Supplementary Data for “Accounting for Differential Variability in Detecting Differentially Methylated Regions”

Investigation of the distance limits to define clusters

We investigated the relationship between the choice of distance limit, the maximal distance between two neighboring sites to be included in a cluster, and the distribution of the difference in the combined signals scores (before smoothing) between neighboring CpG sites using TCGA BRCA data of tumor and normal-adjacent tissues.

Within each chromosome, we ordered the combined signal scores by their genomic locations and calculated the differences in the combined signals scores between neighboring CpG sites. We then change the distance limits from 300 bp to 2,000 bp (300 bps, 500 bps, 700 bps, 1,000 bps, 1,500 bps, and 2,000 bps) and plotted the distribution of the differences for neighboring CpG sites whose distance is less than the specified distance limit (Figure S1). We found that the mean and SD in the difference in combined signal scores between neighboring CpG sites increases as the distance limits increases. In the developed algorithm, users could choose other distance limits as an option.

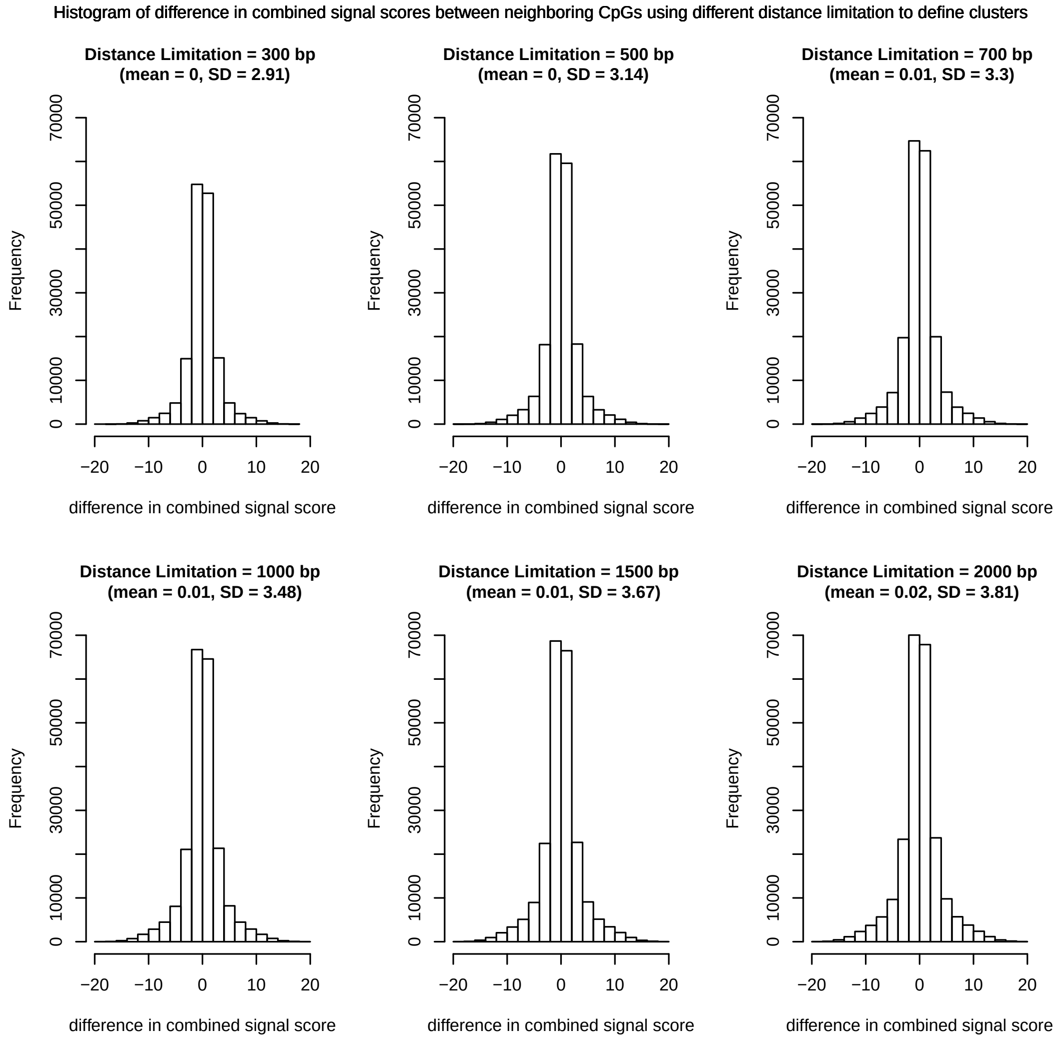

Figure S1. Histogram of the difference in the combined signal scores for neighboring CpG sites with the choice of difference distance limits.

