Supplementary Materials for "The interplay between somatic copy number aberration, DNA methylation, and gene expression"

Contents

A Supplementary Methods

A.1 Data preparation

A.1.1 Somatic copy number aberration (SCNA)

We downloaded from the TCGA data portal level 3 SCNA data (generated using the Affymetric 6.0 array) (`https://tcga-data.nci.nih.gov/tcga/`). For most patients, there were four files: `*.hg18.seg.txt`, `*.nocnv_hg18.seg.txt`, `*.hg19.seg.txt*`, and `*.nocnv_hg19.seg.txt`. We used the results from file `*.nocnv_hg19.seg.txt` because this file contained a more recent genome coordinate (hg19) and removed the effect of germline copy number variation (nocnv). The level 3 SCNA data are represented as segmental mean values across the genome. Each segment usually covers a large genomic region of tens or hundreds of genes. We recorded the copy number measurement for each gene as the segmental mean of the segment where the gene is located. Gene location information was extracted from file `TCGA.hg19.June2011.gaf`, which was downloaded from `http://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle/outputs/TCGA.hg19.June2011.gaf`.

A.1.2 DNA methylation data

We used DNA methylation data from the Illumina 450k array (HM450) or Illumina 27k array (HM27). For either platform, the level 3 DNA methylation data were downloaded from the TCGA data portal. Each patient has one file containing the beta-values. We transferred the beta-values (β) to m-values ($m = \log[\beta/(1-\beta)]$) for all computation completed in this paper. For each cancer type, we removed probes with more than 5% of missing values.

For all cancer types except glioblastoma, the data were generated using the HM450 platform. For glioblastoma, the first cancer type collected by the TCGA, the data were generated using either the HM450 or HM27 platform. We combined the data derived from the two platforms, but only used data on common probes. We only used the probes that are shared between HM450 and HM27 arrays. We calibrated the difference of these two platforms using the following procedures. For all probes that were shared by the two platforms, we calculated the median values (at m-value scale) across all samples. Let n be the total number of probes shared by HM450 and HM27. Denote the methylation of the j-th HM27 probe in the i-th individual as y_{ij}. Denote the median methylation of the j-th probe from HM450 and HM27 by x_j and y_j, respectively. We fit a loess regression $\hat{y}_j = f(x_j)$.

Then, we adjusted the median methylation of the HM27 platform to match the median methylation of the HM450 platform based on this loess fit: $y'_{ij} = y_{ij} - (\hat{y}_j - x_j)$. This procedure is done separately for probes with 1, 2, 3, 4, 5, and > 5 CpGs using the R function `loess`. We kept the default smoothing parameters (span of 0.75 and degree of 2) and set parameter surface="direct".

When summarizing the results with respect to different classes of methylation probes (e.g., CpG island, genomic location with respect to a gene), we used the annotation HM450 array probes in file `HumanMethylation450_15017482_v1-2.csv`, which was downloaded from the Illumina website (`http://support.illumina.com/downloads/humanmethylation450_15017482_v1-2_product_files.html`). We further selected those probes with hg19 location (`Genome_Build="37"`) and without SNP within 10bp (`nfo$Probe_SNPs_10=""`) or beyond 10bp (`info$Probe_SNPs=""`). In total, 393,401 of 486,428 probes within this annotation file survived these filters.

A.1.3 Gene expression data

For glioblastoma, we used gene expression data from the HG133 array. For all other cancer types, we used gene expression data from the RNAseqV2 pipeline. Specifically, we extracted "raw_count" from the level 3 data file named `*.rsem.genes.results`. Many genes may not be adequately expressed for downstream analysis. For each type of cancer, we selected those genes whose 75th percentile for expression (measured by raw counts) is larger than 20 for the following analysis. Using this criterion, we identified ∼15,000 expressed genes. We subsequently log-transformed these raw counts after dividing them by a measurement of read-depth. For each sample, one can measure the read-depth using several metrics, such as the total number of reads or 75th percentile for expression across all genes. We used the latter because it is more robust to the influence of genes with extremely high expression [1].

A.1.4 Demographic data

Data on race/ethnicity are often missing from the TCGA data set. We estimated race by applying principal component analysis (PCA) on the genotype data [2] using the following steps.

1. Downloaded raw data (CEL files of Affymetrix 6.0 arrays) from the TCGA data portal. These are controlled access data; we obtained our permission through dbGAP

data access application process.

2. Used the `apt-geno-qc` command (which is part of the Affymetrix power tools) for quality control (QC). Those samples with `contrast.qc>0.4` and `qc.call.rate.all>0.8` were retained for the next step.

3. Used the `apt-probeset-genotype` command to call genotypes with the option `-a birdseed-v2`. Removed samples with low call rates if there were any such cases.

4. Removed SNPs with more than 5% missing values, and swaped the SNP strand to obtain the genotype from the forward strand.

5. Pruned SNPs. Specifically, calculated the pair-wise R^2 for all SNP pairs within a sliding window of 50 SNPs and step size of 5 SNPs, and removed one of a pair of SNPs if the R^2 was larger than 0.1.

6. Performed PCA using the TCGA samples together with HapMap samples.

7. Chose cutoffs for PC1 and PC2 to select Caucasian samples by manual inspection.

8. Performed PCA again only for those selected Caucasian samples. Selected a few top PCs as demographic covariates that described population stratification.

In addition to genotype PCs, we included age and gender (when patients of both genders are included) as covariates in our association analysis. Age and gender information were obtained from clinical data that were also downloaded from the TCGA data portal.

A.2 Evaluation of the association between SCNA, DNA methylation, and gene expression

We evaluated the association between SCNA, DNA methylation, and gene expression for each cancer type separately using linear regression. We performed all three pair-wise association studies, i.e., SCNA vs. gene expression, SCNA vs. DNA methylation, and DNA methylation vs. gene expression. For each association study, we included several technical and demographic covariates, as well as tumor subtype if available. The technical covariates were indicators for batches, including tissue sites and plates. The demographic covariates were sex, age, and a few PCs from genotype data that account for possible population stratification. Known cancer subtypes are available for some cancer types. In addition, when we evaluated the association between DNA methylation and gene expression, we also

accounted for copy number changes. Specifically, we first regressed the expression of each gene with its copy number and obtained residuals. Then, we used the residuals to asses the association between DNA methylation and gene expression. The specific models and the set of covariates that we used are presented in the following session titled Supplementary Results.

The sample size varies across cancer types, ranging from 100 to 400 . Both gene expression and copy number were measured at the gene level, and DNA methylation was measured for each probe of the methylation array. For each cancer type except glioblastoma, there were approximately 15,000 genes and 400,000 methylation probes. Because older platforms are used for glioblastoma, the number of genes was about 9,100 and the number of methylation probes was about 18,500. The amount of available data presented a huge computational burden. For example, to study the association between DNA methylation and gene expression, we needed to evaluate 15,000 genes $\times$ 400,000 methylation probes = 6 billion models. We used R package MatrixEQTL [3] for this computation, which takes about 2 hours to evaluate 6 billion models.

B Supplementary Results

B.1 Breast Cancer (BRCA)

B.1.1 Sample selection

Figure S1 shows the sample selection process for breast cancer patients. We selected samples from patients who were Caucasian females and whose samples included data on SCNA, DNA methylation, and gene expression. We further selected samples with purity information [4] in order to demonstrate the effect of purity on the association between gene expression and DNA methylation. Specifically, the purity information was obtained from file `pancan12.sample_info.txt`, which was downloaded from `https://www.synapse.org/#!Synapse:syn1710466/version/2`. Breast cancer subtypes can be defined by gene expression [5]; we obtained subtype information from the laboratory of Dr. Charles Perou.

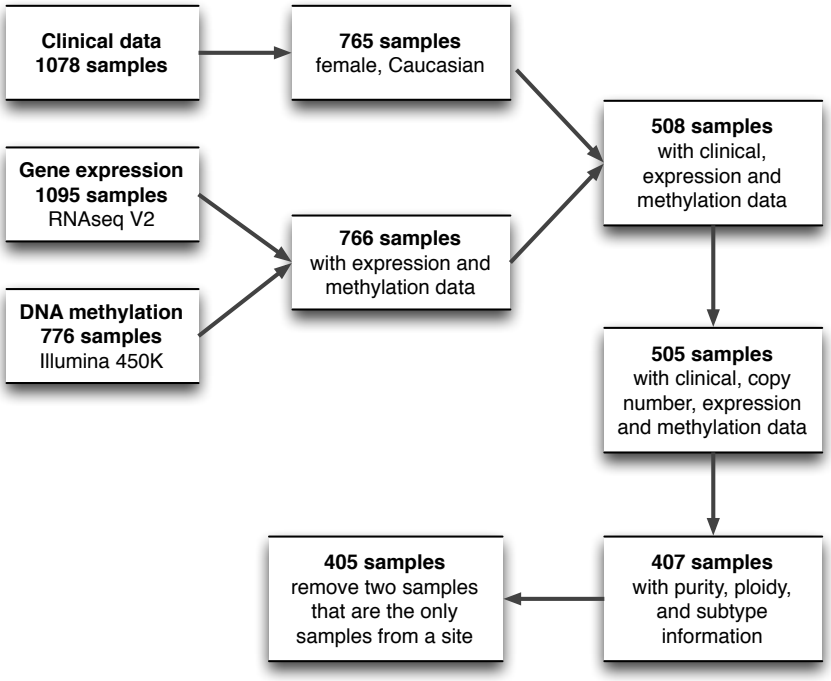

Figure S1: Sample selection for breast cancer.

B.1.2 Genotype PCA

We followed the steps described in section A.1.4 to prepare genotype data and ran PCA on the genotype data twice. In the first round, we ran PCA for TCGA BRCA samples

together with HapMap samples and selected samples from Caucasian females (Figure S2A-B). In the second round, we ran PCA again but only for those selected Caucasian samples. We selected the top 3 PCs as demographic covariates for other analyses.

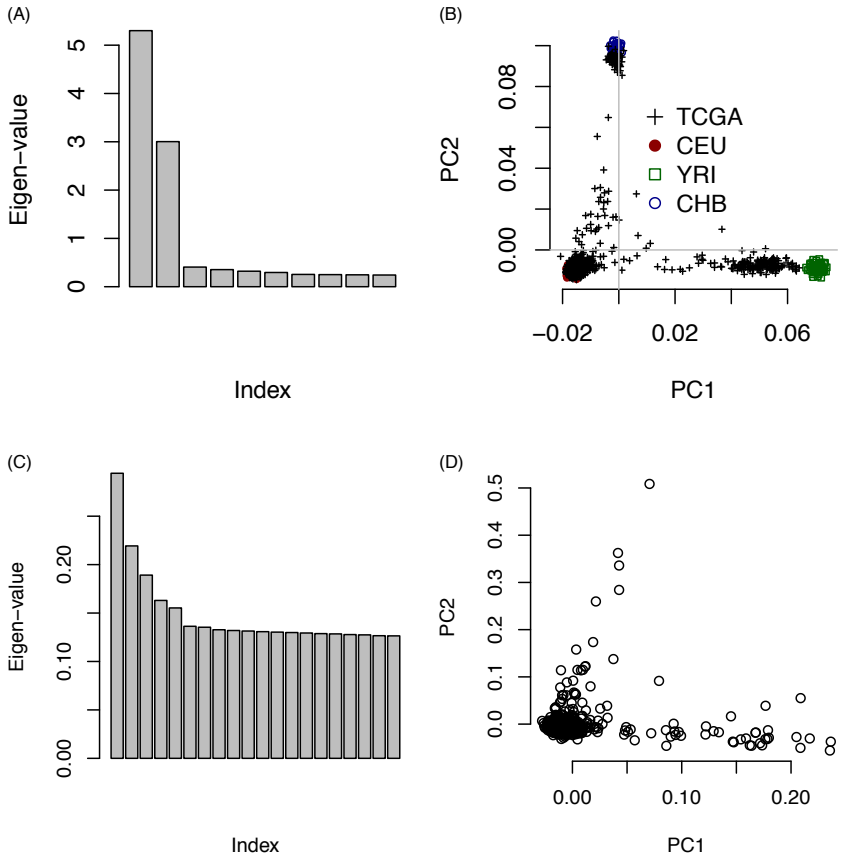

Figure S2: (A-B) Top eigenvalues and scatter plots for PC1 vs. PC2 for PCA of TCGA BRCA samples together with HapMap samples. We arbitrarily define the Caucasian samples as those with PC1 < 0 and PC2 < 0. (C-D) Top eigenvalues and scatter plots of PC1 vs. PC2 for selected Caucasian samples.

B.1.3 Batch effects and demographic covariates.

We evaluated the effects of batch variables and demographic covariates using linear regression where each genomic feature is the response variable and all batch effects and demographic variables are covariates. For breast cancer patients, we considered two batch variables (tissue site and plate) and 4 demographic covariates (age and top 3 genotype PCs). Both tissue site and plate have apparent influence on DNA methylation and gene expression, and age are associate with all three types of genomic data (Figure S3). The effect of genotype PCs is weak (results not shown). The distribution of p-values should be uniform if few genomic features are associated with one covariate and if the genomic features are independent. The p-value distribution for SCNA vs. tissue site or plate shows skewness to the right, which is most likely because the SCNA across genes is highly correlated.

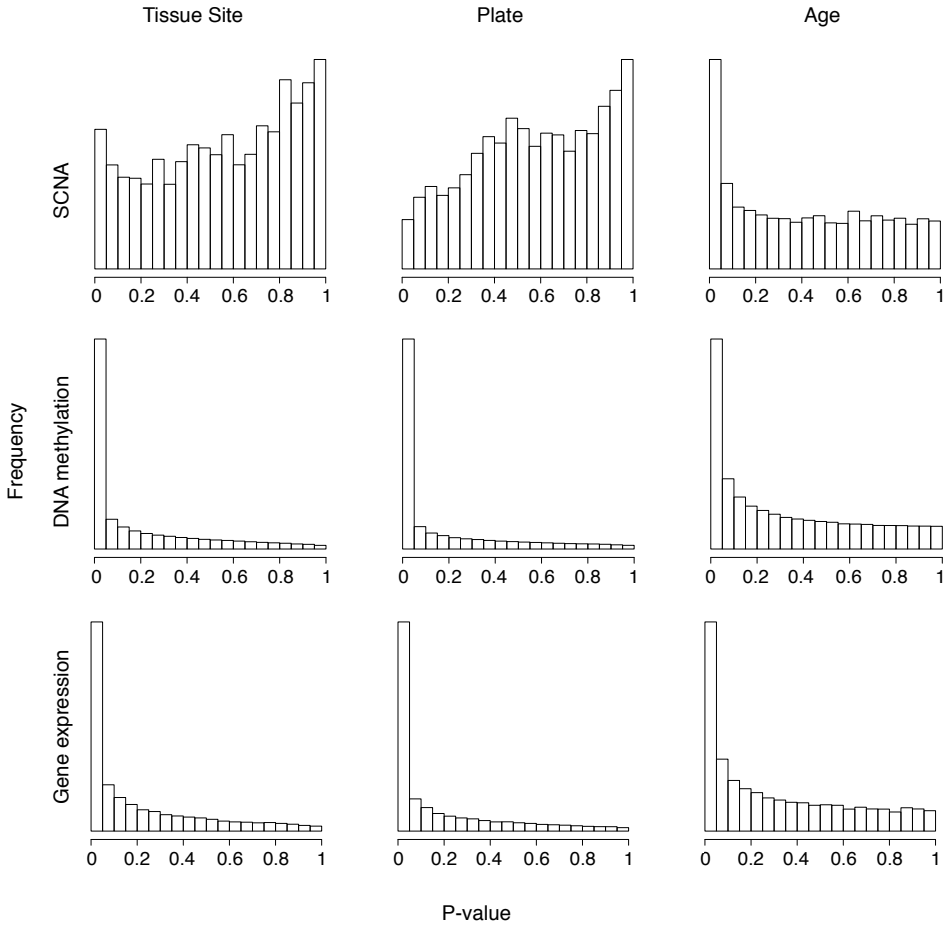

Figure S3: Distribution of p-values.

B.1.4 Summary of SCNA

The SCNA data (mean values over genomic segments) are not available in certain genomic locations (e.g., regions close to the centromere or either end of a chromosome), and thus, the SCNA information for some genes is missing. By keeping genes with non-missing values in at least 80% of samples, we have SCNA data for 19,662 genes. After removing the effects of tissue site, plate, and age (by taking residuals of linear regression against these variables), we performed PCA on the SCNA data. From the PC plots, we can see that breast cancer subtypes has a strong influence on SCNA (Figure S4). A heat map of SCNA data (Figure S5) shows characteristic SCNA in the breast cancer genome, e.g., application of chromosome arm 1q and 8q.

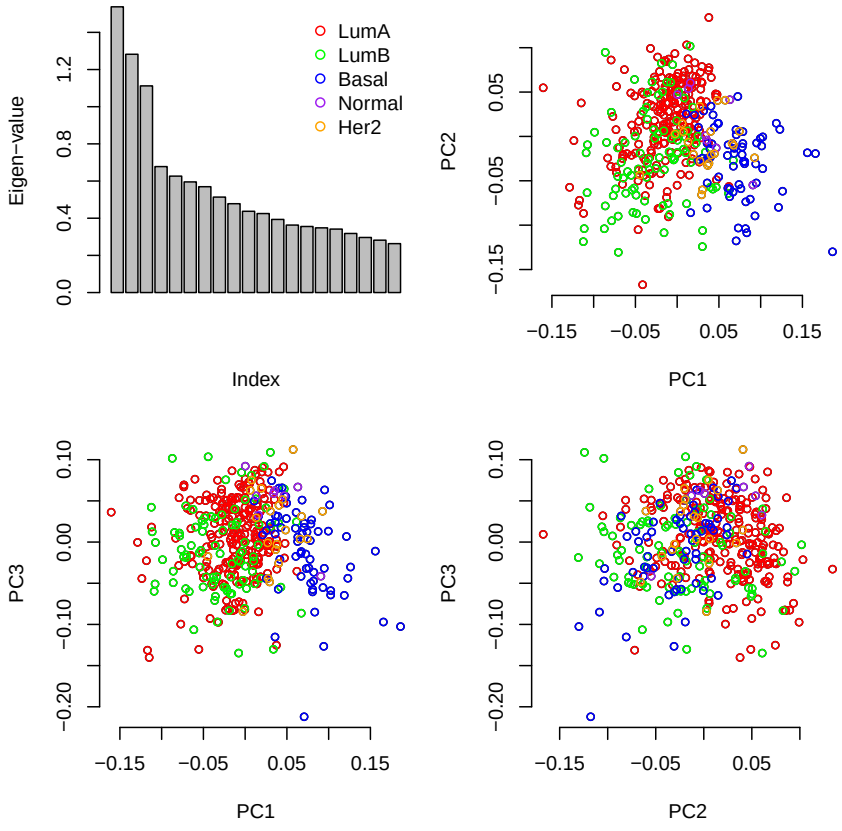

Figure S4: Results of PCA analysis on the SCNA data

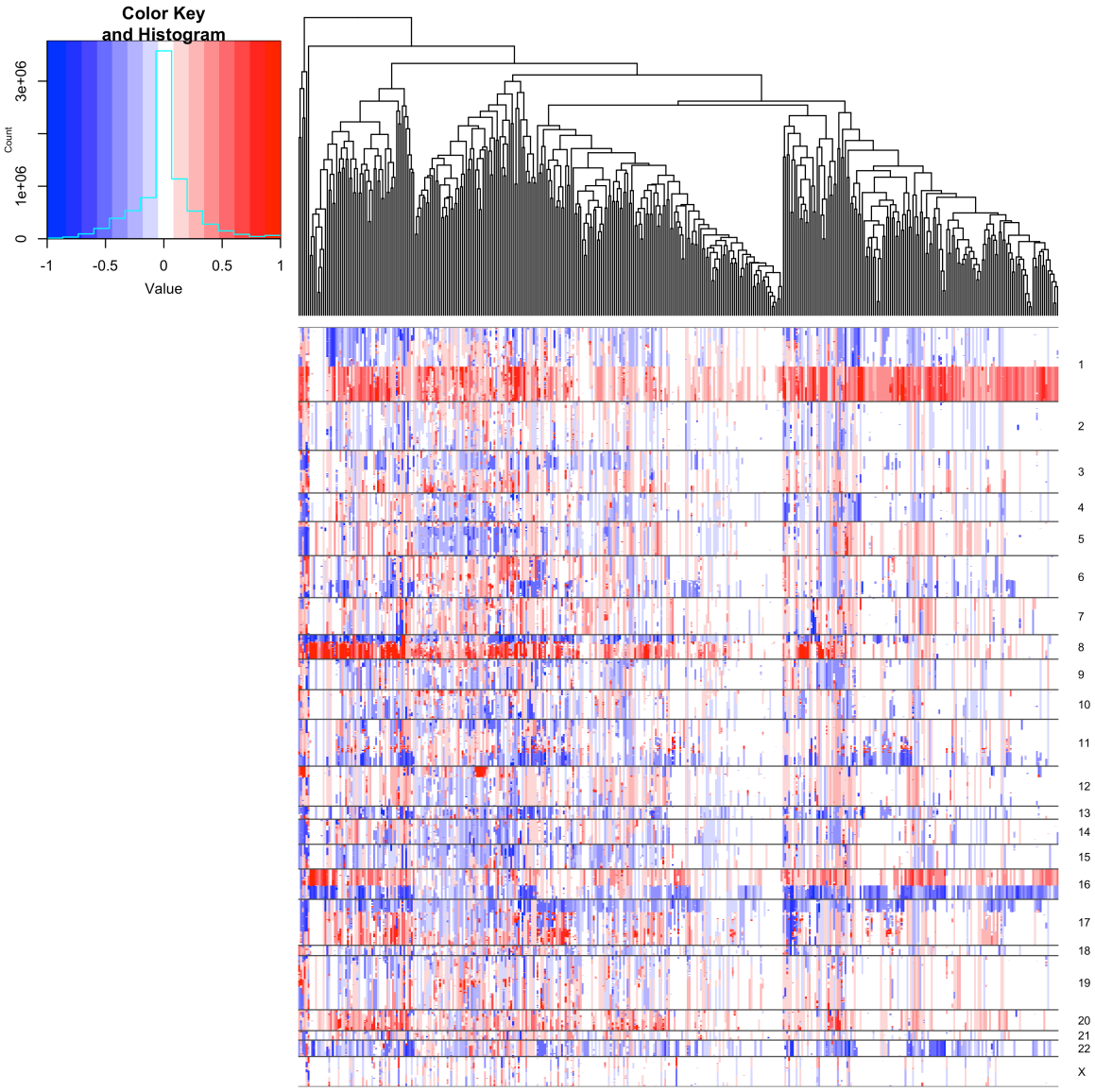

Figure S5: Heatmap for SCNA data. Each row is one gene, and each column is one sample. The genes are ordered by chromosome location. SCNA values are truncated at -1 and 1. The color key in the top left corner shows the correspondence between SCNA values and color. The histogram overlaid on the color key shows the number of observations with certain SCNA values.

B.1.5 Summary of DNA methylation data

We first removed those methylation probes with more than 5% of missing values, which left 394,498 of 485,577 probes for the following analysis. The raw data are in the scale of beta-value. We transferred the beta-values (β) to m-values ($m = \log[\beta/(1-\beta)]$). A summary of the mean values and standard deviations for the 394,498 probes is shown in Figure S6 (A-C). After removing the effects of tissue site, plate, and age (by taking residuals of linear regression against these variables), we performed PCA analysis. It is clear that breast cancer subtypes have strong influence of DNA methylation (Figure S6 (D-E)).

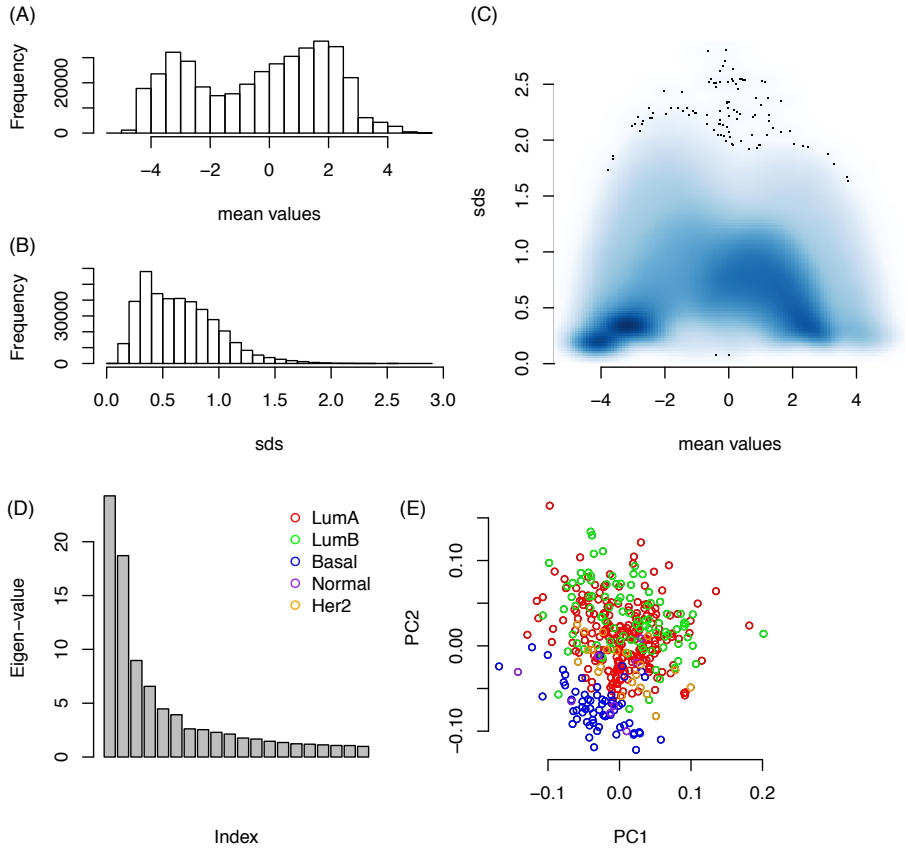

Figure S6: (A-B): The distributions of mean values and standard deviation (sd) of methylation probes. (C): A smooth scatter plot of mean vs. sd for all probes. The color density reflects the number of data points in specific areas. (D-E): Results of the PCA analysis.

