

The direct/inverse interplay among 5-hydroxymethylcytosine (5hmC)-related biological pathways in gastrointestinal pan-cancer study

Sayyed Sajjad Moraveji

University of Tehran

Samane Khoshbakht

University of Tehran

Majid Mokhtari

University of Tehran

Hosein Lanjanian

Shahid Beheshti University of Medical Sciences

Ali Masoudi-Nejad (✉ amasoudin@ut.ac.ir)

University of Tehran

Research Article

Keywords: epigenetics marker, Gene Expression

Posted Date: August 26th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-842553/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

The direct/inverse interplay among 5-hydroxymethylcytosine (5hmC)-related biological pathways in gastrointestinal pan-cancer study

Sayyed Sajjad Moraveji¹, Samane Khoshbakht¹, Majid Mokhtari¹, Mahdiah

Salimi², Hosein Lanjanian³, Ali Masudi-Nejad^{1,3,4,*}

1. Laboratory of Systems Biology and Bioinformatics (LBB), Department of Bioinformatics, Kish International Campus, University of Tehran, Kish Island, Iran
2. Department of medical genetics, institute of medical biotechnology, National Institute of Genetic Engineering and Biotechnology (NIGEB), Tehran, Iran
- 3.
4. Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran, amasoudin@ut.ac.ir

Abstract

5-hydroxymethylcytosine (5-hmC) as an epigenetics marker has significant impacts on cancer progression. Identification of preserved 5-hmC-related subnetworks in pan-cancer studies could lead to a better understanding of gastrointestinal (GI) cancers insights. Here, we conducted a network-based analysis on 5-hmC values of GI cancers, including colon, gastric, pancreatic cancers, and healthy donors. The co-5-hmC network was reconstructed using the weighted gene coexpression network (WGCNA) method. The hierarchical clustering method was implemented to detect pan-cancer-related modules/subnetworks. The preservation of modules was assessed using another dataset. Modules were functionally enriched, and biological pathways were visualized using the ConsensuspathDB. A 5-hmC predictive model was determined using the elastic network classifier to distinguish cancer patients and healthy individuals. To assess the efficiency of the model the recursive operating characteristics (AUC) curve was computed using the 5 cross-fold validation and an external dataset as well. Three pan-cancer-related subnetworks were detected preserved in another dataset. The main biological pathways were the cell cycle, apoptosis, and extracellular matrix (ECM) organization. The direct association between the cell cycle and ECM, the inverse association between apoptosis and ECM organization, and the inverse association between the cell cycle and ECM organization were detected for the 5-hmC marker in GI cancers. The AUC of 92% (0.73-1.00) was detected for the predictive model. In conclusion, the intricate association among biological pathways of ECM organization, Cell cycle, and apoptosis in GI cancers might be the consequence of epigenetics aberration; such findings could be beneficial in precision medicine using liquid biopsy in early stages.

1 | Introduction

Circulating free DNAs (cfDNA) are free DNA fragments found in plasma, serum, or other body fluids. Physiological events, such as apoptosis or micrometastasis are the origin of cfDNAs in the body [1]. Previous studies showed that cfDNA level is 2-3-fold higher in cancer patients compared with healthy individuals [2]. The potential of cfDNA in cancer monitoring has emerged since the advent of liquid biopsy [3]. In cancer studies, cfDNAs are widely investigated for genomics and epigenomics patterns due to the fact that they are representative of the primary tumor, as well as their non-invasive nature; therefore, they could be feasible biomarkers in the clinic for cancer detection and more precise treatments [1]. Among several cfDNA studies, the 5-hydroxymethylcytosine (5hmC) is a specific DNA modification used in cancer monitoring, however, there is less attention in pan-cancer studies [4, 5]. Therefore, we intended to investigate 5hmC-related gene clusters in gastrointestinal cancers.

5hmC, an epigenetic mark, is originated from 5-methylcytosine which is a cytosine modification found in DNA [6]. 5hmC plays a significant role in biological processes, such as gene expression and pathogenesis of cancer [7]. The presence of cytosine modifications, such as 5hmC in gene bodies is related to transcription level; however, this mechanism is currently unknown [7]. Inasmuch as cfDNAs indicate the primary tumor characteristics, therefore, detecting genes related to the increase/decrease of 5hmC could result in finding the mechanism of such cytosine modifications.

