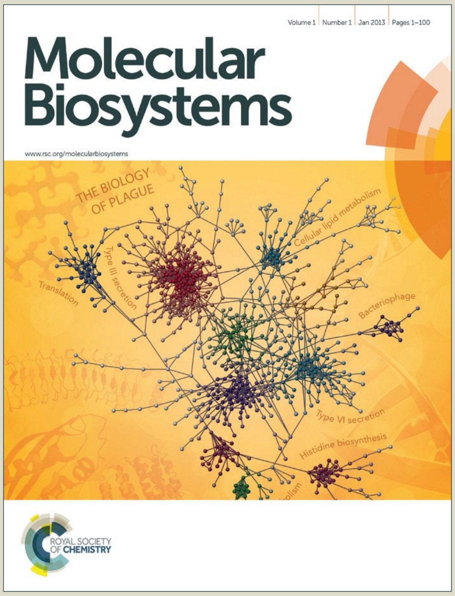


Molecular BioSystems

PAPER

Cell cycle related genes up-regulated in human colorectal development predict the overall survival of late-stage colorectal cancer patients†

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Ning An,[‡a] Xue Yang,[‡a] Yueming Zhang,[‡b] Xiaoyu Shi,[a] Xuexin Yu,[c] Shujun Cheng,[*a] Kaitai Zhang[*a] and Guiqi Wang[*b]

A tumor can be perceived as a special ''organ'' that undergoes aberrant and poorly regulated organogenesis. Embryonic development and carcinogenesis share striking similarities in their cellular behavior and underlying molecular mechanisms. This intimate association makes embryonic development a viable reference model for studying cancer thereby circumventing the potentially misleading complexity of tumor heterogeneity. Therefore, on the basis of global expression profile, the genes simultaneously activated (up-regulated in terms of expression profile) or suppressed (down-regulated) in both embryonic development and cancer stage, probably contain profound information upon the molecular mechanism of cancer. In this study, the Affymetrix expression profile of 1593 colorectal cancer samples was downloaded from Gene Expression Omnibus. The 1396 differentially expressed probes were robustly obtained using 660 colorectal normal and cancer samples, of which the expression pattern was analyzed in our human colorectal developmental data. All these 1396 probes were classified into 27 distinct patterns based on their expression patterns during the developmental process. By means of gene set enrichment analysis, we collected 393 V probes simultaneously up-regulated in both development and carcinogenesis and 207 A probes down-regulated in both. Functional enrichment analysis indicated that V probes were significantly related to cell cycle regulation. Notably, 28 cell-cycle related probes within V probe group were found to be significantly associated with overall survival of Stage III/IV patients (GSE17536 cross validation, n=96, *p*=5.70e-03; GSE29621, n=36, *p*=1.70e-03; GSE39084, n=38, *p*=0.05; GSE39582, n=264, *p*=0.047; GSE17537, n=36, *p*=5.90e-03).

Introduction

Although remarkable progress has been made, understanding the intricate molecular mechanism of colorectal cancer (CRC) was enormously hindered by tumor heterogeneity[1]. Therefore, some novel model similar with cancer in terms of cell-behavioral and molecular attributes, but intrinsically more organized, is urgently needed.

Emerging studies reported the cellular behavioral similarity between ontogenesis and carcinogenesis, for instance, in the process of epithelial-to-mesenchymal transition (EMT)[2], mesenchymal-to-epithelial transition (MET)[3] and immune-surveillance evasion[4]. The molecular resemblances have been documented between certain malignant tumors and developing tissues on the basis of transcription factor activity[5], regulation of chromatin structure[6] and cellular signaling[7]. Important molecules were reported to play substantial roles in both embryonic development and carcinogenesis: *Ptch1* is a key regulator of embryonic development, whose overexpression could drive skin carcinogenesis[8]. Developmental animal models were used to uncover the complicated molecular mechanisms of carcinogenesis, and a variety of novel and pivotal molecules, pathways and biomarkers were discovered[9-11]. For instance, *Notch1*-signaling pathway, greatly activated during development, is proven to be reactivated in the process of carcinogenesis[12, 13]. In addition, there were some pioneering works discovering that mRNA and microRNA expression profile of cancer could recapitulate the expression pattern of embryonic development samples[10, 14-17]. Based on aforementioned abundant evidences, it is sensibly convincing that a tumor can be viewed as an aberrant organ which acquired the capacity for indefinite proliferation through accumulated strikes, and if it is so, then the principles of embryonic development could be exquisitely explored in order to scoop up exclusive information about cancer[18]. Thus, genes activated (up-regulated comparing to normal tissue in respect to expression profile) or suppressed (down-regulated) simultaneously in both development and cancer stage, probably hold meaningful explanation for the underlying mechanisms of carcinogenesis, and might be intimately associated with clinicopathological parameters.

[a] State Key Laboratory of Molecular Oncology, Department of Etiology and Carcinogenesis, Peking Union Medical College & Cancer Institute (Hospital), Chinese Academy of Medical Sciences, Beijing, China. Email: chengshj@263.net.cn, zhangkt_bingyin@sina.cn.
[b] Department of Endoscopy, Cancer Hospital, Chinese Academy of Medical Sciences, Beijing, China. Email: wangguiq@126.com.
[c] College of Bioinformatics Science and Technology, Harbin Medical University, China.

† Electronic supplementary information (ESI) available.
‡ These authors contributed equally to this work.

Table 1 Colorectal cancer microarray datasets included in the study.

Characteristics	GSE17536	GSE17537	GSE39582	GSE29621	GSE39084
Number	177	55	566	65	70
Year	2009	2009	2013	2014	2014
Country	American	American	France	American	France
Gender					
Male	96	26	310	40	35
Female	81	29	256	25	35
Age					
Mean±SD (years)	65.5±13.1	62.3±14.4	63.0±19.0	NR	59.2±18.3
T status					
T1+T2	NR	NR	57	8	13
T3+T4	NR	NR	486	57	57
N status					
N0	NR	NR	302	32	35
N1	NR	NR	134	25	20
N2	NR	NR	104	7	15
M status					
M0	NR	NR	482	46	48
M1	NR	NR	61	18	22
AJCC stage					
Stage I+II	81	19	297	29	31
Stage III+IV	96	36	265	36	38
Pathologic grade					
G1	16	8	NR	4	NR
G2	134	25	NR	51	NR
G3	27	3	NR	10	NR
AdjCTX					
Yes	NR	NR	233	38	NR
No	NR	NR	316	27	NR

Note: *SD*, standard deviation; *AdjCTX*, whether adjuvant chemotherapy was used; *NR*, not reported.

However, there are conspicuous limitations of above-mentioned researches concerning the relation between embryonic development and carcinogenesis. First, all of these researches used mouse model to simulate the developing process of human being. However, because humans and rodents diverge from each other more than 70 million years ago on the basis of evolutionary history[19], the structural and molecular differences between mice and human are surely not negligible, and the time-point to time-point projection from mouse developing time axis to human's cannot be exactly dovetailed. For example, the regulatory mechanism of human pre-implantation development is not completely the same as mice based on expression profile[20]. Secondly, there are some studies about the transcriptomic analysis of early human embryo[21, 22], however, the expression profile of specific developing organ was not available, and the transcriptomic comparison between embryonic development and cancer was not investigated, until our team's study upon the expression profile of human lung developmental and adenocarcinoma tissues[23].

Our research team dedicated ourselves in discovering the association between embryonic development and carcinogenesis, and the transcriptome of rhesus macaques developing organs were constructed in the previous studies since rhesus macaques are more genetically related to human than mice[24, 25]. In this study, we used meta-analysis to find robust differentially expressed probes (DEPs) in 8 public Affymetrix microarray data sets, and the global expression profile of human embryonic colorectal tissues was then constructed. In accordance with the principles of gastrointestinal developmental biology[26], the appearance of a primitive stratified epithelium was established in from about 8 to 10 postovulatory weeks (PWs); the conversion of this epithelium to a villus architecture with developing crypts was completed after 14 PWs. Therefore, the expression profile of human colorectal development was composed of four critical time points: whole embryos (Bud, the start point of any organ) at 3 to 5 PWs, early embryonic colons (EarlyColon, establishment of primitive stratified epithelium) at 8 to 10 PWs and middle embryonic colons (MiddleColon, completion of epithelium conversion) at 14 to 22 PWs, and normal adult colorectal mucosal tissue (end point of human colorectal development). In the transition from development to normal tissue to cancer, we collected Affymetrix probes with the expression pattern shaped like "V" (termed as "V probes", up-regulated in both development and cancer tissues) or "A" (termed as "A probes", down-regulated in both) according to their expression pattern. Their association with overall survival (OS) was thoroughly investigated in 5 independent Affymetrix microarray data sets.

