TCGA
GBM DNA Methylation Update
Patient Centric Analysis

Houtan Noushmehr

Peter W. Laird

Jan 13, 2012

Rules for DNA methylation Calls/Scores

- 1st step, collapse multiple CpG to 1 gene for each platform (27k, 450k).
- Using Gene Expression Data (1 gene = 1 Exp. Value per sample)
- Merged Samples & Probes (DNA methylation) with Gene Exp for each platform
  - Cataloged probes to gene by annotation provided by Illumina & collected CpGs within 500bp of a known TSS.
- Calculated spearman correlation for all CpG probes to 1 gene expression.
- Among multiple CpG, selected 1 CpG with the lowest correlation value.
- Reduce platforms from N:1, to 1:1.

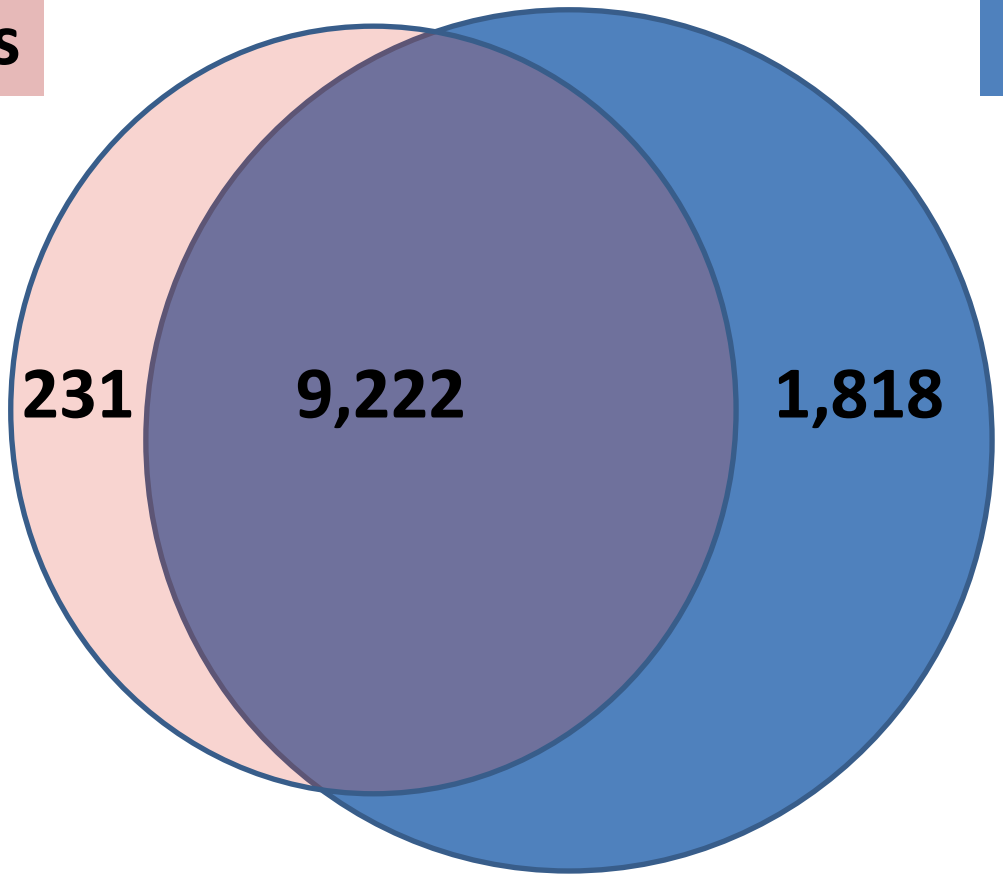
Venn diagram DNA meth vs Gene Exp.

- TODO: add venn diagram showing the number of probes/genes overlapping the two different experiments and the different platforms.

