Exploring Biomarkers of Universal Specificities and Commonalities among Pan-cancer Cohorts in The Cancer Genome Atlas

Jie Wu (jwu@cau.edu.cn)
State Key Laboratory of Agrobiotechnology, China Agricultural University, Beijing, 100193, China.
Deng Lin
Beijing Advanced Innovation Center for Food Nutrition and Human Health, College of Biological Sciences, China Agricultural University
Liandi Jiu
Beijing Advanced Innovation Center for Food Nutrition and Human Health, College of Biological Sciences, China Agricultural University
Qi Liu
State Key Laboratory of Agrobiotechnology, China Agricultural University
Yiqiang Zhao
State Key Laboratory of Agrobiotechnology, China Agricultural University
Zhenglong Gu
Beijing Advanced Innovation Center for Food Nutrition and Human Health, College of Biological Sciences, China Agricultural University

Research Article

Keywords: TCGA, DNAm, instability, biomarker, cancer

Posted Date: February 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-152042/v1

Abstract

Abnormal DNA methylation (DNAm) has been associated with the onset and development of cancer; thus, it is vital to develop an in-depth understanding of DNAm patterns as well as the accompanying variations across multiple cancer cohorts to explore the mechanism of cancer development. The use of DNAm biomarkers holds great promise for the standardization of cancer diagnostics in clinical trials. This research focused on the discovery of universally specific or common biomarkers that would be. First, we explored primary cancers with significantly higher variations compared with normal solid tissues, as well as the heterogeneity in methylation in different cancers. Second, we found 138 differently methylated CpGs (DMCs) with a common methylation trend and eight common differently methylated regions (DMRs) in different cancer cohorts. Third, we found 99 DMCs to distinguish 32 different cancer cohorts in Random Forest (RF) analysis because of the specificity mechanism, but each DMC still had high instability. Additionally, the most important 10 DMCs among the 99 DMCs affected the survival time of cancer individuals. By employing a comprehensive analysis of DNAm patterns and variations in cancers, our results could facilitate the development of biomarkers that are universally specific and common features among DNAm profiles across multiple cancer cohorts. Our findings also provide new insights for the prevention and treatment of cancer.

Introduction

DNA methylation (DNAm) is considered as one of the primary interfaces between the genome and the environment. The environment can induce global and gene-specific DNAm alternations [1, 2], which indicates that epigenetic variation itself can be disease-causing or might contribute to the human phenotypes [3, 4]. Our biological systems employ robust defense mechanisms against environmental perturbations to maintain normal physiological functions and physical health [5-7]. DNA hypermethylation has the potential to silence the tumor-suppressor genes; hypermethylated CpG island (CGI) are frequently located at gene promoter regions, which can identify a specific feature of cancer [8]. Global DNA hypomethylation influences cancer development by inducing chromosomal instability and global loss of imprinting [9, 10]. Genome-wide hypomethylation generally occurs within repetitive transposable DNA elements, such as the LINE-1 or short interspersed nucleotide elements (SINE, or Alu) in several cancers. LINE-1 hypomethylation has been inversely associated with microsatellite instability (MSI) and/or CpG island methylator phenotype (CIMP) [11, 12].

Although somatic mutations play a significant role in cancer development, recent studies have revealed that DNAm contributes to oncogenesis by turning off tumor suppressor genes or activating oncogenes [13-15]. Since epigenetic alterations promote the early onset and manifestation of events leading to malignant transformations, epigenetic events are more promising cancer risk markers compared with gene mutations [16]. Additionally, advancements in genomic technologies have led to the identification of various specific epigenetic alterations as potential clinical biomarkers for cancer

patients. These biomarkers are being increasingly used in cancer detection, diagnosis, prediction, prognosis, and monitoring; DNAm might be a fertile ground for search of other biomarkers and clinical assessment of other diseases [4, 17].

The combination of genome-wide DNAm instability and genetic instability might facilitate or accelerate tumor progression [18, 19]. An easily detectable, highly stable biomarker in the DNAm instable state would be suitable for clinical use. Therefore, the establishment and maintenance of DNAm patterns, including both hyper- and hypomethylation, are crucial for normal cellular function and developmental processes [20]. Also, these patterns are highly heterogeneous at different life stages [21] and between different tissue cohorts[22] , which increases their complexity as well as their importance in carcinogenesis.

TCGA project accelerated our understanding of the molecular basis of cancer through the application of genomic technologies. A comprehensive analysis of DNAm patterns and variations in cancers facilitated the development of biomarkers that could be applied across multiple cancer cohorts and provided new insights about the prevention and, thereby, reduction of cancer-related mortality.

Results

Overall DNAm pattern in cancers

Since there were a higher number of primary cancers than normal solid tissues in the TCGA database assessed on the Illumina Infinium Human Methylation 450 BeadChip array, we selected 10 cancer cohorts that had more than 20 matched normal solid tissue samples [BLCA, BRCA, HNSC, KIRC, KIRP, LIHC, LUAD, LUSC, THCA, and UCEC], resulting in 4501 primary cancers and 598 normal solid tissues.

Next, we projected all 5,099 samples using the *T*-distributed stochastic neighbor embedding (*t-SNE*) method to reveal the overall DNAm pattern. The primary cancers were significantly more clustered with the corresponding normal samples from the same tissues, irrespective of cancer cohorts, confirming the distinct pattern of cell-of-origin in cancer (Fig 1A). Next, we screened gene promoters across 22 chromosomes to reveal DNAm variables. The hypermethylated CGI were frequently located at gene promoters and presented a state of instability in primary cancers, which was usually associated with suppression of tumor suppressor genes and incomplete differentiation, directly linking DNAm changes to oncogenic transformation (Fig 1B). We also plotted the average DNAm level in the gene region. Consistent with previous findings, in normal solid tissues, the gene-body regions exhibited higher methylation levels compared with 5' upstream regions. It supported the hypothesis that genomic regions involved in active transcription were hypomethylated to facilitate the accessibility of transcription factors (TFs). The methylation level peaked at the first exon near the transcription start site (TSS) and the transcription termination site (TTS), and then there was a sharp decrease in TTS (Fig 1C). The distribution of methylation levels was very similar across different cancer cohorts. For the 10 cancer cohorts, although we did not observe a significant decrease in methylation levels of the gene region (i.e.,

5'UTR, coding region, 3'UTR), the distribution of methylation levels was generally similar but in a different shape to that in normal solid tissues. Specifically, the peaks at the first exon near TSS or TTS were almost absent in cancers, suggesting the transcriptional “barriers” were significantly weakened. Conversely, the methylation level in the 5' upstream regions was elevated in cancers (Fig 1D). We hypothesized that it constituted a mechanism of blocking post-transcriptional regulation in primary cancer. Similar to the well-known genomic instability of cancer, we found that the patterns of methylation levels were not that highly accordant with each other as observed in the normal solid tissues, suggesting methylation heterogeneity in different cancers.

Patterns of methylation instability in cancers

Genomic instability is a hallmark of cancer; it includes a variety of DNA alterations, including point mutations, indels, structural variations, or chromosome rearrangements. Since the previously observed methylation heterogeneity could contribute to genomic instability through the methylation-induced somatic mutations; therefore, we investigated if the DNAm of cancers also displayed patterns of instability.