Simulation studies for case-control designs

We adapted the proposed new DMR detection method for case-control designs and conducted simulation studies parallel as for matched case-control designs in the main text to evaluate the type I errors and the performance. We compared the performance of the new method with those DMR detection methods that consider 1) mean signals only using two-sided two-sample t-test, the bump hunting method [1], the modified bump hunting method which divide the regression coefficient estimates from the original bump hunting method by their standard errors; and 2) variance signals only using one-sided F-test, where we applied the same smoothing step and the same significance assessment step.

Simulation setup

To simulate DNA methylation measures of tumor and normal tissues, we assume logit transformed methylation measures of sample s follows a multivariate normal distribution

$$M_{s,k} \sim N_{l_k}(\boldsymbol{\mu}, \Delta^T \Sigma \Delta) \tag{1}$$

where l_k is the size of the k-th cluster, and the mean vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_{l_k})^T$ and diagonal matrix $\Delta = \text{diag}(\sqrt{\delta_1}, ..., \sqrt{\delta_{l_k}})$ controls the mean and variance signals. Σ is a variance-covariance matrix ($l_k \times l_k$) considering correlations among l_k CpG sites within the k-th pre-defined cluster. Here we assume an $AR(1)$ correlation with correlation coefficient ρ, i.e., $\Sigma_{mn} = \sigma \times \rho^{|m-n|}$. We set $\rho = 0.5$ in the simulation studies similarly as in matched case-control designs, and set $\sigma = 0.25$. In each simulation, we generated logit transformed methylation values of 10,000 sites from 100 cancer patients and 100 normal controls., where the genomic locations of these 10,000 sites are the first 10,000 sites of Chromosome 1 on the

Illumina 450K array.

Since DNA methylation measures are known to be associated with variables such as age [2, 3] and gender [4], we work on methylation residuals after adjusting for such confounders. We investigated type I errors to examine if using methylation residuals controls potential spurious DMRs due to unbalanced distribution of confounders, such as gender. More specifically, we set 50% of cancer patients to be female while only 20% of normal controls to be female. We simulated 10 spurious DMRs each having 10 CpG sites, within which we set $\mu = 1$ for both tumor and normal tissues in the female group, while $\mu = 0$ for both tumor and normal tissues in the male group. For all other sites, we set $\mu = 0$ for tumor and normal tissues in both gender groups. We considered two scenarios where we applied the new method on: 1) methylation residuals obtained from regressing logit transformed methylation values on gender using linear models, and 2) logit transformed methylation values directly ignoring gender. We conducted 1000 simulations in each scenario.

In sections to evaluate the performance of the new method, we assume confounders are already accounted for when methylation residuals are used. We simulated 10 true DMRs with different region sizes varying from 3 to 15 CpG sites, and we considered scenarios when each CpG site in the true DMRs has 1) mean signals only, 2) variance signals only, and 3) both mean and variance signals. For all other null sites, we set $\mu = 0$ and $\sigma = 0.25$. For each simulation scenario, we conducted 1000 simulations.

Simulation results

The type I errors of the new method that considers both mean and variance signals, the two-sample t-test that considers mean signals only, and the F-test that considers variance signals only were all well controlled at 0.039, 0.023 and 0.059 when applied to methylation residuals. When applied to the methylation measures ignoring the gender effect, the type I errors were all inflated at 1.000, 1.000 and 0.997. The type I errors of bump hunting [1] and modified bump hunting that directly adjust for gender effect were both well controlled at 0.041. The region size was set at $L \geq 3$ CpG sites.

The ROC curves from the setting with 10 true DMRs having different region sizes are shown in Figure S2. Similar patterns as in matched case-control designs are observed.