Results

DEP collection and its transcriptomic features in colorectal development data

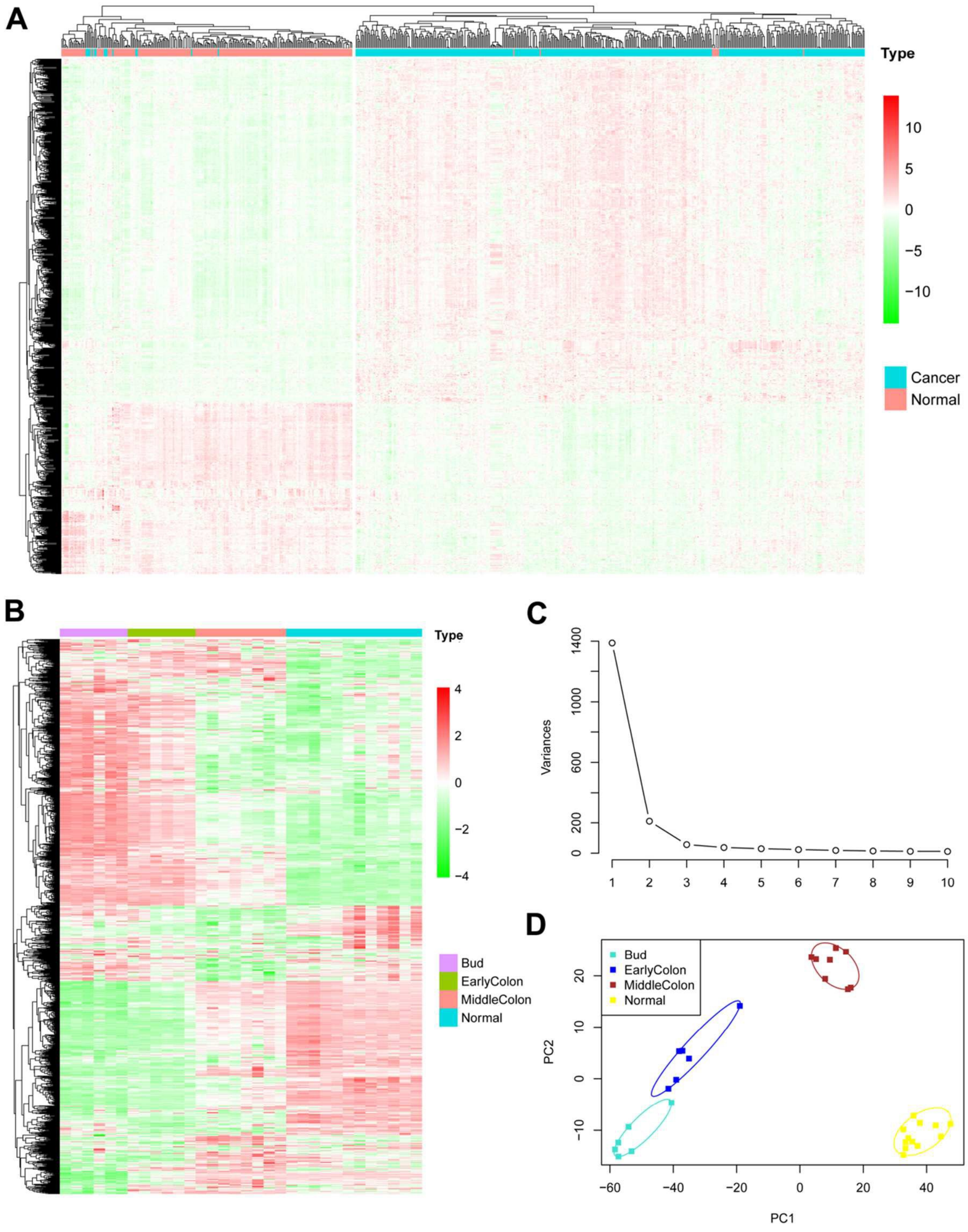


Fig. 1 (A) Heatmap of 1396 DEPs in NC superset. Rows represented DEPs, and columns represented CRC patients. UCA was conducted to cluster samples and probes. (B) Heatmap of 1123 DEGs (mapped from 1274 DEPs) in our human colorectal development samples. UCA was used to cluster DEGs with similar expression pattern during colorectal development. (C) Screeplot of the PCA. PCA was performed with 1123 DEGs in 4 sequential stages of human colorectal development, and the screeplot showed the variances against the number of the principal component. (D) PCA plot of human colorectal development samples with 1123 DEGs. The colorectal developmental samples clustered tightly within each developmental stage, whereas the different stages were distinctly separate. Note: UCA represents unsupervised clustering algorithm; PCA represents principle component analysis.

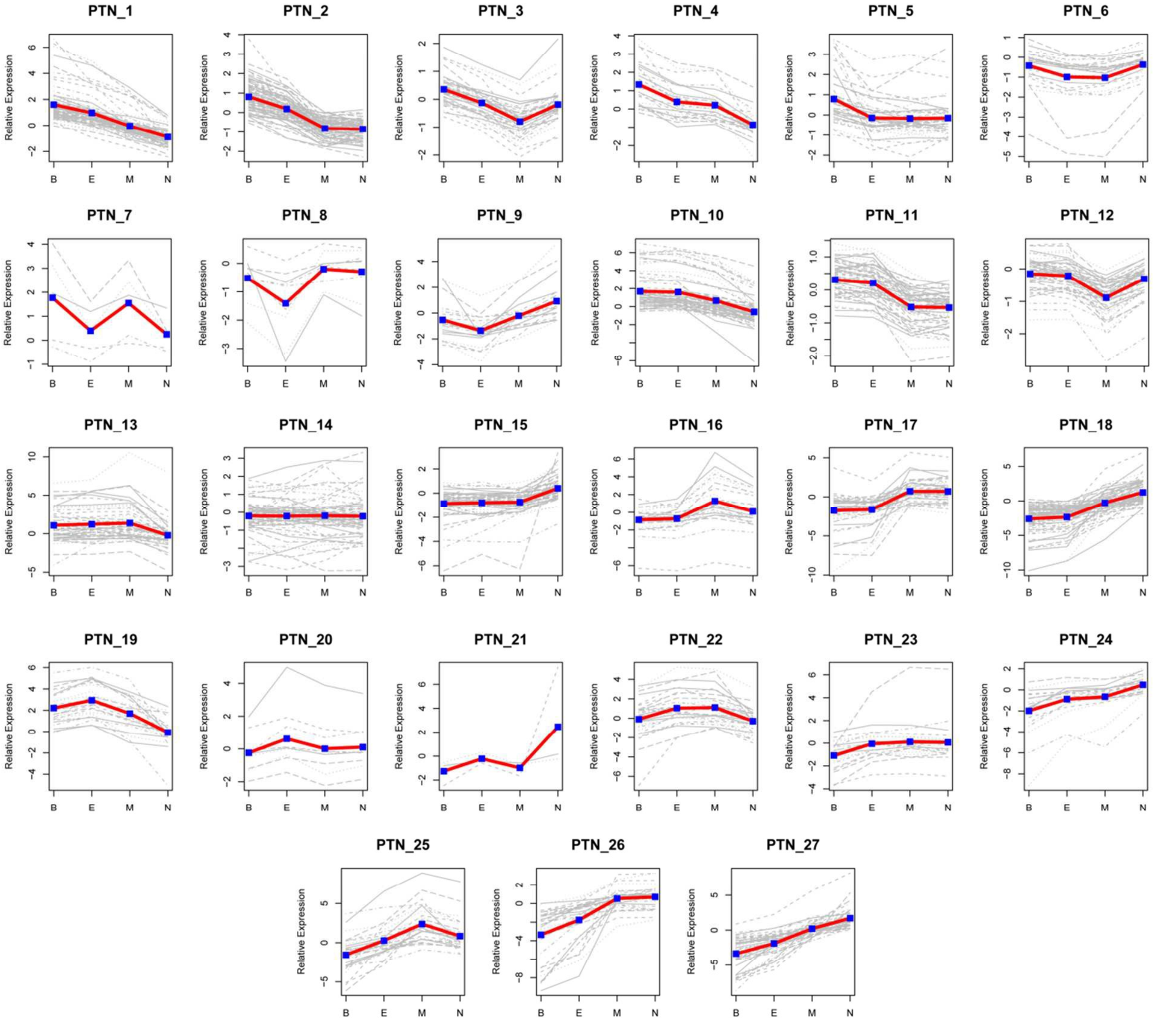


Fig. 2 Clustering 1274 DEPs into 27 modules based on their expression pattern during colorectal development. The time points in the transition from Bud (referred to as B) to EarlyColon (referred to as E) to MiddleColon (referred to as M) to Normal (referred to as N) stages during development were plotted on the *x*-axis, and the normalized gene expression level in every module was plotted on the *y*-axis. Note: each gene is depicted with a grey dot line, and general pattern in each module was highlighted with red line and blue points, calculated by averaging gene expression in each time point.