For considering the 10 cancer cohorts jointly, we first computed the coefficient of variation (CV) across all primary cancers and normal solid tissues, respectively, for a single CpG for each cancer cohort, followed by a paired *t-test* (*P*-value < 0.05 for 10 cancer cohorts) to access the statistical significance between the two CVs. We found that the CV values of primary cancers were significantly higher than that of normal solid tissues (Fig 2A). As expected, there were variations in the overall DNAm levels by Levene's test per CpG per the cancer cohort, and this result further confirmed that DNAm presented different variations in different cancer cohorts (Fig 2B). For each gene set, we used the same method as above and obtained similar results (Figs 2C-D). To visualize the relative methylation variables in primary cancers compared with normal solid tissues, we plotted the *Z*-score normalized $\log_2$FC for each CpG in the form of a histogram. The observed distribution was significantly biased toward high or low values to the expected standard normal distribution (Kolmogorov–Smirnov test, *P*-value < 0.05 for all 10 cancer cohorts), indicating substantial methylation variations in all cancer cohorts (Fig 2E for BLCA and S1 Fig for other nine cancer cohorts). Next, we respectively fitted the Gaussian distribution in means of each CpG of primary cancers and means of each CpG of normal solid tissues to visualize the relative methylation instability in primary cancers. We found that the variance of means of each CpG of primary cancers was significantly higher than that of normal solid tissues (Fig 2F for BLCA and S2 Fig for other nine cancer cohorts). Similarly, most primary cancers exhibited high numbers of both hyper- and hypomethylated whole-genome CpGs (Fig 2G for BLCA and S3 Fig for other nine cancer cohorts). We referred to these primary cancers, characterized by DNA methylation instability (DMI), to distinguish them from cases where only hyper- or hypomethylation occurred. We scored the extent of DMI for each primary cancer by combining the percentages of hyper- and hypomethylated loci and found that most of the individual primary cancers were located in more than 25 % of hyper- and hypomethylated loci, which concluded that there existed a different level of instability in most individual primary cancers.

Patterns of common methylation in cancers

Using the Wilcoxon rank-sum test for all 10 cancer cohorts, we tested the DMCs between primary cancers and normal solid tissues, with the threshold set to $|\log_2 FC| \geq 1$ and *P-value*< 0.05. The DMCs ranged from 1,860 in THCA to 28,587 in UCEC. Among the DMCs, we found that 138 CpGs were shared by all 10 cancer cohorts with the same DNAm tendency (hypo- or hypermethylation) (Fig 3A and S1 Table). When the tissue specificity was removed, primary cancers were more clustered and significantly separated from corresponding normal samples from the same tissue, confirming the presence of a common pattern in different primary cancers (Fig 3B). Similarly, most of the 138 CpGs preferred hypermethylation in primary cancers and hypomethylation in normal solid tissues, and when the tissue specificity was removed, different primary cancers exhibited significant commonality (Fig 3C). MAGMA was used for functional enrichment analysis on 138 CpGs, and we found that nuclear chromosome segregation was the most enriched category in GO analysis (Fig 3D) and dorsoventral axis formation was the most enriched categories in KEGG analysis (Fig 3E).

Compared with scattered DMCs, excessive DMCs within a short distance were found to be more biologically relevant. Thus, we scanned the genome for the hypo-DMCs in DMC-enriched regions. Instead of using a fixed window size, we selected top DMCs as seeds and extended the region in both directions until the score of the region, which was defined as average -log(*p*), was a preset threshold, i.e., 95% quantile of the genome. We performed 1000 simulations by shuffling the *P*-value of each DMC in the whole genome to identify the significance. We observed significantly more DMRs in the real genome compared to simulations (*Empirical distribution*, *P-value* < 0.001) (Fig 4A and in S4 Fig for other nine cancer cohorts). It indicated that in most cancer cohorts, the DMR could be explained by excessive hypomethylation in cancers. To further validate the clustering tendency, we performed a Run test for randomness. A run is defined as a series of consecutive DMCs that show a higher or lower methylation level in primary cancers. As expected, we observed significantly fewer runs than that from a random process, confirming continuous hypomethylation in cancers across the genome (S2 Table).

For DMRs of six different cancer cohorts (BLCA, HNSC, LIHC, LUAD, LUSC, and UCEC) that ensured enough sample cohorts and acquired a larger intersection (Fig 4B), we found eight common DMRs in cancer (S3 Table). MAGMA was used to annotate the hallmark, GO, and KEGG in eight common DMRs, and the hallmark enrichment indicated that the formation of angiogenesis and coagulation was common in cancer progression, and these were also associated with DNA repair and fatty acid metabolism (Fig 4C). Functional enrichment analysis performed on eight common DMRs found motor activity as the most enriched category in GO analysis (Fig 4D) and found cardiac muscle contraction as the most enriched category in KEGG analysis (Fig 4E).

Patterns of differential methylation in cancers

Next, we investigated the differences in methylations in each cancer cohort and explored the feasibility of automated cancer cohort discrimination with cancer-specific DMCs. For real-world applications, too many

predictors would be cost-inefficient and might cause overfitting. To avoid that, the RF algorithm was used to produce a thinned set of DMCs by iteratively removing less discriminative DMCs until reaching the top 10 DMCs in each cancer cohort (S4 Table). Using the derived set of 10 most informative DMCs, we achieved an accuracy, sensitivity, and specificity of 99% by the RF classifier for separating primary cancers and normal solid tissues for all 10 cancer cohorts. Unsupervised hierarchical clustering of the thinned set of DMCs revealed distinct methylation profiles between primary cancers and corresponding normal solid tissues in the top 10 DMCs (Fig 5A for BLCA and S5 Fig for the other nine cancer cohorts). The data was also used for visualization of the *t*-SNE projection (Fig 5B for BLCA and S6 Fig for the other nine cancer cohorts), which agreed with the results of heatmaps.

We further optimized the final model using binary logistic regression with backward elimination for DMCs selection using the reference cohorts. The top 10 DMCs models in each cancer cohort were subsequently applied to the independent validation cohort for evaluating the diagnostic performance. The risk-scores for individual patient in each of cancer cohorts were calculated as follows (for example BLCA: Logit (*P*) = - (2.67 × cg01279413) - (1.93 × cg01425409) - (1.80 × cg03304660) - (1.05 × cg04181701) - (1.43 × cg06804091) - (1.29 × cg07709358) - (1.56 × cg08313382) - (1.16 × cg11912215) - (2.43 × cg13357249) - (3.18 × cg16489193) + 7.25. S5 Table shows the other cancer cohorts.)

Next, we explored whether the DMCs in the thinned set could also be used to simultaneously distinguish 32 different cancer cohorts. Again, the RF classifier was used to perform the multi-classification. We included 99 DMCs (10 DMCs from 10 cancer cohorts and removed a duplicate cg07274716) in the classification model, with 8423 primary cancers from 32 cancer cohorts. An accuracy of 97.9% was achieved, which was 3.8% higher compared with the classification model using a full DMC set (94.1%). The classification result was also used for visualization in *t*-SNE projection in 3D space, which clearly showed that primary cancers could be clustered using 99 DMCs (Fig 5D). Figure 5E is 180 degrees horizontal rotation of Fig 5D. Moreover, the RF was used for iterative computation of 99 DMCs in 32 cancer cohorts and acquired more important 10 DMCs, using the top 10 DMCs with 91.2% accuracy. For each of the top 10 DMCs, they all significantly affected the survival time during 4,000 days. For example, cg13324103 showed hypermethylation in primary cancers (Fig 5A) and hypermethylation corresponded to lower survival rates in corresponding cancer cohorts (Fig 5F for cg13324103 and S7 Fig for the other nine DMCs).