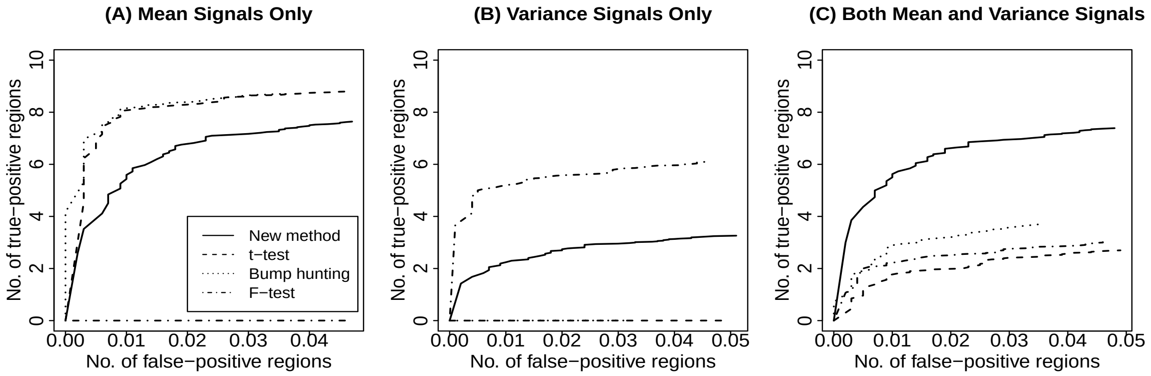

Figure S2. ROC curves from simulation studies when 10 true DMRs have different region sizes varying from 3 to 15 CpG sites with: (A) mean signals only; (B) variance signals only; and (C) both mean and variance signals. DMRs were defined as regions with minimum region size $L \geq 3$ CpG sites.

Real data application

Table S1. Significant DMRs Identified in the TCGA KIRC Data (160 matched pairs)

DMRs ($L^a \geq 3$)	New method	Wilcoxon signed-rank test	Paired t-test	DMRcate	Probe Lasso	Pitman-Morgan test	KS test [b]
Total No. of DMRs (Total No. of DMR-covered CpG sites)	2164 (30558)	146 (4162)	2953 (33786)	23332 (153192)	7996 (41195)	1457 (25070)	1697 (28954)
Mean(SD) number of CpG sites per DMR	14 (7)	29 (7)	11 (6)	7(5)	5 (5)	17 (10)	17 (9)
Mean(SD) number of base pairs per DMR	2575 (1962)	5007 (2489)	2074 (1464)	1207 (1076)	772 (1166)	3226 (2139)	3239 (2127)
No. of overlapping DMRs[c]	-	146	1716	2164	1185	1292	1332

[a] L: minimum region size, i.e., minimum number of CpG sites

[b] Kolmogorov Smirnov test

[c] No. of overlapping DMRs: a DMR identified by the new method is considered to overlap if this DMR has any overlap with DMRs identified by each comparing method.

Table S2. 7 Cancer-Related Genes Identified in the Top 10 Ranked DMRs in TCGA KIRC Data[a]

Cancer	Gene
Breast Cancer	*MCF2L2* [5]
Colorectal Cancer	*GAD2* [6]
Endometrial Carcinoma	*PAX2* [7]
Hepatocellular Carcinoma	*DCAF4L2* [8]
Melanoma	*GPR98* [9]
Pancreatic Cancer	*FOXL1* [10]
Stomach Cancer	*RIMS2* [11]

a: There are 10 genes in the top 10 ranked DMRs out of 100 significant DMRs that were uniquely identified by the new method (compared with all five competing methods except for DMRcate), and 7 genes were previously reported to be cancer-related.

Table S3. 11 Cancer-Related Genes Identified in the Top 10 Ranked DMRs in GEO BRCA Data (Tumor vs. Normal-adjacent) [a]

Cancer	**Gene**
Breast Cancer	*CBX8* [12], *NXPH1* [13]
Colorectal Cancer	*VIM* [14], *WNT1* [15]
Gastric Cancer	*FOXD3* [16], *RASGRF1* [17]
Lung Cancer	*C6orf176* [18]
Ovarian Cancer	*HIST1H3G* [19], *HIST1H2BI* [20], *VCAN* [21]
Prostate Cancer	*GFRA1* [22]

[a]: There are 12 genes in the top 10 ranked DMRs out of 37 significant DMRs that were uniquely identified by the new method (compared with all five competing methods except for DMRcate), and 11 genes were previously reported to be cancer-related.

