Supplemental Methods

Data acquisition, filtering and normalization:

The TCGA data were downloaded as level 3 beta values. Although these data were pre-processed, they were processed by TCGA in a consistent manner using standard pipeline with a minimum normalization. These datasets have masked out data from potentially less reliable probes across the whole dataset as well as data from probes with a large non-detection probability. Quote TCGA data description: "Probes having a common SNP (common SNP is a SNP with Minor Allele Frequency > 1% as defined by the UCSC snp135common track) within 10bp of the interrogated CpG site or having 15bp from the interrogated CpG site overlap with a REPEAT element (as defined by RepeatMasker and Tandem Repeat Finder Masks based on UCSC hg19, Feb 2009) are masked as NA across all samples, and probes with a non-detection probability (P-value) greater than 0.05 in a given sample are masked as NA on that chip." We have further filtered out probes that had a large non-detection probability in more than 5% of the samples. This resulted in the overall reduction of the number of usable probes from 485,512 to 395,874. The GEO data were acquired as beta values provided by the original data depositors and, where available, the methylated and unmethylated signals were downloaded and the beta values were calculated using formula MetSig/(MetSig+UnmetSig+100). A batch correction was not applied, since some of the individual GEO studies did not have available unprocessed raw data, or batch information was not available, and there was some more serious heterogeneity in the data (different studies, different data levels) than just batch effects. However, since all these data are heterogeneous due to different laboratories that generated the data and different level of processing, some correction of these differences had to be applied before performing our analysis. Also a reduction of differences between data distributions of type I and type II probes was necessary. The major differences between the studies and also between the type I and type II probes within samples were positions of the modes of the two extreme peaks of the data distribution. There is published normalization method (BMIQ algorithm)[1] to deal with this discrepancy within the samples. Generally it adjusts dynamic range of beta values – positions of the modes of methylated and unmethylated peaks. We have chosen to use custom modified BMIQ algorithm for the data normalization. The BMIQ algorithm was, in a similar principal, successfully used previously to combine data from a large number of studies and two different platforms (Illumina HumanMethylation27 and Illumina HumanMethylation450 – for the probes covered on both platforms)[2]. In our case the BMIQ algorithm was modified to use external golden standard based on distribution of type I probes from a large set of normal TCGA samples to separately correct the distribution of type I and type II probes within each sample. We have used median of type I probes of normal samples from TCGA BRCA cohort as a golden standard and all the samples from every TCGA

cohort and GEO study were normalized using this external standard. This way the algorithm performed between array normalization (adjustment of the dynamic range of the data of individual samples) that did not require all the samples to be loaded in memory at the same time like in case of e.g. quantile or cyclic loess normalizations. The normalization was applied separately on type I and type II probes within each sample and this way it also adjusted for differences between the two probe chemistries, regardless whether the sample was previously normalized or derived from raw signals.

Determination of DMRs

Each of 23 TCGA cancer type cohorts that had any respective normal samples available was used to find out DMRs for respective cancer type. These DMRS will in a subsequent step serve as a substrate to search for marker regions. Several of the TCGA cancer types had low numbers of normal samples, which could result in less reliable DMR calls (false positives), however subsequent filtering of these regions against large set of normal sample cohorts should eliminate such regions from marker pools. The normalized beta values for individual CpG probes were first converted to M values using formula log2(beta/(1-beta)); the M values have distribution that is closer to normal distribution than the distribution of the beta values and therefore are more suitable for the statistical testing[3]. The limma package[4] was applied to data from individual 395,874 probes to determine differentially methylated CpGs. Genomic positional information of the CpGs covered by individual 395,874 probes was added and overlapping pairs of 2 consecutively covered CpGs up to 500 bp apart were evaluated for differential methylation. The 500 bp distance is large enough to frequently have another CpG covered within and still small enough to be frequently same/similar methylation status. This eliminated data from singleton probes that did not have “neighbors” within 500 bp and thus reduced the amount of usable probes to 291,604. This sacrificed substantial fraction of the data in favor of larger robustness and reliability of the remaining data since each marker candidate is based on the data from at least two individual probes that both had to have large difference in DNA methylation. The pairs of consecutively covered CpGs within 500 bp window that had mean difference in particular tumor cohort from respective normal reference of 0.4 beta or greater were used as basis for DMR calls. The mean 0.4 beta difference was chosen empirically after testing a range from 0.3 to 0.5. Consecutive CpG pairs that passed this filter were clustered and each cluster (DMR) was used as a marker candidate regions for further filtering.