B.1.6 Summary of gene expression data

We started with raw counts of 20,531 genes, and first filtered out 208 genes without location information (based on file `TCGA.hg19.June2011.gaf` as described in Section 1.1). Among the remaining 20,323 genes, we selected those genes with 75th percentile of expression larger than 20 and ended up with 15,843 genes (Figure S7A). Next, we adjusted the expression by read-depth. For each of the 400 samples, we calculated the total number of RNA-seq reads and the 75th percentile across all 15,843 genes. These two measurements of read-depth show good correlation (Figure S7B). Finally, the PCA results, as expected, show that tumor subtypes has a strong influence on gene expression (Figure S7C-D).

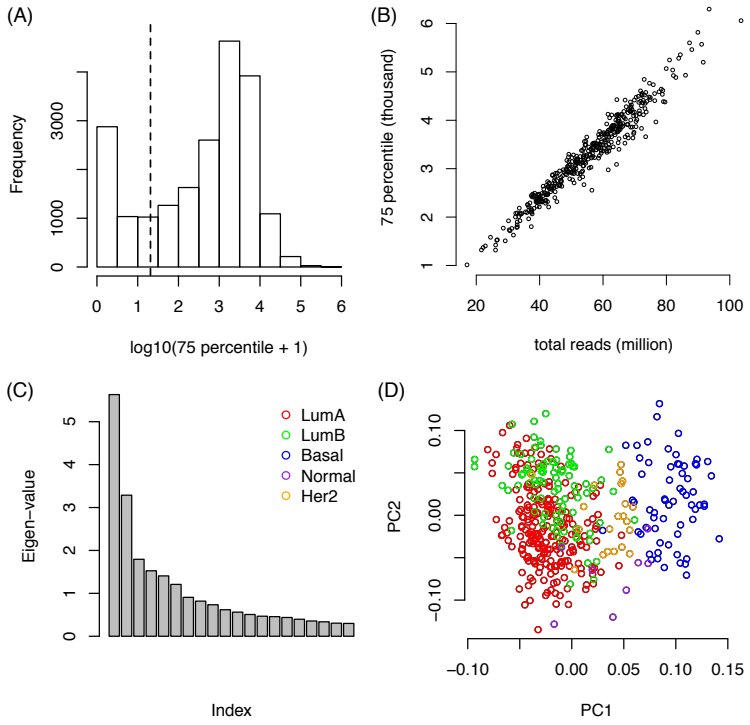

Figure S7: (A): For each gene, we calculated the 75th percentile of its expression across all samples. This histogram summarizes the $\log_{10}$(75th percentile + 1) of 20,323 genes. The vertical line is $\log_{10}(20 + 1)$. (B): This scatter plot has 400 points with each point corresponding to a sample. The x-axis is the total number of reads per sample, and the y-axis is the 75th percentile of gene expression across all the genes within a sample. (C-D): Results of PCA on log-transformed and read-depth-corrected expression data after removing tissue site, plate, age, and genotype PC effects.

B.1.7 Association between SCNA and DNA methylation

SCNA data are available for 19,535 genes. Based on aforementioned filtering, DNA methylation data are available at 394,498 probes. We examined the association of all 394,498 $\times$ 19,535 pairs of DNA methylation probes and SCNA measurements. We denote the methylation of the j-th methylation probe by M_j and denote the SCNA of the k-th gene by C_k. We fit a linear model $E(M_j) = \beta_0 + \beta_1 C_k + \sum_{l=1}^{39} X_l \eta_l$, where the 39 covariates (X_l for l=1, ..., 36) include 14 indicators for `tissue sites`, 21 indicators for `plates`, 1 covariate for `age`, and 3 covariates for genotype PCs. We calculated the p-value for hypothesis testing of $\beta_1 = 0$ vs. $\beta_1 \neq 0$, and we highlighted in Figure S8(A) those p-values that are smaller than 10^{-20}. Next, we fit a model with breast cancer subtypes as additional covariates. As shown in Figure S8(B), many significant p-values disappear after accounting for tumor subtypes (Figure 1B). Furthermore, accounting for purity does not really change the results (Figure 1B in main text vs. Figure S8(B)).

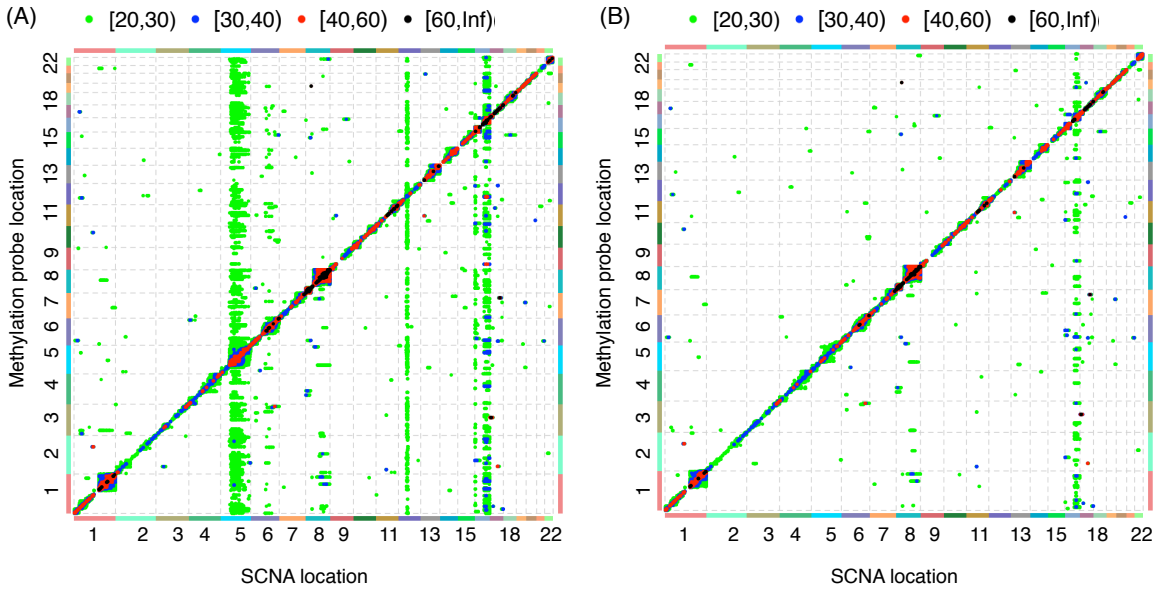

Figure S8: Association between DNA methylation and SCNA after accounting for batch effects, age, and genotype PCs (A) or after accounting for batch effects, age, genotype PCs, and tumor subtypes (B). Each point in this plot indicates the association between the copy number of one gene (x-axis) and the DNA methylation of one probe (y-axis). The color of each point indicates the range of its corresponding $-\log_{10}$(p-value), as shown in the legend at the top of the panels.

We have further studied the hotspot on chromosome 16 where DNA methylation of hundreds of CpGs from multiple chromosomes are associated with SCNA on chromosome 16. At p-value cutoff 10^{20}, there are 19,320 CM (copy number vs. methylation) associations involving 134 CpGs that are not located on chr16 and SCNA of 617 genes that are located on chr16. Among the top 42 genes whose SCNA are associated with more than 90 of the 134 CpGs at p-value cutoff 10^{-20}, 15 of them have CE (SCNA versus gene expression of the same gene) association p-value smaller than 10^{-20} (Table S1).

geneSymbol	chr	start	end	strand	CE p-value
DYNC1LI2	chr16	66754799	66785525	-	1e-78
ATP6V0D1	chr16	67471917	67515089	-	6.9e-58
NAE1	chr16	66836782	66864879	-	9.4e-58
CTCF	chr16	67596464	67673087	+	4.6e-50
C16orf70	chr16	67143866	67182440	+	1.1e-49
CBFB	chr16	67063050	67134956	+	1.8e-47
GFOD2	chr16	67708437	67753273	-	4.5e-44
CES2	chr16	66968347	66978992	+	5e-43
FAM96B	chr16	66965959	66968320	-	8.3e-42
TMEM208	chr16	67261016	67263182	+	4.4e-36
ACD	chr16	67691416	67694718	-	4.8e-36
PDP2	chr16	66914436	66921855	+	2.4e-33
FAM65A	chr16	67562754	67580688	+	1.1e-31
CKLF	chr16	66586466	66600154	+	1.3e-31
TK2	chr16	66543352	66584315	-	4.3e-28

Table S1: Fifteen chr16 genes whose SCNA are associated its own expression as well as at least 90 CpGs that are not located on chr16.

Among this short list of 15 genes, CTCF is one of the likely candidates that are responsible for this CM (SCNA versus methylation) hospot. To check wether those 134 CpGs are located around CTCF binding sites, we downloaded all the CTCF binding sites annotated by CTCFBSDB 2.0 [6] (`http://insulatordb.uthsc.edu/download/CTCFBSDB_all_exp_sites_Sept12_2012.txt.gz`). This file includes CTCF binding sites of different species and different genome build. We focus on the 10,453,868 (potentially overlapping) binding sites annotated to human hg19, from 70 types of tissues or cell lines. We take the union of these binding sites after filtering out those binding sites longer than 350 bps. This filter removes ∼17% of the binding sites, but reduce the total coverage of CTCF binding sites form 933Mb (28.8% of the genome) to 177Mb (5.5% of the genome). Even if without filtering, 123 (91.8%) of 134 CpGs are located within CTCF binding sites, which is significantly larger than 28.8% expected by chance (binomial p-value 0).

B.1.8 Association between SCNA and gene expression

By the aforementioned filtering, gene expression data are available for 15,843 genes. We examined the association of all 15,843 $\times$ 19,535 pairs of gene expression data and SCNA measurements. We used a linear model similar to the one for the methylation vs. SCNA analysis. Denote the i-th gene by Y_i and the k-th SCNA measurement by C_k. We fit a linear model $E(Y_i) = \alpha_0 + \alpha_1 C_k + \sum_{l=1}^{39} X_l \zeta_l$, where the 39 covariates account for the effects of tissue sites, plates, age, and genotype PCs. We calculated the p-value for hypothesis testing of $\alpha_1 = 0$ vs. $\alpha_1 \neq 0$ and highlighted in Figure S9(A) those p-values that are smaller than 10^{-20}. Next, we fit a model with breast cancer subtypes as additional covariates. As shown in Figure S9(B), many significant p-values disappear after accounting for tumor subtypes.

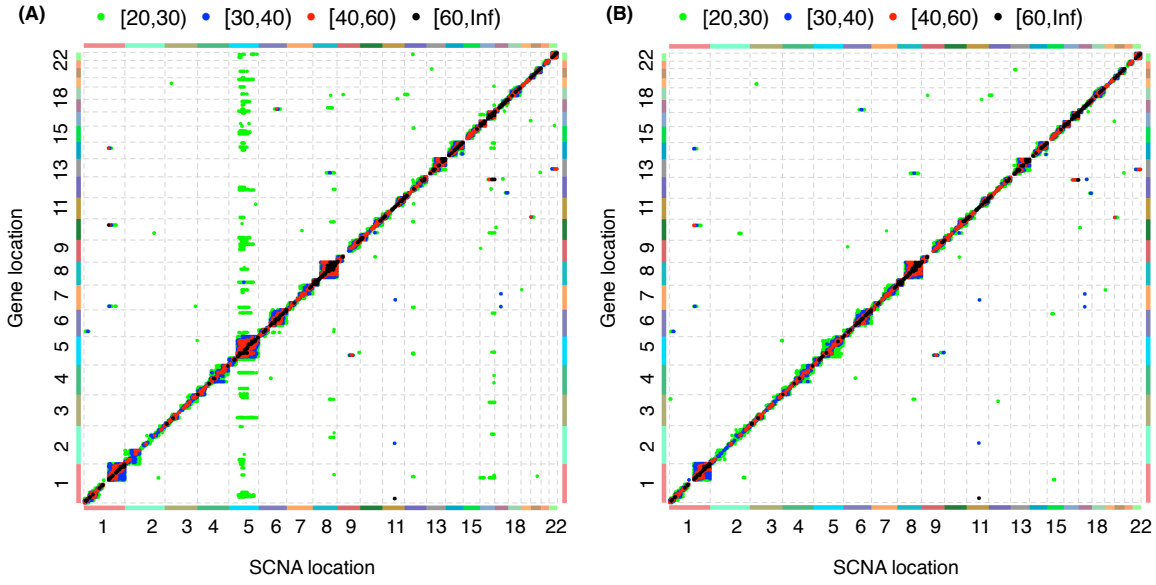

Figure S9: Association between gene expression and SCNA after accounting for batch effects, age, and genotype PCs (A) or after accounting for batch effects, age, genotype PCs, and tumor subtypes (B). Each point in this plot indicates the association between the copy number of one gene (x-axis) and the expression of one probe (y-axis). The color of each point indicates the range of its corresponding $-\log_{10}$(p-value), as shown in the legend at the top of the panels.

B.1.9 Association between DNA methylation and gene expression

We conducted a series of analyses of the association between gene expression and DNA methylation. Each analysis assessed the association of $15,843 \times 394,498 \approx 6250$ million methylation-expression (ME) pairs, while accounting for different sets of covariates. The results of the following four sets of analyses are illustrated in different figures.

1. Batch effects (tissue sites and plates), age, and genotype PCs (Figure S11A).

2. Batch effects, age, genotype PCs, and SCNA (Figure S11B). Ideally, for each ME pair of the i-th gene (Y_i) and the j-th methylation probe (M_j), we fit a model $E(Y_i) = \gamma_0 + \gamma_1 M_j + \gamma_2 C_i + \sum_l X_l \kappa_l$, where C_i is the SCNA of the i-th gene. However, we cannot use matrixEQTL R package to fit such a model because the set of covariates varies from each ME pair due the presence of C_i. We employ an approximation here to regress Y_i on C_i first and obtain the residuals, denoted by $\tilde{Y}_i$. Then, we fit the following model using matrixEQTL: $E(\tilde{Y}_i) = \gamma_0 + \gamma_1 M_j + \sum_l X_l \kappa_l$. In the following analysis, whenever SCNA is accounted, we replace Y_i with $\tilde{Y}_i$.

3. Batch effects, age, genotype PCs, SCNA, and tumor subtypes (Figure 3A in the main text).

4. Batch effects, age, genotype PCs, SCNA, tumor subtypes, and purity estimates by ABSOLUTE [7](Figure S11C).

From these figures, it is apparent that there is a huge number of pair-wise associations between DNA methylation and gene expression (with a large number of horizontal and vertical bands) even after we correct for batch effects, age, genotype PCs, SCNA, and tumor subtypes. However, after adding the purity estimate as a covariate, the number of ME associations reduces dramatically (Figure S11C). Table S2 summarizes the total number of ME associations for a set of 9 models. We adopt a very liberal definition of local ME association; an ME association is referred to as local if the corresponding gene and methylation probe are located at the same chromosome. The percentage of local ME associations can be used as an informative, albeit indirect, measure of the accuracy of the results. Because gene expression is more likely to be affected by nearby DNA methylations, we expect the percentage of local ME associations to be high. However, using the set of 9 models listed in Table S2, the only situation in which the percentages of local ME association is larger than 50% is when the purity estimate by ABSOLUTE is included in the model and we

apply a p-value cutoff of 10^{-100} (Table S3).

models	-$\log_{10}$(p-value)							total
	< 100	(90,100]	(80,90]	(70, 80]	(60, 70]	(50, 60]	(40, 50]	
baseline	2916	6546	19209	50209	127424	323954	843602	1373860
cn	1827	4406	13877	39117	103951	278637	748815	1190630
cn, subtype	1604	3943	12291	35180	93227	239308	604242	989795
cn, subtype, purity	6	17	146	725	2682	10439	39076	53091

Table S2: The number of ME associations found by different models. The baseline model includes batch effects (tissue sites and plates), age, and genotype PCs. All other models are built from this baseline model by adding covariates: cn (SCNA), tumor subtypes, and the purity estimated by ABSOLUTE.

models	-$\log_{10}$(p-value)							total
	< 100	(90,100]	(80,90]	(70, 80]	(60, 70]	(50, 60]	(40, 50]	
baseline	0.086	0.069	0.073	0.067	0.066	0.064	0.062	0.043
cn	0.073	0.066	0.067	0.063	0.061	0.062	0.06	0.037
cn, subtype	0.076	0.063	0.069	0.064	0.06	0.06	0.061	0.038
cn, subtype, purity	1	0.235	0.13	0.103	0.087	0.083	0.075	0.032

Table S3: The proportion of ME associations that are located on the same chromosome.

We have mentioned in the main text that we apply normal quantile transformation to both DNA methylation and gene expression data. If non-linear methylation-expression relation is common, we expect to see that most association strengthes to be stronger after normal quantile transformation.

Most of the remaining off-diagonal signals can be removed by the following steps. After accounting for batch effects, age, genotype PCs, SCNA, and tumor subtypes (Figure 3A in the main text), we first selected all 137,149 ME pairs with association p-values smaller than 10^{-60} and corresponding gene and DNA methylation probes located on different chromosomes. For each gene expression or methylation measurement in these 137,149 ME pairs, we took the residuals of linear regression against batched effects, age, genotype PCs, and tumor subtypes, scaled the residuals to have mean 0 and standard deviation 1, averaged each pair, and thus obtained a data matrix of $137,149 \times 405$. Each row of this matrix corresponds to one ME pair, and each column corresponds to a sample. We performed PCA using this data matrix. After adding the top PC as a covariate in the ME association studies, most off-diagonal signals are removed (Figure S11D).

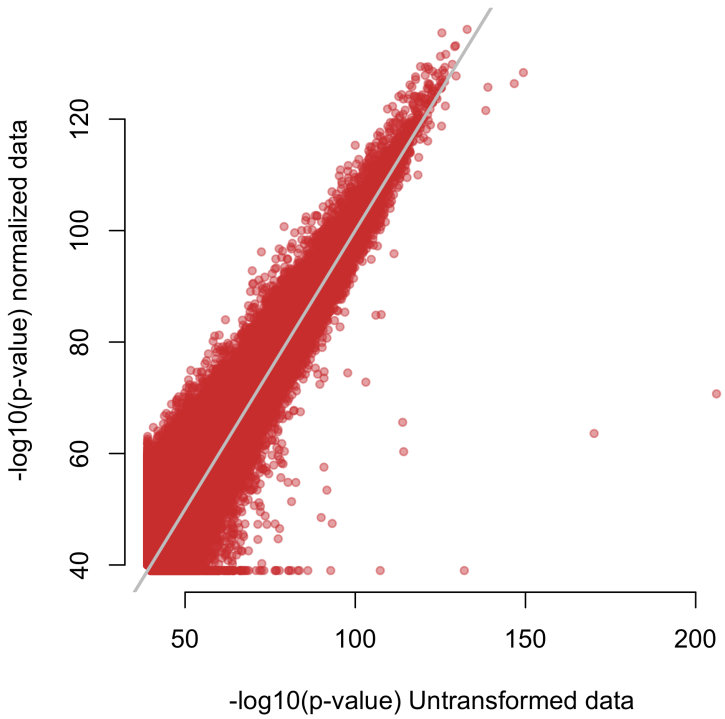

Figure S10: $-\log_{10}$(p-value) for methylation-expression associations before (x-axis) versus after (y-axis) normal quantile transformation. Here we consider the model that account for batch effects, age, genotype PCs, SCNA, and tumor subtypes.

However, if one focuses on those associations with more liberal p-values, e.g., with p-value cutoff 10^{-30} instead of 10^{-60}, many off-diagonal signals emerge (Figure S12A). By including top seven PCs calculated from the top ME pairs, most of the off-diagonal signals are removed (Table S4 and Figure S12B-D). In addition, around 90% of ME associations with p-values smaller than 10^{-30} are local ME associations (i.e., the corresponding gene and DNA methylation probe are located at the same chromosome).

models	-$\log_{10}$(p-value)					total
	< 70	(60, 70]	(50, 60]	(40, 50]	(30, 40]	
PC1	43 (0.98)	110 (0.56)	905 (0.23)	5797 (0.14)	35599 (0.09)	42454 (0.10)
PC1-2	44 (1.00)	58 (0.98)	311 (0.56)	1984 (0.30)	14971 (0.15)	17368 (0.18)
PC1-3	49 (1.00)	63 (0.98)	396 (0.47)	2520 (0.27)	17974 (0.14)	21002 (0.16)
PC1-4	51 (1.00)	70 (0.97)	422 (0.48)	2691 (0.26)	19040 (0.14)	22274 (0.16)
PC1-5	49 (1.00)	79 (0.87)	515 (0.42)	2937 (0.23)	19488 (0.13)	23068 (0.16)
PC1-6	49 (1.00)	66 (1.00)	198 (0.99)	601 (0.90)	3090 (0.53)	4004 (0.62)
PC1-7	48 (1.00)	70 (1.00)	190 (0.99)	560 (0.98)	1701 (0.91)	2569 (0.93)

Table S4: The number (proportion of local associations) of ME associations after removing SCNA effects from gene expression and accounting for the covariates of batch effects, age, tumor subtypes, and certain number of ME PCs derived from highly correlated ME pairs.

The eigenvalue of the first PC is dramatically larger than all other eigenvalues (Figure S13), and the first PC is strongly associated with the purity estimate by ABSOLUTE. This suggests that tumor purity is a major factor underlying the strong associations between gene expression and DNA methylation. The 2nd to the 7th PCs explain the association of many ME pairs; however, they are not correlated or only weakly correlated with purity (Figure S13-S14).

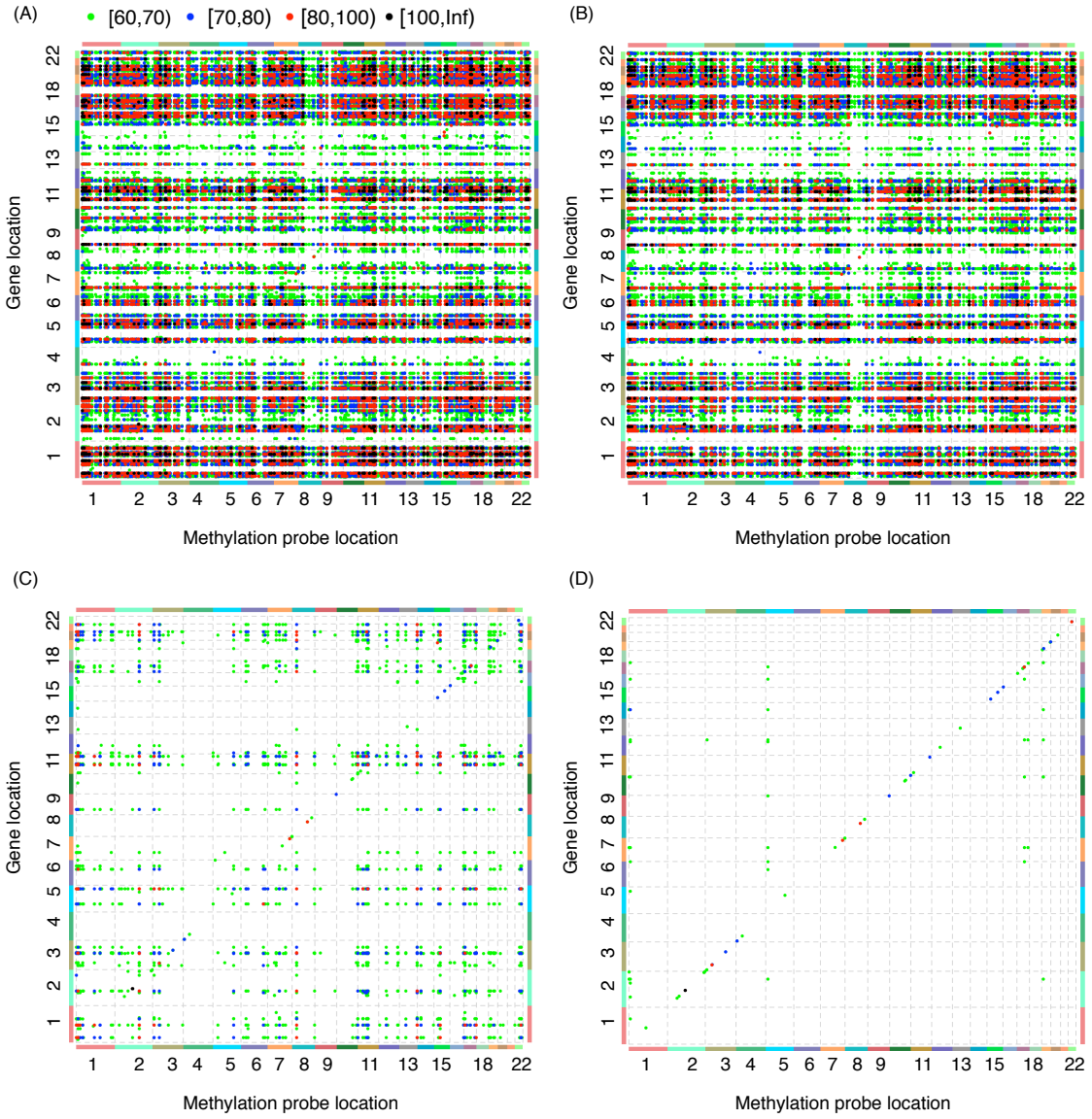

Figure S11: (A) Association between gene expression and DNA methylation after accounting for batch effects, age, and genotype PCs. (B) Results after accounting for batch effects, age, genotype PCs, and SCNA. (C) Results after accounting for batch effects, age, genotype PCs, SCNA, tumor subtypes, and purity estimates by ABSOLUTE. (D) Results after accounting for batch effects, age, SCNA, tumor subtypes, and 1st PC from correlated methylation-expression (ME) pairs.