A significant number of primary cancers exhibited a high frequency of both relative hyper- and hypomethylation in 99 DMCs (Fig 5A for BLCA and S2 Fig for other nine cancer cohorts). We scored the extent of DMI in one of 99 DMCs by combining the percentage of DNAm loci offsetting median > 0.5 or < 0.5, which revealed that 99 DMCs could classify 32 primary cancers, but each DMC still exhibited high instability (Fig 5C). We also used CV divided by the median as a measure of stability for 99 DMCs and found that higher stability (< median) in 99 DMCs exhibited significantly longer survival than the instability patients (Fig 5G for BLCA and S8 Fig for other nine cancer cohorts).

Discussion And Conclusion

Several previous studies have shown that DNAm is continuously influenced by both internal and external environments, such as smoking or aging [23-25]. These environmental contributors are known to act together to cause a diseased state [26]. Evidence has shown that changes in the internal and external environment can cause global and gene-specific DNAm alternation [27]. The DNAm repair mechanism tries to maintain genomic stability [28]. However, long-term exposure to stimulation causes severe DNAm disorder and, in turn, the onset of cancer [24, 29]. It is necessary to investigate DNAm patterns across multiple cancer cohorts to understand the mechanisms of DNAm in tumorigenesis. For a healthy methylation pattern, Hachiya et al. investigated the relationship between DNAm status and reference intervals based on comprehensive DNAm profiles covering approximately 24 million CpGs. In healthy monocytes, 80.1% of CpGs were hypermethylated, 11.3% were hypomethylated, and 8.6% were intermediately methylated [30]. For the aging population, while both hypermethylation and hypomethylation are known to occur with age, the tendency for hypermethylation is more [30, 31]. For cancer tissues, previous research has shown an interesting pattern, including global hypomethylation and local hypermethylation [32]. Gene silencing by hypermethylation of the promoter is known to occur in most of the cancer cohorts, and the silencing of tumor suppressor genes accelerates tumor progression. Genome-wide hypomethylation might be involved in reviving gene repression in genomic regions that are frequently off in normal cells, this may lead to the activations of proto-oncogenes, followed by increased genomic instability and ultimately carcinogenesis [13].

Recent studies have provided extensive data on the common and specific patterns in pan-cancer [33]. However, these studies did not focus on the discovery of global biomarkers that would be universally specific or common and would have significant features with the DNAm profiles. Epigenetic markers associated with specific diseases are considered as emerging biomarkers for the diagnosis and prediction of treatment response and prognosis in many diseases. In this study, the TCGA project was used to accelerate our understanding of the molecular basis of cancer using genomic technologies. We found DNAm patterns with universal specificities and commonalities among DNAm profiles across multiple cancer cohorts. For example, we explored angiogenesis and coagulation function in common DMRs of most cancer cohorts, which are inducing factors for several cancers. Additionally, we explored a few DMCs that could be used to accurately distinguish between primary cancers and normal solid tissues, as well as different cancer cohorts with high accuracy because of the specificity mechanism, which could be used as an important biomarker in clinical trials.

The aberrant methylation of genes, mainly CpGs, and the hyper- and hypomethylation of intergenic and repetitive elements that suppress tumorigenesis, are known to occur early in cancer development and progressively increase, leading to the malignant phenotypes[34] . By exploring hyper- and hypomethylated CpGs, we also confirmed that hypermethylated CpG islands were frequently located at gene promoter regions, with a certain amount of hypomethylation. The DNAm instability of the promoter region caused whole-genome instability, inducing several types of cancer (Fig 1B). We found that the methylation levels peaked at the first exon near TSS and TTS in primary cancers and normal solid tissues. On the contrary, compared with primary cancers, normal solid tissues showed steeper peaks, DNAm in the first exon has been shown to suppress gene expression; however, in the cancer cells, methylation pattern almost

disappeared, which supported the transformation of cells to neoplastic in the first exon, suggesting the transcriptional “barriers” were significantly weakened. (Figs 1C-D).

Recent studies have revealed that genomic instability is one of the drivers of carcinogenesis and has been called a “facilitating characteristic” that helps generate the hallmarks of cancer [35-37]. We further confirmed DNAm instability across all CpGs or genes, using CV in primary cancers and normal solid tissues, respectively. We found that primary cancers were highly disordered compared with the normal solid tissues in different cancer cohorts (Figs 2A-D), same as in genes (Figs 2E-F). Additionally, we scored the extent of DMI in each primary cancer by combining the percentages of hyper- and hypomethylated loci and found that most of the individual primary cancers were located in > 25 % of hyper- and hypomethylated loci, indicating that a different level of instability was present in most of the individual primary cancers (Fig 2G).

Next, we clustered all CpGs in 10 cancer cohorts and showed that biologically similar cancers shared common CpGs (Fig 1A). For example, KIRC and KIRP were clustered together and shared more similar DNAm mechanisms [33]. We also analyzed common DMCs with 138 CpGs in 10 cancer cohorts (Fig 3A) and DMR with eight common regions in six cancer cohorts (Fig 4B). Previous studies have suggested that DMRs are more strongly correlated with gene expression [38]. We used MAGMA to annotate the 138 CpGs in 10 cancer cohorts and eight common DMRs with six cancer cohorts and found their common function, such as hallmarks of eight common DMRs, were associated with angiogenesis and coagulation (Fig 4C).

Different cancers are known to have different dependencies [39-41] . A series of specific DMC β values, also known as single cancer biomarkers, can identify a specific feature of cancer. Therefore, we used RF classifier to classify primary cancer and normal solid tissue. We used iterations of RF to find some important DMCs to reduce DMCs from 396,065 to 10 and found that these top 10 DMCs achieved the same accuracy compared with all DMCs in the dichotomy of single cancer cohorts. Additionally, we used these 99 DMCs to analyze multi-classification. Instead of adopting total DMCs directly, we got rid of the tissue specificity and were able to precisely distinguish the cancer cohorts. As we continued to iterate with RF, we found 10 more important DMCs in 99 DMCs; every DMCs had an impact on the survival time during 4000 days (Fig 5F). Interestingly, we also found that 99 DMCs could classify 32 primary cancers, but each DMC still had high instability (Fig 5C).

Thus, using a comprehensive analysis of DNAm patterns and variations in cancers, our results provided new insights into a common pattern of aberrant DNAm variations, which could facilitate the development of biomarkers based on universal specificities and commonalities in the pan-cancer analysis. A uniform implementation of the DNAm biomarkers holds great promise for the standardization of cancer diagnostics across clinical trials. We expect that the principle of using DNAm signatures as part of combined cancer therapy could improve diagnostic accuracy not only in tumor pathology but also serve as a blueprint in other fields of pathology.

Materials And Methods

TCGA data

We downloaded the DNAm datasets from The Cancer Genome Atlas (TCGA) data portal (https://tcga-data.nci.nih.gov/tcga) [42, 43]. Only methylation data from the Illumina Infinium 450K platform were used, which included 482,421 CpGs. We deleted CpGs with ³ 30% missing values in all samples. Finally, we selected 9,169 methylated samples, including 8,423 primary cancers and 746 normal solid tissues.

We used the DNAm data of the primary 10 cancer cohorts with corresponding normal solid tissues, including bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), thyroid carcinoma (THCA), and uterine corpus endometrial carcinoma (UCEC) (4501 primary cancers and 598 normal solid tissues).

In the multi-classification analysis, we included adrenocortical carcinoma (ACC), brain lower-grade glioma (LGG), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), kidney chromophobe (KICH), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), mesothelioma (MESO), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), sarcoma (SARC), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), testicular germ cell tumors (TGCT), thymoma (THYM), uterine carcinosarcoma (UCS), and uveal melanoma (UVM).