Genes Overlap 27K and 450K Platform

27K
9,453 Genes

450K
11,040 Genes



Data available for analysis

- 27K: 279 samp X 9,453 CpG:Gene
- 450K: 74 samp X 11,040 CpG:Gene

Rules for DNA methylation Calls/Scores

- 2nd step, create cut-offs.
- Using spearman correlation rho (GeneExp vs DNA methylation) we labeled each gene as:

SNC (Strongly Negatively Correlated)

< -0.5

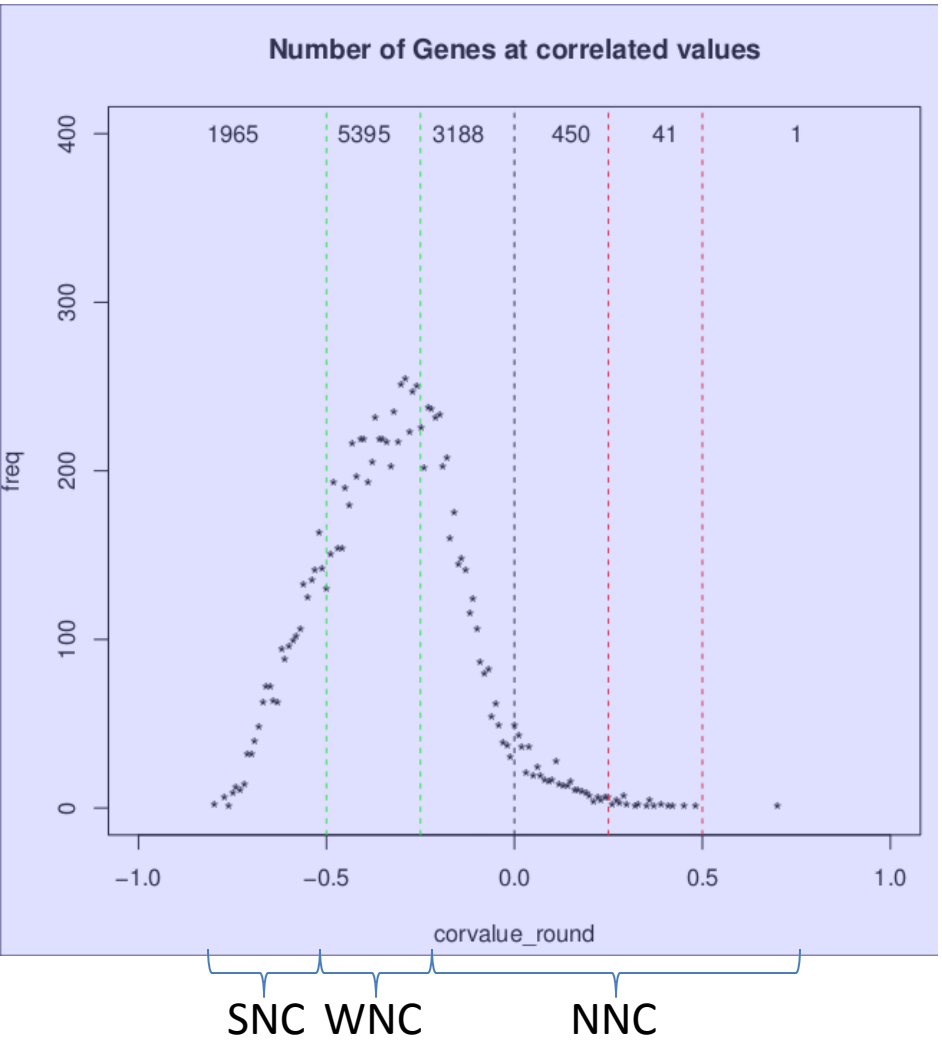
WNC (Weakly Negatively Correlated)