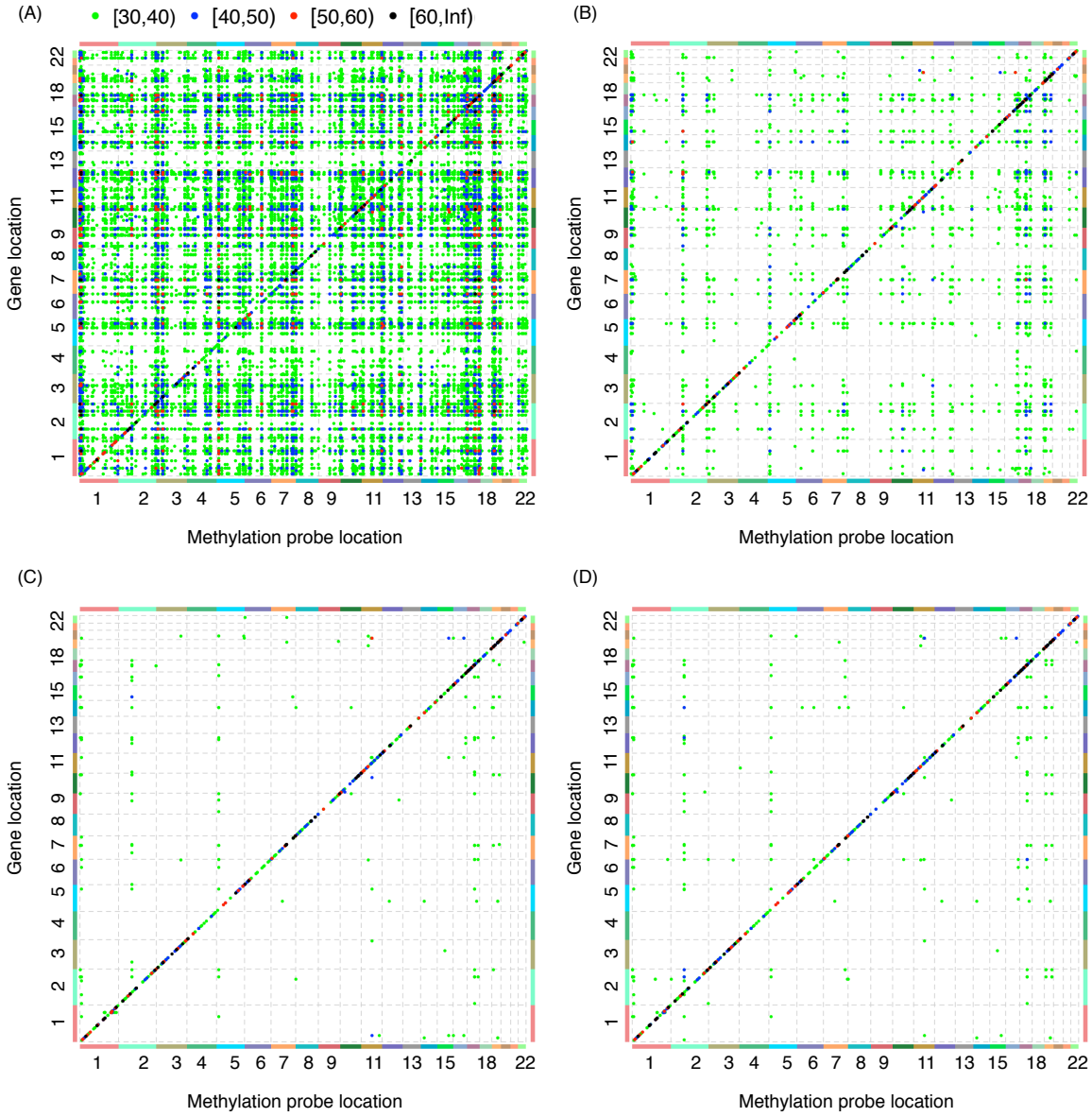

Figure S12: Association between gene expression and DNA methylation after accounting for batch effects, age, tumor subtype, and increasing number of PCs from correlated methylation-expression (ME) pairs: (A) first PC, (B) first to third PCs, (C) first to fifth PCs, and (D) first to sixth PCs. Note that the p-value cutoff used in this figure is much more liberal than the cutoff used in Figure S11.

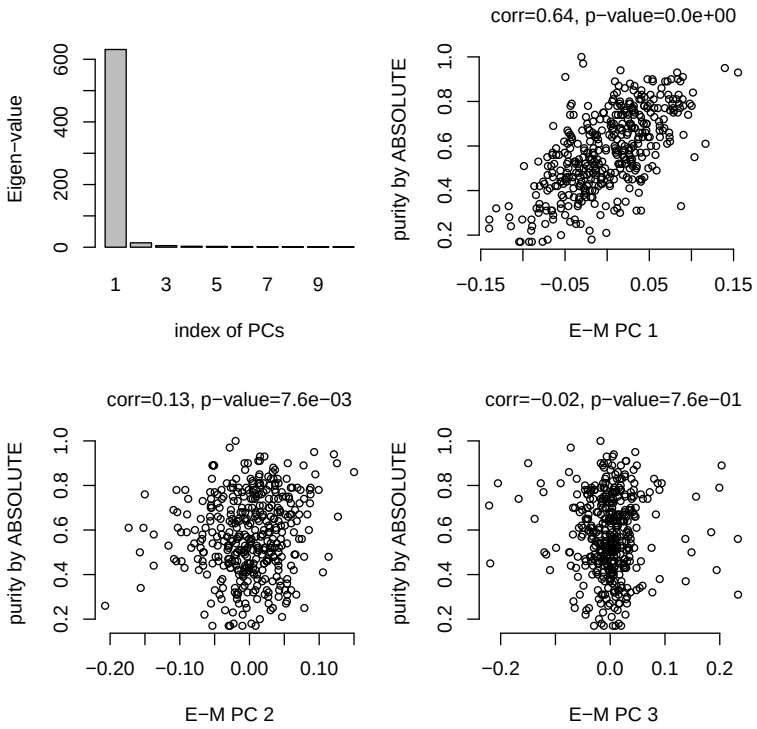

Figure S13: A barplot of eigenvalues from PCA on the average of highly correlated ME pairs, and scatter plots for the top three ME PCs vs. purity by ABSOLUTE.

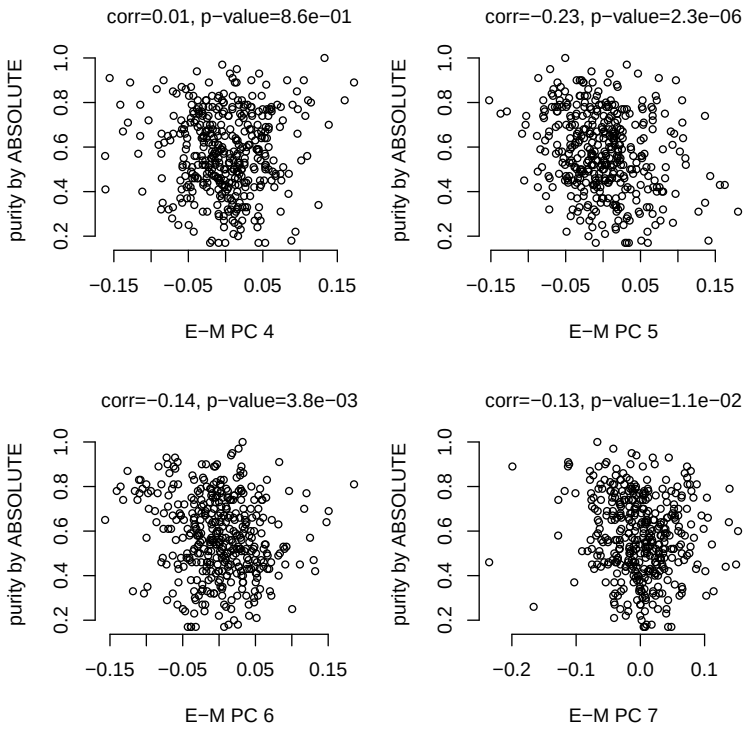

Figure S14: The scatter plots for the 4th to the 7th ME PCs vs. purity by ABSOLUTE.

We also were interested in examining the robustness of the local E-M association when the correction by E-M PCs (i.e., purity correction) is ignored. Using the model that accounts for batch effects, age, SCNA, and tumor subtypes (referred to as model M0), we can identify 28,096 local ME associations, defined as p-values smaller than 10^{-40} and the distance between the starting position of the corresponding gene and DNA methylation probe is smaller than 1Mb. After accounting for 7 ME PCs (referred to as model M1), we identified 42,077 local ME associations at a more liberal p-value cutoff of 10^{-10}. Take the union of these two sets of ME pairs, we obtained 68,570 ME pairs. Figure S16 shows the scatter plot of the p-values for these ME pairs under models M0 and M1. This figure can be a bit misleading, however, because the majority of the points lie at the bottom of the figure when the p-value on y-axis is truncated at 10^{-10}. Nevertheless, it is still clear that the p-value with and without the purity correction is only consistent for a subset of ME pairs. In fact, among the 28,096 local ME associations without the purity correction, 26,493 (94.3%) have p-values larger than 10^{-10} after the purity correction. In other words, the vast majority of local ME associations are simply due to the fact that the corresponding gene expression and DNA-methylation are both associated with purity or underlying cell type compositions.

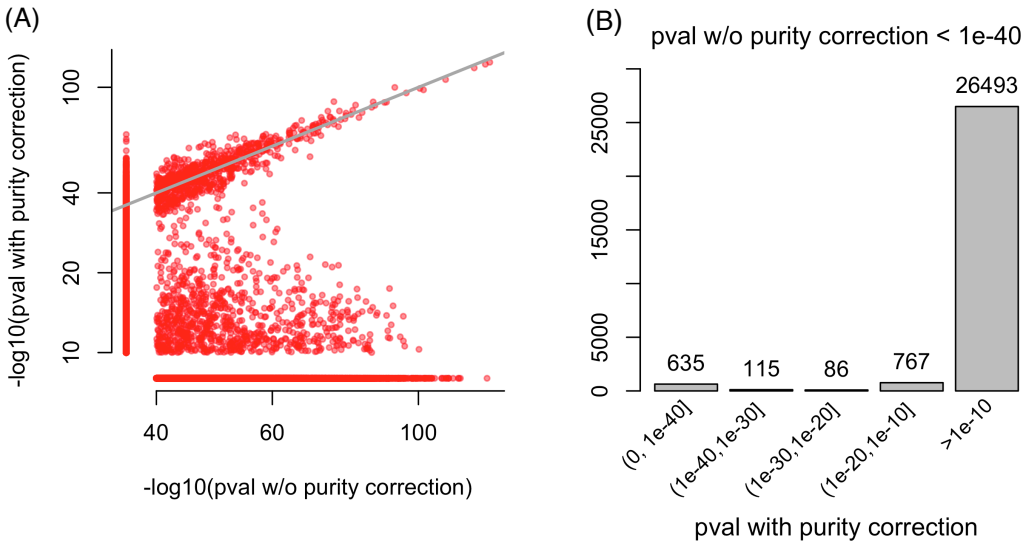

Figure S15: (A) Comparison of the p-value for 68,570 ME pairs that are local ME associations either with correction for latent factors (purity correction) by top 7 ME PCs (p-value $< 10^{-10}$) or without purity correction (p-value $< 10^{-40}$). More stringent p-value cutoff was used for the analysis without purity correction simply because the number of findings was too large. The p-values are truncated at $< 10^{-40}$ and $< 10^{-10}$ for x and y axis, respectively. (B) The distribution of p-values with the purity correction for those ME pairs whose p-values without purity correction are smaller than $< 10^{-40}$.

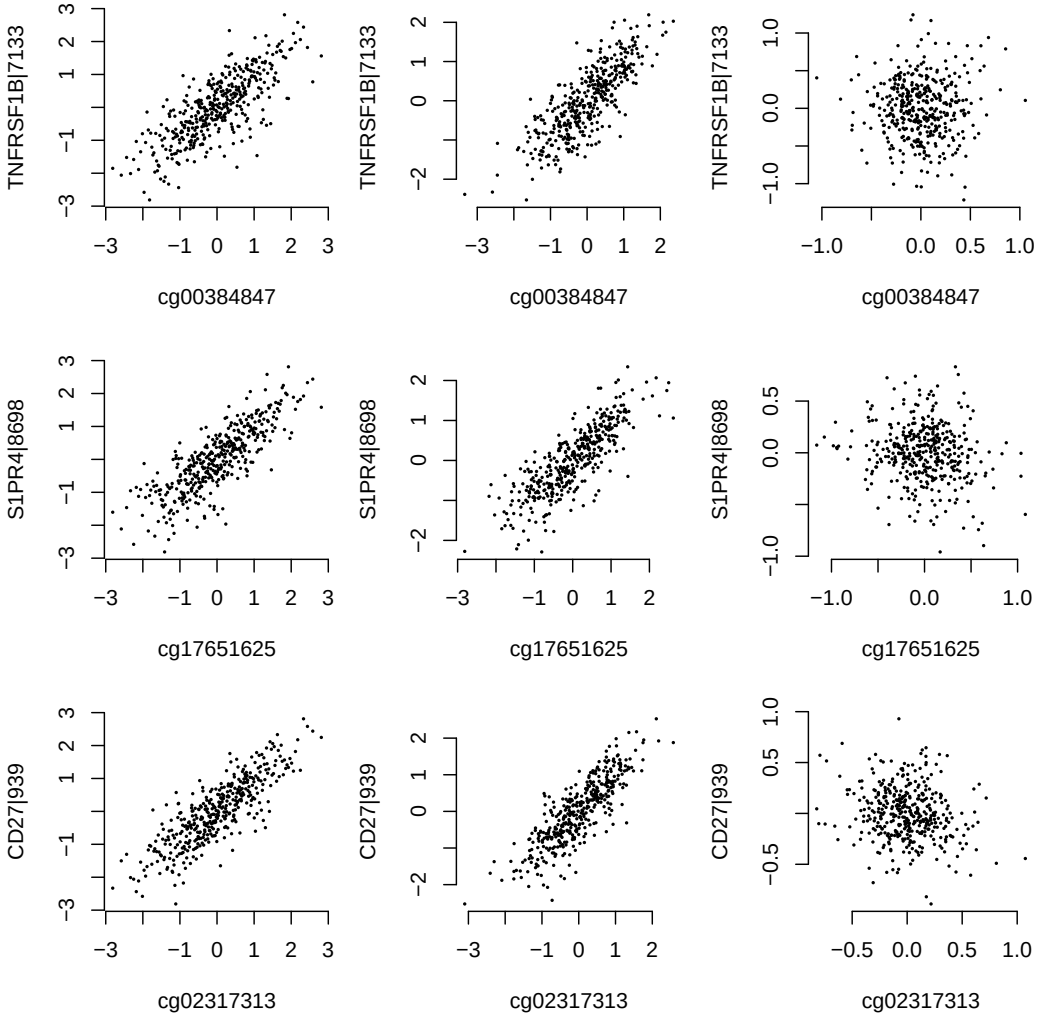

Figure S16: Illustration of the associations of randomly selected three methylation-expression pairs with p-value smaller than 1e-80 before applying ME PCs, and p-value larger than 1e-10 after applying ME PCs. Each row corresponds to one methylation-expression pair. The 1st column shows the original data. The 2nd column are the residualized data after accounting for batch effects, age, genotype PCs, SCNA, and tumor subtypes. The 3rd column are the residualized data that account for all the covariates used for the 2nd column plus 7 ME PCs.

B.1.10 Hot genes and hot methylation probes

We define "hot genes" and "hot methylation probes" using the results of the ME associations accounting for batch effects, age, SCNA, and tumor subtypes. We define those genes with at least 100 ME associations with p-values smaller than 10^{-60} as hot genes. Among 182 such hot genes, 173 are annotated at DAVID (https://david.ncifcrf.gov). By querying those 173 genes using DAVID tools, we found that they are significantly enriched in functional categories related to immune system (Table S5). This enrichment of immune-related genes is robust to the cutoff to define hot genes. The expression of all 182 genes is negatively associated with tumor purity estimates by ABSOLUTE (Figure S17).

Term	Count	Percent	Expected Percent	P-Value	Benjamini
GO:0006955 immune response	54	31.2	5.10	1.75E-33	2.10E-30
GO:0002684 positive regulation of immune system process	36	20.8	1.76	2.35E-31	1.41E-28
GO:0050865 regulation of cell activation	31	17.9	1.29	6.36E-29	2.55E-26
GO:0046649 lymphocyte activation	32	18.5	1.47	1.52E-28	4.56E-26
GO:0045321 leukocyte activation	34	19.7	1.79	1.94E-28	4.66E-26
GO:0051249 regulation of lymphocyte activation	29	16.8	1.09	2.86E-28	5.73E-26
GO:0002694 regulation of leukocyte activation	30	17.3	1.23	3.26E-28	5.61E-26
GO:0001775 cell activation	35	20.2	2.12	3.23E-27	4.85E-25
GO:0050867 positive regulation of cell activation	25	14.5	0.82	8.73E-26	1.17E-23
GO:0002696 positive regulation of leukocyte activation	24	13.9	0.78	8.94E-25	1.07E-22

Table S5: The functional enrichment of 173 hot genes. The columns "Count" and "Percent" are the number and percentage, respectively, of the 173 genes that belong to a certain functional category. The column "Expected %" is the expected percentage based on the annotation of all genes. The "P-value" and "Benjamini" columns show the significance level before and after multiple testing correction, respectively.

We also define the hot genes based on effect size (regression coefficient) rather than p-value. As shown in Figure S18, there is a monotone relationship between p-value and regression coefficient, though the the effect sizes for a specific p-value may have considerable variation, due to variation of effect sizes of other covariates, such as batch effects, tumor subtype, and SCNA. We selected 225 hot genes as those with at least 100 associations with regression coefficients larger than 0.75. Among 225 such hot genes, 220 are annotated at DAVID (https://david.ncifcrf.gov). By querying those 220 genes using DAVID tools, we found they are significantly enriched in functional categories related to immune system (Table S6). This enrichment of immune-related genes is robust to the cutoff to define hot genes..

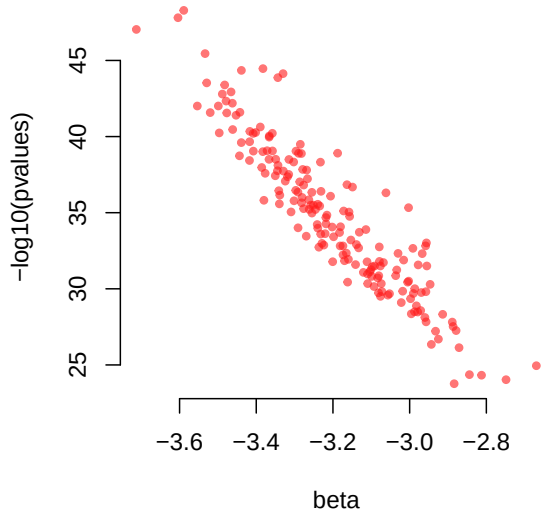

Figure S17: For each of the hot genes, we assess its association with tumor purity with a linear model. This scatter plot shows the regression coefficients versus $-\log_{10}$(p-values) of 182 hot genes.

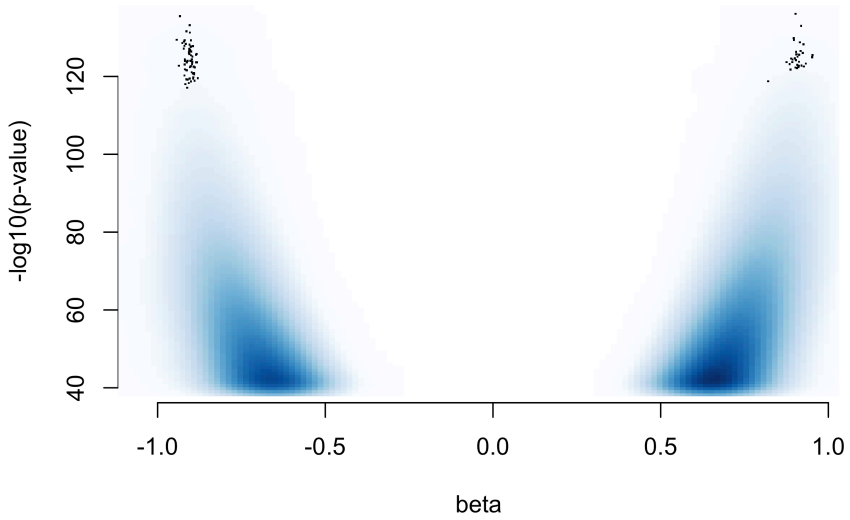

Figure S18: A scatter plot of the regression coefficients versus $-\log_{10}$(p-values) of 182 hot genes.

Term	Count	Percent	Expected Percent	P-Value	Benjamini
GO:0006955 immune response	70	32.9	5.10	3.41E-45	4.33E-42
GO:0046649 lymphocyte activation	41	19.2	1.47	1.39E-37	8.85E-35
GO:0045321 leukocyte activation	43	20.2	1.79	1.13E-36	4.77E-34
GO:0001775 cell activation	44	20.7	2.12	9.69E-35	3.07E-32
GO:0042110 T cell activation	31	14.6	0.93	1.46E-30	3.71E-28
GO:0002684 positive regulation of immune system process	38	17.8	1.76	1.97E-30	4.16E-28
GO:0050865 regulation of cell activation	34	16.0	1.29	5.21E-30	9.45E-28
GO:0051249 regulation of lymphocyte activation	32	15.0	1.09	1.05E-29	1.67E-27
GO:0002694 regulation of leukocyte activation	33	15.5	1.23	1.99E-29	2.80E-27
GO:0050867 positive regulation of cell activation	27	12.7	0.82	2.34E-26	2.97E-24

Table S6: The functional enrichment of 220 hot genes that were selected based on regression coefficients.

We define hot methylation probes as those that are associated with at least 30 genes with p-values smaller than 10^{-60}. There are 1,480 such hot methylation probes. Figure S19 provides a summary of these 1480 methylation probes with respect to a few categories.

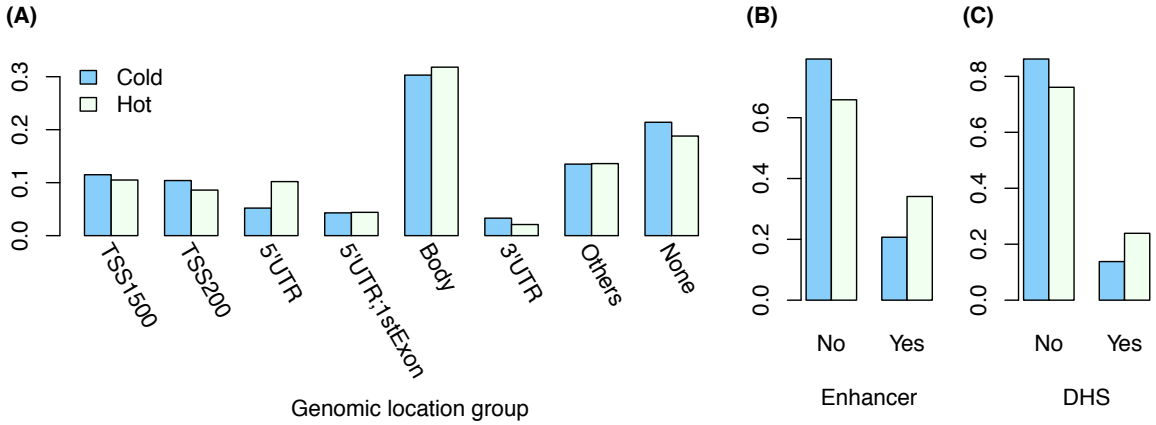

Figure S19: Location of hot and cold (i.e., non-hot) methylation probes with respect to genomic features (A), enhancer (B), and DHS (DNase Hypersensitive Site) (C).

B.1.11 eQTMs

We define expression Quantitative Trait Methylation probes (eQTMs) using the results of the ME associations accounting for batch effects, age, SCNA, tumor subtypes, and 7 ME PCs. There are 2,332 methylation probes that are associated with at least one gene with p-values smaller than 10^{-30}. Figure S20 summarizes the location of these 2,332 methylation probes with respect to several grouping variables.

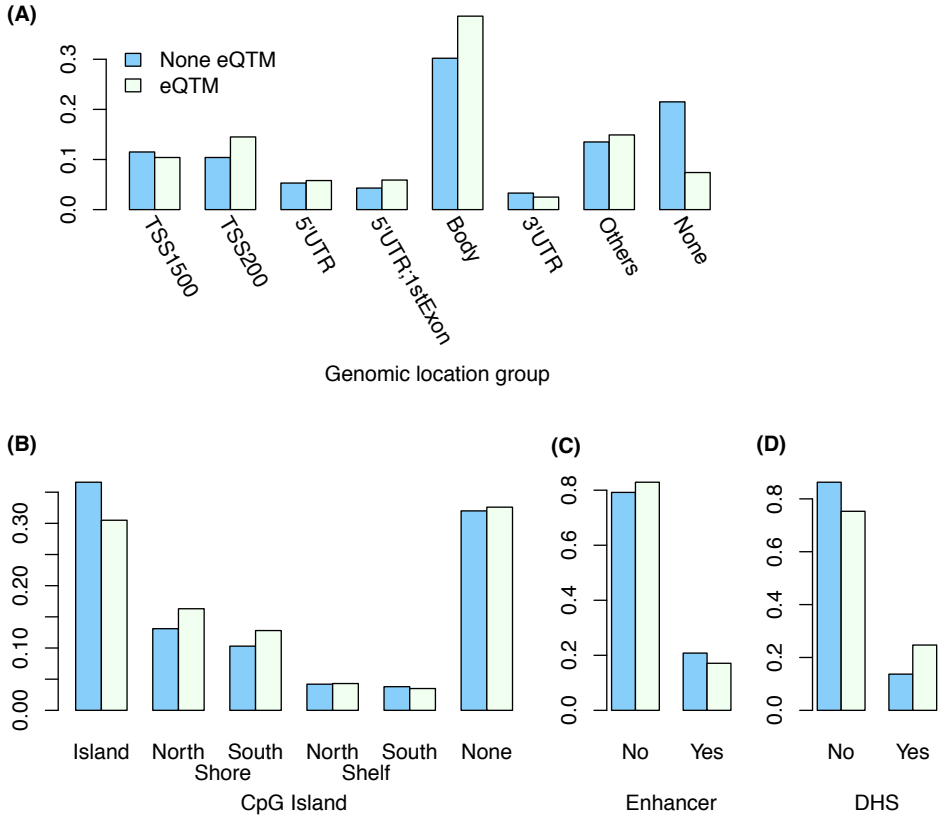

Figure S20: Location of eQTMs and non-eQTMs with respect to genomic features (A), CpG island (B), enhancer (C), and DHS (DNase Hypersensitive Site) (D).

B.2 Other cancer types

B.2.1 Sample selection and characterization

We studied five cancer types in addition to breast cancer: colon adenocarcinoma (COAD), acute myeloid leukemia (LAML), brain lower grade glioma (LGG), glioblastoma multiforme (GBM), and prostate adenocarcinoma (PRAD). Following an approach similar to that for the breast cancer data, we used PCA analysis on germline genotype data to identify Caucasian samples and restricted our analysis to those samples. For prostate adenocarcinoma, only male patients were included; patients of both sexes were included in the other cancers. In addition, we excluded samples if they were the sole representative of a specific plate or tissue site.

We included tumor subtypes as covariates for all cancers other than PRAD, for which we did not find clear definition of tumor subtypes.

- Following an approach similar to that used for the TCGA COAD study [8], we partitioned the COAD samples into two subtypes: hypermutated tumors vs. non-hypermutated tumors. We obtained the number of somatic mutations per patient from a pan-cancer study [4] (file `PANCAN12.mutation.whitelist.maf` downloaded from `https://www.synapse.org/#!Synapse:syn1695324`) and used a cutoff of 1000 somatic mutations per sample. Samples with more than 1000 somatic mutations were considered hypermutated.