Differential promoter

We screened the CpGs in promoter regions (-1500 ~ +500 around TSS) that were re-annotated using ENSEMBL Release GRCh38 and extracted based on $|\log_2 FC| \geq 1$. The *P*-value < 0.05, based on Wilcoxon rank-sum test, was regarded as the differential promoter.

DMI score

We defined hyper- and hypomethylation per *loci* using $\log_2 FC_{loci} = \log_2(loci) - \log_2(\textit{mean of normal sample loci})$, and normal solid tissues were removed in ANOVA analysis when *F*-value > 0.1 and *P*-value > 0.05. Hyper- and hypomethylation *loci* were defined $|\log_2 FC_{loci}| \geq 1$.

The DNA methylation instability (DMI) score, which was designed to capture the concomitant increase in both hyper- and hypomethylation events, was defined as the of *hyper- (H)* and *hypo- (h)* methylation frequencies:

$$DMI = (1 + \beta^2)\frac{Hh}{\beta^2 H + h} \qquad (1)$$

We used β = 2 in our analyses to compensate for the asymmetric distribution of hyper- and hypomethylation frequencies [44]. The enrichment of each DNAm alteration with a DMI score was studied both within each individual primary cancer or single DMC. In addition, for single DMC analysis, for one of 99 DMCs, we also used median segmentation to evaluate relative *hyper(H')*- and *hypo(h')*- methylation. The other methods were same as above.

DMRs and DMCs annotation

We performed a MAGMA annotation (http://ctg.cncr.nl/software/magma) to understand the perturbation in biological system function by common genetic variation and database source from MSigDB (https://www.gsea-msigdb.org/gsea/msigdb) (hallmark, GO, and KEGG). This model examines the relationship between genes and the trait of interest compared with other genes in the same dataset; the multi-marker associations are often easy to detect using linkage disequilibrium (LD) between markers[45-47] . First, we acquired 180 single nucleotide polymorphisms (SNPs) mapped from DMRs using hg38 locus. Next, we calculated gene *P*-values from the DMRs-SNPs data for the MAGMA gene analysis around the transcription start and stop sites. Similarly, 102 SNPs mapped from DMCs were used for the MAGMA gene analysis.

Survival analysis

Kaplan-Meier survival analysis was done to assess the relationship of the signature scores with overall survival. These analyses were performed with the R package "survival" [52]. The hyper- and hypomethylation per *loci* were performed in specific CpGs [using $\log_2 FC_{loci} = \log_2(loci) - \log_2$ (*mean of normal sample loci*)]. For each cancer cohort, the patients were divided into two groups, based on hyper- and hypomethylation state.

Also, we also used hyper- and hypomethylation per *loci* to calculate the CV value via 99 DMCs. In each cancer cohort, the patients were divided into two groups: stability tendency and instability tendency by splitting the median of the CV value.

Classification and feature selection

CpGs with ³ 30% missing values in either primary cancers or normal solid tissues were excluded for the classification analysis. The RF algorithm was used as the primary approach for classification in this study [48]. We considered the class-imbalance problem between primary cancers and normal solid tissues. A synthetic minority oversampling technique (SMOTE) [49] was adopted to eliminate the class-imbalance problem, with 5-fold cross-validation [50].

First, we used the RF algorithm to initially train the CpG features. Features with an important value of 0 were deleted, and the reserved features were trained, and the same procedure was repeated until the number of features with an important value of 0 was less than 1000. Next, the features whose important value was less than 1.00e-7 were deleted, and the reserved features were trained, and the same procedure

was repeated until the number of features whose important value was less than 1.00e-7 was less than 1000. Then, the features whose important value was less than 1.00e-6 were deleted, and reserved features were trained, and the same procedure was repeated until the number of features whose important value was less than 1.00e-6 was less than 1000. The iteration continued until the final ten most important CpGs were obtained.

Binary logistic regression

Here, we used the binary logistic regression model as the second approach for classification [51]. The reference cohorts adopted the resultant dataset from the RF classifier. The risk-scores for individual patients in each cohort of cancer were acquired by calculating the regression coefficient and the intercept. The sigmoid function was defined as follows: $g(z)=1/(1+e^{-z})$, where $z(x)=\theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_n X_n = \theta^T X$.

Declarations

Acknowledgments

We thank the support of the high-performance computing platform of the State Key Laboratory of Agrobiotechnology.

The author declares no competing interests.

References

1. Czamara D, Eraslan G, Page CM, Lahti J, Lahti-Pulkkinen M, Hämäläinen E, Kajantie E, Laivuori H, Villa PM, Reynolds RM *et al*: **Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns**. *Nature Communications* 2019, **10**(1):2548.
2. Eilis H, Olivia K, Karen S, Joe B, Wong CCY, Belsky DW, Corcoran DL, Louise A, Moffitt TE, Avshalom C: **Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins**. *PLOS Genetics* 2018, **14**(8):e1007544-.
3. Birney E, Smith GD, Greally JM: **Epigenome-wide Association Studies and the Interpretation of Disease -Omics**. *Plos Genetics* 2016, **12**(6):e1006105.
4. Muhammad, Ahsan, Weronica, E., Ek, Mathias, Rask-Anderse, Torgny, Karlsson, Allan: **The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases**. *Plos Genetics* 2017.
5. Yafremava LS, Monica W, Suravi T, Arshan N, Minglei W, Mittenthal JE, Gustavo C-A: **A General Framework of Persistence Strategies for Biological Systems Helps Explain Domains of Life**. *Frontiers in Genetics* 2013, **4**:16.
6. Serban M, Green S: **Biological robustness**; 2020.

7. An J, Li J, Wang Y, Wang J, Li Q, Zhou H, Hu X, Zhao Y, Li N: **A Homeostasis Hypothesis of Avian Influenza Resistance in Chickens**. *Genes (Basel)* 2019, **10**(7).
8. Sandhu DS, Shire AM, Roberts LR: **Epigenetic DNA hypermethylation in cholangiocarcinoma: potential roles in pathogenesis, diagnosis and identification of treatment targets**. *Liver Int* 2008, **28**(1):12-27.
9. Flavahan WA, Gaskell E, Bernstein BE: **Epigenetic plasticity and the hallmarks of cancer**. *Science* 2017, **357**(6348).
10. Klein Hesselink EN, Zafon C, Villalmanzo N, Iglesias C, van Hemel BM, Klein Hesselink MS, Montero-Conde C, Buj R, Mauricio D, Peinado MA *et al*: **Increased Global DNA Hypomethylation in Distant Metastatic and Dedifferentiated Thyroid Cancer**. *J Clin Endocrinol Metab* 2018, **103**(2):397-406.
11. **Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers**. *Gastroenterology* 2015, **149**(5):1204-1225.e1212.
12. Ehrlich M: **DNA hypomethylation in cancer cells**. *Epigenomics* 2009, **1**(2):239-259.
13. Sheaffer KL, Elliott EN, Kaestner KH: **DNA Hypomethylation Contributes to Genomic Instability and Intestinal Cancer Initiation**. *Cancer Prev Res (Phila)* 2016, **9**(7):534-546.
14. Kim H, Wang X, Jin P: **Developing DNA methylation-based diagnostic biomarkers**. *J Genet Genomics* 2018, **45**(2):87-97.
15. Ya W, Teschendorff AE, Martin W, Shuang W: **Accounting for differential variability in detecting differentially methylated regions**. *Briefings in bioinformatics* 2019, **2019,20(1)**:47-57.
16. Vedeld HM, Goel A, Lind GE: **Epigenetic biomarkers in gastrointestinal cancers: The current state and clinical perspectives**. *Seminars in Cancer Biology* 2017:S1044579X17301815.
17. Wang C, Chen L, Yang Y, Zhang M, Wong G: **Identification of potential blood biomarkers for Parkinson's disease by gene expression and DNA methylation data integration analysis**. *Clinical Epigenetics* 2019, **11**(1).
18. Zhou S, Treloar AE, Lupien M: **Emergence of the Noncoding Cancer Genome: A Target of Genetic and Epigenetic Alterations**. *Cancer Discov* 2016, **6**(11):1215-1229.
19. Rowlatt A, Hernández-Suárez G, Sanabria-Salas MC, Serrano-López M, Rawlik K, Hernandez-Illan E, Alenda C, Castillejo A, Soto JL, Haley CS *et al*: **The heritability and patterns of DNA methylation in normal human colorectum**. *Hum Mol Genet* 2016, **25**(12):2600-2611.
20. Benton MC, Johnstone A, Eccles D, Harmon B, Hayes MT, Lea RA, Griffiths L, Hoffman EP, Stubbs RS, Macartney-Coxson D: **An analysis of DNA methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss**. *Genome Biol* 2015, **16**(1):8.
21. Reiter JG, Baretti M, Gerold JM, Makohon-Moore AP, Daud A, Iacobuzio-Donahue CA, Azad NS, Kinzler KW, Nowak MA, Vogelstein B: **An analysis of genetic heterogeneity in untreated cancers**. *Nat Rev Cancer* 2019, **19**(11):639-650.