-0.5 to -0.25

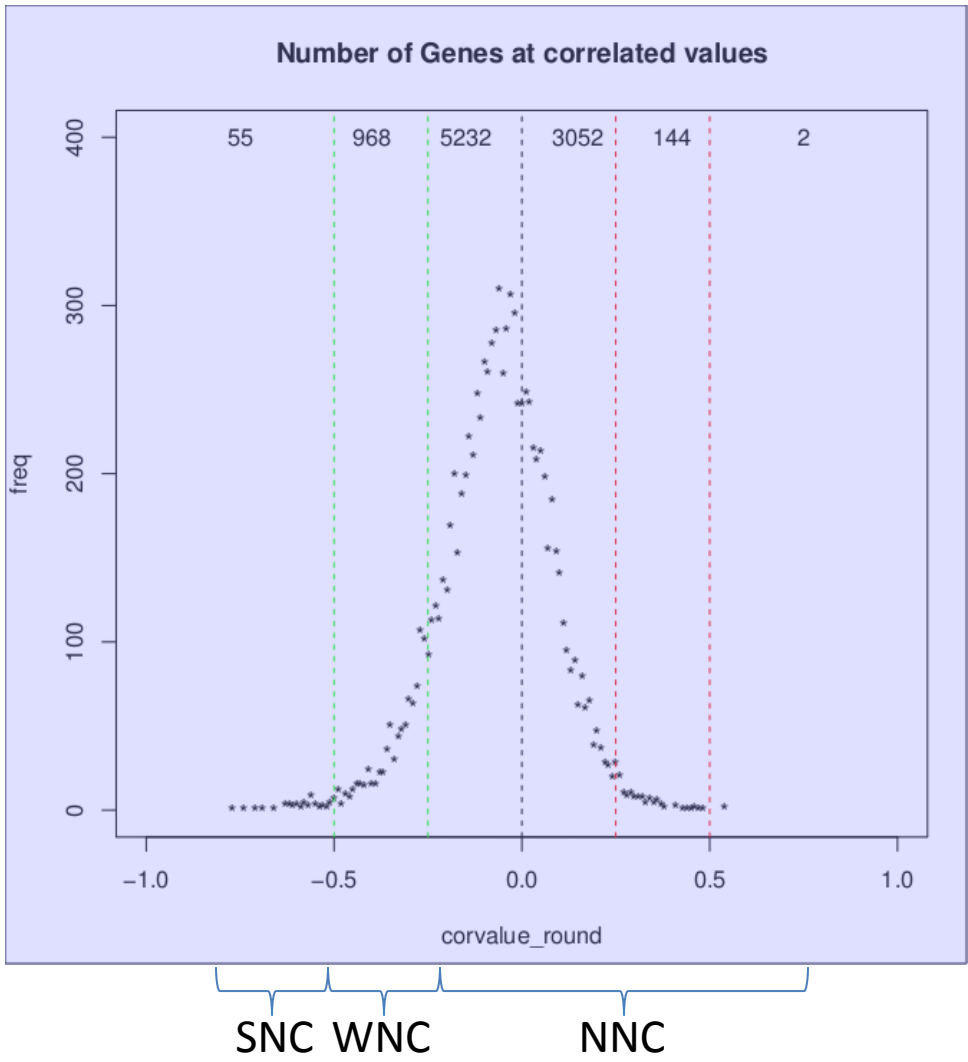
NNC (No Negative Correlation)

> -0.25

450K


Number of Genes at correlated values
1965
5395
3188
450
41
1
freq
0
100
200
300
400
−1.0
−0.5
0.0
0.5
1.0
corvalue_round
SNC
WNC
NNC

27K


Number of Genes at correlated values
55
968
5232
3052
144
2
freq
0
100
200
300
400
−1.0
−0.5
0.0
0.5
1.0
corvalue_round
SNC
WNC
NNC

Rules for DNA methylation Calls/Scores

- 3rd step,
- We took the 10th and 90th percentile beta value across tumor samples for each gene per platform and labeled it:
- **T10, T90.**
- We did the same for normal tissues
- **N10, N90.**
  - When analyzing the 27K platform, we used 4 non-tumor brains
  - When analyzing the 450K platform, we used 24 samples across 3 different normal tissues (72 samples):

    breast, kidney and lung

    Specifically for the 450K, we calculated the 10th, 90th for each tissue type, then calculated the median across gene for each category: 10th, 90th. This is done, so there is no influence on tissue type over another.