Through manual searching of online literatures, 13 Affymetrix datasets were obtained from Gene Expression Omnibus (GEO) database (Fig. S1, ESI†). We used NC superset (8 independent data sets containing colorectal cancer and normal samples, n=660; Table S1) to collect 1396 robust DEPs, including 887 up-regulated DEPs and 509 down-regulated DEPs (Fig. 1A). In our colorectal development data, 1123 differentially expressed genes (DEGs) could be mapped from 1274 DEPs (the genes mapped from the remaining 122 DEPs did not exist in development data). It is visually perceivable that DEGs hold great variance across the samples within distinct developmental stages (Fig. 1B). In addition, principle component analysis (PCA) of 1123 DEGs in development data indicated that the transcriptomic features of colorectal ontogenesis are arranged in a sequential order according to the time axis of development, and the trajectory could be recapitulated by genes which were differentially expressed during carcinogenesis (Fig. 1C-D). Samples clustered tightly within each developmental stage, whereas the different stages were distinctly separated. Thus, the DEGs dysregulated in cancer varied considerably during embryonic development, and could recapitulate the developing trajectory of human colorectal development based on PCA analysis, implying the association between the two processes in terms of gene expression profile.

V and A differential pattern existed in human CRC scenario in the transition from development to normal to cancer

All 1274 DEPs (the DEPs which could be mapped to aforementioned 1123 DEGs) were further divided into 27

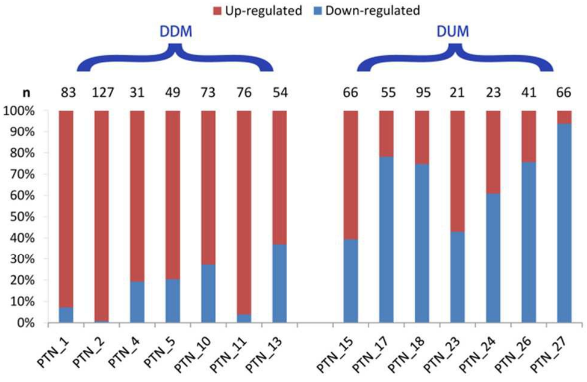


Fig. 3 Percentage analysis of 7 DDMs and 7 DUMs. Module number was shown in *x* axis and corresponding the number of DEPs was shown above each module bar. Note: dark red bar represented the percentage up-regulated in cancer, and blue bar represented down-regulated section.

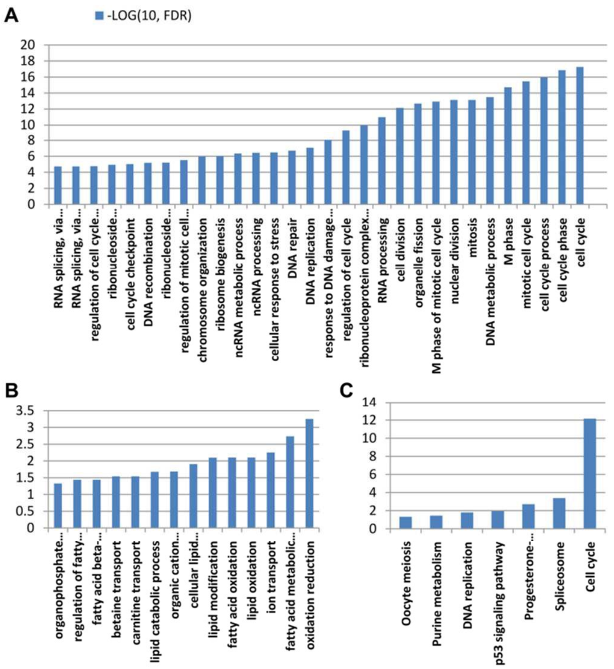


Fig. 4 (A) GO analysis in terms of biological process of V probes, and the bar represents the –log10 transformed FDR value. (B) GO analysis in terms of biological process of A probes. (C) KEGG pathway enrichment analysis of V probes.

modules (PTN_1 ~ PTN_27) based on their distinct expression pattern between successive time points during colorectal development (Fig. 2), and 7 developmental down-regulating modules (DDMs; gradually down-regulating along the developmental time axis, i.e. up-regulated in development samples in comparison to normal samples) and 7 development up-regulating modules (DUMs, containing the probes with expression pattern opposite to DDMs) were collected. The number of DEPs in each DDM or DUM varied within the range between 21 (PTN_23) and 127 (PTN_2). Among DEPs in each DDM or DUM, the percentages of cancer up-regulated DEPs and down-regulated DEPs were calculated, respectively. Percentage analysis indicated DEPs up-regulated in cancer were inclined to be contained in DDMs (up-regulated in development stage), while DEPs down-regulated in cancer tended to be within DUMs (Fig. 3). Gene set enrichment analysis (GSEA) statistically confirmed our former speculation in NC superset (Fig. S2-15, ESI†). In order to exclude a minority of exceptions, DDMs and DUMs which were not significantly enriched in NC superset were discarded from further analysis [false discovery rate (FDR) > 0.001]. Therefore, significant DDMs were composed of PTN_1, PTN_2, PTN_4, PTN_5, PTN_10 and PTN_11; significant DUMs included PTN_17, PTN_18, PTN_26, and PTN_27 (Fig. S2-15, ESI†). Therefore, V probes were defined as the 393 probes (contained in 6 significant DDMs) up-regulated in cancer, and A probes were 207 probes (contained in 4 significant DUMs) down-regulated in cancer. In this way, probes both elevated (V probes) and decreased (A probes) in colorectal embryonic development and carcinogenesis were identified and used for further analysis.

Cell cycle probes differentially expressed in cancer were greatly enriched in V probe group

Gene ontology (GO) biological process and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of V probes and A probes were carried out via the DAVID bioinformatics tool (http://david.abcc.ncifcrf.gov, Fig. 4A-C). GO analysis indicated the probes in V group were greatly related to cell cycle (the first rank, FDR = 5.54e-18), mitosis, cell division, and DNA repair (Fig. 4A). A very small number of GO terms were weakly enriched with A probes (Fig. 4B), including oxidation reduction, fatty acid metabolic process and ion transport, and the FDR values were very close to the verge of insignificance (with the criterion FDR=0.05). KEGG analysis of V probes (Fig. 4C) showed pathway cell cycle ("hsa04110") was most significantly enriched (FDR=6.94e-13), far ahead from the pathway in the second rank ("splicesome", FDR=0.042), while there was no pathway significantly enriched with A probes. The result of enrichment analysis indicated that V probes, instead of A probes, showed enormous functional concentration upon cell cycle regulation, suggesting prognostic probes might be identified within V probe group.