Table S4. Significant DMRs Identified in the GEO BRCA Data (Tumor vs. Normal)

DMRs ($L^a \geq 3$)	**New method**	***t*-test**	**Bump hunting**	**Modified bump hunting**	**F-test**
Total No. of DMRs (Total No. of DMR-covered CpG sites)	830 (15692)	2097 (28384)	683 (11445)	860 (14600)	94 (2537)
Mean (SD) number of CpG sites per DMR	19 (9)	14 (7)	17 (8)	17 (8)	27 (12)
Mean (SD) number of base pairs per DMR	3276 (1955)	2418 (1539)	2898 (1739)	2974 (1754)	4047 (2819)
No. of overlapping DMRs[b]	-	811	498	529	85

[a] *L*: minimum region size, i.e., minimum number of CpG sites

[b] No. of overlapping DMRs: number of DMRs identified by the new method that has any overlap with DMRs identified by each comparing method.

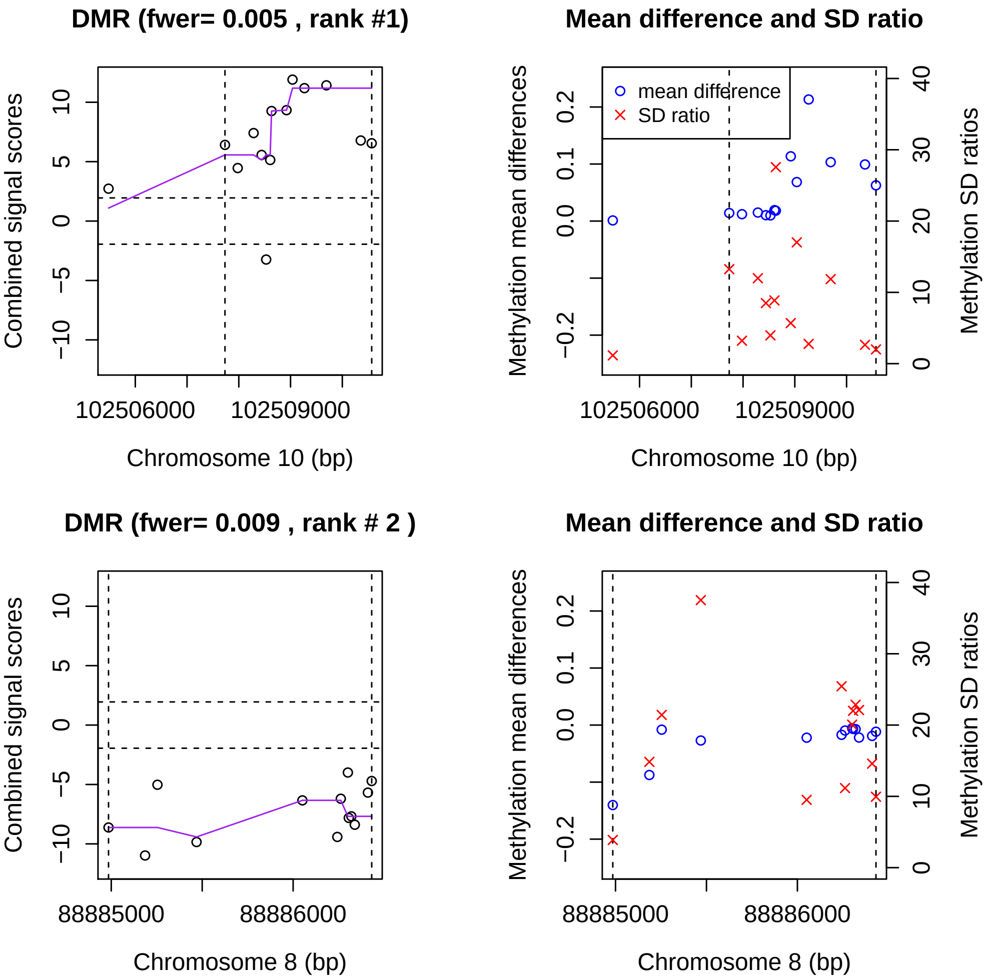

Figure S3. DMR #1 (top row) and #2 (bottom row) located on chromosomes 6 and 11 (out of 170 DMRs) that were identified uniquely by the new method in the TCGA KIRC data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between tumor and normal-adjacent tissues.