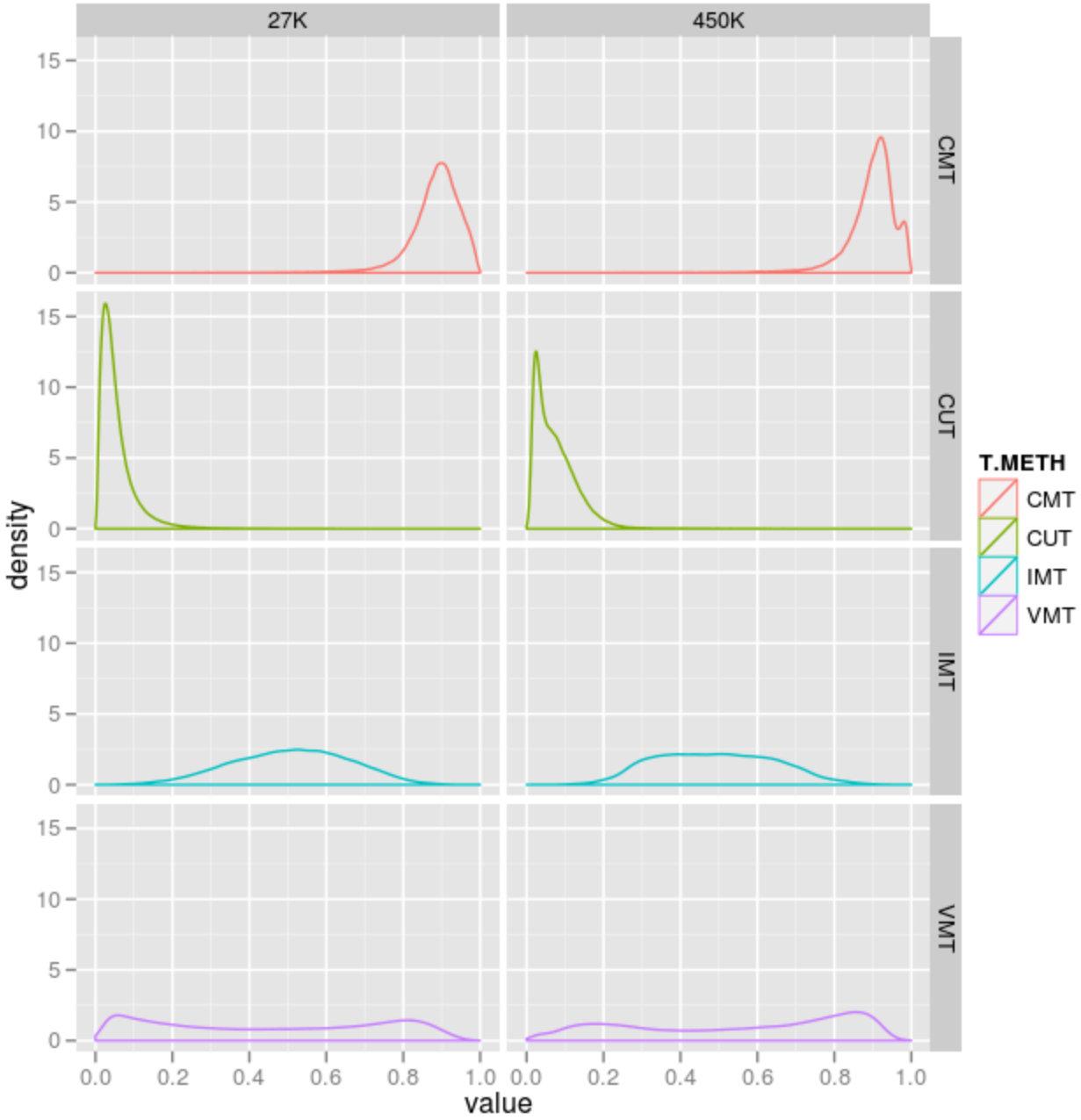
Normal Samples

	Tumor	MatchedNormal27	UnmatchedNormal27	MatchedNormal450	UnmatchedNormal450
1	AML	0	0	0	0
2	BLCA	0	0	11	0
3	LGG	0	0	2	0
4	BRCA	27	0	90	0
5	CESC	0	0	0	0
6	COAD	37	3	0	0
7	GBM	0	4	0	0
8	HNSC	0	0	0	0
9	KIRC	199	0	147	5
10	KIRP	5	0	0	0
11	LIHC	0	0	0	0
12	LUAD	24	0	0	0
13	LUSC	27	0	40	2
14	OV	4	8	0	0
15	PAAD	0	0	0	0
16	PRAD	0	0	0	0
17	READ	5	2	0	0
18	STAD	43	0	2	0
19	THCA	0	0	0	0
20	UCEC	1	1	15	11

Randomly
Selected 24 normals
from each tumor pair
(BRCA, KIRC, LUSC)