V probes and cell cycle probes in V group were intimately associated with clinicopathological variables

We downloaded 124 genes involved in "Cell Cycle" pathway in KEGG database. There were 23 genes in cell cycle pathway which could be mapped from 28 V probes (termed as V cycle probes). Logistic regression analysis was applied to assess the correlations between 3 probe groups (A probes, V probes, and V cycle probes) and 8 clinicopathological variables. As show in Table S2 (ESI†), A probes showed no significant association with any of the 8 variables (*p*<0.01). However, V probes and V cycle probes both showed significant clinical association with American Joint Committee on Cancer (AJCC) stage, AdjCTX (whether adjuvant chemotherapy was used), pathological

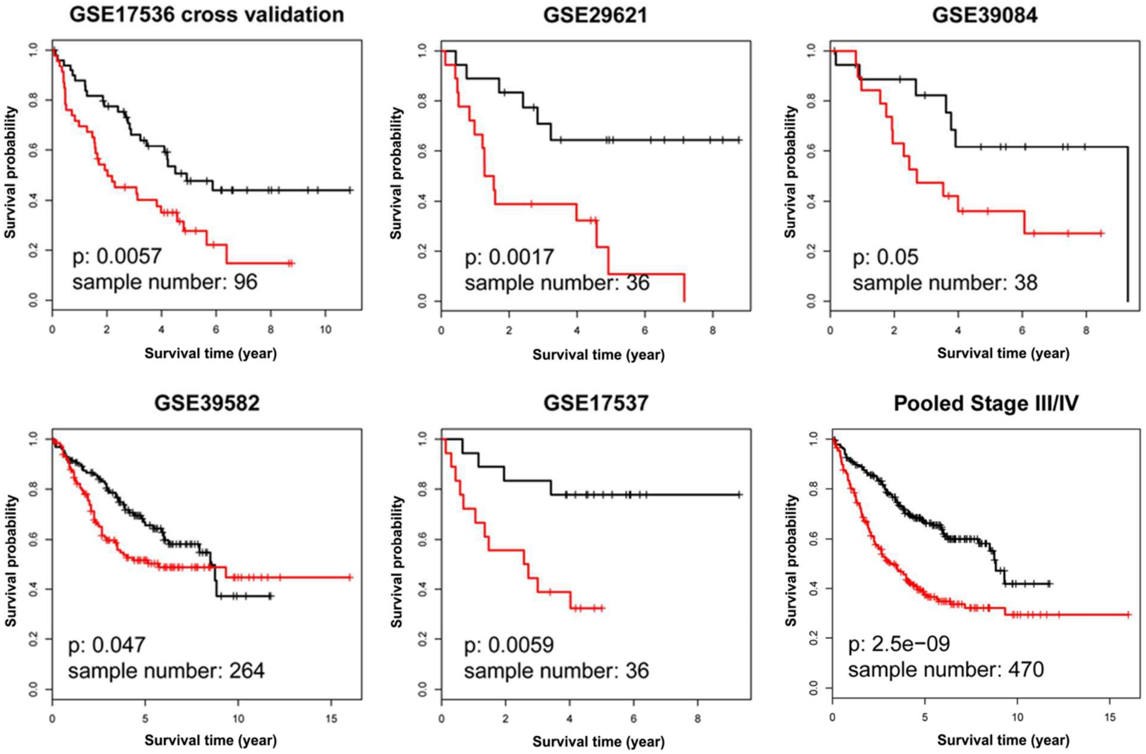
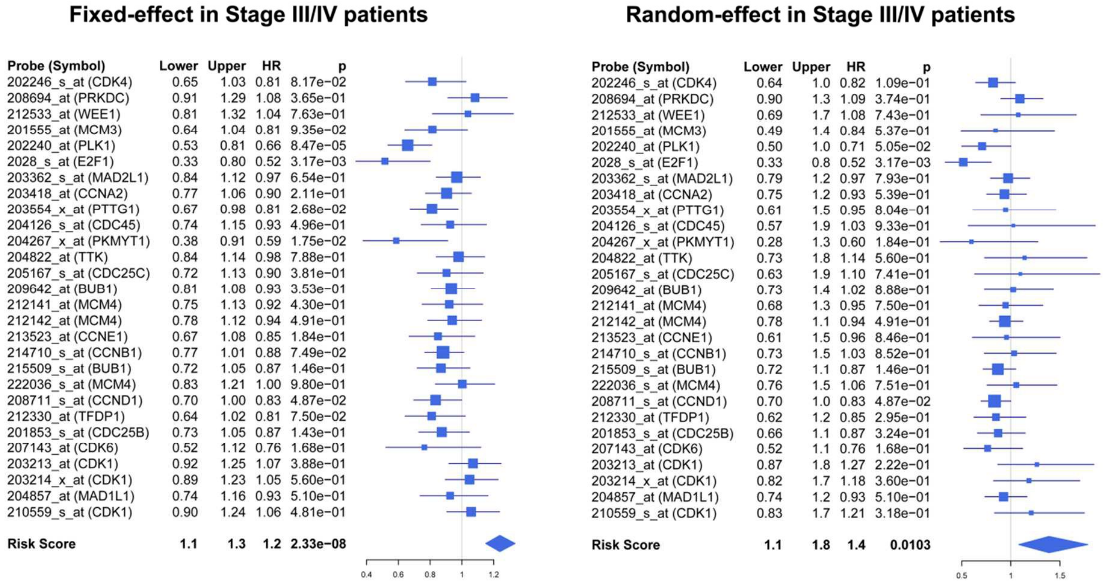


Fig. 5 Kaplan–Meier survival analysis of 28 V cycle probes with Stage III/IV patients in 5 independent data sets of Clinicinfo superset. Survival analysis was performed to discriminate OS between risk score assigned groups in training cohort with 10-fold cross validation, remaining 4 testing cohorts, and pooled samples. Note: in Kaplan–Meier survival analysis, red curve represents the subgroup with higher risk score, and black curve represents lower risk score.


Fixed-effect in Stage III/IV patients

Probe (Symbol)	Lower	Upper	HR	p
202246_s_at (CDK4)	0.65	1.03	0.81	8.17e−02
208694_at (PRKDC)	0.91	1.29	1.08	3.65e−01
212533_at (WEE1)	0.81	1.32	1.04	7.63e−01
201555_at (MCM3)	0.64	1.04	0.81	9.35e−02
202240_at (PLK1)	0.53	0.81	0.66	8.47e−05
2028_s_at (E2F1)	0.33	0.80	0.52	3.17e−03
203362_s_at (MAD2L1)	0.84	1.12	0.97	6.54e−01
203418_at (CCNA2)	0.77	1.06	0.90	2.11e−01
203554_x_at (PTTG1)	0.67	0.98	0.81	2.68e−02
204126_s_at (CDC45)	0.74	1.15	0.93	4.96e−01
204267_x_at (PKMYT1)	0.38	0.91	0.59	1.75e−02
204822_at (TTK)	0.84	1.14	0.98	7.88e−01
205167_s_at (CDC25C)	0.72	1.13	0.90	3.81e−01
209642_at (BUB1)	0.81	1.08	0.93	3.53e−01
212141_at (MCM4)	0.75	1.13	0.92	4.30e−01
212142_at (MCM4)	0.78	1.12	0.94	4.91e−01
213523_at (CCNE1)	0.67	1.08	0.85	1.84e−01
214710_s_at (CCNB1)	0.77	1.01	0.88	7.49e−02
215509_s_at (BUB1)	0.72	1.05	0.87	1.46e−01
222036_s_at (MCM4)	0.83	1.21	1.00	9.80e−01
208711_s_at (CCND1)	0.70	1.00	0.83	4.87e−02
212330_at (TFDP1)	0.64	1.02	0.81	7.50e−02
201853_s_at (CDC25B)	0.73	1.05	0.87	1.43e−01
207143_at (CDK6)	0.52	1.12	0.76	1.68e−01
203213_at (CDK1)	0.92	1.25	1.07	3.88e−01
203214_x_at (CDK1)	0.89	1.23	1.05	5.60e−01
204857_at (MAD1L1)	0.74	1.16	0.93	5.10e−01
210559_s_at (CDK1)	0.90	1.24	1.06	4.81e−01
Risk Score	**1.1**	**1.3**	**1.2**	**2.33e−08**

Random-effect in Stage III/IV patients

Probe (Symbol)	Lower	Upper	HR	p
202246_s_at (CDK4)	0.64	1.0	0.82	1.09e−01
208694_at (PRKDC)	0.90	1.3	1.09	3.74e−01
212533_at (WEE1)	0.69	1.7	1.08	7.43e−01
201555_at (MCM3)	0.49	1.4	0.84	5.37e−01
202240_at (PLK1)	0.50	1.0	0.71	5.05e−02
2028_s_at (E2F1)	0.33	0.8	0.52	3.17e−03
203362_s_at (MAD2L1)	0.79	1.2	0.97	7.93e−01
203418_at (CCNA2)	0.75	1.2	0.93	5.39e−01
203554_x_at (PTTG1)	0.61	1.5	0.95	8.04e−01
204126_s_at (CDC45)	0.57	1.9	1.03	9.33e−01
204267_x_at (PKMYT1)	0.28	1.3	0.60	1.84e−01
204822_at (TTK)	0.73	1.8	1.14	5.60e−01
205167_s_at (CDC25C)	0.63	1.9	1.10	7.41e−01
209642_at (BUB1)	0.73	1.4	1.02	8.88e−01
212141_at (MCM4)	0.68	1.3	0.95	7.50e−01
212142_at (MCM4)	0.78	1.1	0.94	4.91e−01
213523_at (CCNE1)	0.61	1.5	0.96	8.46e−01
214710_s_at (CCNB1)	0.73	1.5	1.03	8.52e−01
215509_s_at (BUB1)	0.72	1.1	0.87	1.46e−01
222036_s_at (MCM4)	0.76	1.5	1.06	7.51e−01
208711_s_at (CCND1)	0.70	1.0	0.83	4.87e−02
212330_at (TFDP1)	0.62	1.2	0.85	2.95e−01
201853_s_at (CDC25B)	0.66	1.1	0.87	3.24e−01
207143_at (CDK6)	0.52	1.1	0.76	1.68e−01
203213_at (CDK1)	0.87	1.8	1.27	2.22e−01
203214_x_at (CDK1)	0.82	1.7	1.18	3.60e−01
204857_at (MAD1L1)	0.74	1.2	0.93	5.10e−01
210559_s_at (CDK1)	0.83	1.7	1.21	3.18e−01
Risk Score	**1.1**	**1.8**	**1.4**	**0.0103**

Fig. 6 Forest plot of 28 V cycle probes with fixed-effect and random-effect model in Stage III/IV patients. Meta-analysis of 28 V cycle probes in 5 independent data set of Clinicinfo superset was conducted, and HR, 95% CI, and corresponding *p* value of each probe and risk score were calculated and plotted in the forest plot for Stage III/IV samples. Note: HR represents hazard ratio; CI represents confidence interval.