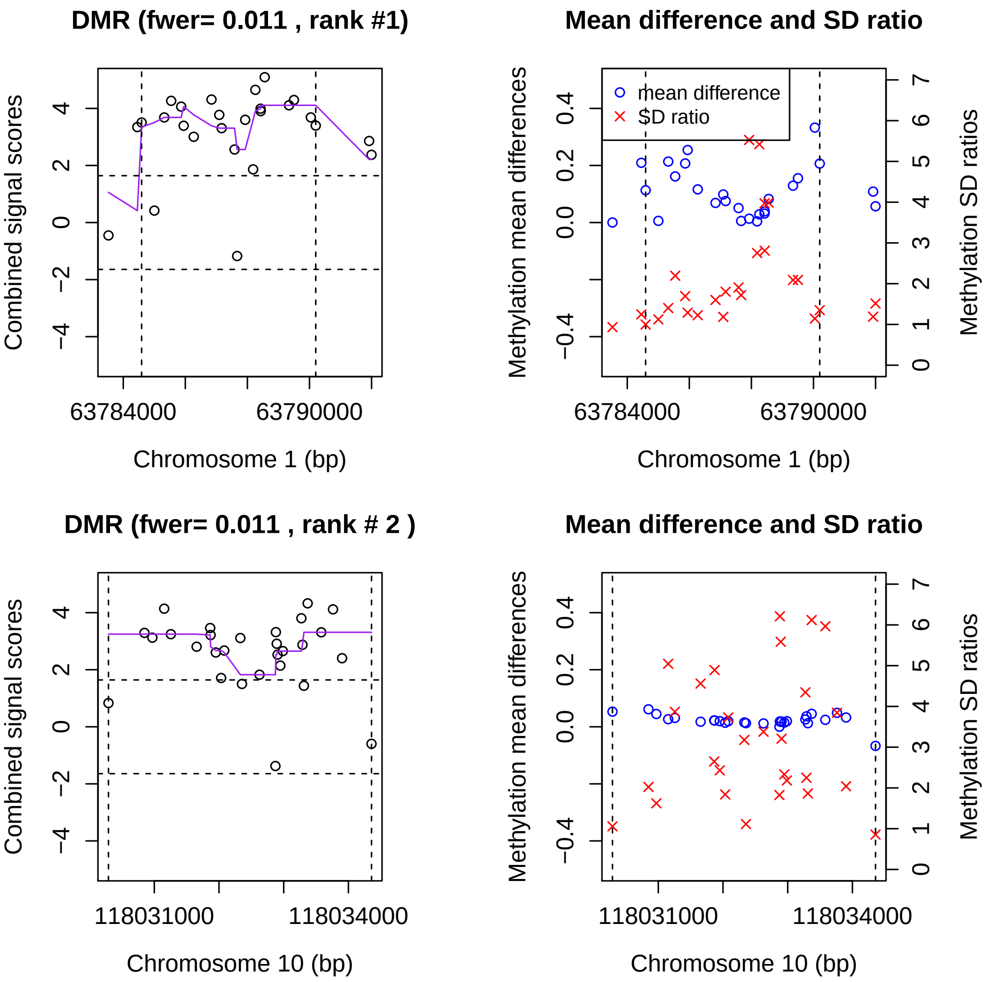

Figure S4. DMR #1 (top row) and #2 (bottom row) located on chromosomes 7 and 4 (out of 89 DMRs) that were identified uniquely by the new method in the GEO BRCA tumor vs. normal-adjacent data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between tumor and normal-adjacent tissues. There are 3 gene, *SGCE*, *PEG10* and *PHOX2B* in these 2 DMRs. *SGCE* was reported to be associated with colorectal cancer [23], *PEG10* was reported to be associated with hepatocellular carcinoma [24], and *PHOX2B* was reported to be associated with neuroblastoma [25]