- LGG can be classified into three subtypes: (1) LGG with IDH1/IDH2 mutation and 1p/19q deletion, (2) LGG with IDH1/IDH2 mutation, but without 1p/19q deletion, and (3) other samples [9]. Because we did not have complete informant about IDH1/IDH2 mutation status, we introduced two additional subtypes: unknown mutation of IDH1/IDH2 with or without 1p/19q deletion.

- The LAML samples were grouped into three groups based on Cyto risks: favorable, intermediate, and poor. We adopted this classification from a previous paper [10], using updated Table S1, which was downloaded from `https://tcga-data.nci.nih.gov/docs/publications/laml_2012/`.

- We classified the GBM samples into four transcriptomic subtypes: proneural, neural, classical, and mesenchymal. In addition, we included as a covariates an indicator for

samples with the glioma-CpG island methylator phenotype (G-CIMP). The covariates information was obtained from a recent TCGA GBM study [11] (Table S7).

- To the best of our knowledge, there is no well-defined and validated subtype for PARD.

In addition to tumor subtypes, we included plate, tissue site, age, sex (if patients of both sexes were included), and genotype PCs as covariates when we assessed the association among SCNA, DNA methylation, and gene expression (Table S7). Note that genotype PCs were calculated using genotype data after we selected the Caucasian samples. The number of of genotype PCs were chosen based on the sizes of eigenvalues.

Cancer Type	Sample Size	Covariates	d.f. of Covariates
BRCA	405	plate, site, 3 genotype PCs, 5 tumor subtypes	42
COAD	160	plate, site, age, sex, 4 genotype PCs, 2 tumor subtypes	27
GBM	243	plate, site, age, sex, 3 genotype PCs, 5 tumor subtypes, and an indicator of methylation platform	33
LAML	134	plate, age, sex, 1 genotype PCs, and 3 tumor subtypes	8
LGG	380	plate, age, sex, 2 genotype PCs, and 5 tumor subtypes	39
PRAD	377	plate, age, and 5 genotype PCs	44

Table S7: Summary of sample size and covariates used for each type of cancer.

B.2.2 The effect of SCNA on DNA methylation and gene expression

For each cancer type, we assessed the association between SCNA and DNA methylation, as well as the associations between SCNA and gene expression, while accounting for all covariates listed in Table S7. As shown in Figures S21-S25, most of these associations are local, except for the associations between SCNA and DNA methylation in PRAD, where there are a few noticeable vertical bands. A vertical band may indicate that the SCNA at a locus affects the methylation of many CpGs or it may be due to some shared unknown confounders, e.g., unknown tumor subtypes.

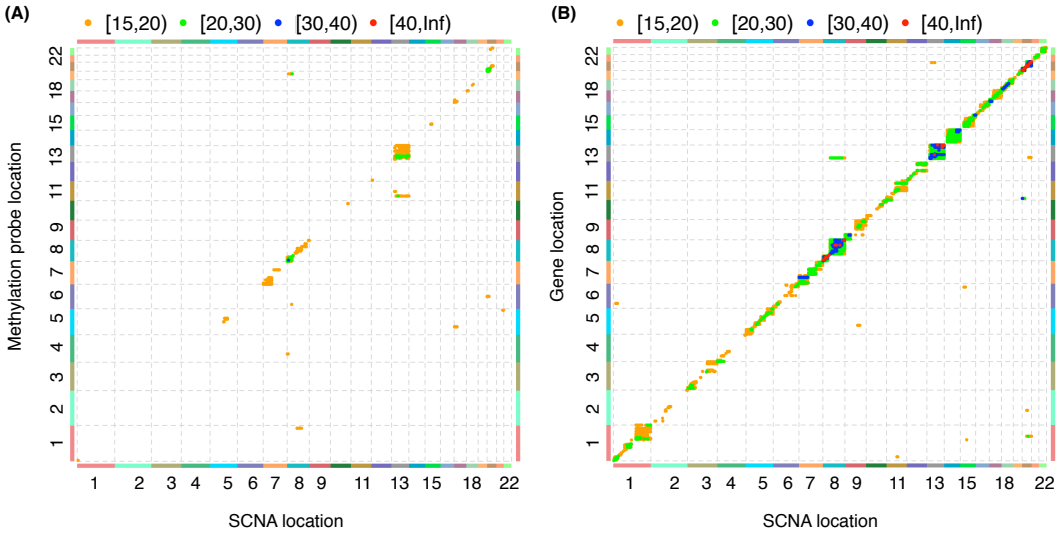

Figure S21: Association between SCNA and DNA methylation (A) and between SCNA and gene expression (B) in TCGA colon adenocarcinoma (COAD) samples.

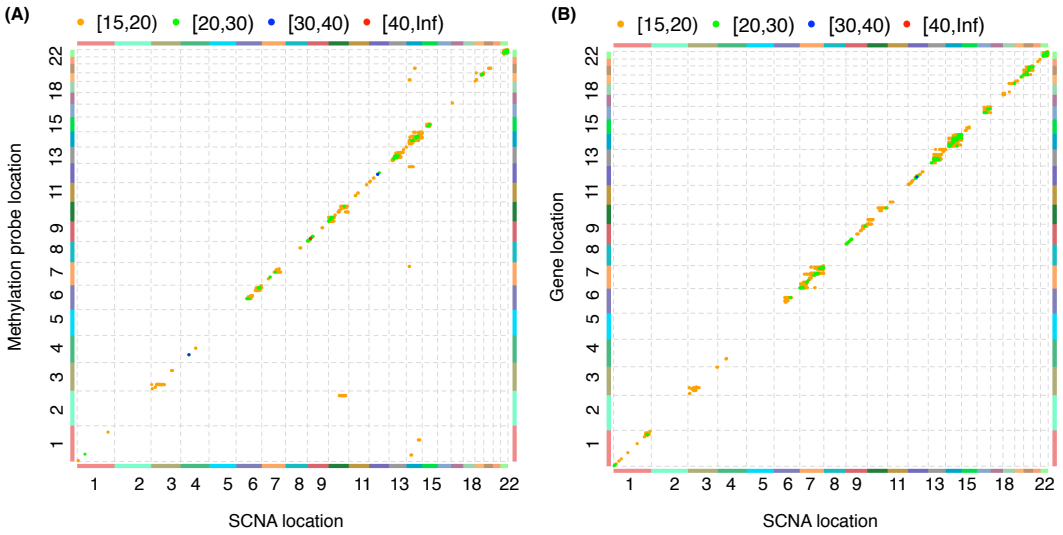

Figure S22: Association between SCNA and DNA methylation (A) and between SCNA and gene expression (B) in TCGA glioblastoma multiforme (GBM) samples.

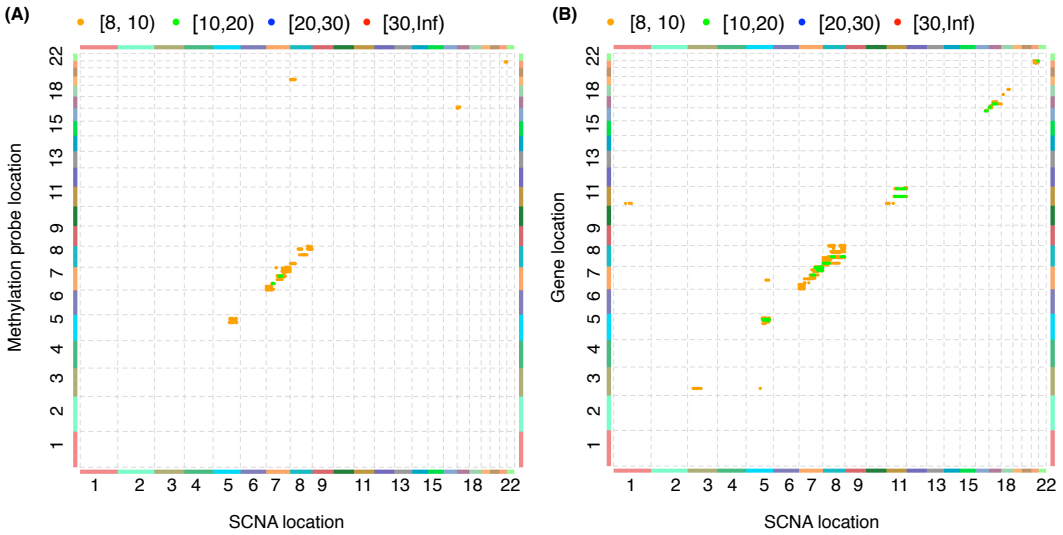

Figure S23: Association between SCNA and DNA methylation (A) and between SCNA and gene expression (B) in TCGA acute myeloid leukemia (LAML) samples. There are few SCNA events in LAML samples and thus few associations between SCNA and DNA methylation or gene expression.

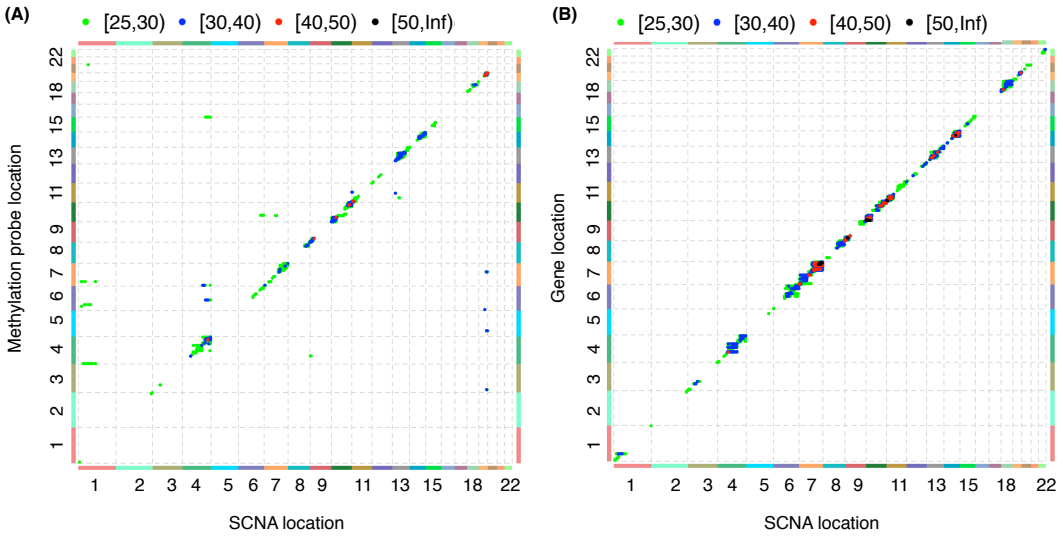

Figure S24: Association between SCNA and DNA methylation (A) and between SCNA and gene expression (B) in TCGA lower grade glioma (LGG) samples.

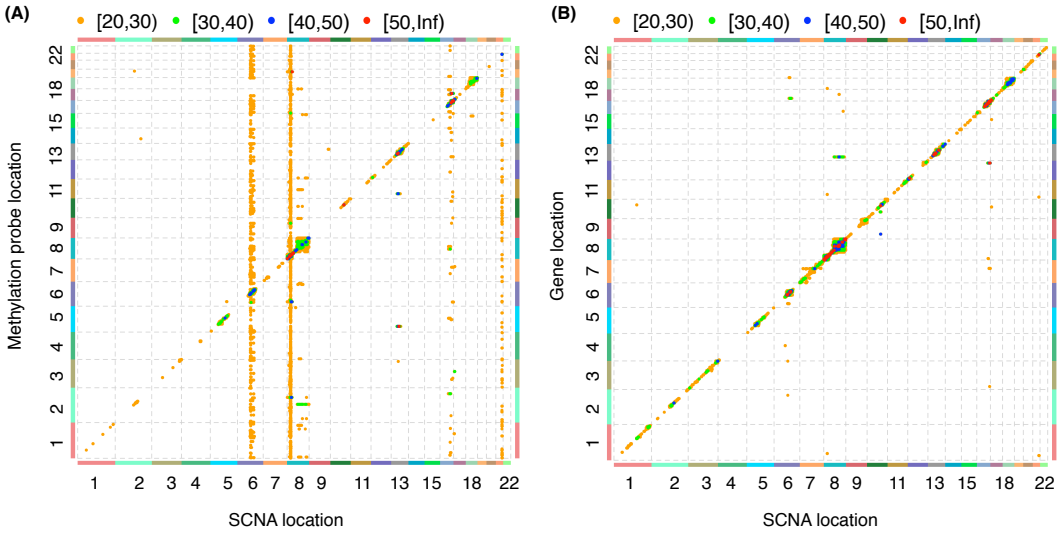

Figure S25: Association between SCNA and DNA methylation (A) and between SCNA and gene expression (B) in TCGA prostate adenocarcinoma (PRAD) samples.

The vast majority of the associations between SCNA and gene expression are positive and same-chromosome associations (Table S8). In contrast, the association between SCNA and DNA methylation is either positive or negative. Similar to the observation for breast cancer, most of the negative associations between SCNA and DNA methylation occur for methylation probes on CpG islands (Figures S26-S30).

	BRCA	COAD	GBM	LAML	LGG	PRAD
p-value threshold	10^{-20}	10^{-20}	10^{-10}	10^{-10}	10^{-20}	10^{-20}
# of eQCNs	877,954	104,608	176,948	21,589	188,243	114,137
% of same-chr eQCNs	99.52%	99.66%	99.99%	98.77%	99.75%	99.35%
% of positive association	99.99%	100.0%	99.997%	98.90%	99.91%	99.86%

Table S8: Summary of the association between SCNA and gene expression. The abbreviation eQCN stands for "expression Quantitative Trait Copy Number alterations".

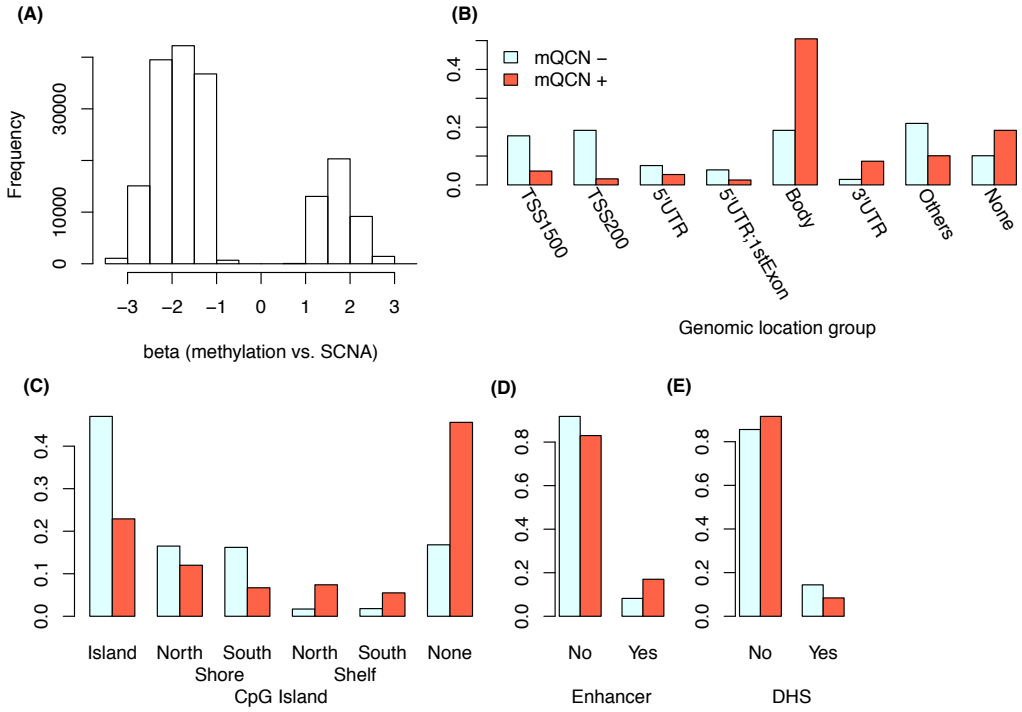

Figure S26: The distribution of the regression coefficients for methylation Quantitative trait Copy Number alterations (mQCN) (A) and summary of those methylation probes with positive or negative associations with SCNA (B-E) in COAD samples.

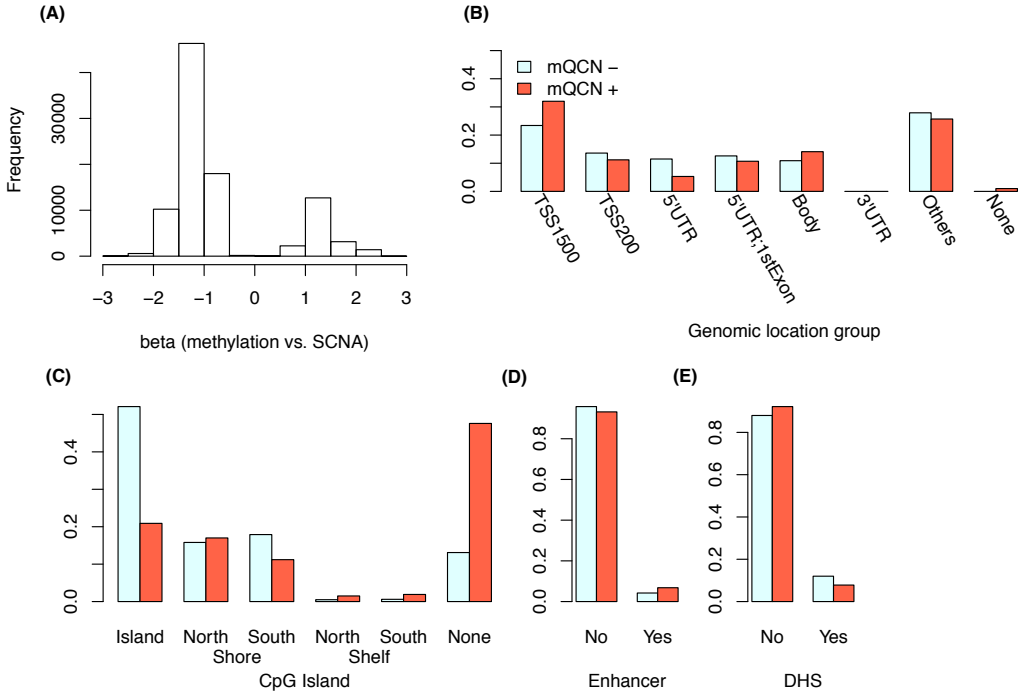

Figure S27: The distribution of the regression coefficients for methylation Quantitative Trait Copy Number alterations (mQCN) (A) and summary of those methylation probes with positive or negative associations with SCNA (B-E) in GBM samples.

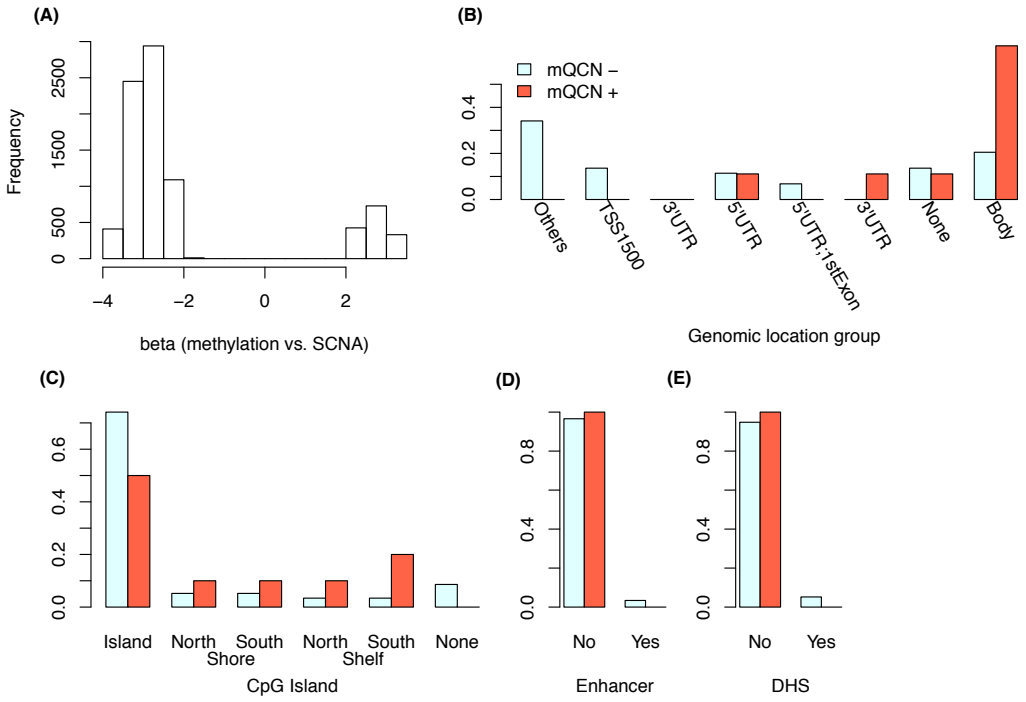

Figure S28: The distribution of the regression coefficients for methylation Quantitative Trait Copy Number alterations (mQCN) (A) and summary of those methylation probes with positive or negative associations with SCNA (B-E) in LAML samples.

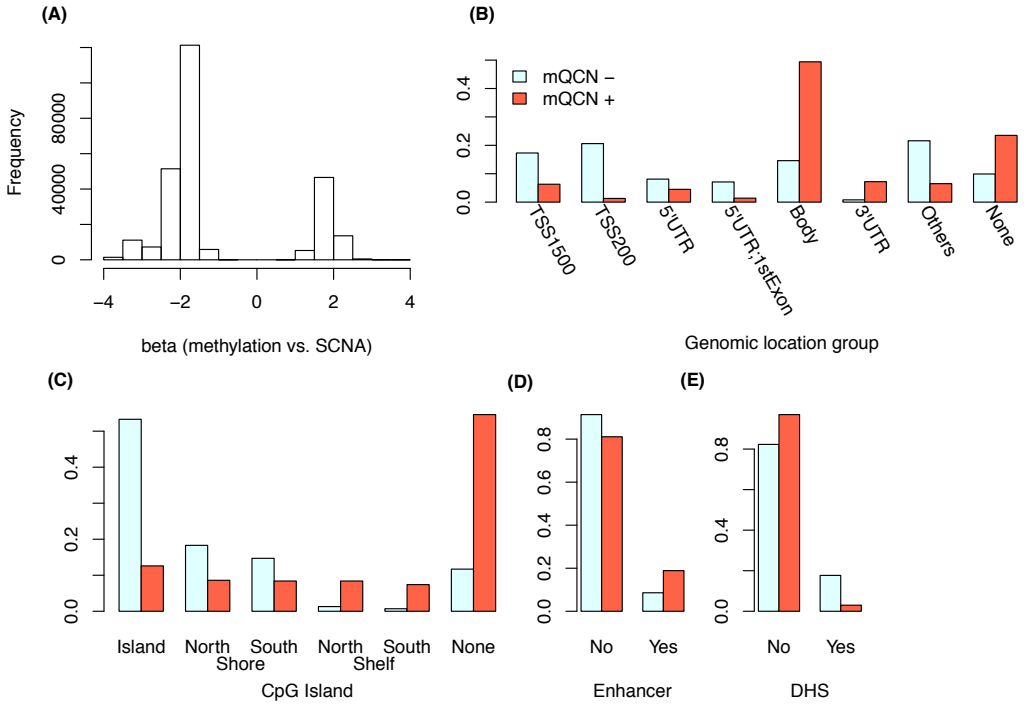

Figure S29: The distribution of the regression coefficients for methylation Quantitative Trait Copy Number alterations (mQCN) (A) and summary of those methylation probes with positive or negative associations with SCNA (B-E) in LGG samples.

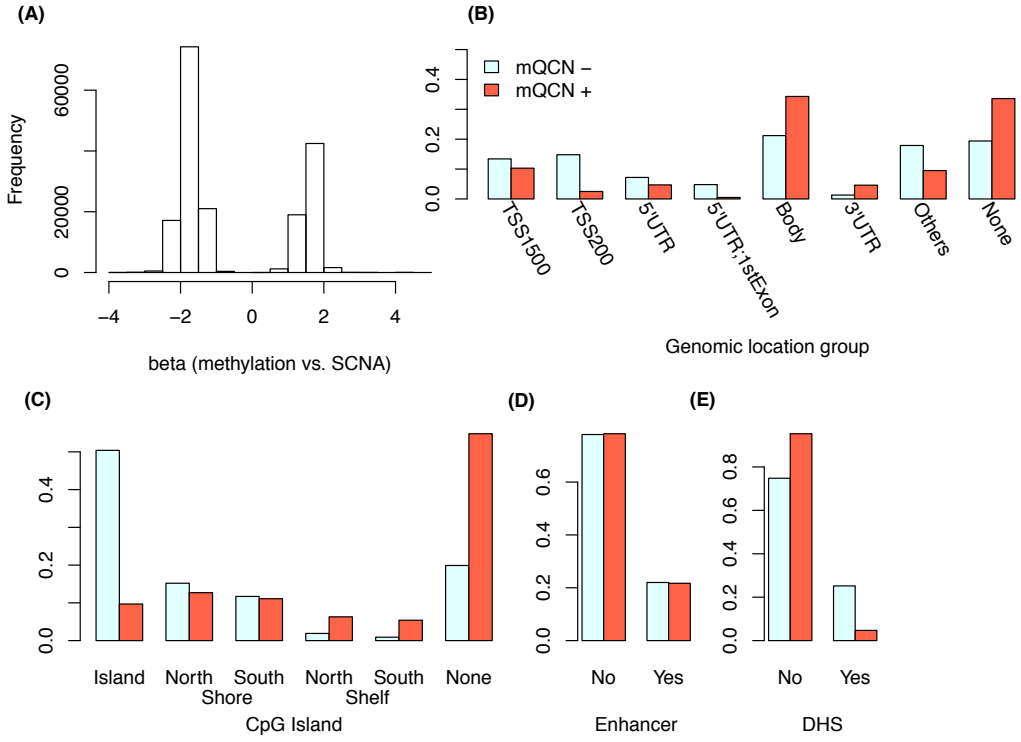

Figure S30: The distribution of the regression coefficients for methylation Quantitative Trait Copy Number alterations (mQCN) (A) and summary of those methylation probes with positive or negative associations with SCNA (B-E) in PRAD samples.