22. Zhang B, Zhou Y, Lin N, Lowdon RF, Hong C, Nagarajan RP, Cheng JB, Li D, Stevens M, Lee HJ *et al*: **Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm**. *Genome Res* 2013, **23**(9):1522-1540.
23. Bell CG, Beck S: **The epigenomic interface between genome and environment in common complex diseases**. *Brief Funct Genomics* 2010, **9**(5-6):477-485.
24. Feil R, Fraga MF: **Epigenetics and the environment: emerging patterns and implications**. *Nat Rev Genet* 2012, **13**(2):97-109.
25. Amit I, Winter DR, Jung S: **The role of the local environment and epigenetics in shaping macrophage identity and their effect on tissue homeostasis**. *Nat Immunol* 2016, **17**(1):18-25.
26. Rappaport SM, Smith MT: **Environment and Disease Risks**. *Science* 2010, **330**(6003):460-461.
27. Feil R, Fraga MF: **Epigenetics and the environment: Emerging patterns and implications**. *Nature Reviews Genetics* 2012, **13**(2):97-109.
28. Schär P, Fritsch O: **DNA repair and the control of DNA methylation**. *Fortschritte Der Arzneimittelforschungprogress in Drug Researchprogrès Des Recherches Pharmaceutiques* 2011, **67**:51-68.
29. Yan, V., Sun: **The Influences of Genetic and Environmental Factors on Methylome-Wide Association Studies for Human Diseases**. *Current Genetic Medicine Reports* 2014, **2**(4):261-270.
30. Jei Kim J-YK, Jean-Pierre J. Issa: **Aging and DNA methylation**. *Current Chemical Biology* 2009, **3**(1):1-9.
31. Archana, Unnikrishnan, Willard M, Freeman, Jordan, Jackson, Jonathan D, Wren, Hunter, Porter: **The role of DNA methylation in epigenetics of aging**. *Pharmacology & therapeutics* 2018.
32. Jones PA, Issa JP, Baylin S: **Targeting the cancer epigenome for therapy**. *Nat Rev Genet* 2016, **17**(10):630-641.
33. Xiaofei Y, Lin G, Shihua Z: **Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns**. *Briefings in Bioinformatics* 2016(5):761.
34. Okugawa Y, Grady WM, Goel A: **Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers**. *Gastroenterology* 2015, **149**(5):1204-1225.e1212.
35. Zhang Y, Wendte JM, Ji L, Schmitz RJ: **Natural variation in DNA methylation homeostasis and the emergence of epialleles**. *Proc Natl Acad Sci U S A* 2020, **117**(9):4874-4884.
36. Pikor L, Thu K, Vucic E, Lam W: **The detection and implication of genome instability in cancer**. *Cancer Metastasis Rev* 2013, **32**(3-4):341-352.
37. Tubbs A, Nussenzweig A: **Endogenous DNA Damage as a Source of Genomic Instability in Cancer**. *Cell* 2017, **168**(4):644-656.
38. Xu J, Chen G, Hermanson PJ, Xu Q, Sun C, Chen W, Kan Q, Li M, Crisp PA, Yan J *et al*: **Population-level analysis reveals the widespread occurrence and phenotypic consequence of DNA methylation variation not tagged by genetic variation in maize**. *Genome Biol* 2019, **20**(1):243.

39. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM *et al*: **Defining a Cancer Dependency Map**. *Cell* 2017, **170**(3):564-576.e516.
40. Boehm JS, Golub TR: **An ecosystem of cancer cell line factories to support a cancer dependency map**. *Nat Rev Genet* 2015, **16**(7):373-374.
41. Eifert C, Powers RS: **From cancer genomes to oncogenic drivers, tumour dependencies and therapeutic targets**. *Nat Rev Cancer* 2012, **12**(8):572-578.
42. **Comprehensive molecular portraits of human breast tumours**. *Nature* 2012, **490**(7418):61-70.
43. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV *et al*: **An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics**. *Cell* 2018, **173**(2):400-416.e411.
44. Saghafinia S, Mina M, Riggi N, Hanahan D, Ciriello G: **Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors**. *Cell Rep* 2018, **25**(4):1066-1080.e1068.
45. de Leeuw CA, Mooij JM, Heskes T, Posthuma D: **MAGMA: generalized gene-set analysis of GWAS data**. *PLoS Comput Biol* 2015, **11**(4):e1004219.
46. de Leeuw CA, Neale BM, Heskes T, Posthuma D: **The statistical properties of gene-set analysis**. *Nat Rev Genet* 2016, **17**(6):353-364.
47. de Leeuw CA, Stringer S, Dekkers IA, Heskes T, Posthuma D: **Conditional and interaction gene-set analysis reveals novel functional pathways for blood pressure**. *Nat Commun* 2018, **9**(1):3768.
48. Wu J, Zhao Y: **Machine learning technology in the application of genome analysis: A systematic review**. *Gene* 2019, **705**:149-156.
49. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA: **Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?** *Brief Bioinform* 2013, **14**(3):315-326.
50. Blagus R, Lusa L: **SMOTE for high-dimensional class-imbalanced data**. *BMC Bioinformatics* 2013, **14**:106.
51. Levy JJ, O'Malley AJ: **Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning**. *BMC Med Res Methodol* 2020, **20**(1):171.