Table 2 Univariate and multivariate analyses of overall survival (Cox regression model) in 121 Stage III/IV patients.

Factors	Univariate Cox regression		Multivariate Cox regression	
	HR (95% *CI*)	***P***	**HR (95% *CI*)**	***P***
Age	1.021 (1.002~1.041)	**0.030**	1.015 (0.994~1.038)	0.166
Gender (Male/Female)	1.101 (0.680~1.782)	0.696	-	-
Stage (IV/III)	3.688 (2.244~6.063)	**2.091e-07**	3.579 (2.109~6.072)	**2.285e-06**
Grade (III/II+I)	1.858 (1.058~3.265)	**0.042**	2.300 (1.280~4.135)	**0.005**
Risk score	1.311 (1.212~1.419)	**7.245e-08**	1.231 (1.134~1.336)	**6.737e-07**

Note: In Clinicinfo superset, 121 Stage III/IV CRC patients with definite information of age, gender, AJCC stage, pathological grade and OS information were used in univariate and multivariate Cox regression analyses to evaluate the independence of 28 V cycle probes. Significant *p* values were in bold (*p*<0.05).

grade, tumor size and lymph node invasion, whereas V cycle probes were also significantly related to distant metastasis. Notably, the strong association between V cycle probes and clinicopathological variables indicated that these 28 probes could probably monitor the continuous deterioration of CRC, and hold profound prognostic information.

Validation of 28 V cycle probes' prognostic value via Cox regression model

The Clinicinfo superset (containing 5 Affymetrix CRC datasets with OS information, n=933, Table 1) was used to evaluate the OS predicting ability of these 28 V cycle probes. We used GSE17536 as training cohort to train Cox regression model with 28 V cycle probes, and then used the constructed model to evaluate the risk score of patients in test cohorts. Kaplan–Meier survival analysis indicated the risk score calculated in Stage I/II patients was not significantly associated with OS in both self-cross validation and 4 individual test cohorts (Fig. S16, ESI†), while patients with higher risk score in Stage III/IV patient groups tended to live significantly shorter than those with lower risk score (GSE17536 cross validation, n=96, *p*=5.70e-03; GSE29621, n=36, *p*=1.70e-03; GSE39084, n=38, *p*=0.05; GSE39582, n=264, *p*=0.047; GSE17537, n=36, *p*=5.90e-03; Fig. 5). The ability of risk score to discriminate OS was satisfactory in all stage samples except in GSE39582 and GSE17537 (GSE17536 cross validation, n=177, *p*=0.03; GSE29621, n=65, *p*=5.20e-04; GSE39084, n=70, *p*=0.011; GSE39582, n=562, *p*=0.99; GSE17537, n=55, *p*=0.22; Fig. S17, ESI†), which is probably caused by the disturbance of early-stage portion. We then pooled all the samples whose OS information were available within Clinicinfo superset (n=929), and survival analysis (Fig. 5; Fig. S16-17, ESI†) also implied that the expression of V cycle probes could predict OS in Stage III/IV samples (n=470, *p*=2.50e-09, Fig. 5), rather than Stage I/II samples (n=454, *p*=0.095, Fig. S16, ESI†). Meta-analysis of 28 V cycle probes and risk score in 5 Clinicinfo datasets also confirmed the result of survival analysis with both fixed-effect model and random-effect model (Fig. 6; Fig. S18-19, ESI†). Additionally, we collected 121 Stage III/IV patients in Clinicinfo superset with definite information of OS, age, gender, stage and grade to evaluate the independence of the prognostic factors with Cox regression analysis (Table 2). The result indicated that the risk score was an independent prognostic factor for Stage III/IV patients [hazard ratio (*HR*): 1.231; 95% confidence interval (*CI)*: 1.134~1.336; *p*=6.737e-07]. Based on aforementioned analyses, these 28 V cycle probes were significantly associated with the OS of late-stage (Stage III/IV) CRC patients, rather than early-stage (Stage I/II) patients.

Discussion

In the last decade, high throughput technologies greatly facilitated numerous researches upon cancer etiology. Recent expression profiling datasets are in lack of consistent results between the studies due to different technological platforms and lab protocols[27, 28], which could probably compromise the accuracy and robustness of the whole meta-analysis. In addition, the relatively small number of sample size and noisiness of microarray data might cause the inconsistency of biological conclusions. To address these challenges, we collected 13 microarray data sets (n=1593) from GEO database with 22,277 common probes to discover robust DEPs and their significant clinical relevance.

The reason why so many cellular behaviors and transcriptomic features are shared by embryonic development and carcinogenesis intrigues worldwide heated debates. Generally speaking, there are two kinds of voices. Cancer may regain the power of excessive territorial expansion, migration and invasion via mutational and epigenetic changes, and these properties are also highly characterized during normal developmental stage[29-31]. The second guess is that tumors originate in either tissue stem cells or their immediate progeny through diverging from tightly regulated normal developing pathway, and thereby tumors possess characteristics shared with embryonic cells[32]. The existence of cancer stem cell has been seemingly proven, especially in hematopoietic and colorectal system[33, 34]. Although the definite reason is still unclear, what we do know is that the expression of certain pivotal genes shows synchronized differentiation pattern in embryonic development and carcinogenesis. For example, the activity of *ENAH*, a very important molecule in breast cancer transformation and invasiveness, decreases during mammary gland development, yet reloaded in breast tumors[35]. *VICKZ* was thought to be essential to generate and stabilize the transformed phenotype. Its expression disappears from virtually all tissues soon after birth; however, it is expressed or amplified in at least 12 different kinds of cancers[36]. Therefore, the genes up-regulated or down-regulated simultaneously in both cancer and development tissues are probably functionally core genes promoting the process of tumor formation.