B.2.3 Association between DNA methylation and gene expression

Following our approach to study the association between DNA methylation and gene expression in breast cancer, we conducted similar analyses for the other five types of cancers. Specifically, we first assessed genome-wide pair-wise associations between DNA methylation and gene expression (i.e. ME associations), while accounting for a set of standard covariates (Table S7). Then, we selected the top 100-200k ME associations. For each gene expression or DNA methylation in these top pairs, we took residuals from linear regression against all known covariates. Then, we averaged the residuals of E and M within each ME pair and performed PCA on the resulting data matrix, where each row corresponds to an ME pair and each column corresponds to a sample. Finally, we assessed the association between DNA methylation and gene expression after including 7 PCs in the model. The choice of top 7 PCs was arbitrary. It can been seen from the following Figures S31-S35 that 7 PCs are enough to remove most of the off-diagonal associations for GBM, LAML, and LGG, but some off-diagonal associations remain for COAD and PRAD. In fact, the results across cancer types are not directly comparable due to different sample sizes and different p-value cutoffs for ME associations. However, findings from all cancer types support the conclusion that the vast majority of the ME associations can be explained by a few dimensions.

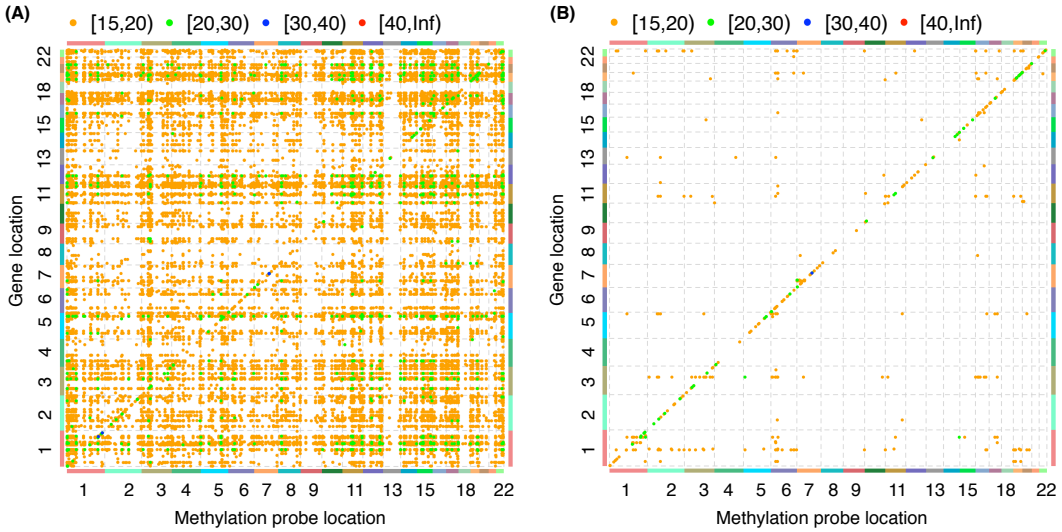

Figure S31: Association between DNA methylation and gene expression in COAD samples. (A) Results when accounting for the covariates listed in Table S7. (B) Results when we further account for 7 ME PCs.

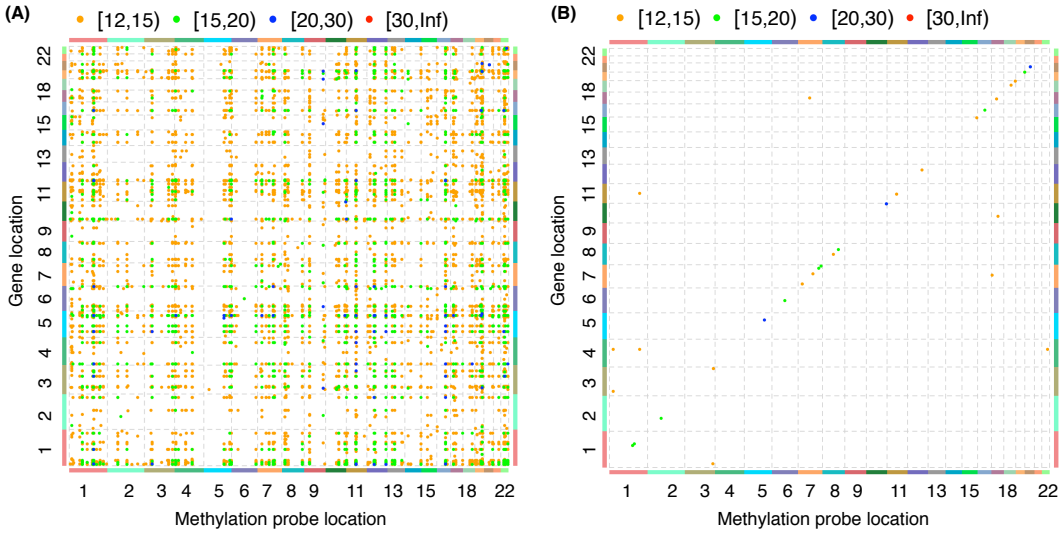

Figure S32: Association between DNA methylation and gene expression in GBM samples. (A) Results when accounting for the covariates listed in Table S7. (B) Results when we further account for 7 ME PCs.

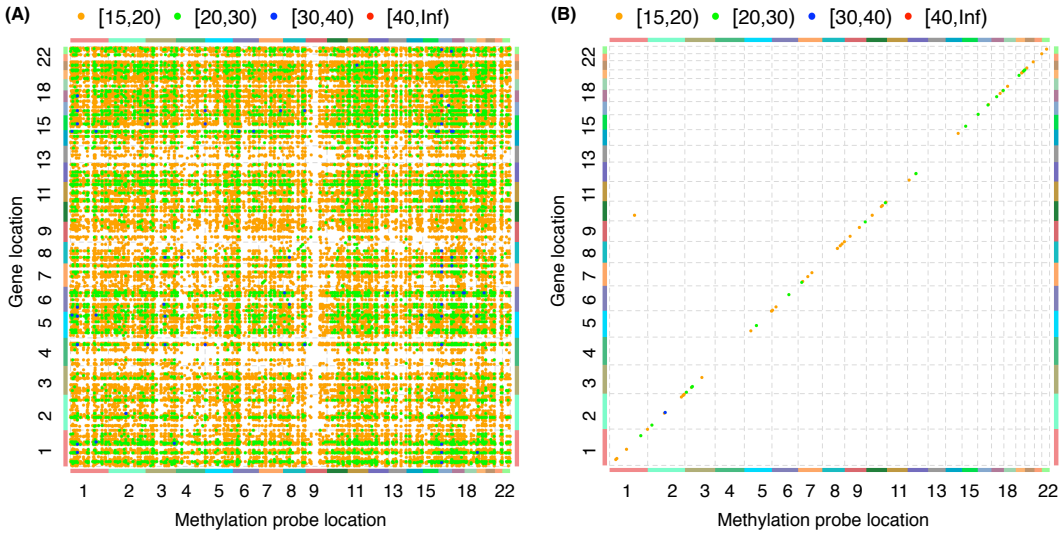

Figure S33: Association between DNA methylation and gene expression in LAML samples. (A) Results when accounting for the covariates listed in Table S7. (B) Results when we further account for 7 ME PCs.

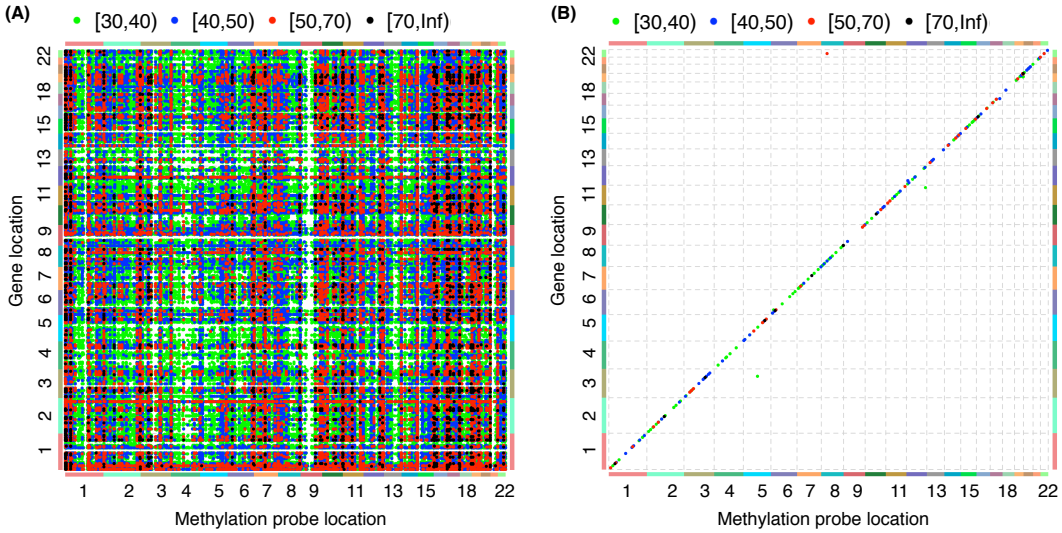

Figure S34: Association between DNA methylation and gene expression in LGG samples. (A) Results when accounting for the covariates listed in Table S7. (B) Results when we further account for 7 ME PCs.

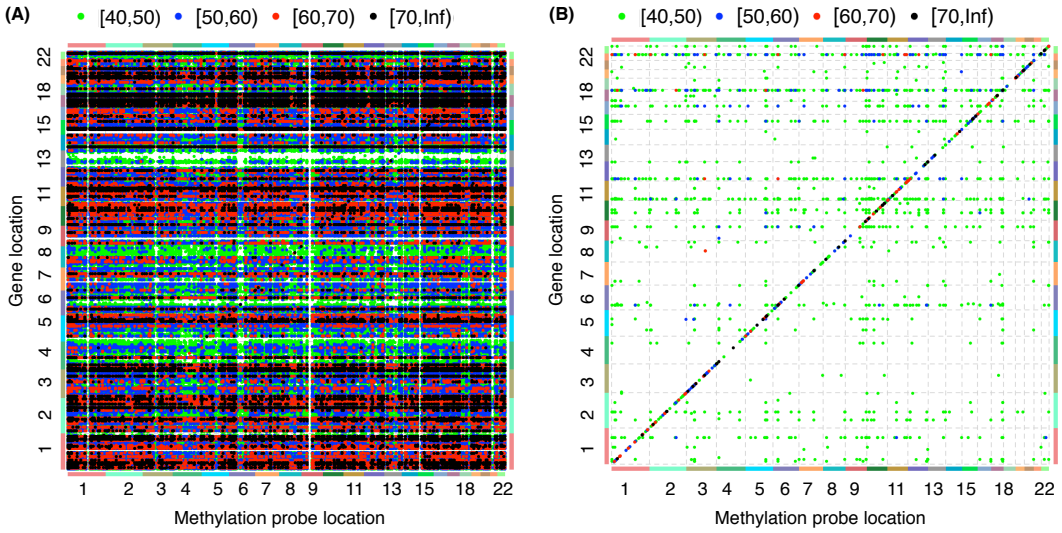

Figure S35: Association between DNA methylation and gene expression in PRAD samples. (A) Results when accounting for the covariates listed in Table S7. (B) Results when we further account for 7 ME PCs.

B.2.4 Characterization of hot genes

For GBM, we define hot genes as genes that are associated with at least 20 methylation probes at a p-value cutoff 10^{-15}. For other cancer types, we define hot genes as genes that are associated with at least 100 methylation probes at specific p-value cutoffs. We used a different cutoff for GBM because we only included those methylation probes that are shared by the HM27 and HM450 arrays. These p-value cutoffs and number of (annotated) hot genes are listed in Table S9. A gene is annotated if it corresponds to an Ensembl gene ID, which can be recognized by the DAVID functional annotation tool (`https://david.ncifcrf.gov/summary.jsp`).

	BRCA	COAD	GBM	LAML	LGG	PRAD
p-value threshold	10^{-60}	10^{-15}	10^{-15}	10^{-15}	10^{-40}	10^{-60}
total # of ME associations	146,245	42,523	1,631	89,329	145,791	251,316
# of hot genes	182	126	25	212	421	213
# of hot genes with annotation	173	120	24	199	393	205

Table S9: Summary of the ME associations.

For each cancer type, we queried hot genes at DAVID functional annotation tool (`https://david.ncifcrf.gov/summary.jsp`).

Term	Count	Percent	Expected Percent	P-Value	Benjamini
GO:0046649 lymphocyte activation	23	19.2	1.47	4.19E-21	3.84E-18
GO:0045321 leukocyte activation	23	19.2	1.79	3.09E-19	1.42E-16
GO:0001775 cell activation	23	19.2	2.12	1.22E-17	3.73E-15
GO:0042110 T cell activation	16	13.3	0.93	4.49E-15	1.02E-12
GO:0005886 plasma membrane	57	47.5	29.55	1.69E-11	2.11E-09
GO:0002684 positive regulation of immune system process	16	13.3	1.76	5.15E-11	9.45E-09
GO:0006955 immune response	24	20.0	5.10	1.04E-10	1.58E-08
GO:0051249 regulation of lymphocyte activation	13	10.8	1.09	2.82E-10	3.69E-08
GO:0002694 regulation of leukocyte activation	13	10.8	1.23	1.07E-09	1.22E-07
GO:0009897 external side of plasma membrane	13	10.8	1.33	1.55E-09	9.69E-08

Table S10: The functional enrichment of 120 annotated hot genes from the COAD data. The columns "Count" and "Percent" are the number and percentage, respectively, of genes that belong to certain functional categories. The column "Expected %" is the expected percentage based on all annotated genes. "P-value" and "Benjamini" indicate the significance before and after multiple testing correction.

Term	Count	Percent	Expected Percent	P-Value	Benjamini
GO:0006955 immune response	11	45.8	5.10	6.79E-09	2.93E-06
GO:0006952 defense response	10	41.7	4.55	4.80E-08	1.04E-05
GO:0006954 inflammatory response	8	33.3	2.40	1.70E-07	2.45E-05
GO:0009611 response to wounding	8	33.3	3.92	4.55E-06	4.91E-04
GO:0002443 leukocyte mediated immunity	5	20.8	0.64	5.49E-06	4.74E-04
GO:0002684 positive regulation of immune system process	6	25.0	1.76	1.54E-05	1.11E-03
GO:0002252 immune effector process	5	20.8	0.99	3.18E-05	1.96E-03
GO:0050778 positive regulation of immune response	5	20.8	1.07	4.33E-05	2.34E-03
GO:0002449 lymphocyte mediated immunity	4	16.7	0.52	1.21E-04	5.80E-03
GO:0002250 adaptive immune response	4	16.7	0.57	1.61E-04	6.93E-03

Table S11: The functional enrichment of 24 annotated hot genes from the GBM data.

Term	Count	Percent	Expected Percent	P-Value	Benjamini
GO:0030097 hemopoiesis	19	9.5	1.74	1.08E-10	1.27E-07
GO:0048534 hemopoietic or lymphoid organ development	19	9.5	1.92	5.28E-10	3.09E-07
GO:0002520 immune system development	19	9.5	2.04	1.38E-09	5.39E-07
GO:0042101 T cell receptor complex	7	3.5	0.09	1.61E-09	3.28E-07
GO:0005886 plasma membrane	76	38.2	29.55	1.07E-08	1.09E-06
GO:0006955 immune response	27	13.6	5.10	3.18E-08	9.30E-06
GO:0002694 regulation of leukocyte activation	14	7.0	1.23	3.53E-08	8.27E-06
GO:0030217 T cell differentiation	10	5.0	0.48	3.54E-08	6.91E-06
GO:0050865 regulation of cell activation	14	7.0	1.29	6.60E-08	1.10E-05
GO:0042110 T cell activation	12	6.0	0.93	1.37E-07	2.01E-05

Table S12: The functional enrichment of 199 annotated hot genes from the LAML data.

Term	Count	Percent	Expected Percent	P-Value	Benjamini
GO:0044456 synapse part	35	8.9	1.92	1.00E-17	2.97E-15
GO:0045202 synapse	39	9.9	2.78	7.51E-16	1.15E-13
GO:0019226 transmission of nerve impulse	37	9.4	2.59	2.69E-15	4.36E-12
GO:0043005 neuron projection	37	9.4	2.68	7.97E-15	7.91E-13
GO:0007268 synaptic transmission	32	8.1	2.20	1.94E-13	1.59E-10
GO:0006836 neurotransmitter transport	18	4.6	0.61	1.22E-12	6.64E-10
GO:0008021 synaptic vesicle	17	4.3	0.59	8.88E-12	6.59E-10
GO:0030136 clathrin-coated vesicle	19	4.8	1.03	8.50E-10	5.05E-08
GO:0007269 neurotransmitter secretion	11	2.8	0.25	1.28E-09	5.22E-07
GO:0044459 plasma membrane part	91	23.2	17.24	2.36E-09	1.17E-07

Table S13: The functional enrichment of 393 annotated hot genes from the LGG data.

Term	Count	Percent	Expected Percent	P-Value	Benjamini
GO:0044459 plasma membrane part	68	33.2	17.24	9.15E-13	1.73E-10
GO:0005886 plasma membrane	92	44.9	29.55	4.74E-12	4.48E-10
GO:0050867 positive regulation of cell activation	13	6.3	0.82	4.96E-09	6.57E-06
GO:0031226 intrinsic to plasma membrane	42	20.5	9.51	6.06E-09	3.82E-07
GO:0005887 integral to plasma membrane	41	20.0	9.29	1.04E-08	4.93E-07
GO:0002696 positive regulation of leukocyte activation	12	5.9	0.78	3.41E-08	2.26E-05
GO:0045058 T cell selection	7	3.4	0.14	4.93E-08	2.17E-05
GO:0007155 cell adhesion	27	13.2	5.17	9.70E-08	3.21E-05
GO:0022610 biological adhesion	27	13.2	5.18	9.98E-08	2.64E-05
GO:0050865 regulation of cell activation	14	6.8	1.29	1.05E-07	2.33E-05

Table S14: The functional enrichment of 205 annotated hot genes from the PRAD data.

B.2.5 Characterization of hot methylation probes and eQTMs

From the results of the ME association analyses, two types of methylation probes are of interest: hot methylation probes that are associated with a number of genes and gene expression quantitative trait methylation sites (eQTMs) that are associated with gene expression after accounting for ME-PCs. Figures S36-S40 summarize these two types of methylation probes across cancer types. The conclusions are consistent across cancer types: hot methylation probes are less likely to be located at CpG islands. eQTMs can be divided into two classes: those that are positively or negatively associated with the corresponding genes. Those with negative associations with gene expression (`eQTM-`) tend to located at transcription starting sites (TSSs), and those with positive associations with gene expression (`eQTM+`) tend to located on gene bodies.

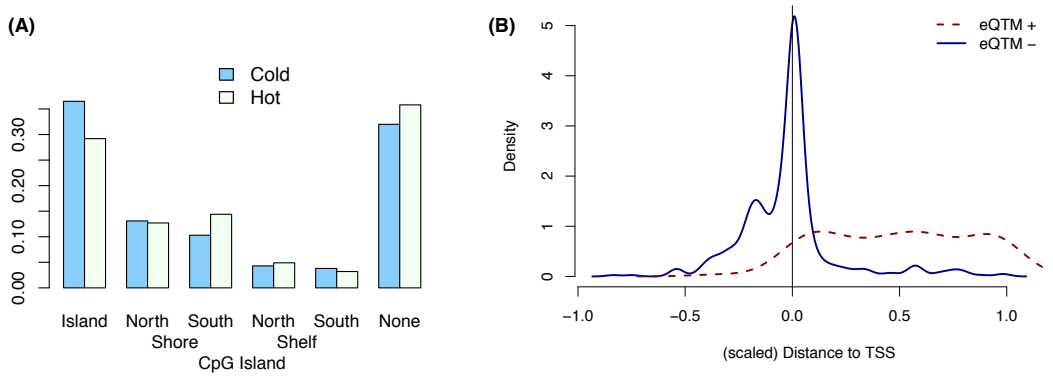

Figure S36: (A) Genomic location of hot/cold methylation sites with respect to CpG islands. (B) Genomic location of eQTMs with respect to their associated genes in COAD samples. eQTM+ / eQTM - indicate methylation probes with positive/negative associations with the corresponding genes. The distance to the left and right side of origin has been scaled such that 0 means transcription starting site (TSS), -1 means 1000bp upstream, and 1 means the end of the gene body.

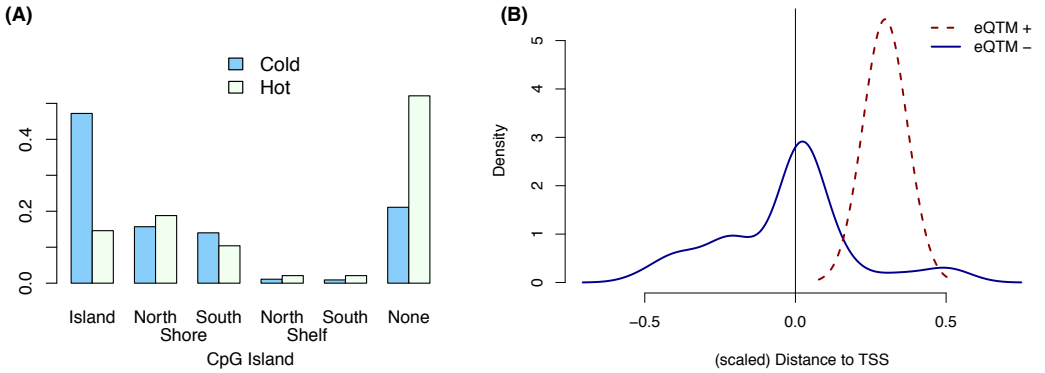

Figure S37: (A) Genomic location of hot/cold methylation sites with respect to CpG islands. (B) Genomic location of eQTMs with respect to their associated genes in GBM samples. The density curve in panel (B) should be interpreted with caution because there are only 40 eQTMs, with 38 being `eQTM-` and 2 being `eQTM+`.

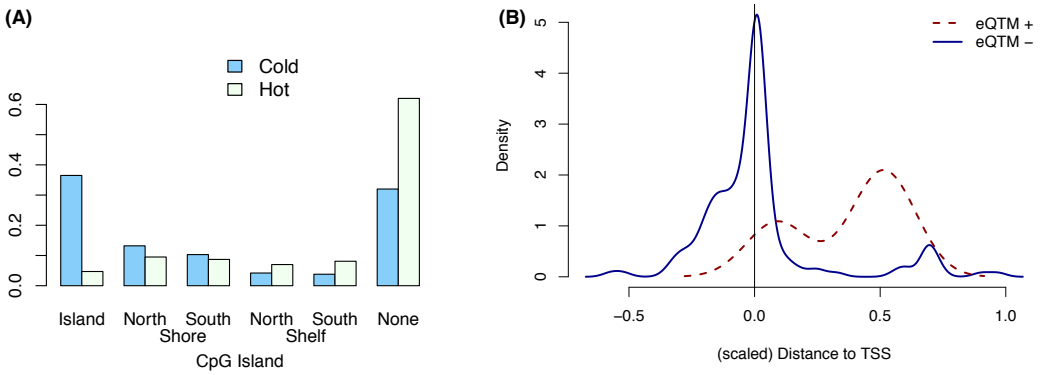

Figure S38: (A) Genomic location of hot/cold methylation sites with respect to CpG islands. (B) Genomic locations of eQTMs with respect to their associated genes in LAML samples.

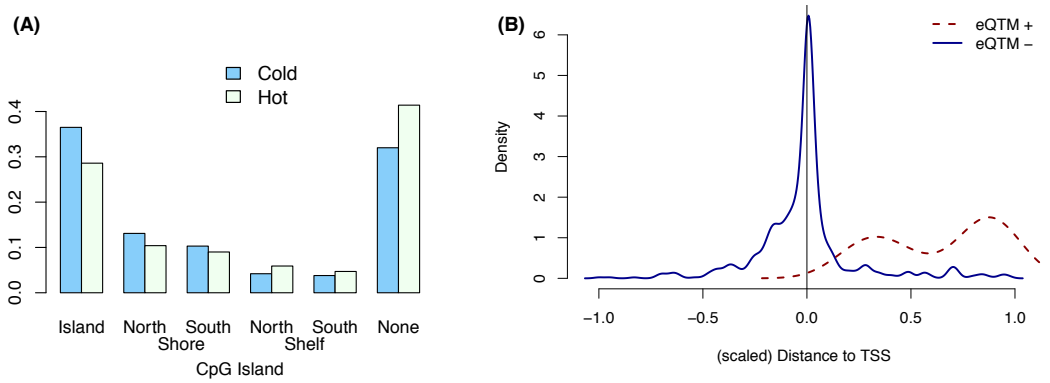

Figure S39: (A) Genomic location of hot/cold methylation sites with respect to CpG islands. (B) Genomic location of eQTMs with respect to their associated genes in LGG samples.

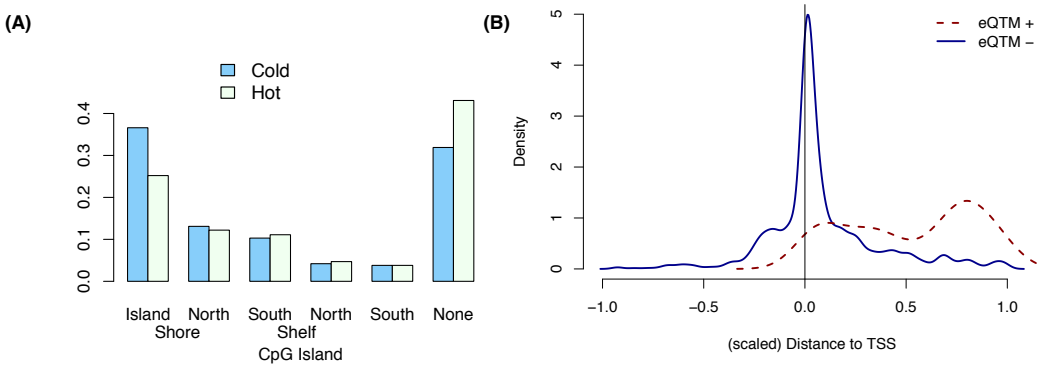

Figure S40: (A) Genomic location of hot/cold methylation sites with respect to CpG islands. (B) Genomic location of eQTMs with respect to their associated genes in PRAD samples.