Filtering of DMRs to obtain markers

Marker candidate clusters (DMRs) were filtered against 4 sets of cohorts of normal samples to eliminate regions with tissue specific methylation. During the filtering, the data from all CpGs of the DMR were tested and the best performing CpG in each DMR was used to represent the region. The normal sample cohorts consisted of: 1) CaTyNT - cohort of respective normal TCGA samples – this was specific for each cancer type; 2) AllNTmed - set of normal TCGA samples that consisted of medians of methylation values of all respective normal tissue sets with at least 3 normal samples each – 19 total tissues (BLCA, BRCA, CESC, CHOL, COAD, ESCA, HNSC, KIRC, KIRP, LIHC, LUAD, LUSC, PAAD, PCPG, PRAD, READ, SARC, THCA, UCEC); 3) blood – a large cohort (n=1,388) that consisted of two GEO datasets of the whole blood samples from cancer free subjects (GSE40279 and GSE87571); 4) GEO18 - a set of 18 cohorts of GEO samples (total n=2,189) described in Table S2. Rather than using one parameter with a very stringent cutoff a complex set of multiple parameters with less stringent cutoffs was used. These parameters were obtained using empirical testing of ranges of values and application of this set of filters resulted in a very stringent filtering. The individual reference cohort sets are referred using names listed above. The description of filtering parameters and cutoffs for hypermethylated DMRs was as follows; to pass as a marker the beta values for at least one CpG in hypemethylated DMR had to fulfill all the following criteria: 1) maximum of medians of CaTyNT, AllNTmed and blood references < 0.1; 2) maximum of AllNTmed reference < 1/3 (0.333); 3) maximum of upper whiskers of AllNTmed and blood references < 0.1; 4) upper hinge of CaTyNT reference < 1/6 (0.166); 5) the sum of three AUCs using CaTyNT, AllNTmed and blood references respectively > 2.7; 6) maximum of medians of GEO18 < 1/8 (0.125); 7) maximum of upper hinges of GEO18 < 0.2; 8) maximum of upper whiskers of GEO18 < 0.3; 9) maximum of 95th percentiles of GEO18 < 0.45; 10) at least 2/3 of the cases had to have at least 0.25 beta > maximum (medians of CaTyNT and AllNTmed, and upper whisker of blood) – this is to ensure that a large fraction of tumor samples carries the DMR. In addition a minimum distance of 2.0 kb was applied to consider nearby CpGs as separate markers, in case of CpGs in closer proximity only the better performing one passed the filter. Whiskers and hinges of control cohorts were calculated using boxplot R function with the default settings. The hypomethylated DMRs were filtered in the same fashion, except the cutoff values were 1-(cutoff for the hyper DMRs) and the respective quantiles were also flipped 1-(quantile for hyper). The sum of AUCs cutoff for hypo markers was >2.7, same as in the case of hyper markers.

CpG/marker enrichment analysis

All the probes of the Illumina HumanMethylation450 were annotated using the information about the histone H3lysine27trimethylation domains in ES cells as described[5]. These domains are known to be a signature of polycomb regulation. About ¼ of all Illumina HumanMethylation450 covered CpGs fell within these regions or within 500 bp. The abundance of these polycomb associated probes in the 291,604 good probes that have neighbors within 500 bp and therefore could become marker regions if fulfilled differential methylation criteria was found (universe), and then within the individual DMR or marker region probes as subsets. The hypergeometric test[6] was used to test, if there is a significant enrichment of polycomb associated loci within DMRs and markers sets. Similarly for non-coding gene association testing, all the probes of the Illumina HumanMethylation450 were first annotated in a similar way as described in methods for marker regions, but using RefSeq accession numbers instead of gene symbols. Then, based on the presence of NM_ or NR_ in the accession number each probe was assigned to be associated with coding gene, non-coding gene, both types of genes or none. Using this information for all good paired 291,604 probes (universe) and probes within marker sets as subsets, the hypergeometric test was used to test, if there is a significant enrichment of non-coding RNA associated CpGs within marker sets.

Supplemental references

1. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 2013; 29:189-96.

2. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol 2013; 14:R115.

3. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 2010; 11:587.

4. Smyth GK. Limma: linear models for microarray data. In: Gentleman R CV, Huber W, Irizarry R, Dudoit S, ed. Bioinformatics and Computational Biology Solutions using R and Bioconductor. New York: Springer, 2005:397-420.

5. Vrba L, Munoz-Rodriguez JL, Stampfer MR, Futscher BW. miRNA gene promoters are frequent targets of aberrant DNA methylation in human breast cancer. PLoS One 2013; 8:e54398.

6. Fury W, Batliwalla F, Gregersen PK, Li W. Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. Conf Proc IEEE Eng Med Biol Soc 2006; 1:5531-4.