5hmC in cfDNAs are widely studied as promising biomarkers in cancer studies due to the non-invasive nature and rapid turnaround utility. Studying such feasible markers in a pan-cancer study could lead to detecting biomarkers or predictive models to classify normal/disease samples, which could be used as an integral part of the management of several cancers [8, 9].

There are several methods to reveal biological mechanisms of 5hmC enriched genes; amongst the network-based methods are promising to show clustered 5hmC enriched genes which could activate specific biological pathways; To detect such pathways could result in more precise feasible treatments and deeper insights in cancer biology. Moreover, using predictive models for cancer detection plays an essential part in the clinic. Finally, assessing the 5hmC markers in cfDNAs from two perspectives of biological insights and predictive practical models could be of great interest to both scientists and physicians.

In this study, we aimed to identify co-5hydroxymethylcytosine (co-5hmC) cluster genes related to gastrointestinal cancers. The values of 5hmC for hepatocellular carcinoma, pancreatic cancer, colon cancer, and gastric cancer were used. The differentially hydroxymethylated genes were detected, and the co-5hmC network was reconstructed using the weighted gene co-expression analysis (WGCNA). Next, hierarchical clustering was implemented to cluster the network and detect modules. Finally, the Pearson correlation between the first principal component of each module and the cancer trait was computed. The gene significance was used as a feature selection index to select the most cancer-related genes. After selecting the cancer-related modules, the elastic net classification method was conducted for them. The best model with the highest AUC was selected as the final predictive model with k-fold cross-validation. Moreover, to know the biological pathways related to 5hmC in GI cancers, modules were functionally enriched using the ConsensusPathDB webserver. To show the power of the module in differentiating the normal and cancer cfDNAs, we implemented a clustering method on samples.

2 | Materials and Methods

2.1 | Data, and preprocessing

The GSE81314 and GSE89570 were downloaded from the national center for biotechnology information (NCBI). The 5-hydroxymethylcytosine (5hmC) data were used for downstream analyses. GSE81314 includes 5hmC methylation value of healthy donors, lung cancer, hepatocellular carcinoma, pancreatic cancer, breast cancer, colon cancer, glioblastoma, and gastric cancer; in which, breast, lung, and glioblastoma cancers were filtered out. The data of GI cancers were included in the study. The GI-related cancers, such as gastric, colon, and healthy individuals were used in our study for validation part.

For preprocessing part, genes with not available (NA) values were filtered out. Next, the coefficient of variation was computed for every gene, and the first quartile of genes that had minimum variation among samples was filtered out. After preprocessing, differentially hydroxymethylated Genes (DHMGs) were detected at probe level using the limma package in R (FDR<0.05, Benjamini Hochberg adjustment). The significant methylated genes were computed for every cancer type separately and also altogether. The significant DHMGs were combined to reconstruct the co-methylated network.

2.2 | Network reconstruction and gene signature detection

We aimed to detect the 5hmC GI-related modules/subnetworks in circulating free DNAs; therefore, after preprocessing and detecting DMGs, the co-hydroxymethylated network was reconstructed using the weighted gene co-expression network analysis (WGCNA) approach. Since our data were spread around zero and they were small values, we did not implement the scale-free part of the WGCNA approach (soft thresholding power = 1). Then, the topological overlap matrix was computed, and hierarchical clustering was implemented to detect modules. The parameters of minimum module size and the deepSplit parameter were set at 30 and four,

respectively. To have GI-related modules, the first principle of every module was computed as module eigengenes, then the Pearson correlation coefficients (PCC) between module eigengenes and cancer trait (cancer=1/healthy=0) were identified. Finally, the modules with PCC higher than 0.5 and significant p-values were selected as GI-related modules.

To detect biological pathways and biological processes related to GI-cancer modules, we implemented overrepresentation analyses using ConsensusPthDB webserver (p-value cutoff = 0.01, minimum overlap = 2).

2.3 | Preservation

The reproducibility of modules is of great importance. The reproducibility could be checked by preservation statistics. To investigate the reproducibility of detected modules, we implemented the preservation procedure defined in the WGCNA package. Since the validation datasets for preservation analysis were limited, we computed the preservation statistics for the external dataset of GSE89570; Moreover, we randomly selected 75%, 50%, and 25% of samples of GSE81314 to run the pattern preservation analyses.