Rules for DNA methylation Calls/Scores

- 4^{th},
- Using the 10^{th}, 90^{th} cut-offs for tumor and normals we defined each gene per platform per tissue type (normal and tumor) according to the following rules:
  - If percentile 90 < 0.25 → **CUN or CUT** (constitutively unmethylated in normal or tumor)
  - If percentile 10 > 0.75 → **CMN or CMT** (constitutively methylated in normal or tumor)
  - If percentile 10 > 0.25 && percentile 90 <0.75 → **IMN or IMT** (intermediate methylated in normal or tumor)
  - Else → **VMN or VMT** (variably methylated in normal or tumor


27K
450K
CMT
CUT
IMT
VMT
density
value
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
T.METH
CMT
CUT
IMT
VMT

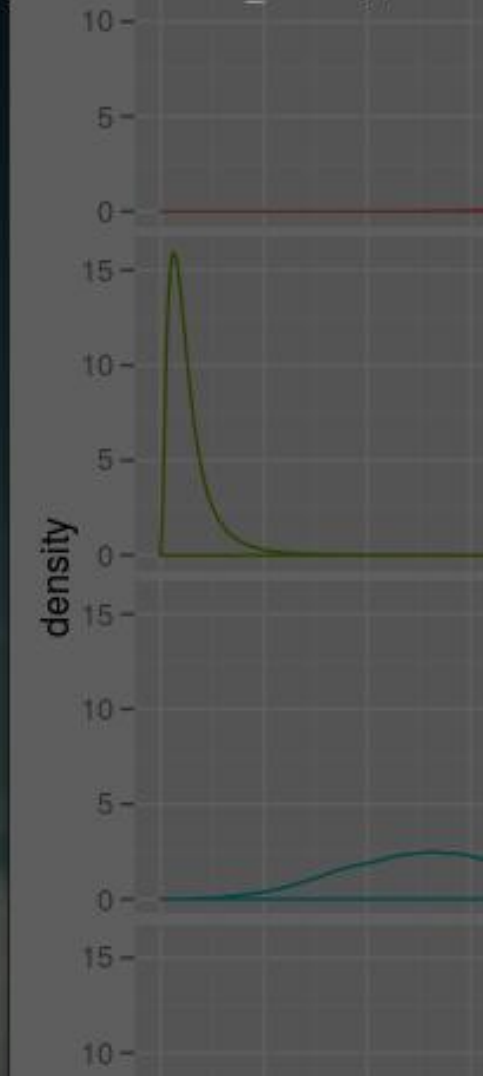
Rules for DNA methylation Calls/Scores

- 5th step,
- Each gene could exist as a pair of one of the combinations
- 48 combinations could exist (3x4x4)

Correlation	**Normal Methylation**	**Tumor Methylation**
SNC	CUN	CUT
WNC	CMN	CMT
NNC	VMN	VMT
	IMN	IMT

450K (74 samples x 11,040 genes) Results

```
> table(genenames.450[,"N.METH"],genenames.450[,"T.METH"], genenames.450[,"correlation_call"])
, ,  = NNC

      CMT  CUT  IMT  VMT
  CMN 556    0    1  240
  CUN   0  781    1   97
  IMN   1    0   35   97
  VMN  71   14   47  597

, ,  = SNC

      CMT  CUT  IMT  VMT
  CMN 777    0    7  268
  CUN   0 1185    2  213
  IMN   9    1   65  160
  VMN 141   38   71  812

, ,  = WNC

      CMT  CUT  IMT  VMT
  CMN 630    0   11  301
  CUN   0 1980    3  338
  IMN   2    1   60  199
  VMN 111   29   95  993
```



27K (279 samples X 9,453 genes) Results

```
table(genenames[,"N.METH"],genenames[,"T.METH"], genenames[,"correlation_call"])
,  = NNC


     CMT  CUT  IMT  VMT
CMN  335    0    2  432
CUN    0 4037    0  599
IMN    7   34  137  973
UNK    0    3   11   68
VMN   41  286   22  826

,  = SNC


     CMT  CUT  IMT  VMT
CMN    3    0    0    7
CUN    0   15    1   44
IMN    0    1   10   66
UNK    0    0    0    0
VMN    0    3    0   29

,  = WNC


     CMT  CUT  IMT  VMT
CMN   16    0    0   29
CUN    0  780    0  199
IMN    2    8   31  207
UNK    0    1    1    7
VMN    3   61    3  113
```