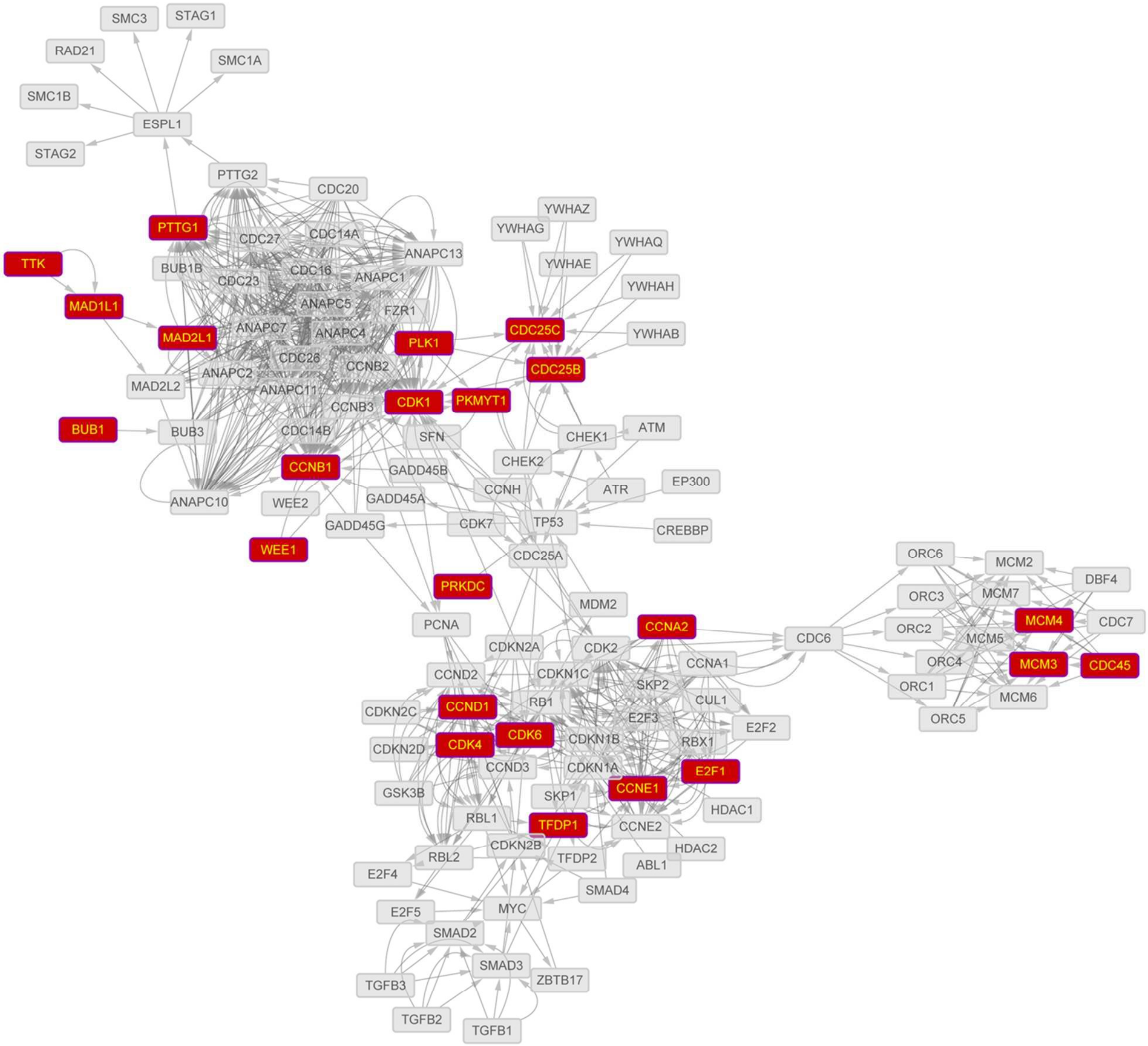


Fig. 7 KEGG cell cycle pathway network. Gene-to-gene regulatory connections of 124 genes involved in "Cell Cycle" pathway were retrieved from KEGG. Note: among these 124 cell cycle genes, 23 genes could be mapped from 28 V probes, which were highlighted in dark red.

In this study, we were aiming to identify Affymetrix expression probes significantly associated with OS, instead of summarized genes. The probe is the most direct unit to measure gene expression value (acting as an expression ruler); however, the summarization process might introduce certain signal noise due to the different transcription level of different probes. Therefore, measuring the expression level of specific prognostic probes is surely a more direct and accurate way to facilitate future clinical implementation, and this strategy was also adopted in previous studies[37, 38]. The limitation of this study is that the human development data was generated with Agilent platform in our previous investigation[39], thus mapping DEPs into 27 patterns could only be achieved by Agilent probe summarization, since Agilent CRC data is quite limited in GEO database.

Meta-analysis was used to discover 1394 robust DEPs in NC superset. The expression of the genes differentially expressed in CRC could depict the trajectory of human colorectal development, suggesting that the genes important for carcinogenesis might also play substantial roles in embryonic colorectal development. We further investigated the differential direction of expression profiles in development and CRC samples. Notably, in most circumstances, genes up-regulated in cancer were generally being down regulated along the development time axis (up-regulated in developmental state comparing with normal stage); genes down-regulated in cancer stage were generally being up regulated along the development time axis (down-regulated in developmental state), which is concordant with our previous investigation upon lung cancer[23], implying that cancer probably hijacks the programs essential for embryonic development to acquire the power of tumor initiation and promotion.

Among 1396 DEPs, there were 32 DEPs which could be mapped to cell cycle genes, and 28 of 32 DEPs belonged to V probe group. The majority of these cell cycle probes dysregulated in cancer consistently show the same expression pattern in development, indicating bouncing the activity of cell cycle genes back to embryonic development status probably plays a very important role in carcinogenesis. These 28 cell-cycle probes (mapping to 23 cell cycle genes, Fig. 7), closely related to a variety of clinicopathological variables, performed exceedingly satisfactory in predicting CRC Stage III/IV patient's OS.

According to AJCC staging system (7th edition)[40], the lesion of early stage CRC (Stage I/II) is relatively contained with neither lymph node invasion nor distant metastasis; when tumor advance to late stage (Stage III/IV) , the involved area is greatly increased; lymph node is invaded (Stage III/IV), and distant organs might be afflicted via distant metastasis (Stage IV). Because of the small size of tumor involvement, Stage I patients only need to receive radical surgery treatment; Stage II patients also need to undertake radical surgery treatment following adjuvant chemotherapy (patients with risk factors) or not (patients without risk factors), to defuse the peril caused by molecularly chaotic tumors[41]. Stage III patients principally should be treated with neoadjuvant hemoradiation therapy followed by surgery with or without adjuvant chemotherapy, and patients with Stage IV CRC are primarily treated with chemotherapy although a selected group of patients can be cured with metastasectomy[42]. Surgical resection of the primary tumor is not beneficial for most of Stage IV patients[43, 44]. Prognostic genes have the ability to predict patient's survival status, probably by means of exerting influence on or reflecting tumor encroachment in the patients. Suppose the tumor is completely removed from the patient, then the expression of these genes might not continue to precisely predict OS, since the persistent influence of the tumor is terminated along with the tumor excision. On account of the massive tumor involvement and potential metastasis of Stage III/IV CRC, surgical excision in late stage patients might not remove the tumor with extensive molecular dysregulation as completely as in early stage patients. Therefore, these prognostic genes might continue manifesting the interaction between the residual neoplasms (or secondary recurrence) and CRC patients, probably explaining the reason why these 28 V cycle probes were only significantly associated with the OS of late stage CRC patients.

Many researches used feature selection algorithms to generate prognostic gene signatures[45-47]. However, we collected these 28 V cycle probes only based on their expression pattern in embryonic development and cancer samples, and its prognostic value was confirmed in 5 independent microarray analyses. This phenomenon was also reported in our previous study upon lung adenocarcinoma[23]. Our present study indicated that this phenomenon might extensively exist across many types of cancer.

Materials and methods

Data collection, preprocessing, and normalization

The expression profile of human colorectal developmental data was constructed in our previous study[39]. Developing colon was obtained from 20 cases of abortion in Maternal & Child Health Care Hospital of Hai Dian. The samples included whole embryos (Bud) at 3 to 5 PWs, early embryonic colons (EarlyColon) at 8 to 10 PWs and middle embryonic colons (MiddleColon) at 14 to 22 PWs. Twelve normal colorectal mucosae samples were collected from patients with hemorrhoids who received surgical excision in the Department of Colon and Rectal Surgery of Beijing Shi Ji Tan Hospital. Purified RNA samples were labeled and hybridized to Agilent 4*44K Whole Human Genome Oligo Microarrays (G4112F) according to the manufacturer's protocol. Raw data were normalized with the GeneSpring GX software, version 11.5 (Silicon Genetics, USA). A total of 41,091 single probes were obtained according to GeneSpring's default setting. The expression value for a particular gene was determined as the median value of all the probes which could be mapped to this gene. Eventually, the expression values of 18,986 genes were obtained. The raw and processed data have been deposited in GEO database with the series accession number GSE71187.

The raw data for 13 human CRC mRNA microarray studies were downloaded from GEO database. The combined data set contained a total of 1593 samples hybridized to probe sets present on both the Affymetrix HG-U133A (with GEO accession number GPL96) and the HG-U133A Plus2 (GPL570) platform. Eight data sets (NC superset, n=660, including 234 normal samples and 426 cancer samples), with accession numbers GSE20916, GSE21510, GSE22598, GSE23878, GSE24514, GSE32323, GSE37364 and GSE41258, were used for identifying DEPs between colorectal cancer and normal tissues; the remaining 5 data sets (Clinicinfo superset, n=933, 929 samples with clear OS information), with accession numbers GSE39582, GSE17536, GSE29621, GSE39084 and GSE17537, contained OS and 8 other types of clinical information, including age, gender, AJCC stage, pathological grade, tumor size, lymph node invasion, distant metastasis and AdjCTX, which were used for the assessment of the clinical relevance of identified probes. In total, 22,277 probes were common in all data sets, and whose expression values were retrieved via robust multi-array average (RMA) algorithm and further quantile normalized using the "affy" Bioconductor package. The ComBat algorithm was utilized to eliminate potential batch. All clinical information was extracted from the original publications.