In this part, two statistics of preservation called $Z_{summary}$ and $Median_{rank}$ were computed. The higher $Z_{summary}$ and lower $Median_{rank}$ indicate higher pattern preservation of modules. The $Z_{summary}$ value higher than 10 indicates preserved patterns of modules, but there is no cutoff for $Median_{rank}$.

2.4 | Association among modules

The association among modules was computed using the PCC for the first principle of each module. The first principle of a module is the representative of the module which has the most information of module/subnetwork.

2.5 | predictive model for detection of GI cancers in cfDNA

Since the detection of GI cancer patients in the early stages is of great interest in the clinic, we determined to find a predictive model for classifying patients. In our study, we defined two labels of cancer/normal for classification. We included our three significant modules of green-yellow, tan, and black as feature space for three different classification models. Since we had limitations for sample size compared to feature space for every model, we aimed to use the elastic net classification model. The elastic net model is a regularized model which is a mixture of the least absolute shrinkage and selection operator (lasso) and ridge methods. There are two parameters of alpha and lambda in the elastic net. The lambda indicates the penalty effect, and the alpha indicates the effects of norm 1 or norm 2 on the elastic model. If the value of the alpha parameter is one, the model will be lasso, while if the value of alpha is zero, the model will be ridge model. In this study, to detect the best alpha and lambda values, we used the `cv.glmnet` and the `cva.glmnet` functions in R that are cross validation-based functions to find the optimum lambda and alpha values. The regularized models, such as elastic net are suitable to reduce overfitting. The area under the curve (AUC) of models was computed for every module, and the model with the highest AUC was selected as the final model. In cross-validation, 75% of samples were used for training the predictive model. Moreover, the AUC was detected in the **external dataset of GSE89570**. The GI-related samples and healthy individuals were included in the validation analysis.

Before implementing the elastic net on modules, we computed the gene significance (GS) for three significant modules, which indicates the correlation between the eigengene vector of the module and the trait (cancer/normal hydroxyl methylation). The higher GS indicates a higher relation to the trait. We used GS as a feature selection method for our classification model (cutoff = 0.5).

3 | Results

In this article, we aimed to detect hydroxymethylated modules to predict GI cancers, and also investigate the biological patterns. At the first step, we implemented the preprocessing on data, then we used the WGCNA method to reconstruct a co-hydroxymethylated network, and finally, the hydroxymethylated modules related to GI cancers were extracted.

After coefficient variation and DHMG filters were implemented on data, 7000 genes and all samples were used for downstream analyses. The network was reconstructed and 14 modules were detected; in which three modules had statistically significant p-values, and the absolute value of Pearson correlation higher than 0.5 ($|cor| > 0.5$). The significant modules were GreenYellow, tan, and black, which include 94, 91, and 759 genes, respectively (Table S1,S2,S3). These modules were functionally enriched, and the top-10 significant pathways were reported in figure 1 (p-value < 0.01). The GreenYellow module indicated the biological pathways, including ‘*cell cycle*’, ‘*Toll-Like Receptors Cascades*’, and ‘*NOD1/2 Signaling Pathway*’ (Fig. 1a). The tan module indicated biological pathways, including ‘*Apoptosis*’, ‘*Adherens junction*’, ‘*DNA Damage Response (only ATM dependent)*’, ‘*Neutrophil degranulation*’, and ‘*BMP receptor signaling*’ (Fig. 1b). The black module indicated the biological pathways, including ‘*RUNX3 Regulates Immune Response and Cell Migration*’, ‘*Extracellular matrix organization*’, ‘*Integrin cell surface interactions*’, and ‘*transcriptional activation of dbpb from mrna*’ (Fig. 1c).