Figures

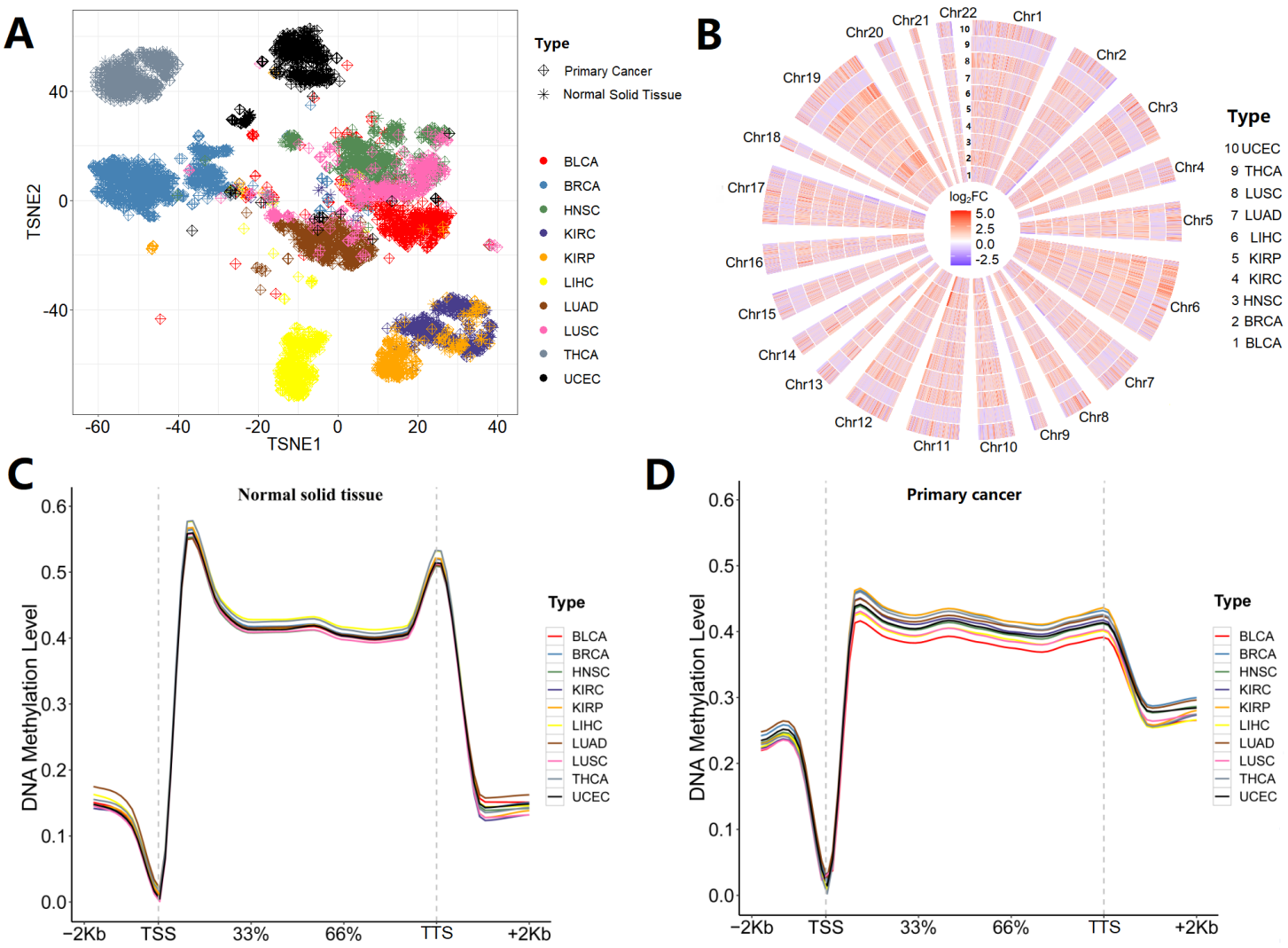

Figure 1

(A) Unsupervised clustering of reference cohort samples (n = 5,099) with all CpGs using t-SNE. Individual samples are color-coded (m = 10) and labeled based on a specific cohort. (B) Hypermethylated CGIs were frequently found to be located at gene promoter regions and presented an unstable state in primary cancers. (C) The mean DNAm levels of normal solid tissues across gene body regions and their 2-kb flanking regions for TSS and TTS. (D) The mean DNAm levels in primary cancers across gene body regions and their 2-kb flanking regions for TSS and TTS.

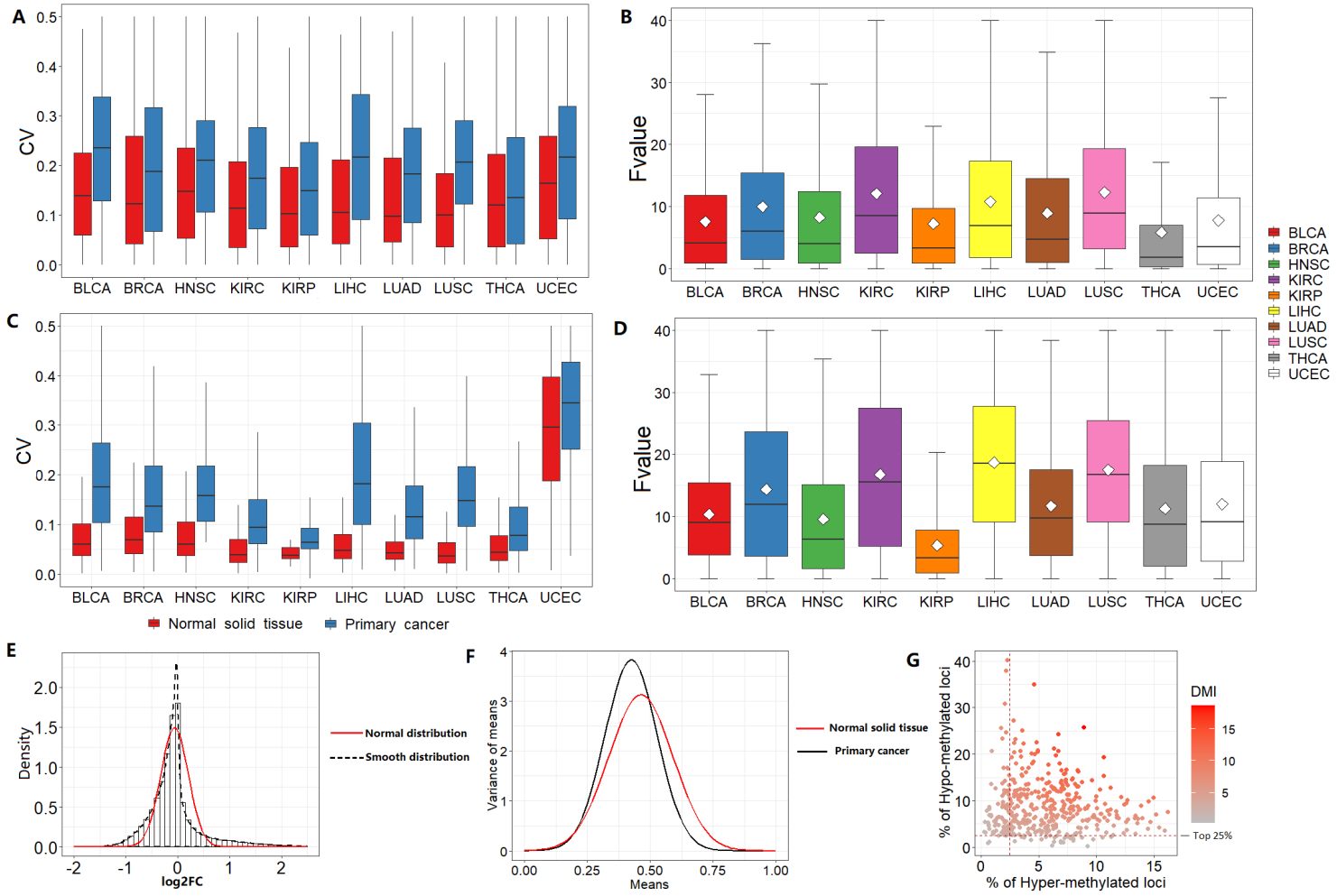

Figure 2

(A) The boxplot shows the median and interquartile range of CV and primary cancer had higher variations than the normal solid tissues in the CpGs matrix. (B) The boxplot shows the CpGs median and interquartile range of the F-value, we observed different levels of DNAm variation in different cancer cohorts. (C) The boxplot shows the median and interquartile range of CV and primary cancers had higher variation than the normal solid tissues in the gene matrix. (D) The boxplot shows that the genes' median and interquartile range of F-value, we observed different levels of DNAm variation in different cancer cohorts. (E) Histogram of Z-score normalized log2FC methylation levels. Red curve shows the expected standard normal distribution, and the broken black curve is a smoothed curve for the observed distribution. (F) The variance of means of each CpG of primary cancers was significantly higher than that of normal solid tissues. (G) The extent of DMI in each sample was calculated by combining the percentages of hypermethylated and hypomethylated loci, and most of the individual primary cancers were located in > 25% of hypermethylated and hypomethylated loci.