Identification of DEPs using meta-analysis

In NC superset, meta-analysis was used to identify DEPs between colorectal cancer and normal tissues. In order to collect DEPs with consistent result, Heterogeneity test was conducted to calculate corresponding *p* value. Inverse variance weighting is used for pooling; fixed-effect estimates and corresponding *p* values were calculated for meta-analyses. Considering the potential influence of single large experiment on the meta-analysis results, leave-one-disease-out cross validation was performed. One dataset within NC superset at a time was excluded and aforementioned meta-analysis method was then applied to the remaining datasets. Only probes with heterogeneity *p* value > 0.1 and fixed-effect *p* value < 1e-10 in both overall result and 8 leave-one-disease-out cross validations were considered as robust DEPs.

Divide DEGs in to 27 modules by expression pattern

DEGs were retrieved through mapping Affymetrix DEPs to gene Entrez identifiers. Gene expression pattern during colorectal development was defined as the direction of differential expression of a particular gene identified with unpaired Student's t-test (*p*<0.05) between consecutive colorectal development stages (Bud, EarlyColon, MiddleColon and Normal stage). There were three transitions among four time points, and for each transition, there were three scenarios [up-regulation (u), down-regulation (d), and no significant change (n)]. Therefore, all the DEGs could be assigned to 3*3 = 27 modules based on their corresponding differential expression pattern. For instance, DEGs with expression pattern "d, d, d" were assigned to the first pattern (hereafter referred to as "PTN_1"), and DEGs with expression pattern "u, u, u" were assigned to PTN_27. The same clustering method was also adopted in our previous research upon lung development samples[23].

Collection of V and A probe sets

Among 27 modules, modules with contradictory differential expressions ("u" and "d" coexisted, like PTN_3 with expression pattern "d, d, u") and PTN_14 (expression pattern "n, n, n") were ruled out from further analysis. As for the remaining 14 modules, modules with at least one "d" were designated as DDMs, and modules with at least one "u" were designated as DUMs. Therefore, 7 DDMs and 7 DUMs were preliminarily collected, containing corresponding DEPs mapped from DEGs in each DDM or DUM. GSEA analysis was conducted for the 14 module DEPs in NC superset, and module DEP sets could not show statistically significant and concordant differences between normal and cancer samples were excluded (FDR<0.001). Thus, DEPs up-regulated in cancer and belonging to GSEA-significant DDMs were termed as "V" probes, since the trajectory from Bud to normal to cancer is similar with the letter "V"; DEPs down-regulated in cancer and belonging to GSEA-significant DUMs were termed as "A" probes.

Assessment of the association of gene expression with clinicopathological variables

The potential associations between gene groups and 8 clinicopathological variables (age, gender, stage, grade, tumor size, lymph node invasion, distant metastasis and AdjCTX) were evaluated in 933 CRC samples of Clinicinfo superset. As for a given variable, samples with clear description of the variable were pooled, and thus a single-column vector called the module eigengene (ME) was calculated by the first principal component following PCA analysis of probe expression level across corresponding samples. As ME captures the majority of total variance, it represents a summary measure for the overall expression status of the whole gene group. Samples with the given variable information were further divided by the median of ME, and logistic regression analyses were conducted between each variable values and ME assigned sample groups.

Validation of gene signature's prognostic value in Clinicinfo superset

In order to assess the prognostic value of the gene signature we obtained (suppose the signature contained n genes), the risk score formula for predicting OS was developed based on a linear combination of the expression level (x_1, x_2, …, x_n) of a given patient weighted by the regression coefficients derived from the Cox regression analysis. GSE17536 was used as training cohort for Cox regression model construction and the remaining 5 Clinicinfo data sets were treated as testing cohorts. The regression coefficient *β* was calculated with training cohort and the same coefficient was further applied to testing cohorts. The risk score *r* for Patient *j* was calculated as follows:

$$r_{j=}\sum_{i=1}^{n}\beta_i\, x_{ij}$$

Ten-fold cross validation was also conducted within training cohort to strengthen the validity of the test. We then divided patients into high-risk and low-risk groups using the median gene signature risk score. Patients with higher risk scores are expected to have significantly poor OS status if the gene signature is closely related to OS. Kaplan–Meier survival analysis and log-rank test were performed to evaluate the prognostic difference between the two risk score assigned groups both in 5 independent data sets and in pooled data set.

Conclusions

In summary, we conducted meta-analysis in NC superset containing 8 microarray data sets to collect robust DEPs in CRC. The expression profile of human colorectal developing tissues in 4 sequential stages was accomplished, and V probes and A probes were obtained based on differential expression patterns during embryonic development and carcinogenesis. Cell-cycle related probes were greatly enriched in V probe group, and they were strongly associated with OS and other clinicopathological variables, suggesting that they held enormous information for clinical practice.

Author contributions

Shujun Cheng, Kaitai Zhang and Guiqi Wang participated in the design and coordination of the study. Ning An, Xue Yang and Yueming Zhang carried out the sample selection, experiments and algorithm construction. Ning An and Xiaoyu Shi performed the microarray experiments. Ning An and Xuexin Yu performed data analysis. Ning An and Xue Yang wrote the manuscript. All authors have read and approved the manuscript and its contents, and are aware of responsibilities connected to authorship.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

We thank the families for their participation in this project. This work was supported by the National High Technology Research and Development Program of China (SS2014AA020801) and the Sci-Tech Development Program of Beijing (D121100004712002).

