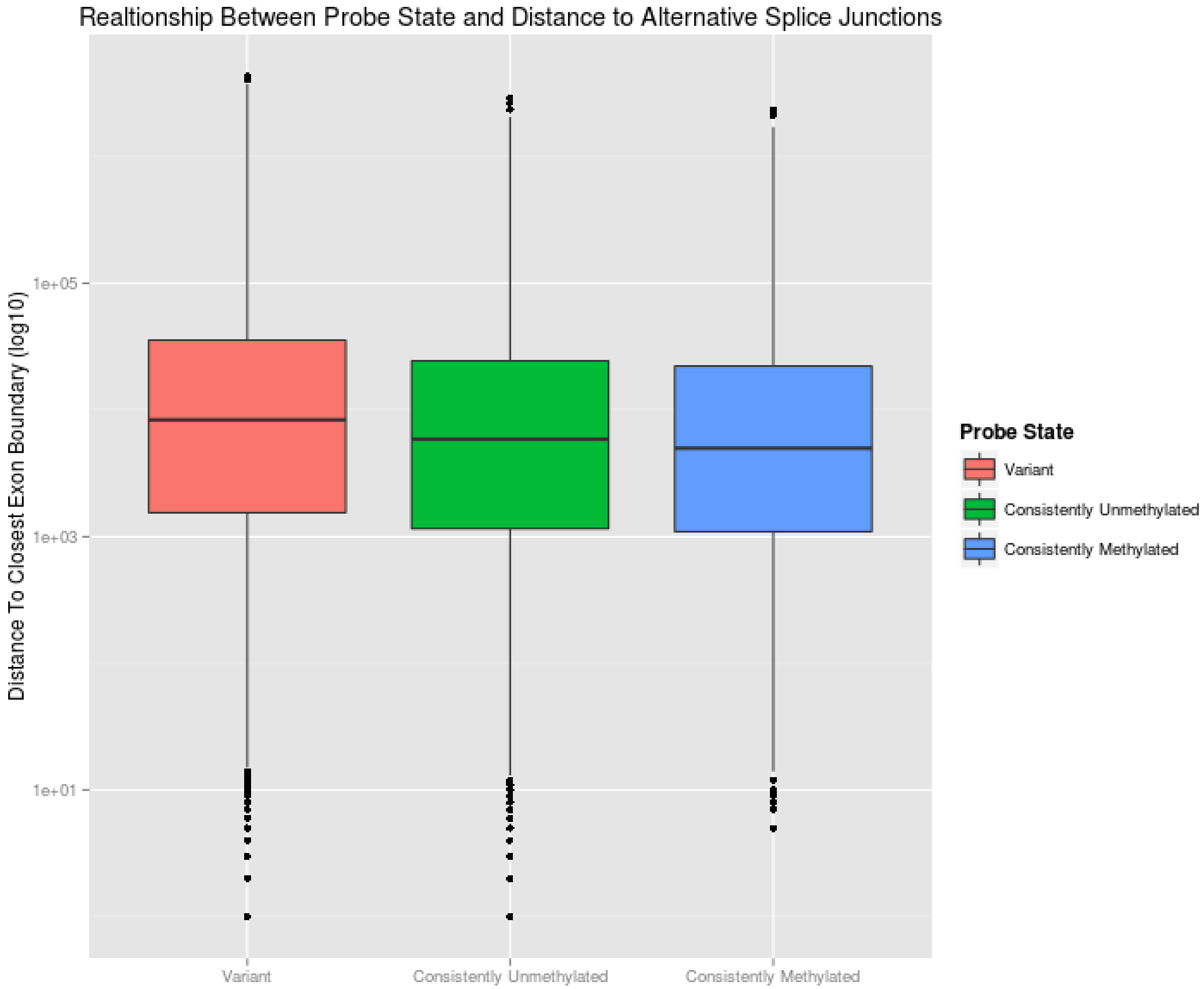

Beta Density Plot
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
Beta

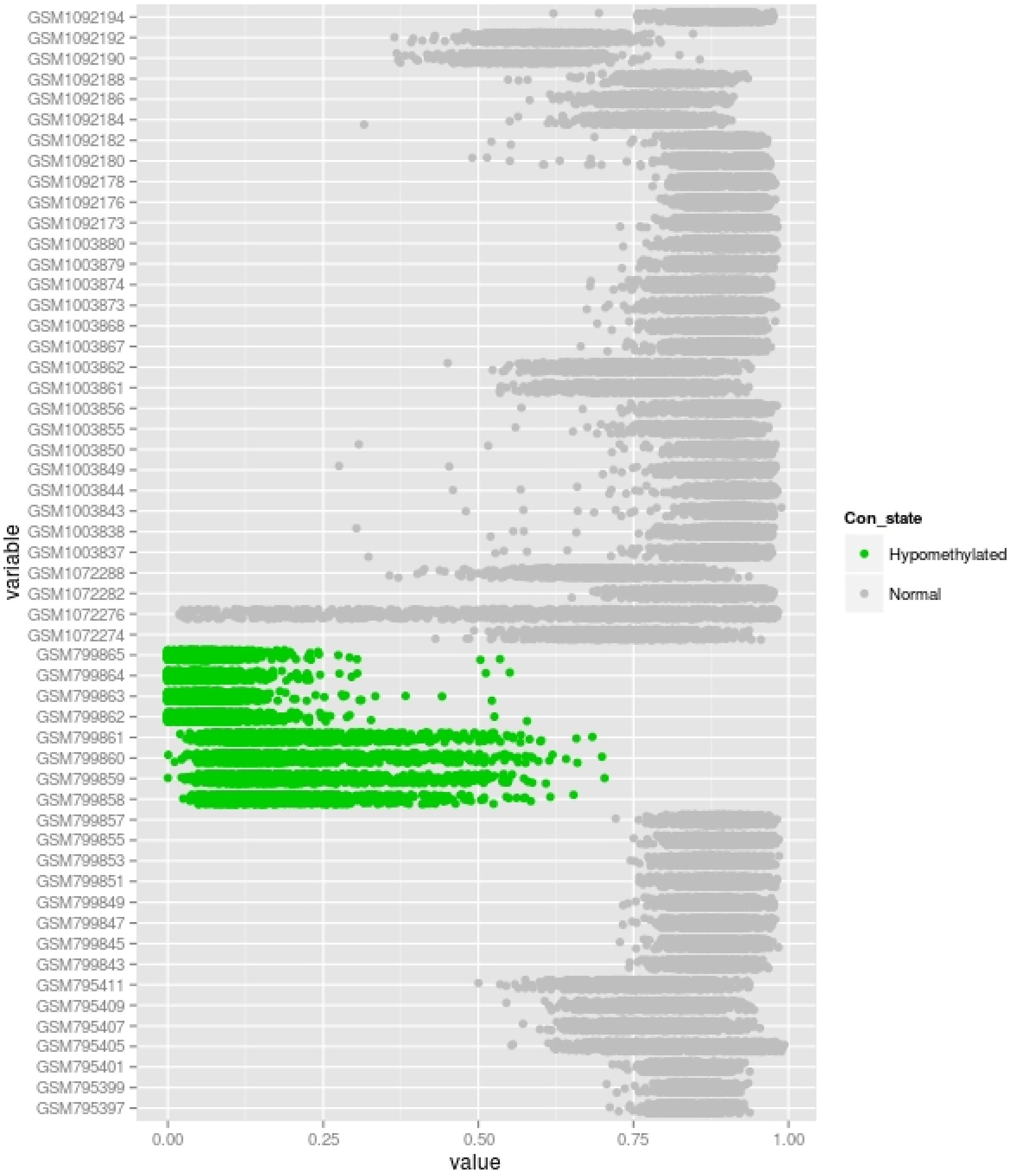

Mean Beta of Regions
Mean Beta
1.00
0.75
0.50
0.25
0.00
All Island Probes (150254 probes)
All Non Island Probes (106394 probes)
Dipped Island Shore (9100 probes)
Flat Island Shore (9046 probes)
Genome Region



Consistent Probes Are Not Evenly Distributed Among Islands

KS test between actual consistent probe distribution among island and mean of ten random distributions $p<0.001$


Realtionship Between Probe State and Distance to Alternative Splice Junctions
Distance To Closest Exon Boundary (log10)
1e+05
1e+03
1e+01
Variant
Consistently Unmethylated
Consistently Methylated
Probe State
Variant
Consistently Unmethylated
Consistently Methylated


GSM1092194
GSM1092192
GSM1092190
GSM1092188
GSM1092186
GSM1092184
GSM1092182
GSM1092180
GSM1092178
GSM1092176
GSM1092173
GSM1003880
GSM1003879
GSM1003874
GSM1003873
GSM1003868
GSM1003867
GSM1003862
GSM1003861
GSM1003856
GSM1003855
GSM1003850
GSM1003849
GSM1003844
GSM1003843
GSM1003838
GSM1003837
GSM1072288
GSM1072282
GSM1072276
GSM1072274
GSM799865
GSM799864
GSM799863
GSM799862
GSM799861
GSM799860
GSM799859
GSM799858
GSM799857
GSM799855
GSM799853
GSM799851
GSM799849
GSM799847
GSM799845
GSM799843
GSM795411
GSM795409
GSM795407
GSM795405
GSM795401
GSM795399
GSM795397
variable
0.00
0.25
0.50
0.75
1.00
value
Con_state
Hypomethylated
Normal