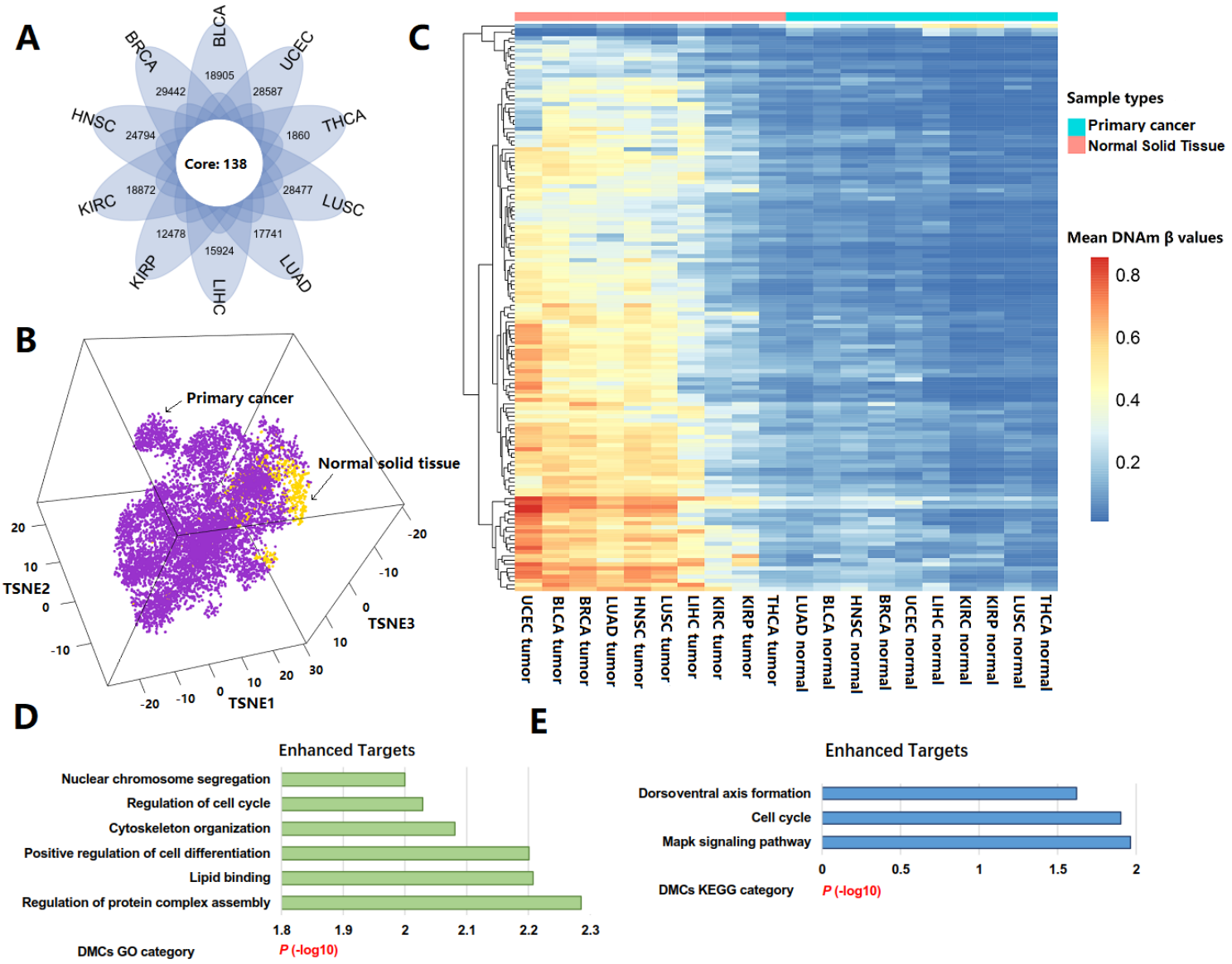

Figure 3

(A) All cancer cohort shared 138 CpGs with the same DNAm tendency (hypo- or hypermethylation) among DMCs. (B) When the tissue specificity was removed, primary cancers were more clustered and significantly separated from the corresponding normal samples from the same tissue, confirming the presence of a common pattern in different primary cancers. (C) The 138 CpGs preferred hypermethylation in primary cancers and hypomethylation in normal solid tissues. When tissue specificity was removed, different primary cancers showed significant commonality. (D) MAGMA was used for functional enrichment analysis on 138 CpGs via GO analysis and (E) KEGG analysis.