3.1 | Module reproducibility

To show the preservation or reproducibility of modules, we extracted two preservation statistics of $Z_{summary}$ and $Median_{rank}$. Both statistics showed the hydroxymethylated GI-related modules were preserved. We implemented the preservation on 25%, 50%, 75 % of our data, and GSE89570 as well (Table S5,S6). The $Z_{summary}$ values of the GreenYellow module were

12, 15, 16, and 18 ($Z_{summary} > 12$ is highly preserved). The $Z_{summary}$ values of the Tan module were 12, 13, 15, and 16. The $Z_{summary}$ values of the Black module were 37, 35, 42, and 40. Since the black module size is much higher than GreenYellow and d tan modules, therefore, we extracted the $Median_{rank}$ statistics which is less sensitive to module size. The higher value of $Median_{rank}$ indicated the higher preservation. The $Median_{rank}$ values for the black module were 10, 11, 11, and 11.5 out of 15. The $Median_{rank}$ values for the GreenYellow module were 7, 12, 12, and 12 out of 15. The $Median_{rank}$ values for the tan module were 10, 10, 10, and 11.5 out of 15.

3.2 | Module association

The association among GI-related modules was extracted using PCC. The association values and their biological processes were illustrated in figure 2.

3.3 | Predictive modules for detection of GI cancers in cfDNA

To have a predictive model for classifying cancer and normal cfDNA, we implemented the elastic net for three hydroxymethylated modules of GreenYellow, tan, and black. To reduce the feature space, we computed the gene significance (GS) for every gene, and genes with GS higher than 0.5 were selected as the input for the elastic net model. In this step, 22 genes were selected as significant features for the elastic net. The model was trained and the best alpha and lambda values were estimated at 0.216 and 0.048, respectively. The AUC was estimated at 92% (Confidence Interval: 0.731-1.000) (Fig. 3a). The final features (11 features) and the estimated coefficients were reported in table S4.

The elastic net for other modules did not reveal high AUC, therefore, the GreenYellow module was selected as the final model. The clustering potential of the GreenYellow module was assessed using the heatmap. The significant module was able to cluster normal/cancer samples well (Fig. 3b).

4 | Discussion

Although 5-Hydroxymethylcytosine signatures in cell-free DNAs provide information about tumors, the co-5-Hydroxymethylcytosine subnetworks/modules, their biological pathways, and the associations of such subnetworks in cfDNAs were not thoroughly investigated. Moreover, 5-hydroxymethylated modules could be a pan-cancer signature. Therefore, they might be employed for a broad range of cancers as diagnostic biomarkers in the clinic using liquid biopsy [10, 11].

Here, we reconstructed the 5-hydroxymethylated network and cluster it to detect the GI-cancer-related subnetworks. The reproducibility of GI cancer-related modules was assessed. The biological pathways of significant modules were detected and visualized. To identify a predictive model for distinguishing GI patients and normal individuals, the GS method and elastic net classifier were implemented for feature selection and model specification, respectively. The cross-validation, heatmap, and AUC were used for model validation.

The biological pathways of three GI cancer-related modules and their associations were assessed in this study (Fig.1,2). Amongst, direct association between cell-cycle arrest, innate immune, and apoptosis pathways were previously investigated in several cancer-related studies [12-14], but, in our study, the direct association between these metastasis-prone biological pathways was detected for aberrant 5-hmC genes in GI cancers (Fig. 2); High and direct association indicate the concurrent activation of cell-cycle, innate immune, and apoptosis pathways through 5hmC aberrations in GI cancers. Inasmuch as the cfDNAs are detectable using liquid biopsy, such findings could be employed in the clinic to provide more precise therapies in the early stages [2, 15, 16]. On the contrary, there is a high and inverse association between the cell-cycle pathway and extracellular matrix organization/regulation of actin cytoskeleton pathways (Fig. 2). In studies conducted by Walker et al. and Pickup et al., they

individually explored several aspects of ECM dysregulation leading to cancer progression and metastasis [17, 18]; They reported the ECM concentration increases and it gets stiffen during tumor development. The stiffened ECM results in increasing cell mobility and reducing the expression of genes that typically inhibit cell-cycle progression and proliferation [17, 18]. In our study, we could detect the inverse association between ECM dysregulation and cell-cycle activity for hydroxymethylated modules of black and greenyellow (Fig. 2), which is consistent with Walker et al. and Pickup et al [17, 18]. Moreover, the gene family of LUX secreted by primary tumor cells is responsible for ECM stiffness and total ECM concentration [17]. Some members of the LUX family, such as LUX, LUX1, and LUX3 were detected in our study. They revealed a low level of 5-hmC in our study. Therefore, such inverse association between the high volume of ECM stiffness and reduction of cell cycle inhibition might be due to aberrant 5-hmC in primary cancer cells, which were detected in cfDNAs.