Correlation	Normal Meth	Tumor Meth
1.SNC	1.CUN	1.CUT
1.SNC	1.CUN	2.VMT
1.SNC	1.CUN	3.IMT
1.SNC	1.CUN	4.CMT
1.SNC	2.VMN	1.CUT
1.SNC	2.VMN	2.VMT
1.SNC	2.VMN	3.IMT
1.SNC	2.VMN	4.CMT
1.SNC	3.IMN	1.CUT
1.SNC	3.IMN	2.VMT
1.SNC	3.IMN	3.IMT
1.SNC	3.IMN	4.CMT
1.SNC	4.CMN	1.CUT
1.SNC	4.CMN	2.VMT
1.SNC	4.CMN	3.IMT
1.SNC	4.CMN	4.CMT
2.WNC	1.CUN	1.CUT
2.WNC	1.CUN	2.VMT
2.WNC	1.CUN	3.IMT
2.WNC	1.CUN	4.CMT
2.WNC	2.VMN	1.CUT
2.WNC	2.VMN	2.VMT
2.WNC	2.VMN	3.IMT
2.WNC	2.VMN	4.CMT
2.WNC	3.IMN	1.CUT
2.WNC	3.IMN	2.VMT
2.WNC	3.IMN	3.IMT
2.WNC	3.IMN	4.CMT
2.WNC	4.CMN	1.CUT
2.WNC	4.CMN	2.VMT
2.WNC	4.CMN	3.IMT
2.WNC	4.CMN	4.CMT
3.NNC	1.CUN	1.CUT
3.NNC	1.CUN	2.VMT
3.NNC	1.CUN	3.IMT
3.NNC	1.CUN	4.CMT
3.NNC	2.VMN	1.CUT
3.NNC	2.VMN	2.VMT
3.NNC	2.VMN	3.IMT
3.NNC	2.VMN	4.CMT
3.NNC	3.IMN	1.CUT
3.NNC	3.IMN	2.VMT
3.NNC	3.IMN	3.IMT
3.NNC	3.IMN	4.CMT
3.NNC	4.CMN	1.CUT
3.NNC	4.CMN	2.VMT
3.NNC	4.CMN	3.IMT
3.NNC	4.CMN	4.CMT

Rules for DNA methylation Calls/Scores

- 5th step cont.
- Need to decide appropriate “call” and “score” labels.

Call	Desc
MG	Methylation gain compared to normal
ML	Methylation loss compared to normal
MT	Methylated in tumor
UT	Unmethylated in tumor
ES	Epigenetically silenced
UC	Unable to make call