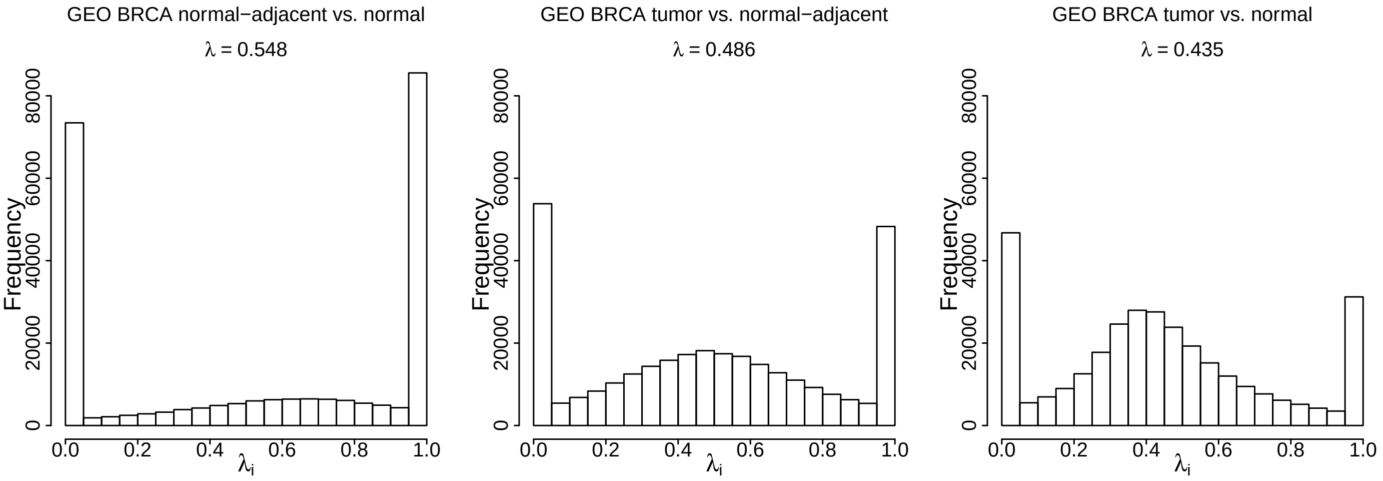

Figure S5. Distributions of genome-wide site-level scale parameter λ_i in the GEO BRCA data. From left to right shows distribution of λ_i in (1) normal-adjacent vs. normal, (2) tumor vs. normal-adjacent, and (3) tumor vs. normal comparisons.

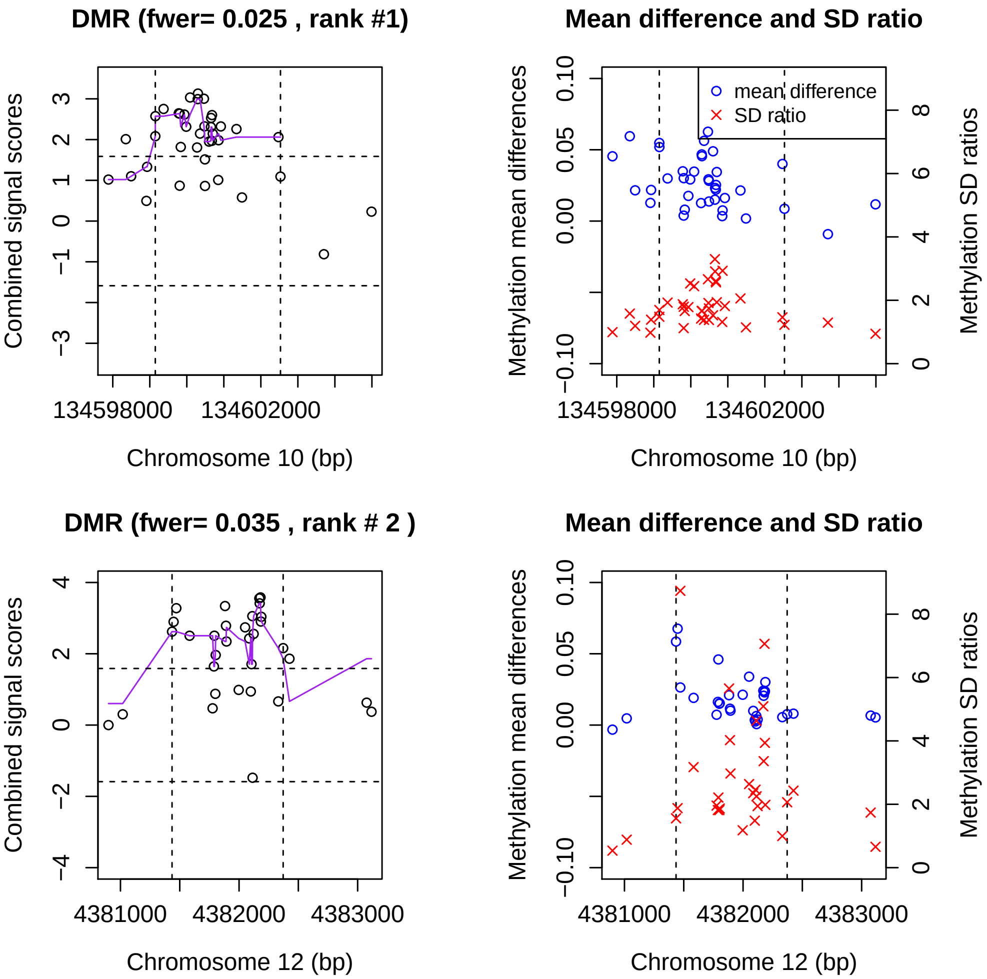

Figure S6. DMR #1 (top row) and #2 (bottom row) located on chromosomes 10 and 12 that were identified uniquely by the new method in the GEO BRCA normal-adjacent vs. normal data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between normal-adjacent and normal tissues.