B.3 Additional results for all cancer types

B.3.1 Magnitude of ME eigenvalues

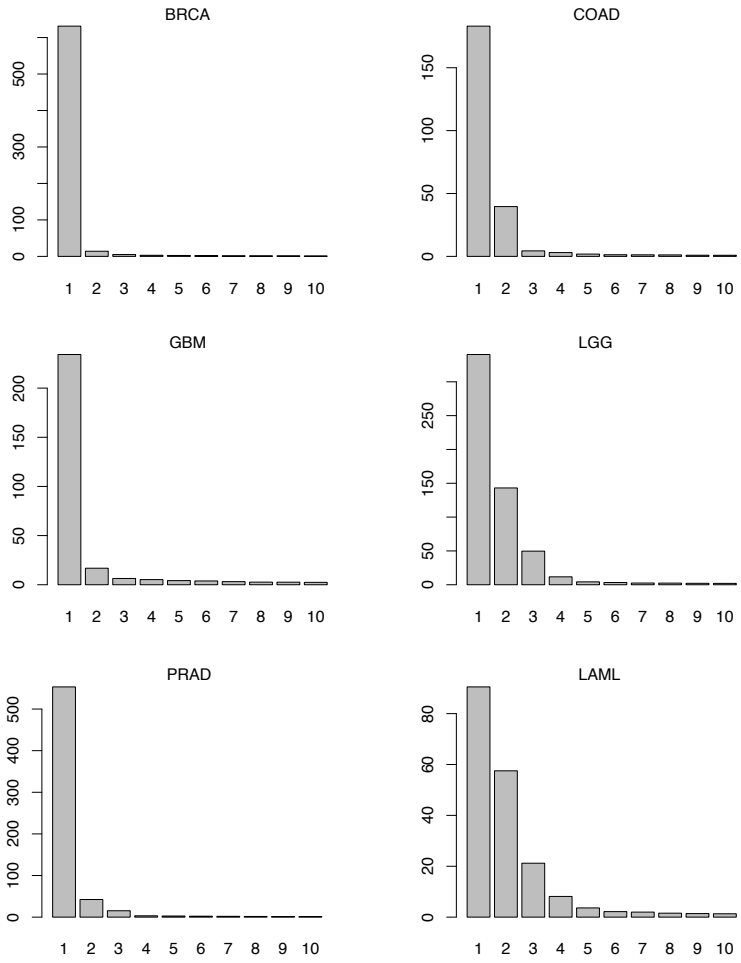

Figure S41: The eigenvalues of the top 10 ME PCs across six cancer types. The size of the eigenvalues varies across cancer types, which implies that the meaning of the corresponding PCs may differ across cancer types.

B.3.2 Association between expression of cell type-specific genes and ME PCs

The following figures show the relative percentage of variance of cell type-specific genes explained by ME PCs. These are relative percentages in the sense that we calculated the R^2 explained by each cell type together with batch effects, demographic variables, and cancer subtypes. The percentage is calculated by dividing the R^2 explained by each PC with the total R^2 explained by all PCs.

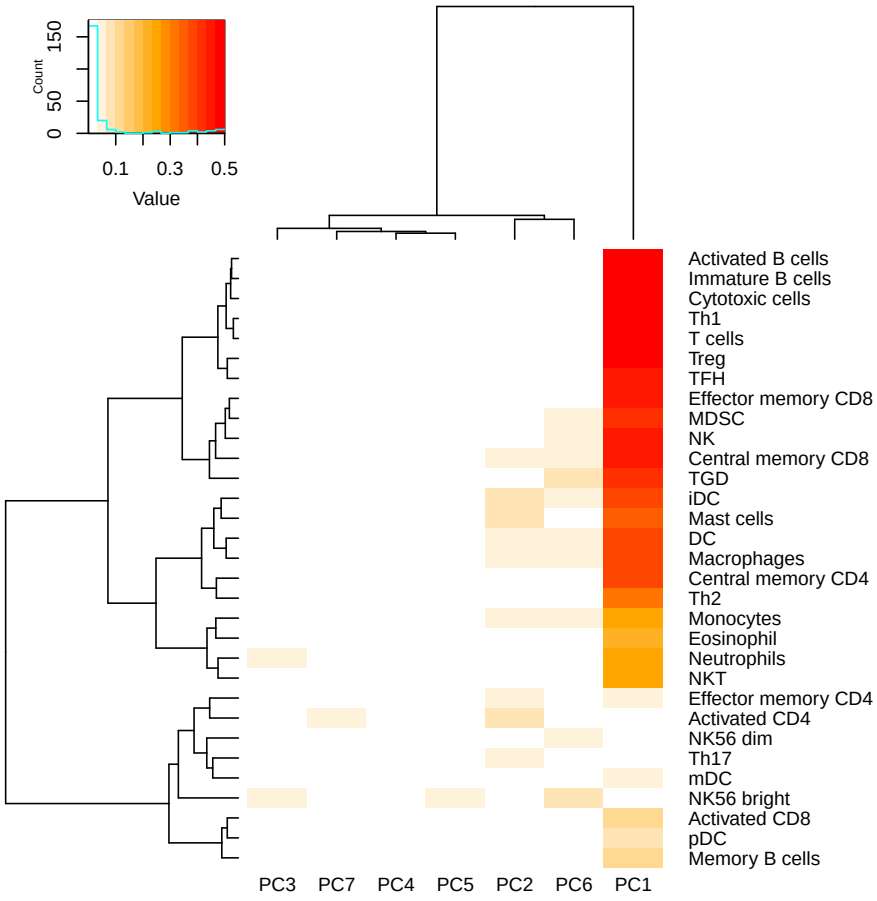

Figure S42: R^2 of cell type-specific gene expression explained by each of the first 7 ME PCs in COAD samples.

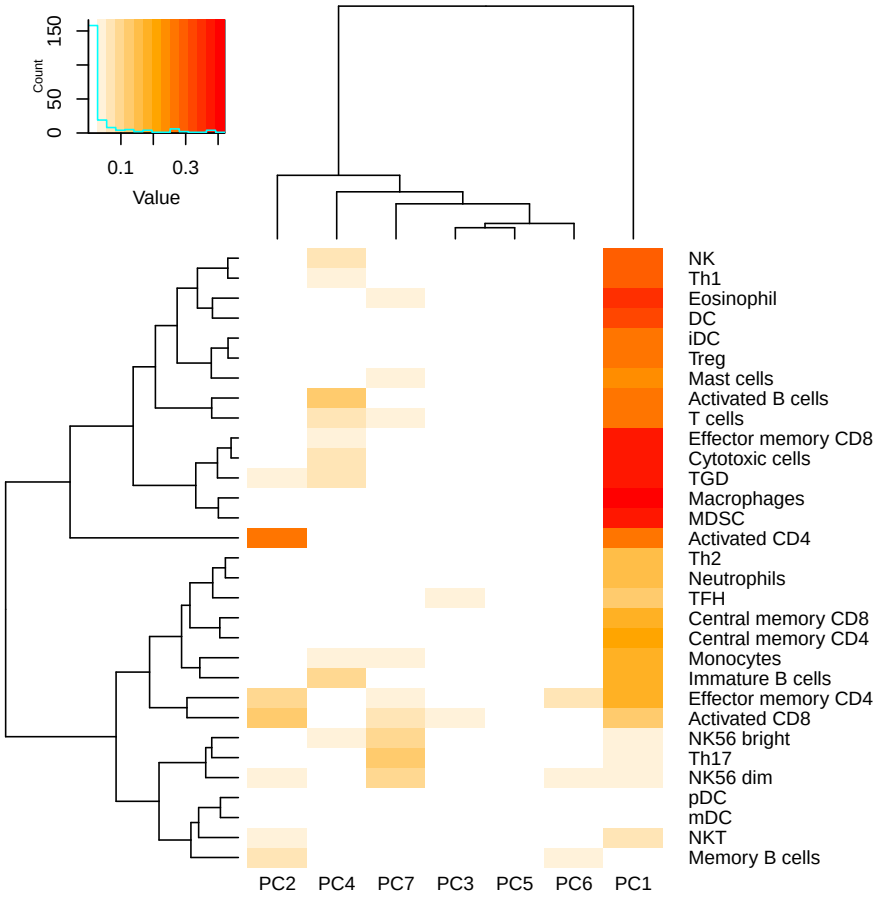

Figure S43: R^2 of cell type-specific gene expression explained by each of the first 7 ME PCs in GBM samples.

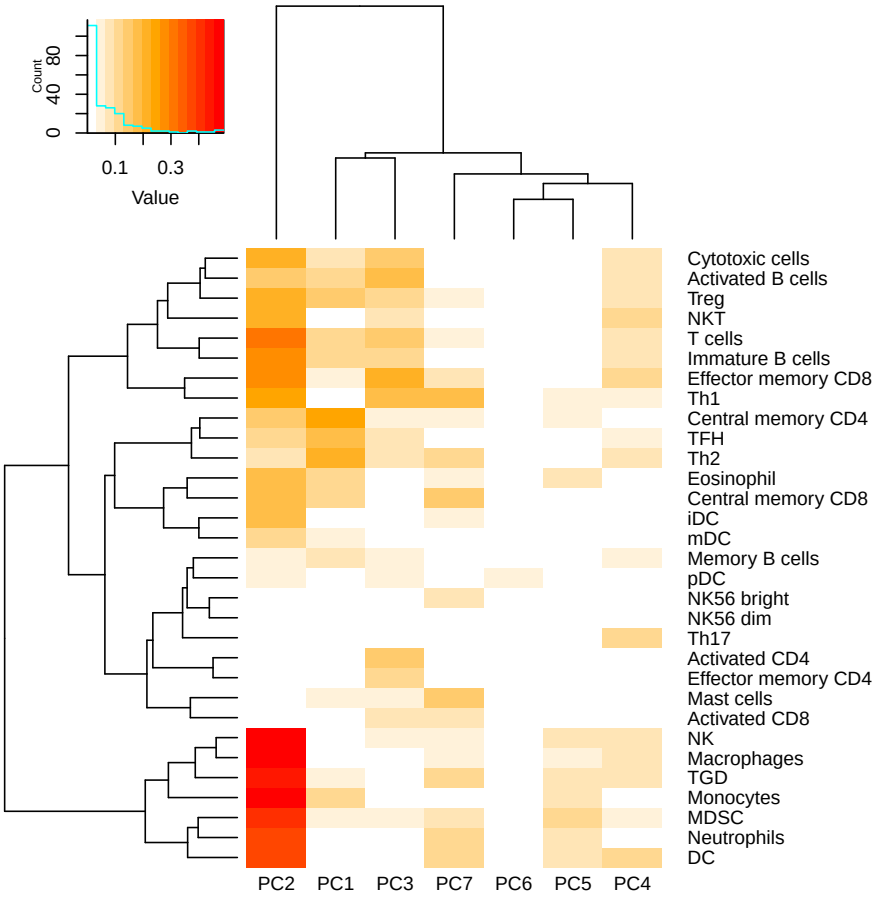

Figure S44: R^2 of cell type-specific gene expression explained by each of the first 7 ME PCs in LAML samples.

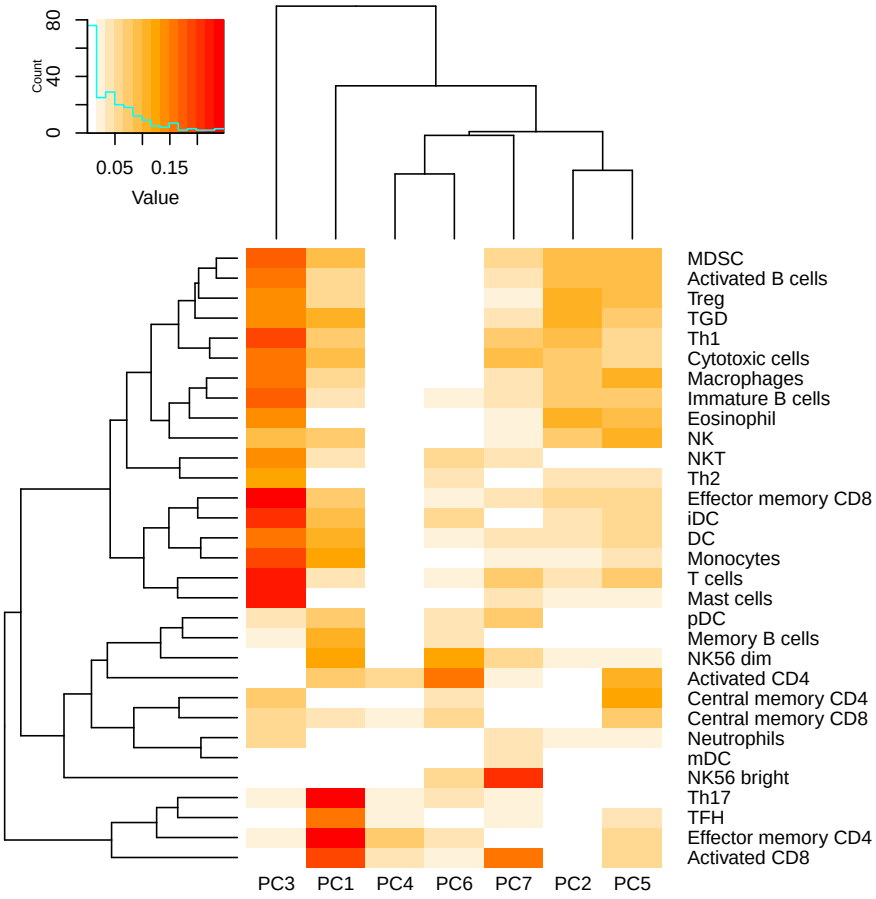

Figure S45: R^2 of cell type-specific gene expression explained by each of the first 7 ME PCs in LGG samples.

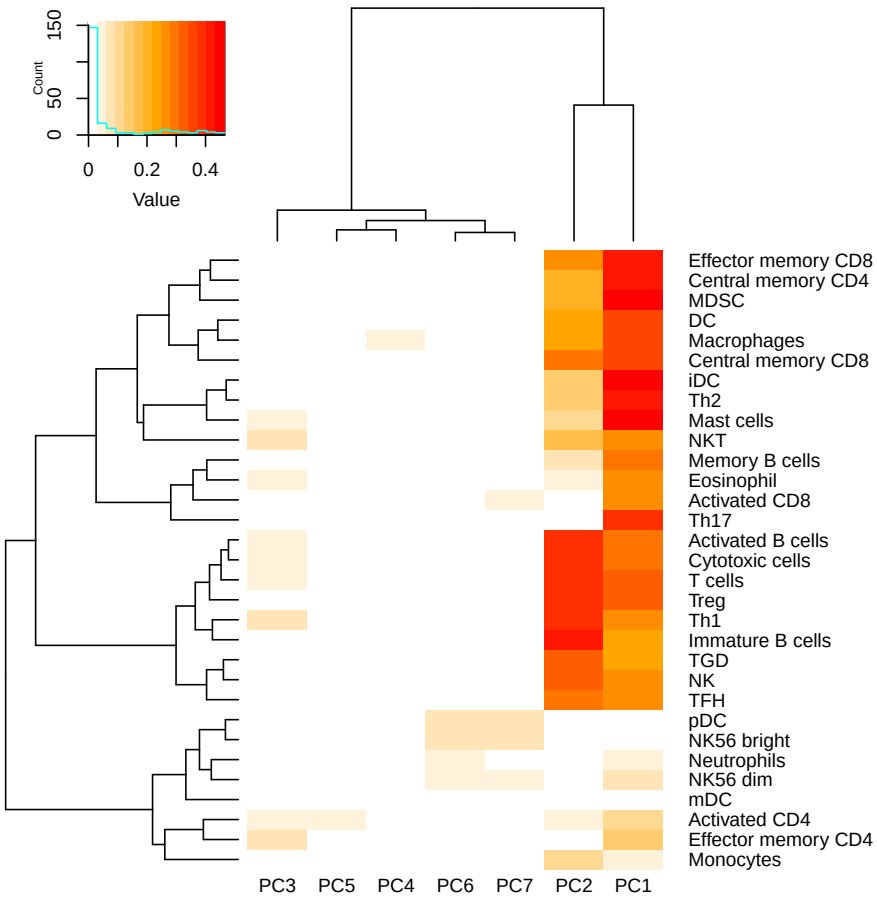

Figure S46: R^2 of cell type-specific gene expression explained by each of the first 7 ME PCs in PRAD samples.

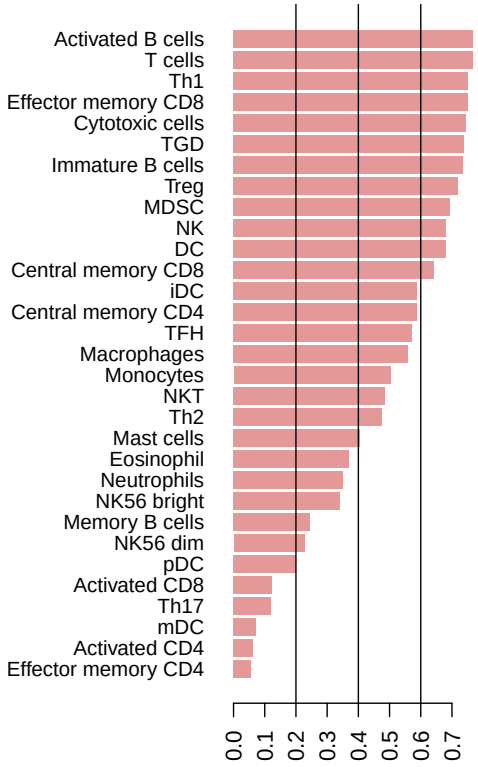

Figure S47: Total R^2 of cell type-specific gene expression explained by each of the first 7 ME PCs in BRCA samples.

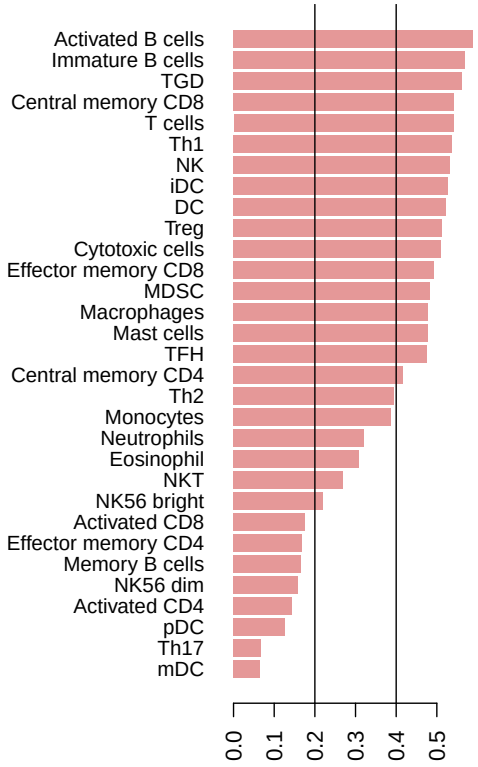

Figure S48: Total R^2 of cell type-specific gene expression explained by the first 7 ME PCs in COAD samples.

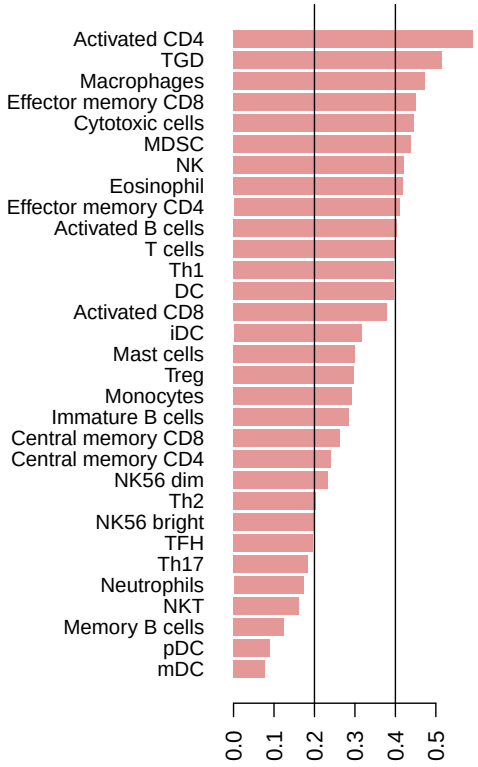

Figure S49: Total R^2 of cell type-specific gene expression explained by the first 7 ME PCs in GBM samples.

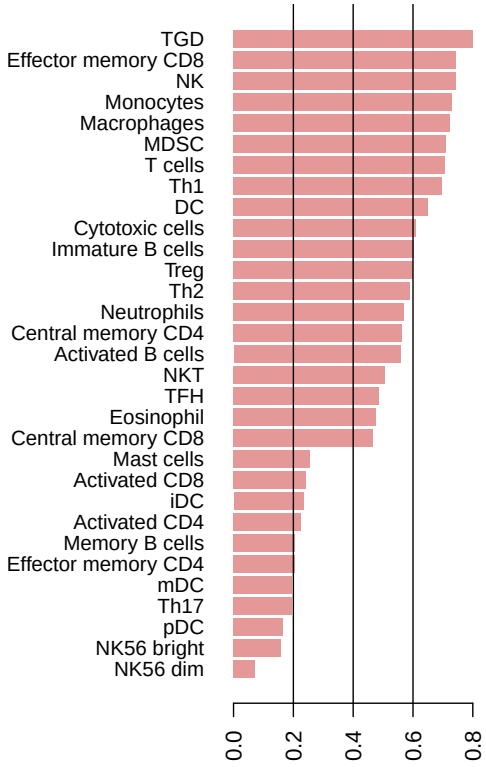

Figure S50: Total R^2 of cell type-specific gene expression explained by the first 7 ME PCs in LAML samples.

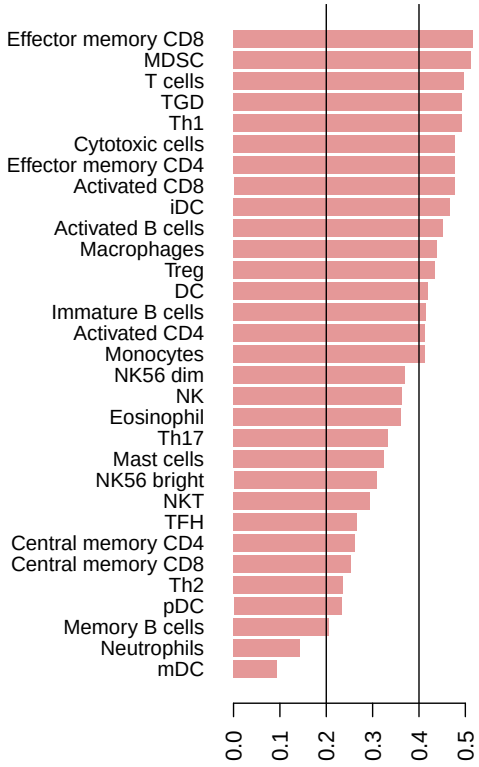

Figure S51: Total R^2 of cell type-specific gene expression explained by the first 7 ME PCs in LGG samples.

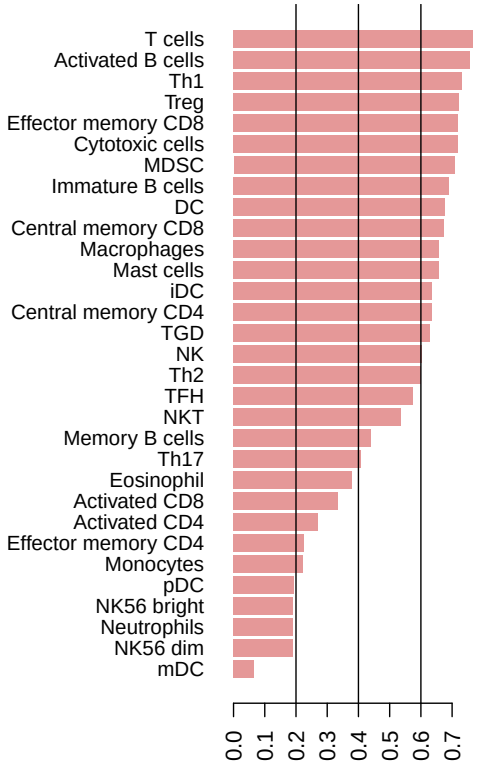

Figure S52: Total R^2 of cell type-specific gene expression explained by the first 7 ME PCs in PRAD samples.

B.3.3 Two-way associations and conditional associations for gene expression, SCNA, and DNA methylation

In the following figures (one for each cancer type), we compare the -$\log_{10}$ p-values for the associations of two types omic data versus the -$\log_{10}$ p-values for conditional associations. The left panel compares SCNA-expression (CE) associations versus CE associations given DNA methylation (CE | M). We examined all the CE pairs with significant associations and choose the local methylation probe that has the most significant association with the SCNA measurement. Both the middle panel and the right panel examine the CM pairs with significant CM associations. The middle panel compares CM associations versus CM | E associations. The right panel compares ME associations versus ME | C associations. All associations and conditional associations are assessed after accounting for the effects of batch effects, demographic variables (e.g., age, gender, genotype PCs), and top 7 ME PCs.