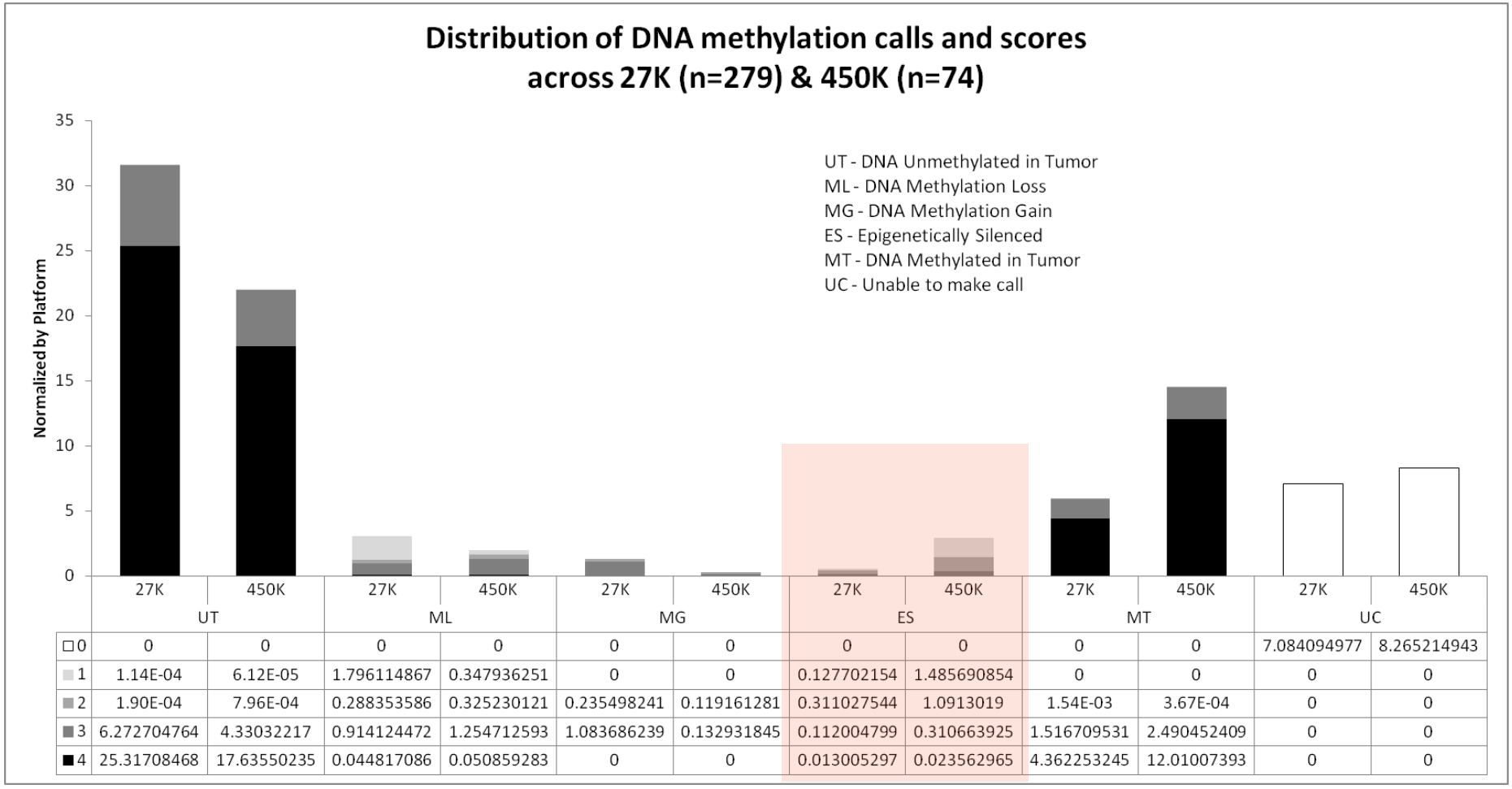
Score	Desc (confidence)
4	High
3	Med-High
2	Med
1	Low
0	No call