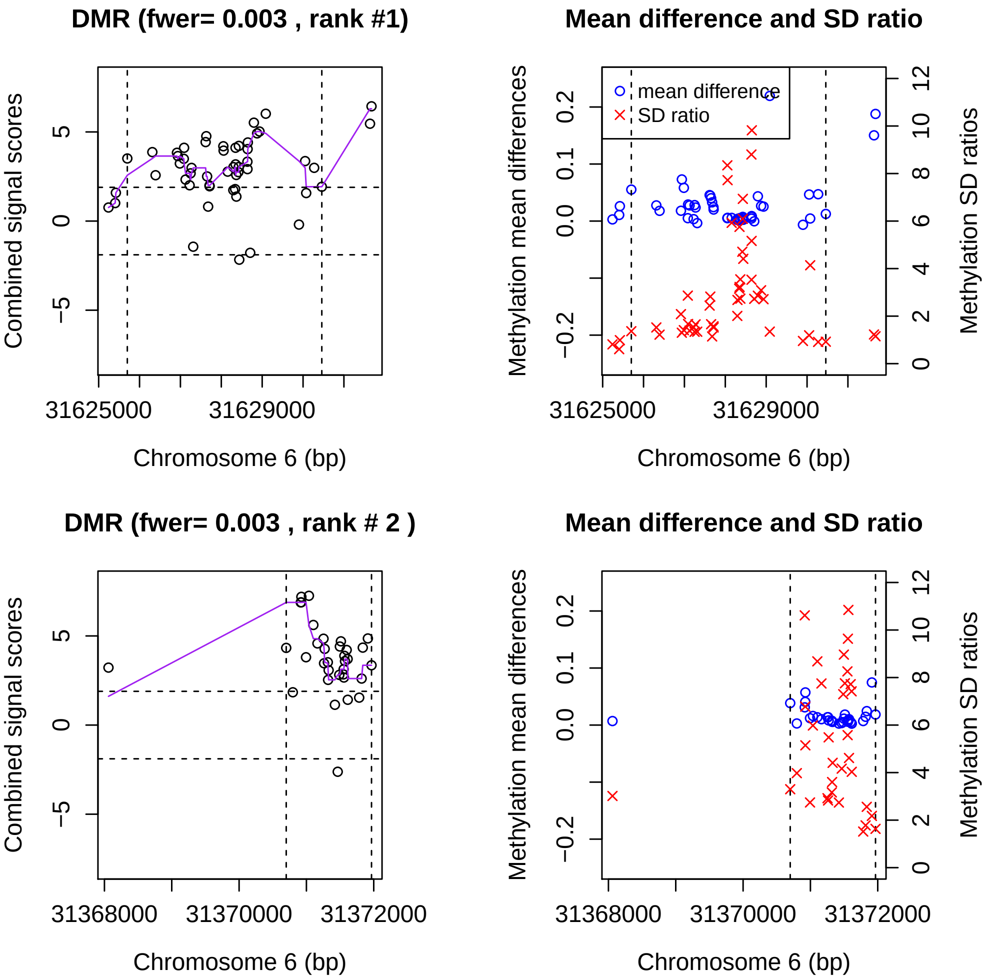

Figure S7. DMR #1 (top row) and #2 (bottom row) located on chromosomes 6 (out of 15 DMRs) that were identified uniquely by the new method in the GEO BRCA tumor vs. normal data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between tumor and normal tissues.

References

1. Jaffe AE, Murakami P, Lee H, et al. Bump hunting to identifying differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, 41(1):200–209, 2012.

2. Christensen BC, Houseman EA, Marsit CJ, et al. Aging and environmental exposures alter tissue-specific dna methylation dependent upon cpg island context. *PLoS Genetics*, 5(8):e1000602, 2009.

3. Teschendorff AE, Menon U, Gentry-Maharaj A, et al. Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research*, 20(4):440–446, 2010.

4. Liu J, Morgan M, Hutchison K, et al. A study of the influence of sex on genome wide methylation. *PLoS One*, 5(4):e10028, 2010.