Of note, there is a wide range of studies, which investigated the role of 5-hmC on the regulation of gene expression in previous studies [4, 19-21], but in this study, due to limitations of lack of gene expression values, we were not able to assess the correlation between dysregulation of gene expression and aberrant 5-hmC values for detected genes in significant modules. Genes, such as SAMD11 and MTHFD2L were reported as diagnostic 5-hmC signatures in previous studies [2, 22]; Their findings are in line with our results for the greenYellow and black modules. Whereas there is no experimental evaluation in our study, the new genes detected in our modules were reported as the new 5-hmC-related biomarkers for GI cancers using the literature mining.

The other significant finding in this study is the high and inverse association between black and tan modules, which indicate apoptosis and ECM organization and pathways, respectively. Of note, signaling properties of ECM, such as the three-dimensional organization of cells and ECM structures have significant impacts on cellular processes, such as proliferation and

apoptosis [23]. Moreover, the DNA damage response pathway, which is significant in the tan module in our study, is the vital determinant of the plasticity of the cellular genome and depends on signaling pathways that regulate apoptosis. Cell adhesion to ECM as a part of ECM organization regulates several of these pathways [23, 24]. Less cell-ECM contact leads to more apoptosis. This finding is in line with our result of two 5-hmC-related modules (Fig. 1,2). we concluded that the activation of such pathways of ECM and apoptosis might relate to aberrant 5-hmC values. Our finding of 5-hmC pathways was not experimentally validated, but their producibility was validated computationally in another dataset; Therefore, they could be used for early-stage treatments after diagnosis by liquid biopsy. Since the cfDNAs mirror the primary characteristics of every patient, they could be an option for precision medicine in the clinic [25].

Despite the 5-hmC-related subnetworks and their associations, the predictive models based on 5-hmC are highly practical for cfDNAs due to the fact that they are detectable using a non-invasive method. Such predictive models could be used in precision medicine as well.

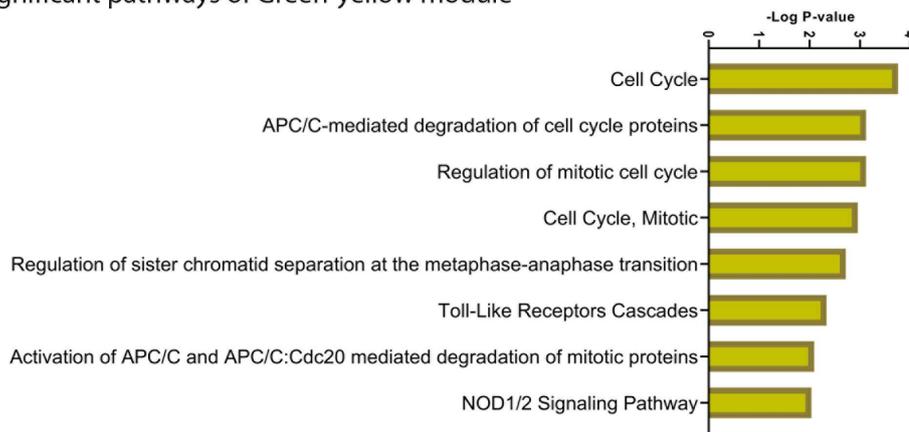
In conclusion, not only are the 5-hmC-related subnetworks, their biological pathways, and the association among them highly efficient for precision medicine but also provide biological insights about 5-hydroxymethylcytosine. Moreover, the predictive models for non-invasive cancer detection methods with a few features are valuable in the clinic.