Correlation	Normal Meth	Tumor Meth	Call if TM < 0.25	Confidence Score if TM < 0.25	Call if 0.25 < TM < 0.75	Confidence Score if 0.25 < TM < 0.75	Call if TM > 0.75	Confidence Score if TM > 0.75
1.SNC	1.CUN	1.CUT	UT	4	ES	1	ES	2
1.SNC	1.CUN	2.VMT	UT	3	ES	3	ES	4
1.SNC	1.CUN	3.IMT	UT	3	ES	2	ES	3
1.SNC	1.CUN	4.CMT	UT	3	ES	3	ES	4
1.SNC	2.VMN	1.CUT	ML	1	UC	0	ES	1
1.SNC	2.VMN	2.VMT	UT	3	UC	0	ES	1
1.SNC	2.VMN	3.IMT	UT	3	UC	0	ES	1
1.SNC	2.VMN	4.CMT	UT	2	UC	0	ES	2
1.SNC	3.IMN	1.CUT	ML	2	UC	0	ES	1
1.SNC	3.IMN	2.VMT	UT	3	UC	0	ES	1
1.SNC	3.IMN	3.IMT	UT	3	UC	0	ES	1
1.SNC	3.IMN	4.CMT	UT	1	UC	0	ES	2
1.SNC	4.CMN	1.CUT	ML	4	ML	3	MT	2
1.SNC	4.CMN	2.VMT	ML	4	ML	3	MT	3
1.SNC	4.CMN	3.IMT	ML	3	ML	2	MT	3
1.SNC	4.CMN	4.CMT	ML	1	ML	2	MT	4
2.WNC	1.CUN	1.CUT	UT	4	MG	2	MG	2
2.WNC	1.CUN	2.VMT	UT	3	ES	2	ES	3
2.WNC	1.CUN	3.IMT	UT	3	ES	1	ES	2
2.WNC	1.CUN	4.CMT	UT	3	ES	2	ES	3
2.WNC	2.VMN	1.CUT	ML	1	UC	0	MT	2
2.WNC	2.VMN	2.VMT	UT	3	UC	0	MT	4
2.WNC	2.VMN	3.IMT	UT	3	UC	0	MT	4
2.WNC	2.VMN	4.CMT	UT	2	UC	0	MT	4
2.WNC	3.IMN	1.CUT	ML	2	UC	0	MT	2
2.WNC	3.IMN	2.VMT	UT	3	UC	0	MT	4
2.WNC	3.IMN	3.IMT	UT	3	UC	0	MT	4
2.WNC	3.IMN	4.CMT	UT	1	UC	0	MT	4
2.WNC	4.CMN	1.CUT	ML	4	ML	3	MT	2
2.WNC	4.CMN	2.VMT	ML	4	ML	3	MT	3
2.WNC	4.CMN	3.IMT	ML	3	ML	2	MT	3
2.WNC	4.CMN	4.CMT	ML	1	ML	2	MT	4
3.NNC	1.CUN	1.CUT	UT	4	MG	2	MG	2
3.NNC	1.CUN	2.VMT	UT	3	MG	3	MG	3
3.NNC	1.CUN	3.IMT	UT	3	MG	2	MG	3
3.NNC	1.CUN	4.CMT	UT	3	MG	3	MG	4
3.NNC	2.VMN	1.CUT	ML	1	UC	0	MT	2
3.NNC	2.VMN	2.VMT	UT	3	UC	0	MT	4
3.NNC	2.VMN	3.IMT	UT	3	UC	0	MT	4
3.NNC	2.VMN	4.CMT	UT	2	UC	0	MT	4
3.NNC	3.IMN	1.CUT	ML	2	UC	0	MT	2
3.NNC	3.IMN	2.VMT	UT	3	UC	0	MT	4
3.NNC	3.IMN	3.IMT	UT	3	UC	0	MT	4
3.NNC	3.IMN	4.CMT	UT	1	UC	0	MT	4
3.NNC	4.CMN	1.CUT	ML	4	ML	3	MT	2
3.NNC	4.CMN	2.VMT	ML	4	ML	3	MT	3
3.NNC	4.CMN	3.IMT	ML	3	ML	2	MT	3
3.NNC	4.CMN	4.CMT	ML	1	ML	2	MT	4

Rules for DNA methylation Calls/Scores

- 6^{th} step,
- Using the table rule (from step 5) as reference, we made “calls” on each gene per sample, see R script for exact steps.
- Basically, each cell in the DNA methylation data matrix (samps X genes) is evaluated.
  - Beta-values < 0.25
  - Beta-values > 0.25 & < 0.75
  - Beta-values > 0.75

27K: 279 samp X 9453 (x2) = 5,274,774
450K: 74 samp X 11040 (x2) = 1,633,920



	UT		ML		MG		ES		MT		UC	
	27K	450K	27K	450K	27K	450K	27K	450K	27K	450K	27K	450K
□0	0	0	0	0	0	0	0	0	0	0	7.084094977	8.265214943
1	1.14E-04	6.12E-05	1.796114867	0.347936251	0	0	0.127702154	1.485690854	0	0	0	0
2	1.90E-04	7.96E-04	0.288353586	0.325230121	0.235498241	0.119161281	0.311027544	1.0913019	1.54E-03	3.67E-04	0	0
3	6.272704764	4.33032217	0.914124472	1.254712593	1.083686239	0.132931845	0.112004799	0.310663925	1.516709531	2.490452409	0	0
■4	25.31708468	17.63550235	0.044817086	0.050859283	0	0	0.013005297	0.023562965	4.362253245	12.01007393	0	0

Data and code available in dropbox

- Finally,
- Merged 27K/450K based on geneID.
- Added data/codes etc. to dropbox.
- See following path: “Dropbox/TCGA_GBM_MS_Writing_Group/Working_Group_SampleManifest/DNA_methylation”
- R file (which contains the entire data matrix with calls, scores), R scripts (most of the code is listed), exported data as tab-deliminated file (for those non R users). “Calls” and “Scores” are separated in two data matrix.
- R file: “TCGA_GBM_DNAMETHYLATION_CALLS_SCORES_20120112_Noushmehr_ver2.rda”