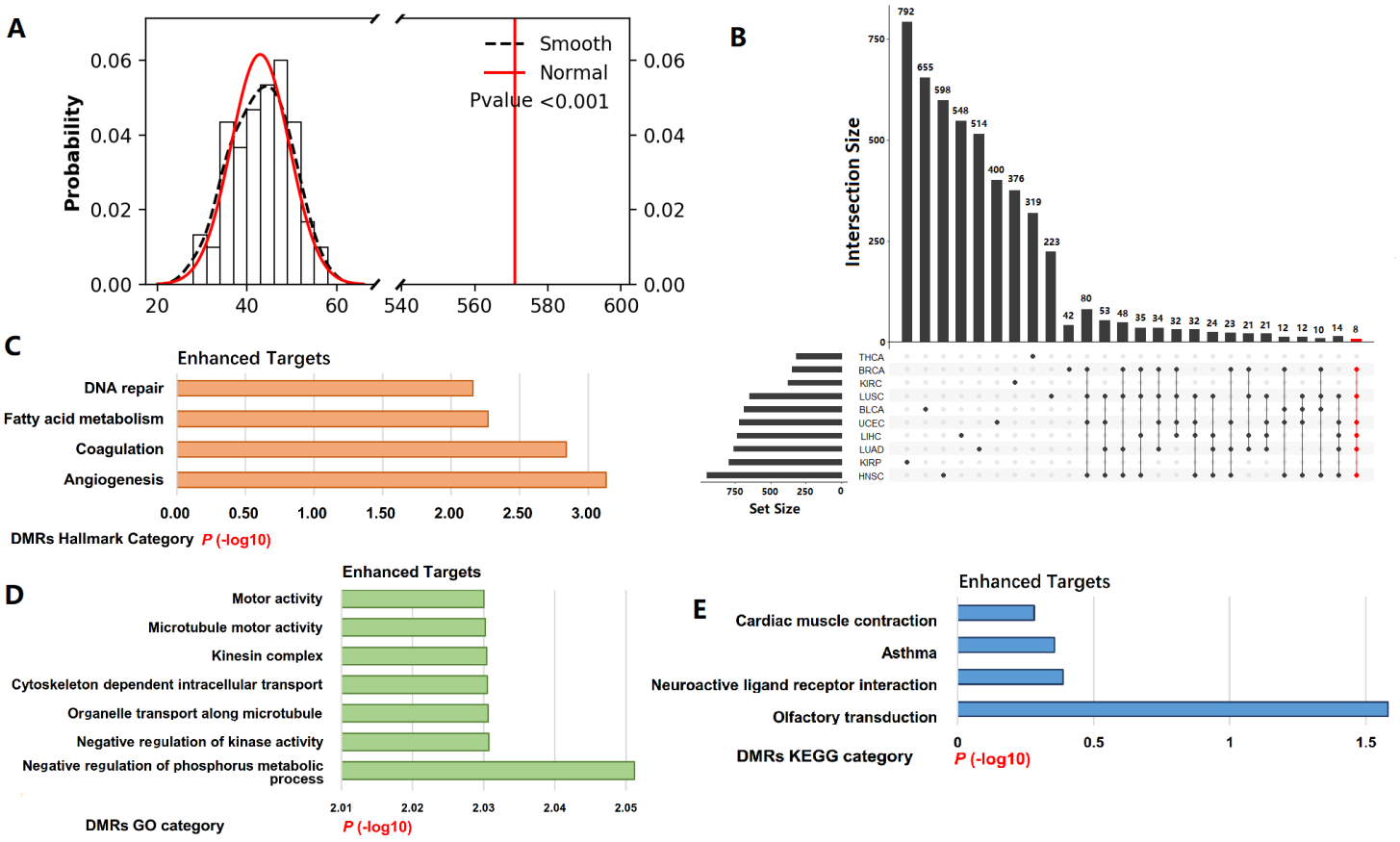

Figure 4

(A) For BLCA, the distribution of the number of DMRs included clusters of 1,000 randomized rearrangement, both upregulated and downregulated. The number of hypo-DMRs found in clusters of the real genomes was indicated by a red straight line, which was significantly higher than the randomized datasets. (B) For DMRs of 6 different cancer cohorts (BLCA, HNSC, LIHC, LUAD, LUSC, and UCEC) to acquire a larger intersection, we found 8 common DMRs in cancer. (C) MAGMA was used to annotate hallmarks in 8 common DMRs. The result shows indicated that the formation of angiogenesis and coagulation in cancer progression were common and were associated with DNA repair and fatty acid metabolism. (D) Functional enrichment analysis was performed on the 8 common DMRs in GO analysis and (E) KEGG analysis.

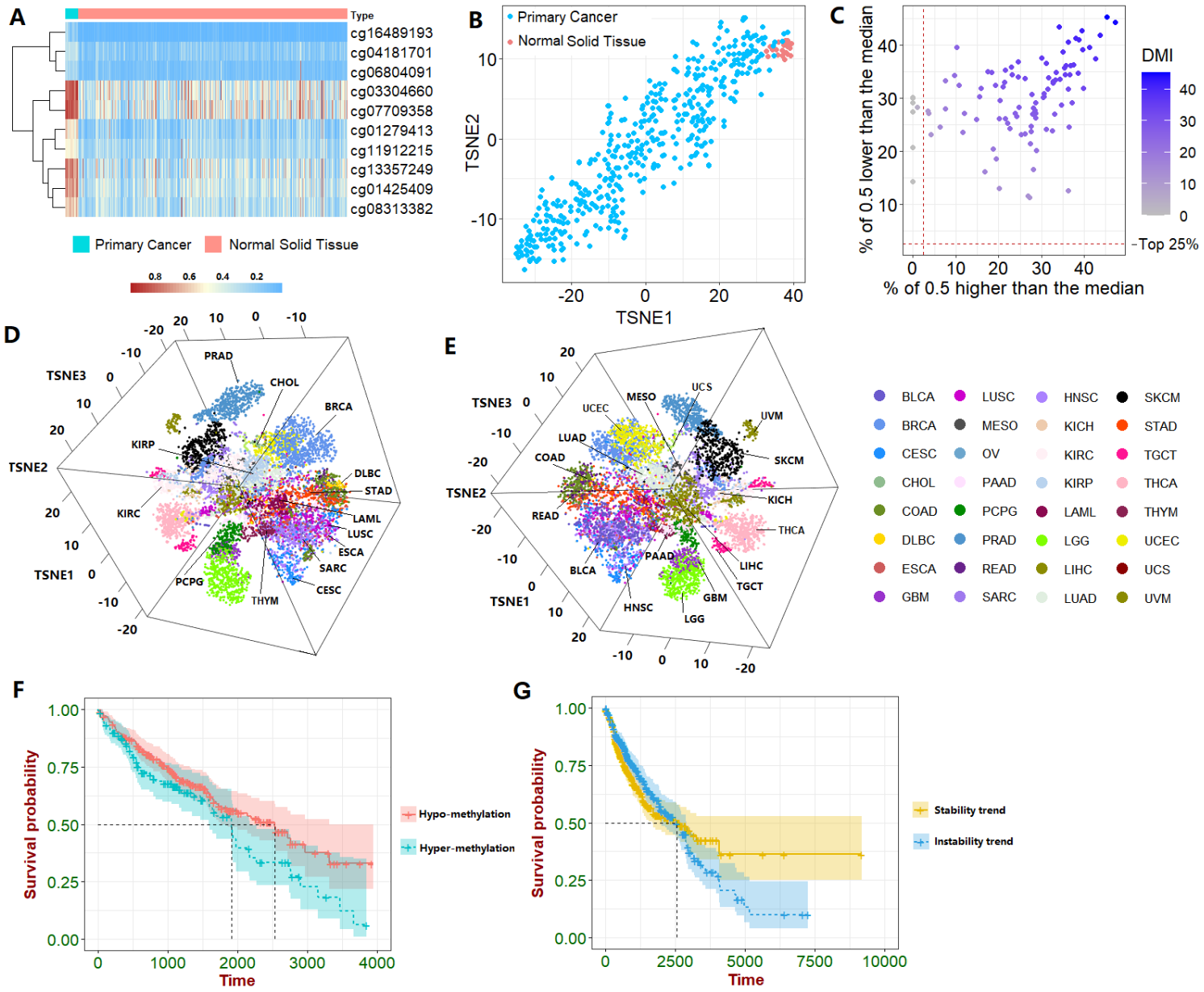

Figure 5

(A) Top 10 CpGs from each cancer cohort are shown in the heatmap. (B) Scatter plots show a clear boundary between primary cancers and normal solid tissues in t-SNE distributions. (C) We scored the extent of DMI in each DMC by combining the percentage of DNA methylation loci offsetting median > 0.5 or < 0.5, which revealed that 99 DMCs could classify 32 primary cancers, but each CpG exhibited high instability. (D) The classification result was also used for visualization in t-SNE projection in 3D space, which showed that primary cancers could be clustered using 99 DMCs. (E) Figure 5D, after 180 degrees horizontal rotation. (F) cg13324103 had important effects on the survival during 4,000 days. For example, cg13324103 showed hypermethylation in the primary cancers, which was associated with lower survival rates in corresponding cancer cohorts. (G) For BLCA, we also used CV divided by the median as a measure of stability for 99 DMCs and found that higher stability (< median) in 99 DMCs exhibited significantly longer survival than instability patients.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementtable.xls
- supplementfigure.docx