References

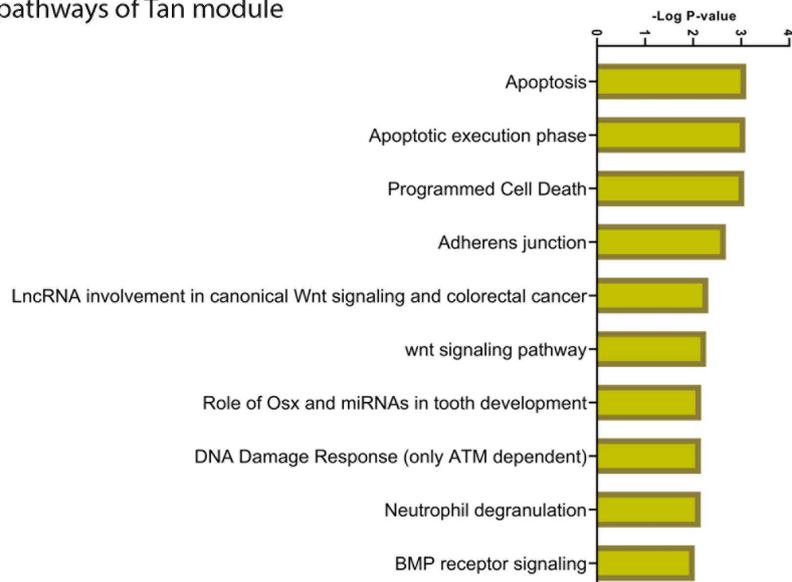
1. Huang, Z.-B., et al., *Cell-free DNA as a liquid biopsy for early detection of gastric cancer*. 2021. **21**(1): p. 1-1.
2. De Mattos-Arruda, L., D. Olmos, and J.J.F.O. Taberero, *Prognostic and predictive roles for circulating biomarkers in gastrointestinal cancer*. 2011. **7**(12): p. 1385-1397.
3. Shahjehan, F., S. Kamatham, and P.M.J.F.i.o. Kasi, *Role of circulating tumor dna in gastrointestinal cancers: update from abstracts and sessions at ASCO 2018*. 2019. **9**: p. 358.
4. Song, C.-X., et al., *5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages*. 2017. **27**(10): p. 1231-1242.
5. Zhang, J. and K.J.B.G. Huang, *Pan-cancer analysis of frequent DNA co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers*. 2017. **18**(1): p. 1-14.
6. Pfeifer, G.P., et al., *The role of 5-hydroxymethylcytosine in human cancer*. 2014. **356**(3): p. 631-641.
7. Vasanthakumar, A. and L.A.J.C.g. Godley, *5-hydroxymethylcytosine in cancer: significance in diagnosis and therapy*. 2015. **208**(5): p. 167-177.
8. Pfeifer, G.P., et al., *5-hydroxymethylcytosine and its potential roles in development and cancer*. 2013. **6**(1): p. 1-9.
9. Zhang, Y., et al., *Pan-cancer circulating tumor DNA detection in over 10,000 Chinese patients*. 2021. **12**(1): p. 1-14.
10. Li, W., et al., *5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers*. *Cell research*, 2017. **27**(10): p. 1243-1257.
11. Vasanthakumar, A. and L.A. Godley, *5-hydroxymethylcytosine in cancer: significance in diagnosis and therapy*. *Cancer genetics*, 2015. **208**(5): p. 167-177.
12. Poole, C.J., et al., *MYC deregulates TET1 and TET2 expression to control global DNA (hydroxy) methylation and gene expression to maintain a neoplastic phenotype in T-ALL*. *Epigenetics & chromatin*, 2019. **12**(1): p. 1-20.
13. Alenzi, F.Q., *Links between apoptosis, proliferation and the cell cycle*. *British journal of biomedical science*, 2004. **61**(2): p. 99-102.
14. Heim, V.J., C.A. Stafford, and U. Nachbur, *NOD signaling and cell death*. *Frontiers in cell and developmental biology*, 2019. **7**: p. 208.
15. Liu, L., et al., *Characterization of Cell Free Plasma Methyl-DNA From Xenografted Tumors to Guide the Selection of Diagnostic Markers for Early-Stage Cancers*. *Frontiers in Oncology*, 2021. **11**: p. 503.
16. Zeng, C., et al., *Towards precision medicine: advances in 5-hydroxymethylcytosine cancer biomarker discovery in liquid biopsy*. *Cancer Communications*, 2019. **39**(1): p. 1-9.
17. Walker, C., E. Mojares, and A. del Río Hernández, *Role of extracellular matrix in development and cancer progression*. *International journal of molecular sciences*, 2018. **19**(10): p. 3028.
18. Pickup, M.W., J.K. Mouw, and V.M. Weaver, *The extracellular matrix modulates the hallmarks of cancer*. *EMBO reports*, 2014. **15**(12): p. 1243-1253.