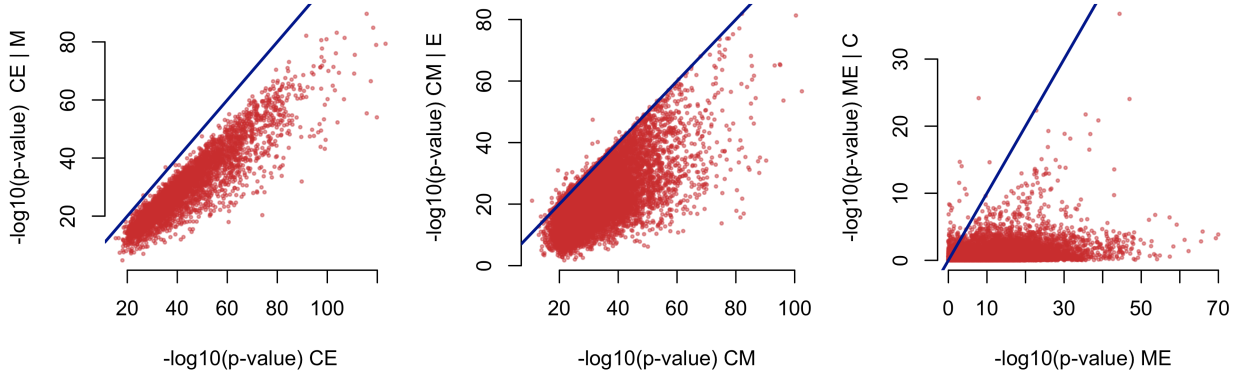

Figure S53: Comparison of associations and conditional associations in BRCA samples

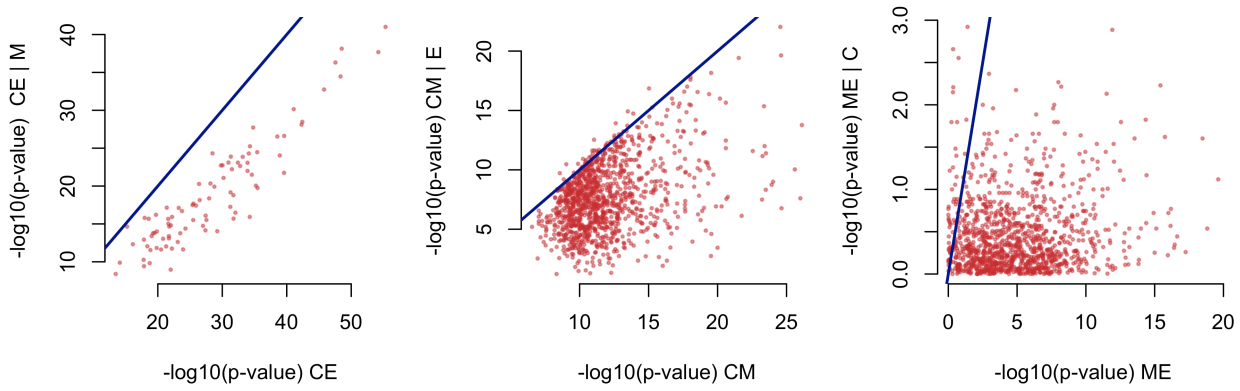

Figure S54: Comparison of associations and conditional associations in COAD samples

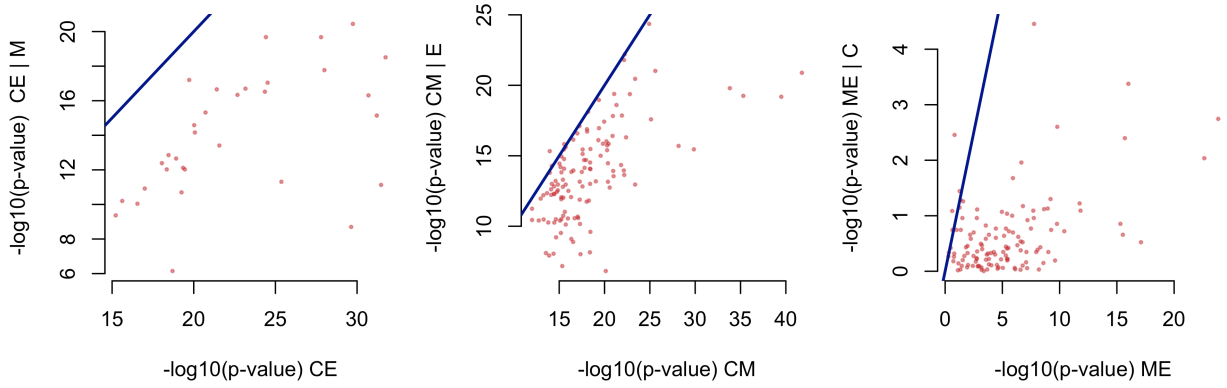

Figure S55: Comparison of associations and conditional associations in GBM samples

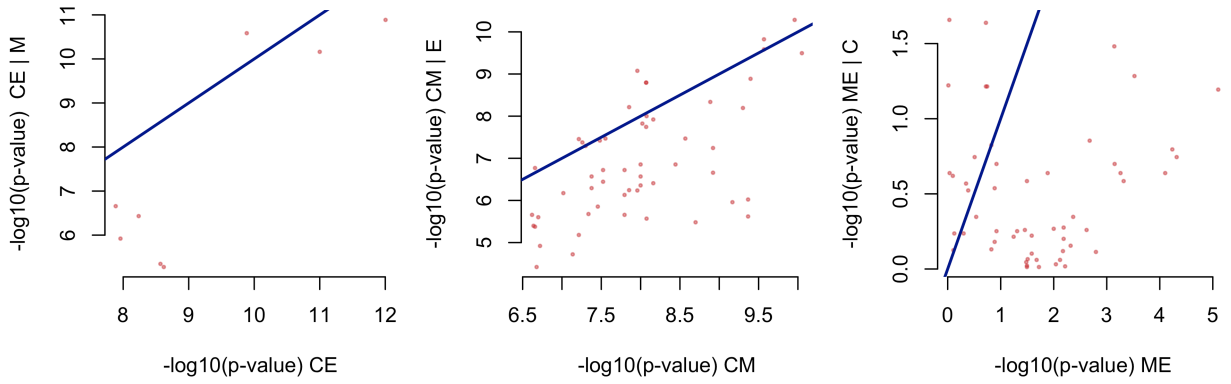

Figure S56: Comparison of associations and conditional associations in LAML samples

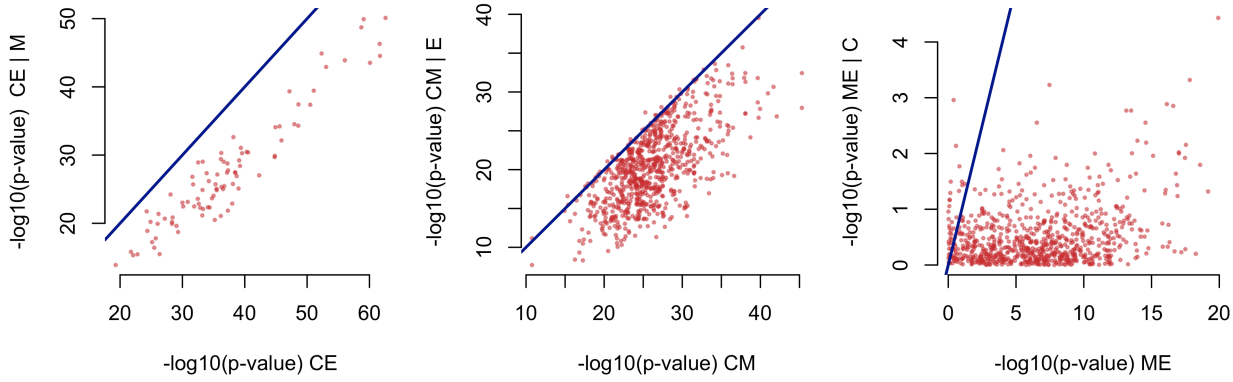

Figure S57: Comparison of associations and conditional associations in LGG samples

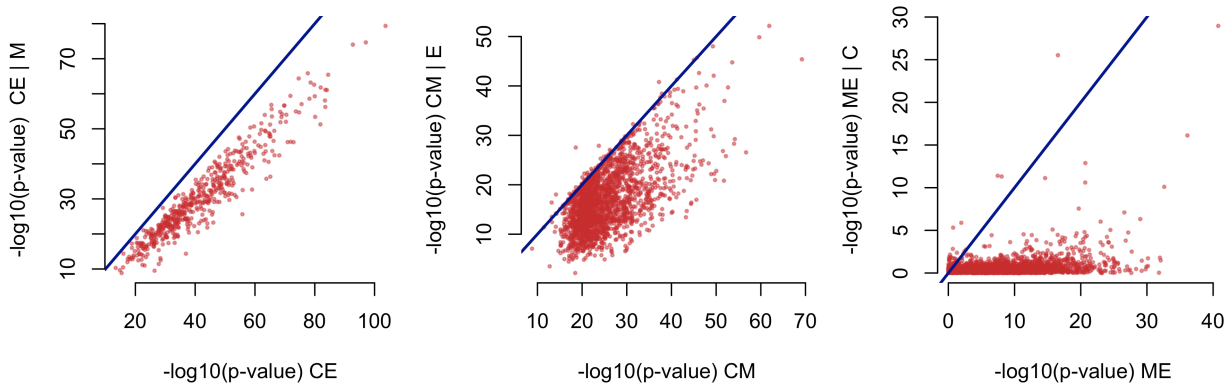

Figure S58: Comparison of associations and conditional associations in PRAD samples

B.3.4 Associations between SCNA and DNA methylation

We first illustrate the SCNA data by heatmaps for all cancer types (the heatmap for BRCA samples has been presented earlier as Supplementary Figure 5) and then plot the average SCNA for each cancer type across the genome, and overlay with labels for the genes with strongest CM associations, up to 1,000 CM associations for readability of the figure.

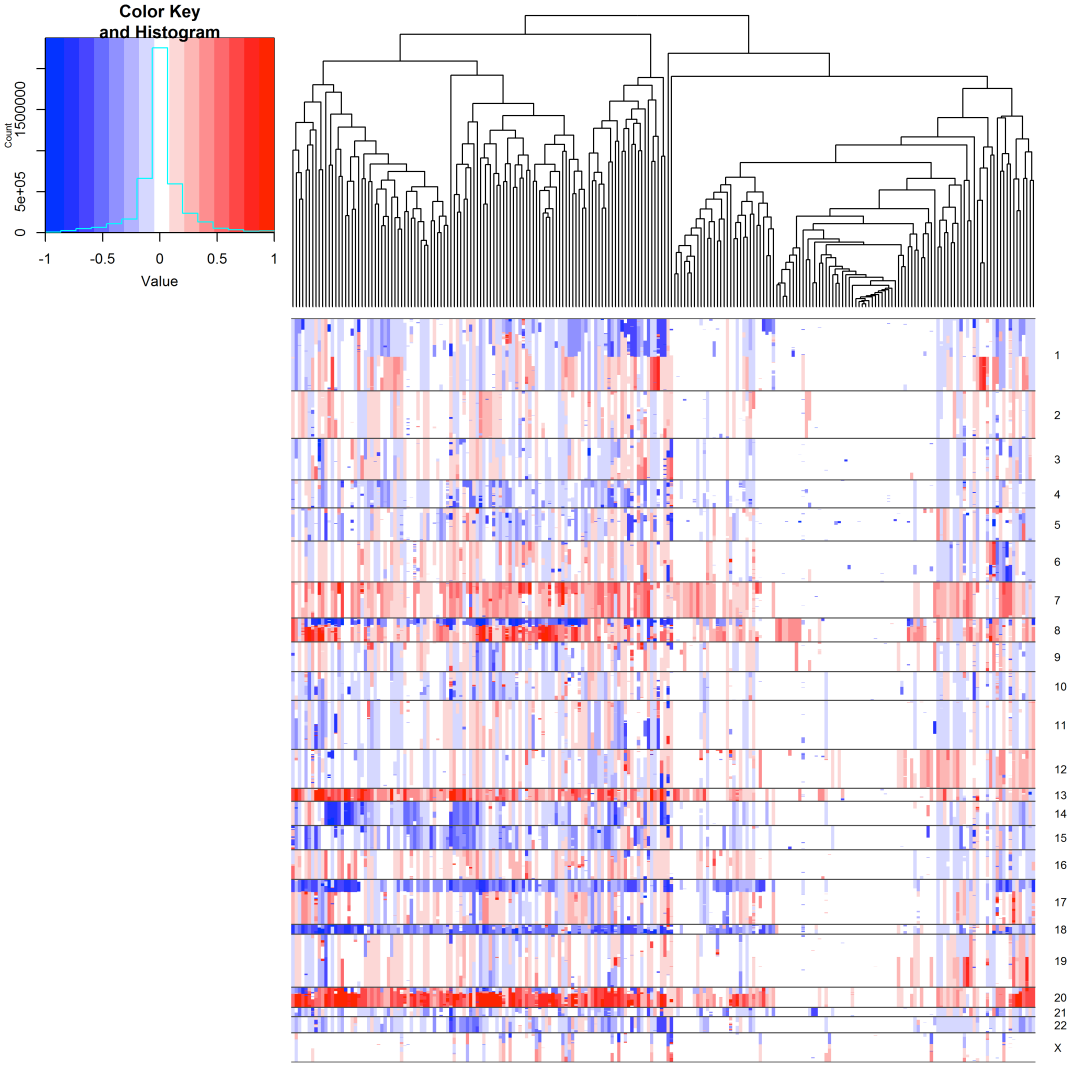

Figure S59: HeatMap of SCNA data for COAD samples. Copy number measurement were truncated by -1 and 1 to maintain the color contrast in the whole figure.

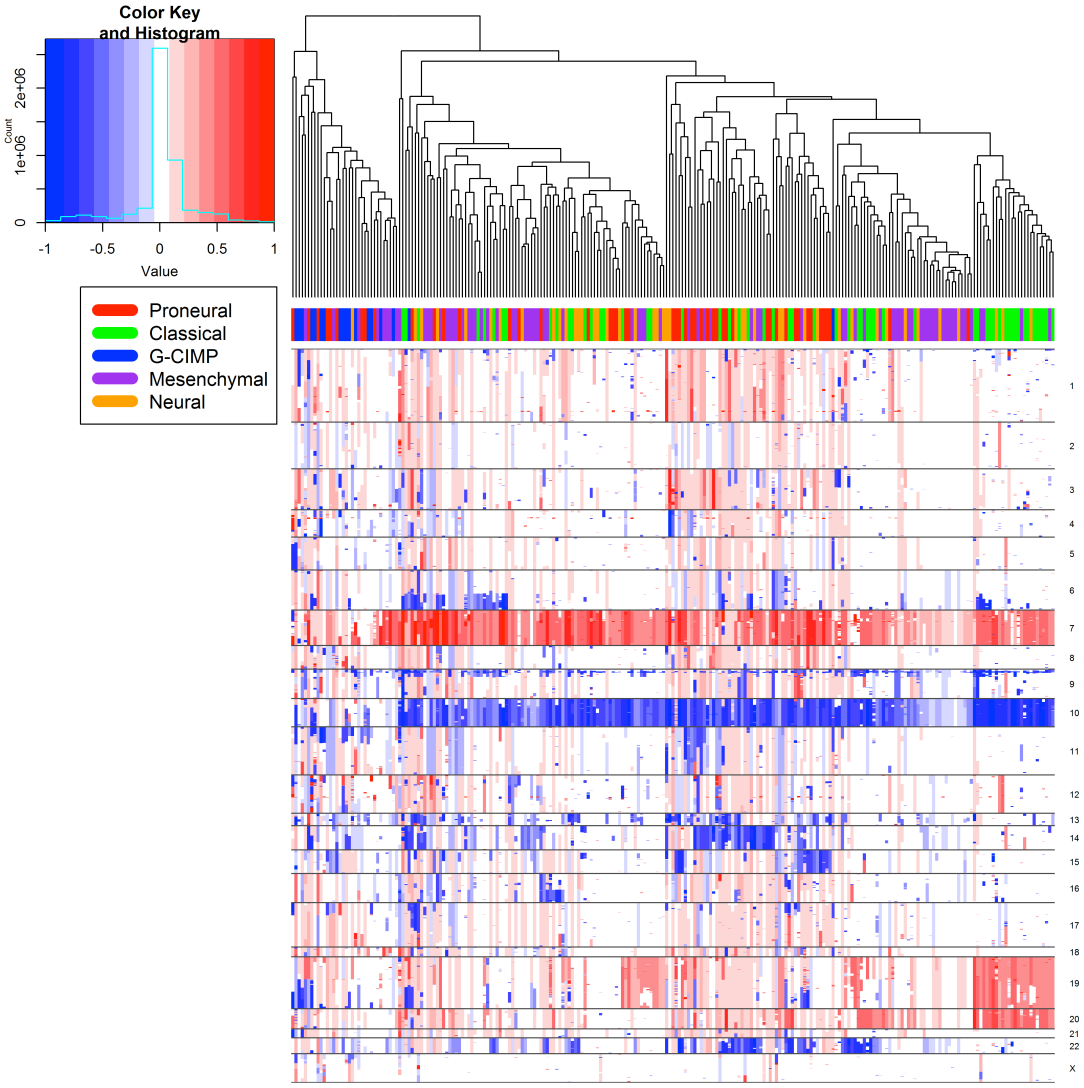

Figure S60: HeatMap of SCNA data for GBM samples

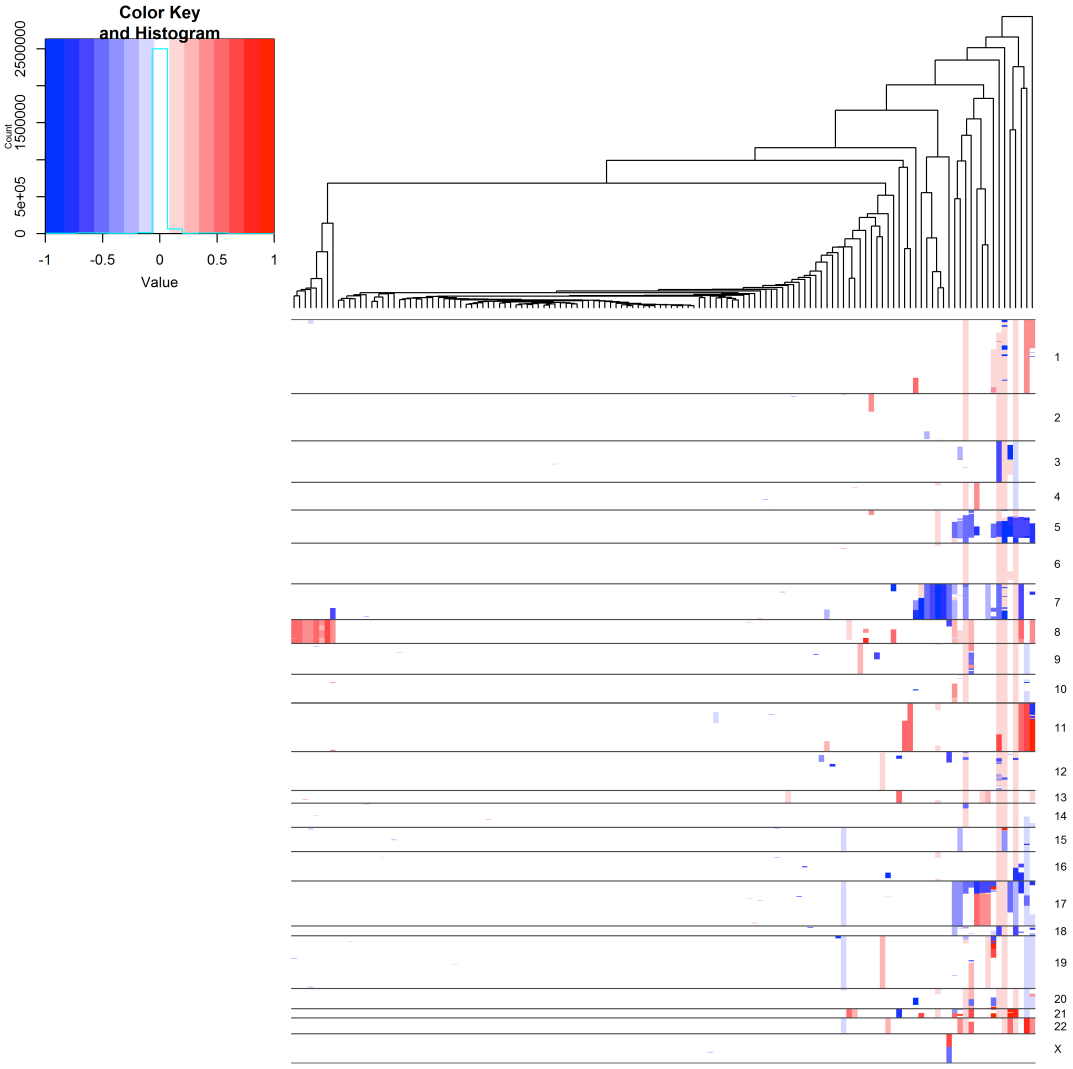

Figure S61: HeatMap of SCNA data for LAML samples

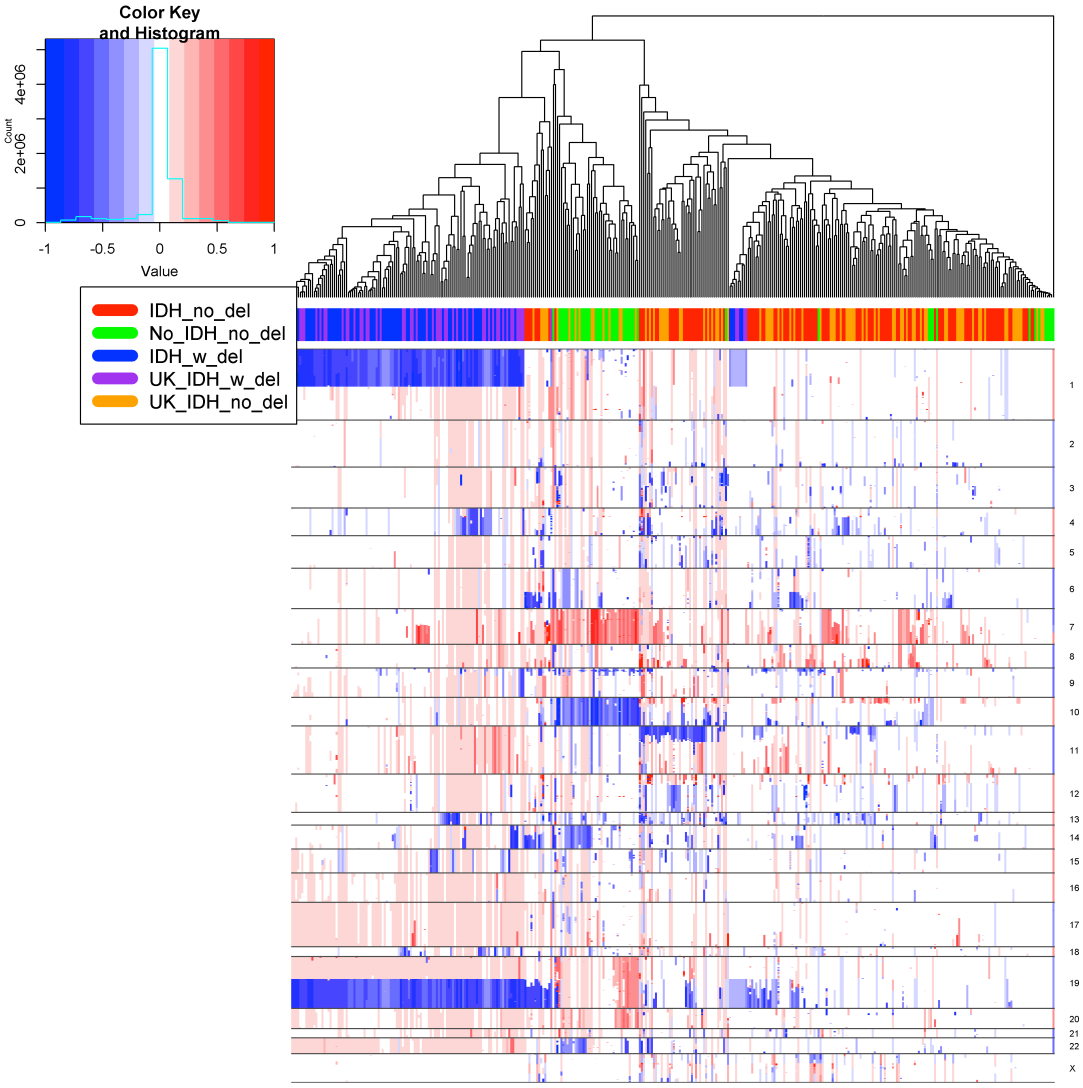

Figure S62: HeatMap of SCNA data for LGG samples

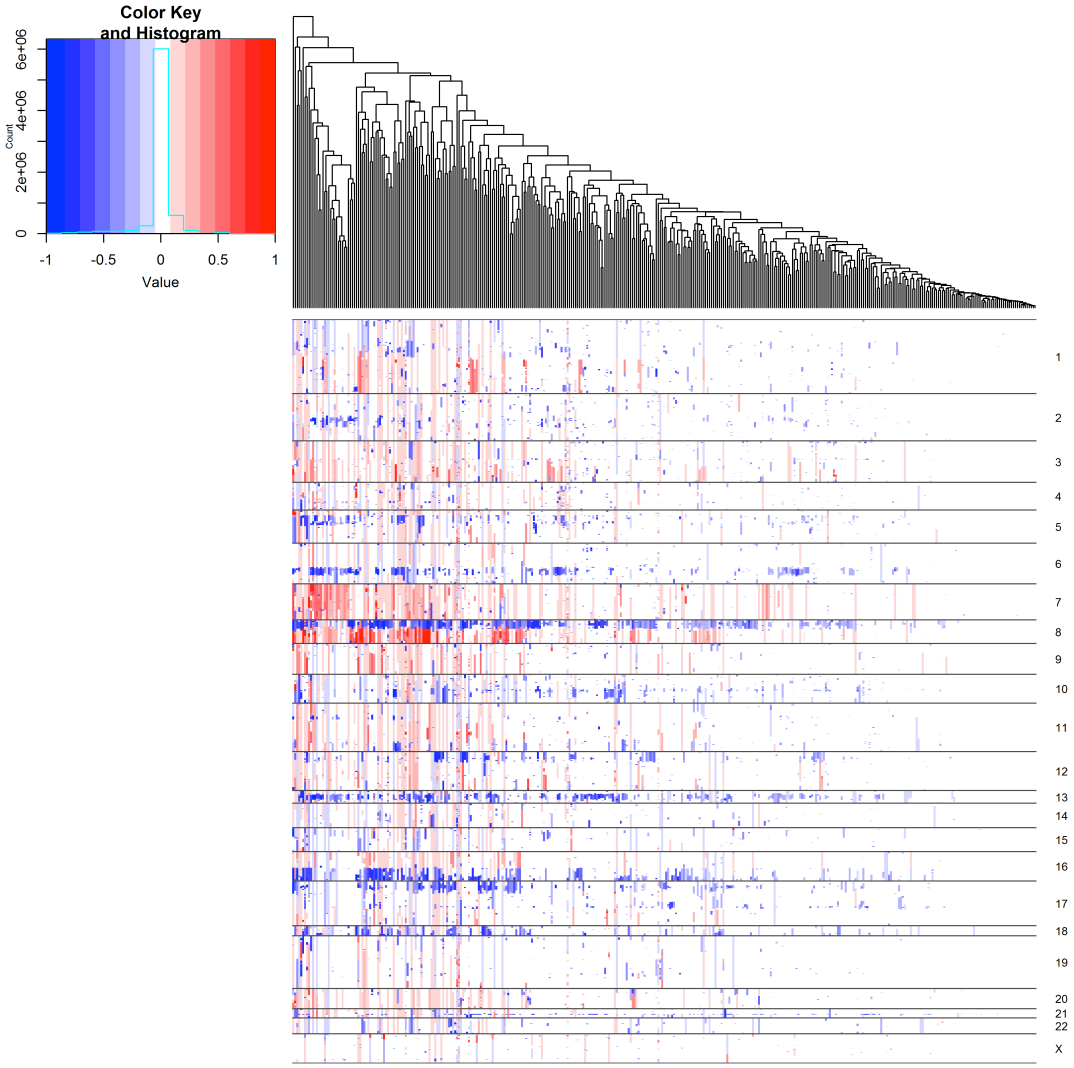

Figure S63: HeatMap of SCNA data for PRAD samples

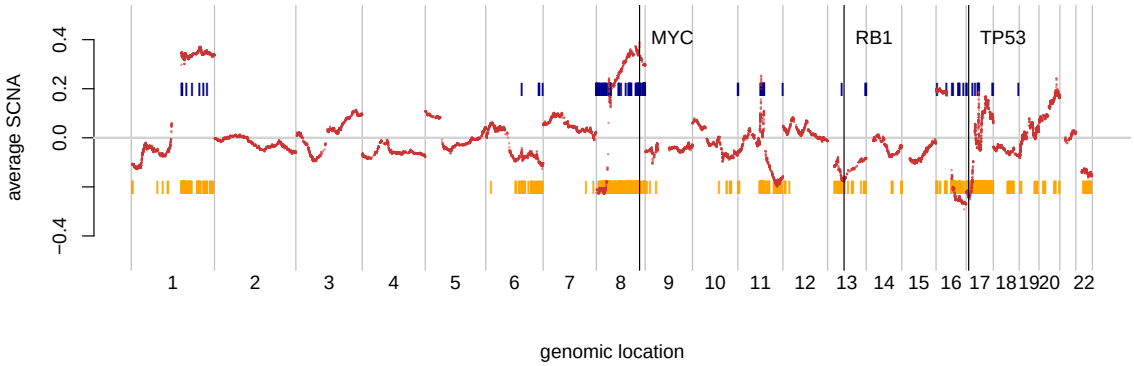

Figure S64: Location of the genes with strongest CM associations in BRCA samples. The red points are average of SCNA signals across all BRCA samples, and the blue/orange bars indicate genes with positive/negative CM associations, respectively. Three known cancer genes with strong CM associations in at least four of six cancer types are labeled.