5. Legendre C, Gooden GC, Johnson K, et al. Whole-genome bisulfite sequencing of cell-free dna identifies signature associated with metastatic breast cancer. *Clinical epigenetics*, 7(1):100, 2015.

6. Li H, Du Y, Zhang D, et al. Identification of novel dna methylation markers in colorectal cancer using mira-based microarrays. *Oncology reports*, 28(1):99–104, 2012.

7. Wu H, Chen Y, Liang J, et al. Hypomethylation-linked activation of pax2 mediates tamoxifen-stimulated endometrial carcinogenesis. *Nature*, 438(7070):981–987, 2005.

8. Song MA, Tiirikainen M, Kwee S, et al. Elucidating the landscape of aberrant dna methylation in hepatocellular carcinoma. *PloS one*, 8(2):e55761, 2013.

9. Harvey KF, Zhang X, and Thomas DM. The hippo pathway and human cancer. *Nature Reviews Cancer*, 13(4):246–257, 2013.

10. Zhang G, He P, Gaedcke J, et al. Foxl1, a novel candidate tumor suppressor, inhibits tumor aggressiveness and predicts outcome in human pancreatic cancer. *Cancer research*, 2013.

11. Ewing AD, Gacita A, Wood LD, et al. Widespread somatic l1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome research*, 25(10):1536–1545, 2015.

12. Lee SH, Um SJ, and Kim EJ. Cbx8 suppresses sirtinol-induced premature senescence in human breast cancer cells via cooperation with sirt1. *Cancer letters*, 335(2):397–403, 2013.

13. Faryna M, Konermann C, Aulmann S, et al. Genome-wide methylation screen in low-grade breast cancer identifies novel epigenetically altered genes as potential biomarkers for tumor diagnosis. *The FASEB Journal*, 26(12):4937–4950, 2012.

14. Costa VL, Henrique R, Danielsen SA, et al. Three epigenetic biomarkers, gdf15, tmeff2 and vim, accurately predict bladder cancer from dna-based analyses of urine samples. *Clinical Cancer Research*, pages clincanres–1312, 2010.

15. He B, Reguart N, You L, et al. Blockade of wnt-1 signaling induces apoptosis in human colorectal cancer cells containing downstream mutations. *Oncogene*, 24(18):3054–3058, 2005.

16. Cheng AS, Li MS, Kang W, et al. Helicobacter pylori causes epigenetic dysregulation of foxd3 to promote gastric carcinogenesis. *Gastroenterology*, 144(1):122–133, 2013.

17. Takamaru H, Yamamoto E, Suzuki H, et al. Aberrant methylation of rasgrf1 is associated with an epigenetic field defect and increased risk of gastric cancer. *Cancer Prevention Research*, 5(10):1203–1212, 2012.

18. Chen Z, Li JL, Lin S, et al. camp/creb-regulated linc00473 marks lkb1-inactivated lung cancer and mediates tumor growth. *The Journal of clinical investigation*, 126(6):2267, 2016.

19. Zhang M and Luo S. Gene expression profiling of epithelial ovarian cancer reveals key genes and pathways associated with chemotherapy resistance. *Genet Mol Res*, 15(1):11, 2016.

20. Hong S, Dong H, Jin L, et al. Gene co-expression network analysis of two ovarian cancer datasets. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on*, pages 269–274. IEEE, 2010.

21. Ghosh S, Albitar L, LeBaron R, et al. Up-regulation of stromal versican expression in advanced stage serous ovarian cancer. *Gynecologic oncology*, 119(1):114–120, 2010.

22. Huber RM, Lucas JM, Gomez-Sarosi LA, et al. Dna damage induces gdnf secretion in the tumor microenvironment with paracrine effects promoting prostate cancer treatment resistance. *Oncotarget*, 6(4):2134, 2015.

23. Ortega P, Moran A, Fernandez-Marcelo T, et al. Mmp-7 and sgce as distinctive molecular

factors in sporadic colorectal cancers from the mutator phenotype pathway. *International journal of oncology*, 36(5):1209, 2010.

24. Ip WK, Lai PBS, Wong NLY, et al. Identification of peg10 as a progression related biomarker for hepatocellular carcinoma. *Cancer letters*, 250(2):284–291, 2007.

25. De Pontual L, Trochet D, Bourdeaut F, et al. Methylation-associated phox2b gene silencing is a rare event in human neuroblastoma. *European Journal of Cancer*, 43(16):2366–2372, 2007.