19. Guler, G.D., et al., *Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA*. Nature communications, 2020. **11**(1): p. 1-12.
20. Tian, X., et al., *Circulating tumor DNA 5-hydroxymethylcytosine as a novel diagnostic biomarker for esophageal cancer*. Cell research, 2018. **28**(5): p. 597-600.
21. Han, D., et al., *A highly sensitive and robust method for genome-wide 5hmC profiling of rare cell populations*. Molecular cell, 2016. **63**(4): p. 711-719.
22. Krushkal, J., et al., *Concerted changes in transcriptional regulation of genes involved in DNA methylation, demethylation, and folate-mediated one-carbon metabolism pathways in the NCI-60 cancer cell line panel in response to cancer drug treatment*. Clinical epigenetics, 2016. **8**(1): p. 1-24.
23. Lukashev, M.E. and Z. Werb, *ECM signalling: orchestrating cell behaviour and misbehaviour*. Trends in cell biology, 1998. **8**(11): p. 437-441.
24. Boudreau, N., et al., *Suppression of ICE and apoptosis in mammary epithelial cells by extracellular matrix*. Science, 1995. **267**(5199): p. 891-893.
25. Morris, V.K. and J.H. Strickler, *Use of Circulating Cell-Free DNA to Guide Precision Medicine in Patients with Colorectal Cancer*. Annual Review of Medicine, 2021. **72**: p. 399-413.

(a) Significant pathways of Green-yellow module



(b) Significant pathways of Tan module



(c) Significant pathways of Black module

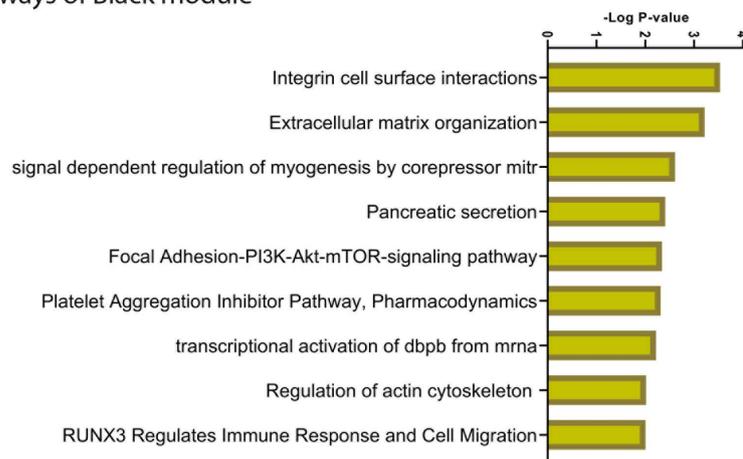


Figure 1 Biological pathways related to three significant for the 5hmC gastrointestinal pan-cancer subnetworks.