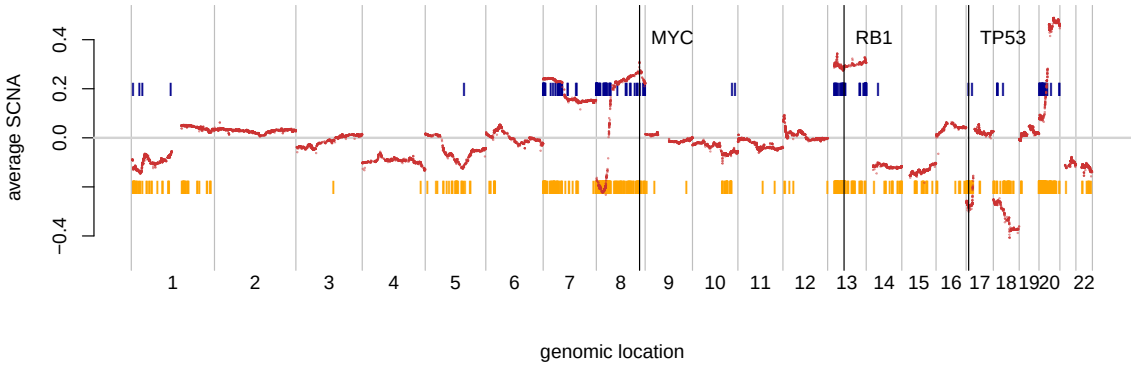

Figure S65: Location of the genes with strongest CM associations in COAD samples.

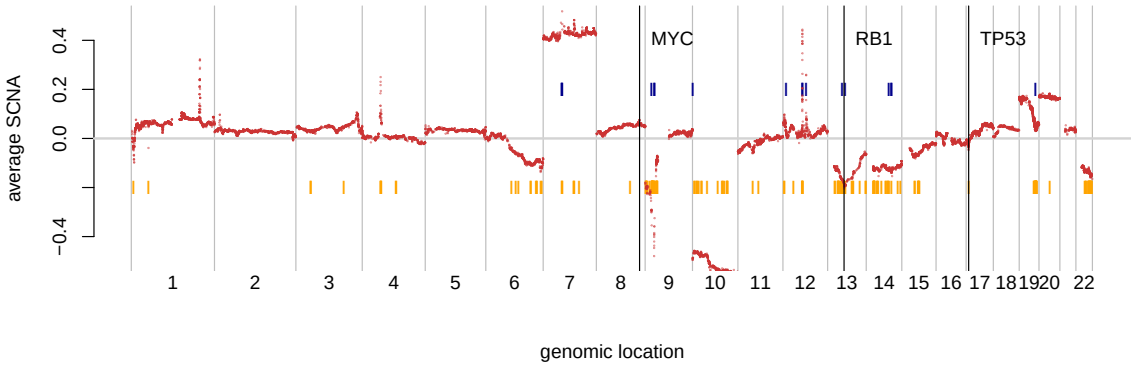

Figure S66: Location of the genes with strongest CM associations in GBM samples.

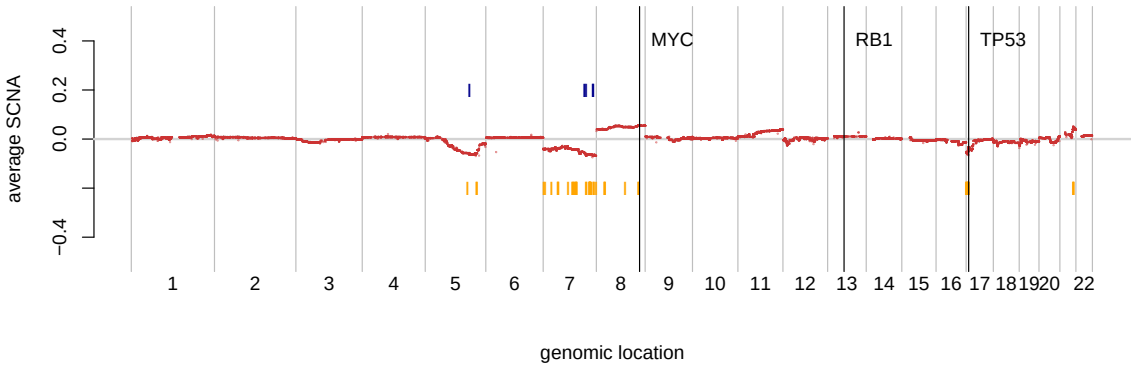

Figure S67: Location of the genes with strongest CM associations in LAML samples.

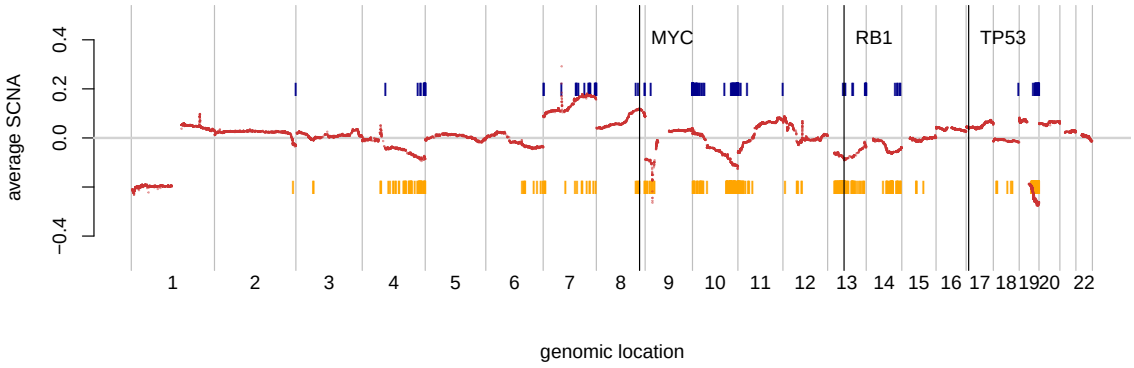

Figure S68: Location of the genes with strongest CM associations in LGG samples.

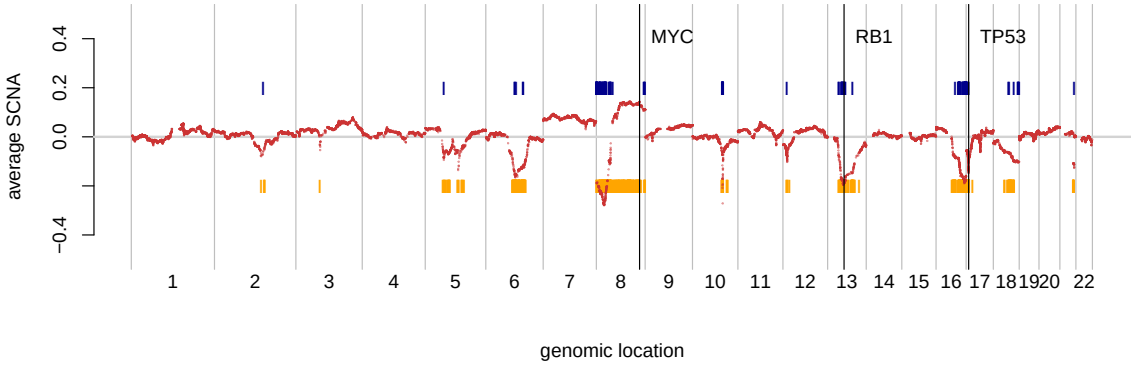

Figure S69: Location of the genes with strongest CM associations in PRAD samples.

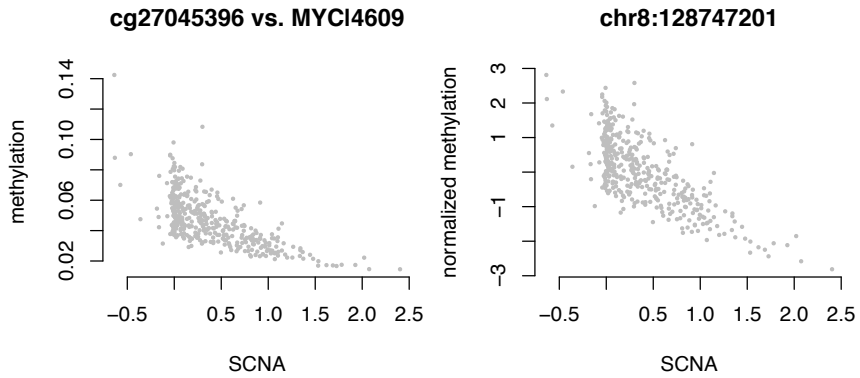

Figure S70: Association between the SCNA of gene MYC (entrez gene ID: 4609, chr8:128747765-128753678) and the DNA methylation of a nearby CpG (cg27045396, chr8:128747201) about 500bp upstream of MYC in BRCA patients. Left panel shows methylation by beta-value in the range of 0 to 1, and the right panel shows the methylation after normal quantile transformation.

B.4 Comparing SCNA and DNA methylation data between tumor and adjacent normal samples in COAD patients.

We downloaded SCNA data of 90 adjacent normal samples from NCI Genomic Data Commons (GDC) data portal legacy archive. We selected adjacent normals based on TCGA sample type "11" from TCGA sample barcode. For example, in barcode TCGA-G4-6298-11A, the sample type is 11. The file format is the same as the files used for tumor samples: segmented copy number data saved in text files with file names `*.nocnv_hg19.seg.txt`. Following the same pipeline for tumor samples, we extracted gene-specific copy number measurement for each sample. By manual examination, there is no SCNA event in these 90 adjacent normal samples.

DNA methylation data (Illumina 450k array) of adjacent normal samples in COAD patients were downloaded from NCI GDC data portal. There were 38 such normal samples. We prepared the data following the same steps as for DNA methylation data from tumor samples. We removed probes that are close to a known SNP, and probes with more than 5% of missing values, and ended up with a data matrix of 332,816 probes for 38 samples. The methylation data from tumor samples is a data matrix of 332,689 probes for 160 samples. More than 99.9% of the probes (332,374 probes) are shared between the two datasets. There are 12 patients with methylation data from both tumor and adjacent normal samples. PCA of these 198 (=38+160) samples shows that they are clustered by tissue of origin rather than patient identity (Figure S71).

For each probe, we quantified the difference between the distribution of DNA methylation in tumor samples versus normal samples by a Kolmogorov-Smirnov (K-S) statistic, which is the maximum distance between the empirical cumulative distribution functions of the two distributions. Larger value of a K-S statistic indicates larger difference of the two distributions. The DNA methylation on those CpGs involved in SCNA-methylation (CM) associations are more similar between tumor and normal samples than the remaining CpGs (Figure S72). Moreover, for those CpGs involved in CM associations, if we only consider the tumor samples without strong evidence of SCNA events, the DNA methylation are even more similar between tumor and normal samples (Figure S72). This observation are also illustrated in a few examples (Figure S72).

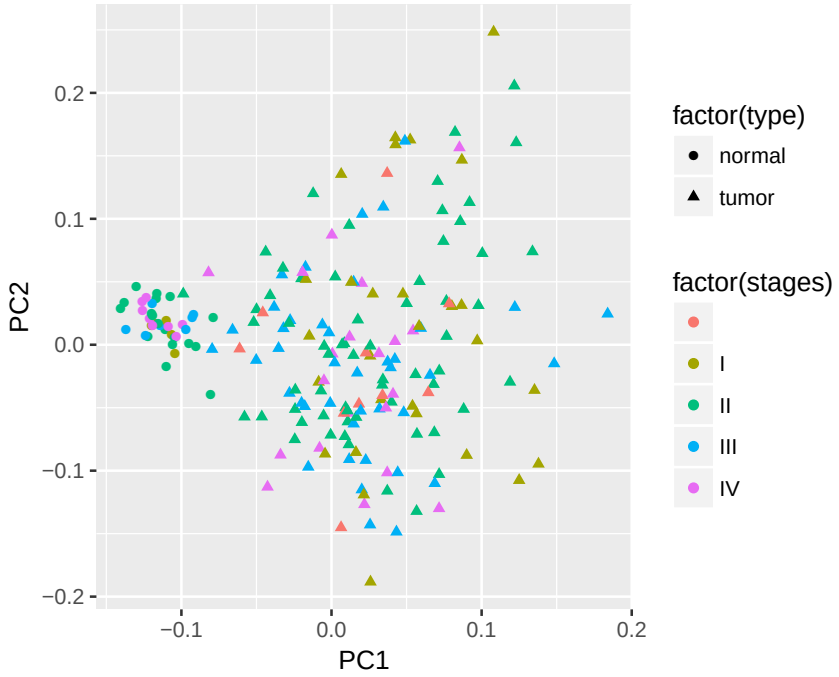

Figure S71: PC1 versus PC2 for the autosome DNA methylation data of 198 samples, including 160 tumor samples and 38 adjacent normal samples.

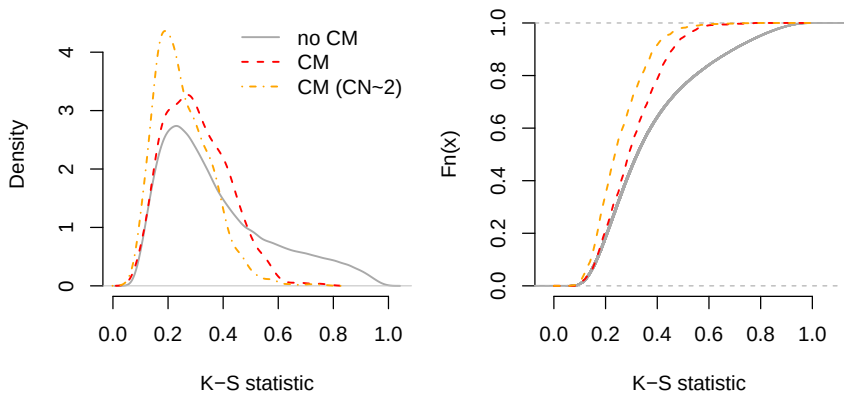

Figure S72: The density (left panel) and cumulative distribution (right panel) of K-S statistics for three comparisons between tumor and paired normal samples. "CM" and "no CM" indicate comparison between all normal samples versus all tumor samples for those CpGs that are involved in CM associations (p-value $< 10^{-10}$) and the remaining ones. "CM (CN$\sim$2)" indicates comparison for those CpGs that are involved in CM associations between all normal samples versus the tumor samples with copy number close to 2 (SCNA measurement from -0.5 to 0.5).

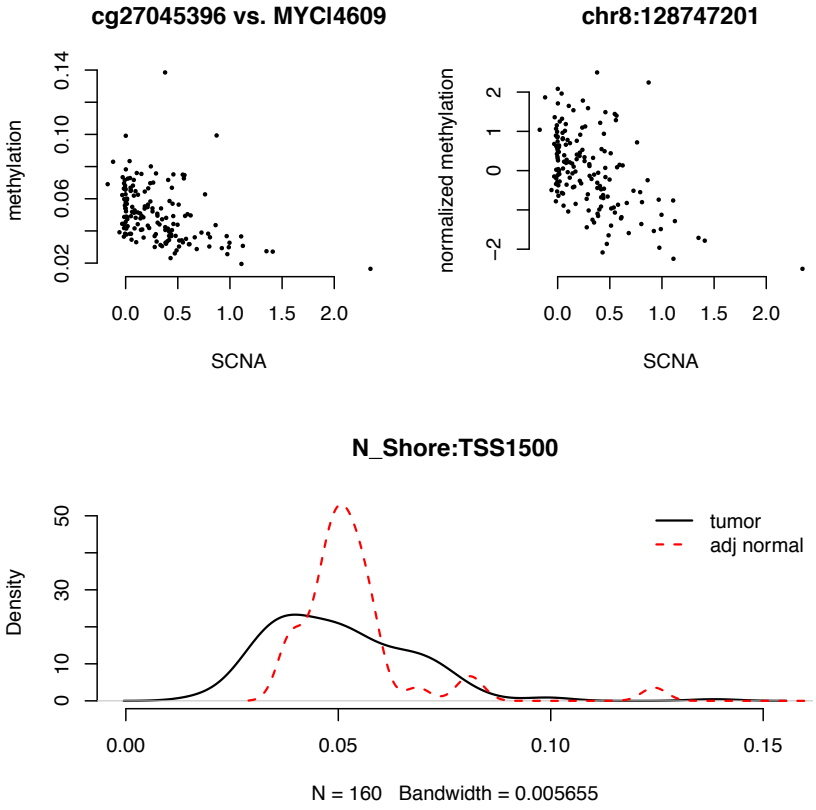

Figure S73: Copy number of gene MYC versus DNA methylation of a nearby CpG. The upper panels show SCNA measurement versus DNA methylation in the scale of beta-value (upper-left panel) or the scale after normal quantile transformation (upper-right panel). The lower panel shows the density plot of DNA methylation of this CpG probe in tumor samples and adjacent normal samples. The header of upper left panel shows CpG probe id and gene symbol | entrez ID. The header of upper right panel and lower panel shows the location of the CpG probe and its annotation.

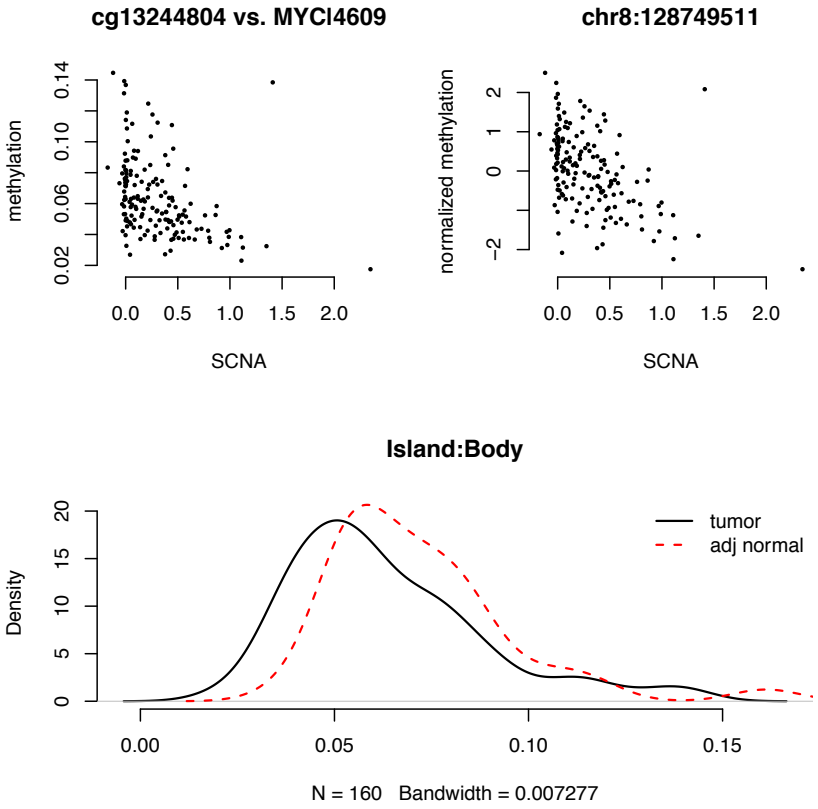

Figure S74: Copy number of gene MYC versus DNA methylation of another nearby CpG.

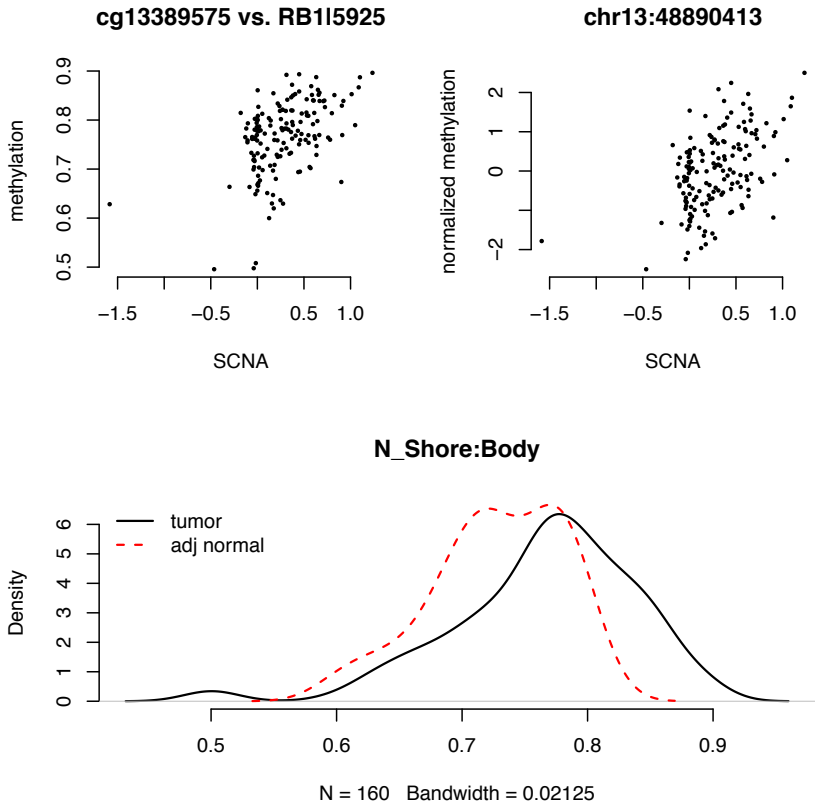

Figure S75: Copy number of gene RB1 versus DNA methylation of a nearby CpG.

B.5 Two-way associations and conditional associations for IDH mutation, DNA methylation, and gene expression

Similar to our analysis for the relations between SCNA, DNA methylation, and gene expression in the previous section, we studied the association and conditional association of IDH mutation, DNA methylation, and gene expression in TCGA LGG samples. All associations and conditional associations are assessed after accounting for the effects of batch effects, demographic variables (e.g., age, gender, genotype PCs), and top 7 ME PCs. For each gene, we considered all the CpGs within the gene body or within 500kb of gene boundaries. The ME associations were assessed using linear regression with all the covariates used in earlier analysis except tumor subtype. There are 58,188 such local ME associations at p-value cutoff 10^{-10}, and they involve 5,483 genes and 33,350 CpGs. For each gene, we chose the local CpG with strongest association, which led to 5,483 ME associations. Then for each of these 5,483 ME pairs, we calculated IDH vs. methylation and IDH vs. gene expression associations, as well as conditional associations: IDH vs. methylation given gene expression and IDH vs. gene expression association given methylation.

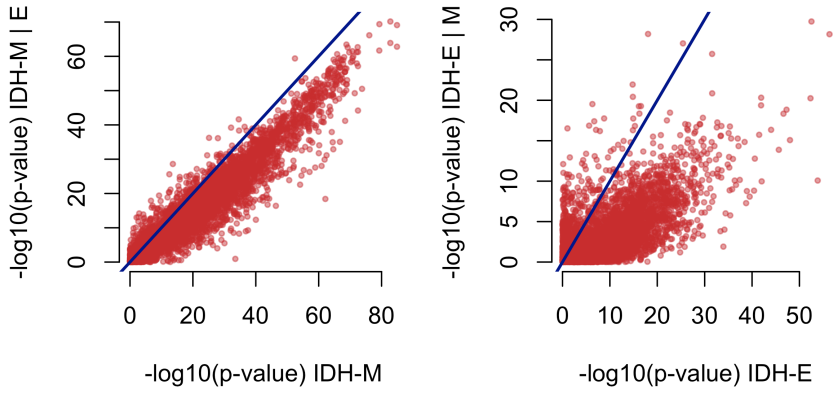

Figure S76: Comparison of associations and conditional associations between IDH and DNA methylation (left panel) and between IDH and gene expression (right panel) in TCGA LGG samples.

References

[1] Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC bioinformatics,* **11**(1), 94.

[2] Patterson, N., Price, A. L., and Reich, D. (2006) Population structure and eigenanalysis.. *PLoS genetics,* **2**(12), e190.

[3] Shabalin, A. A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics,* **28**(10), 1353–1358.

[4] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013) The cancer genome atlas pan-cancer analysis project. *Nature genetics,* **45**(10), 1113–1120.

[5] Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature,* **490**(7418), 61–70.

[6] Ziebarth, J. D., Bhattacharya, A., and Cui, Y. (2012) Ctcfbsdb 2.0: a database for ctcf-binding sites and genome organization. *Nucleic acids research,* **41**(D1), D188–D194.

[7] Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., et al. (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology,* **30**(5), 413–421.

[8] Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature,* **487**(7407), 330–337.

[9] Cancer Genome Atlas Network (2014) Integrative genomic characterization of lower grade gliomas. *Neuro-oncology,* **16**(suppl 3), iii3–iii3.

[10] Cancer Genome Atlas Network (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine,* **368**(22), 2059.

[11] Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., Zheng, S., Chakravarty, D., Sanborn, J. Z., Berman, S. H., et al. (2013) The somatic genomic landscape of glioblastoma. *Cell,* **155**(2), 462–477.