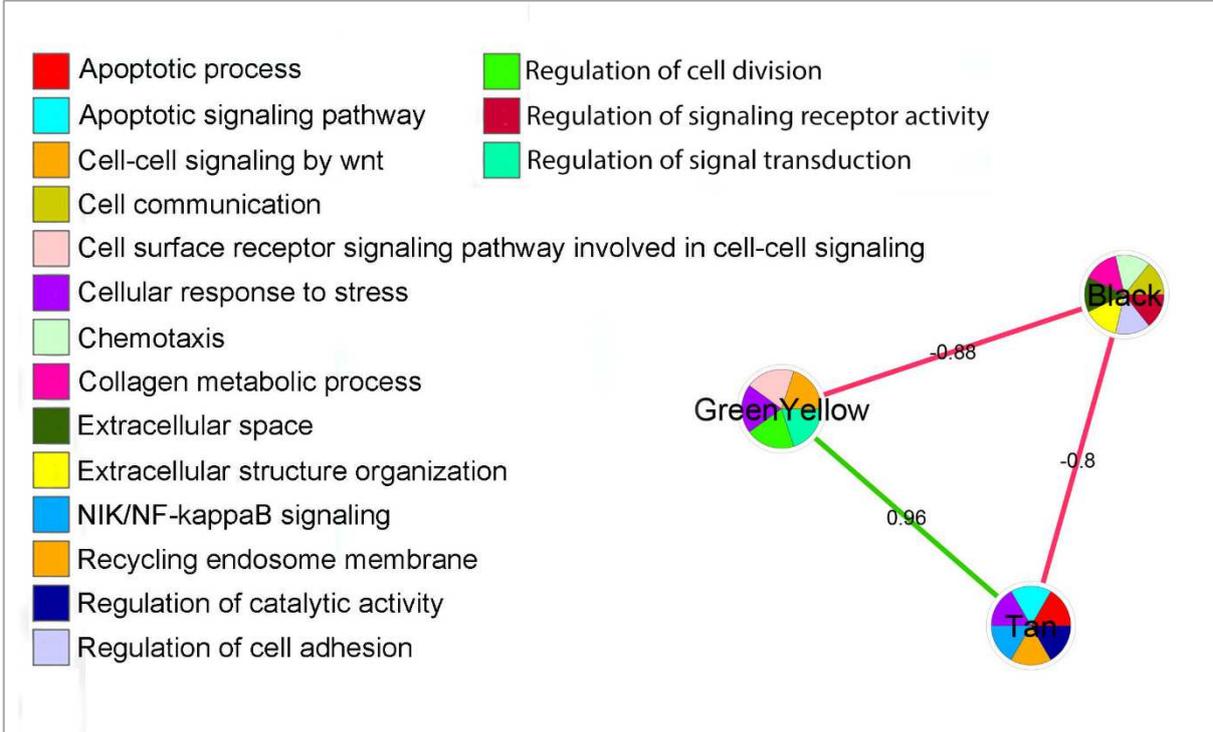
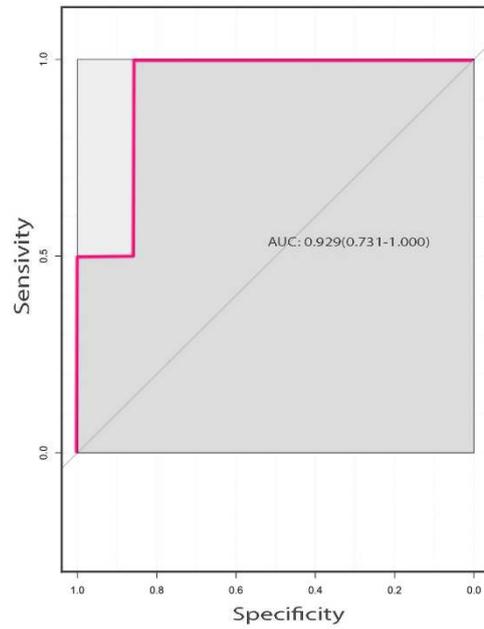


Figure 2 the inverse/direct interplay among the 5hmC gastrointestinal pan-cancer subnetworks

a) AUC for predictive model



b) Heatmap of GreenYellow module

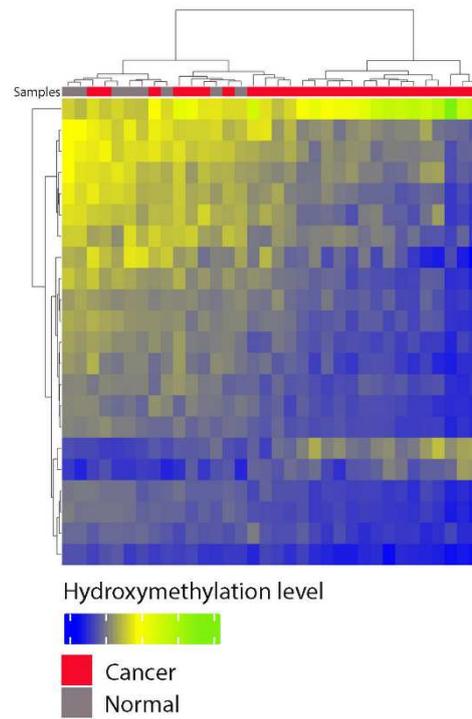


Figure 3 Distinguishing power of the 5hmC gastrointestinal pan-cancer subnetworks.

a) The Area Under The Curve (AUC) GreenYellow subnetwork.

b) The heatmap of the GreenYellow subnetwork.

Supplementary Files

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

- [supplementary.docx](#)