Notes and references

1. R. A. Burrell, N. McGranahan, J. Bartek and C. Swanton, *Nature*, 2013, 501, 338-345.
2. M. A. Nieto, *Science*, 2013, 342, 708-+.
3. A. M. Eastham, H. Spencer, F. Soncin, S. Ritson, C. L. R. Merry, P. L. Stern and C. M. Ward, *Cancer Res.*, 2007, 67, 11254-11262.
4. L. Ridolfi, M. Petrini, L. Fiammenghi, A. Riccobon and R. Ridolfi, *Immunobiology*, 2009, 214, 61-76.
5. K. A. Hartwell, B. Muir, F. Reinhardt, A. E. Carpenter, D. C. Sgroi and R. A. Weinberg, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, 103, 18969-18974.
6. A. Sparmann and M. van Lohuizen, *Nat Rev Cancer*, 2006, 6, 846-856.
7. S. L. Liu, G. Dontu, I. D. Mantle, S. Patel, N. S. Ahn, K. W. Jackson, P. Suri and M. S. Wicha, *Cancer Res.*, 2006, 66, 6063-6071.
8. H. C. Kang, Y. Wakabayashi, K. Y. Jen, J. H. Mao, V. Zoumpourlis, R. Del Rosario and A. Balmain, *J. Invest. Dermatol.*, 2013, 133, 1311-1320.
9. A. T. Kho, Q. Zhao, Z. H. Cai, A. J. Butte, J. Y. H. Kim, S. L. Pomeroy, D. H. Rowitch and I. S. Kohane, *Genes Dev.*, 2004, 18, 629-640.
10. H. Y. Liu, A. T. Kho, I. S. Kohane and Y. Sun, *PLoS Med.*, 2006, 3, 1090-1102.
11. S. Kaiser, Y. K. Park, J. L. Franklin, R. B. Halberg, M. Yu, W. J. Jessen, J. Freudenberg, X. D. Chen, K. Haigis, A. G. Jegga, S. Kong, B. Sakthivel, H. Xu, T. Reichling, M. Azhar, G. P. Boivin, R. B. Roberts, A. C. Bissahoyo, F. Gonzales, G. C. Bloom, S. Eschrich, S. L. Carter, J. E. Aronow, J. Kleimeyer, M. Kleimeyer, V. Ramaswamy, S. H. Settle, B. Boone, S. Levy, J. M. Graff, T. Doetschman, J. Groden, W. F. Dove, D. W. Threadgill, T. J. Yeatman, R. J. Coffey and B. J. Aronow, *Genome Biol.*, 2007, 8.
12. A. D. Rhim and B. Z. Stanger, *Development, Differentiation and Disease of the Para-Alimentary Tract*, 2010, 97, 41-78.
13. H. Hu, L. Zhou, A. Awadallah and W. Xin, *Appl. Immunohistochem. Mol. Morphol.*, 2013, 21, 242-247.
14. M. Hu and R. A. Shivdasani, *Cancer Res.*, 2005, 65, 8715-8722.
15. A. C. Borczuk, L. Gorenstein, K. L. Walter, A. A. Assaad, L. Q. Wang and C. A. Powell, *Am. J. Pathol.*, 2003, 163, 1949-1960.
16. A. T. Kho, Q. Zhao, Z. Cai, A. J. Butte, J. Y. Kim, S. L. Pomeroy, D. H. Rowitch and I. S. Kohane, *Genes Dev.*, 2004, 18, 629-640.
17. M. Monzo, A. Navarro, E. Bandres, R. Artells, I. Moreno, B. Gel, R. Ibeas, J. Moreno, F. Martinez, T. Diaz, A. Martinez, O. Balague and J. Garcia-Foncillas, *Cell Res.*, 2008, 18, 823-833.
18. T. Reya, S. J. Morrison, M. F. Clarke and I. L. Weissman, *Nature*, 2001, 414, 105-111.
19. R. A. Gibbs, J. Rogers, M. G. Katze, R. Bumgarner, G. M. Weinstock, E. R. Mardis, K. A. Remington, R. L. Strausberg, J. C. Venter, R. K. Wilson, M. A. Batzer, C. D. Bustamante, E. E. Eichler, M. W. Hahn, R. C. Hardison, K. D. Makova, W. Miller, A. Milosavljevic, R. E. Palermo, A. Siepel, J. M. Sikela, T. Attaway, S. Bell, K. E. Bernard, C. J. Buhay, M. N. Chandrabose, M. Dao, C. Davis, K. D. Delehaunty, Y. Ding, H. H. Dinh, S. Dugan-Rocha, L. A. Fulton, R. A. Gabisi, T. T. Garner, J. Godfrey, A. C. Hawes, J. Hernandez, S. Hines, M. Holder, J. Hume, S. N. Jhangiani, V. Joshi, Z. M. Khan, E. F. Kirkness, A. Cree, R. G. Fowler, S. Lee, L. R. Lewis, Z. W. Li, Y. S. Liu, S. M. Moore, D. Muzny, L. V. Nazareth, D. N. Ngo, G. O. Okwuonu, G. Pai, D. Parker, H. A. Paul, C. Pfannkoch, C. S. Pohl, Y. H. Rogers, S. J. Ruiz, A. Sabo, J. Santibanez, B. W. Schneider, S. M. Smith, E. Sodergren, A. F. Svatek, T. R. Utterback, S. Vattathil, W. Warren, C. S. White, A. T. Chinwalla, Y. Feng, A. L. Halpern, L. W. Hillier, X. Q. Huang, P. Minx, J. O. Nelson, K. H. Pepin, X. Qin, G. G. Sutton, E. Venter, B. P. Walenz, J. W. Wallis, K. C. Worley, S. P. Yang, S. M. Jones, M. A. Marra, M. Rocchi, J. E. Schein, R. Baertsch, L. Clarke, M. Csuros, J. Glasscock, R. A. Harris, P. Haviak, A. R. Jackson, H. Y. Jiang, Y. Liu, D. N. Messina, Y. F. Shen, H. X. Z. Song, T. Wylie, L. Zhang, E. Birney, K. Han, M. K. Konkel, J. N. Lee, A. F. A. Smit, B. Ullmer, H. Wang, J. Xing, R. Burhans, Z. Cheng, J. E. Karro, J. Ma, B. Raney, X. W. She, M. J. Cox, J. P. Demuth, L. J. Dumas, S. G. Han, J. Hopkins, A. Karimpour-Fard, Y. H. Kim, J. R. Pollack, T. Vinar, C. Addo-Quaye, J. Degenhardt, A. Denby, M. J. Hubisz, A. Indap, C. Kosiol, B. T. Lahn, H. A. Lawson, A. Marklein, R. Nielsen, E. J. Vallender, A. G. Clark, B. Ferguson, R. D. Hernandez, K. Hirani, H. Kehrer-Sawatzki, J. Kolb, S. Patil, L. L. Pu, Y. Ren, D. G. Smith, D. A. Wheeler, I. Schenck, E. V. Ball, R. Chen, D. N. Cooper, B. Giardine, F. Hsu, W. J. Kent, A. Lesk, D. L. Nelson, W. E. O'Brien, K. Prufer, P. D. Stenson, J. C. Wallace, H. Ke, X. M. Liu, P. Wang, A. P. Xiang, F. Yang, G. P. Barber, D. Haussler, D. Karolchik, A. D. Kern, R. M. Kuhn, K. E. Smith, A. S. Zwieg and R. M. G. S. Consortium, *Science*, 2007, 316, 222-234.
20. K. He, H. B. Zhao, Q. S. Wang and Y. C. Pan, *Reprod. Biol. Endocrinol.*, 2010, 8.
21. H. Fang, Y. Yang, C. L. Li, S. J. Fu, Z. Q. Yang, G. Jin, K. K. Wang, J. Zhang and Y. Jin, *Dev. Cell*, 2010, 19, 174-184.
22. H. Yi, L. Xue, M. X. Guo, J. A. Ma, Y. Zeng, W. Wang, J. Y. Cai, H. M. Hu, H. B. Shu, Y. B. Shi and W. X. Li, *FASEB J.*, 2010, 24, 3341-3350.
23. L. Feng, J. M. Wang, B. R. Cao, Y. Zhang, B. Wu, X. B. Di, W. Jiang, N. An, D. Lu, S. H. Gao, Y. D. Zhao, Z. L. Chen, Y. S. Mao, Y. N. Gao, D. S. Zhou, J. Jen, X. H. Liu, Y. P. Zhang, X. Li, K. T. Zhang, J. He and S. J. Cheng, *PLoS ONE*, 2014, 9.
24. C. Zhang, C. Li, Y. Xu, L. Feng, D. Shang, X. Yang, J. Han, Z. Sun, Y. Li and X. Li, *Mol. Biosyst.*, 2015, 11, 1271-1284.
25. F. Li, Y. Xiao, F. Huang, W. Deng, H. Zhao, X. Shi, S. Wang, X. Yu, L. Zhang, Z. Han, L. Luo, Q. Zhu, W. Jiang, S. Cheng, X. Li and K. Zhang, *Mol. Biosyst.*, 2015, DOI: 10.1039/c5mb00474h.
26. R. K. Montgomery, A. E. Mulberg and R. J. Grand, *Gastroenterology*, 1999, 116, 702-731.
27. R. Chen, P. Khatri, P. K. Mazur, M. Polin, Y. Y. Zheng, D. Vaka, C. D. Hoang, J. Shrager, Y. Xu, S. Vicent, A. J. Butte and E. A. Sweet-Cordero, *Cancer Res.*, 2014, 74, 2892-2902.
28. N. C. W. Goonesekere, X. S. Wang, L. Ludwig and C. Guda, *PLoS ONE*, 2014, 9.
29. M. Greaves and C. C. Maley, *Nature*, 2012, 481, 306-313.
30. I. R. Watson, K. Takahashi, P. A. Futreal and L. Chin, *Nat Rev Genet*, 2013, 14, 703-718.
31. M. R. Stratton, *Science*, 2011, 331, 1553-1558.
32. M. S. Wicha, S. L. Liu and G. Dontu, *Cancer Res.*, 2006, 66, 1883-1890.
33. J. E. Dick, *Blood*, 2008, 112, 4793-4807.
34. N. Barker, R. A. Ridgway, J. H. van Es, M. van de Wetering, H. Begthel, M. van den Born, E. Danenberg, A. R. Clarke, O. J. Sansom and H. Clevers, *Nature*, 2009, 457, 608-U119.
35. S. Gurzu, D. Ciortea, I. Ember and I. Jung, *Biomed Res Int*, 2013, 2013, 365192.
36. K. Yaniv and J. K. Yisraeli, *Gene*, 2002, 287, 49-54.
37. M. N. Nguyen, T. G. Choi, D. T. Nguyen, J. H. Kim, Y. H. Jo, M. Shahid, S. Akter, S. N. Aryal, J. Y. Yoo, Y. J. Ahn, K. M. Cho, J. S. Lee, W. Choe, I. Kang, J. Ha and S. S. Kim, *Oncotarget*, 2015.

38. S. Sood, I. J. Gallagher, K. Lunnon, E. Rullman, A. Keohane, H. Crossland, B. E. Phillips, T. Cederholm, T. Jensen, L. J. van Loon, L. Lannfelt, W. E. Kraus, P. J. Atherton, R. Howard, T. Gustafsson, A. Hodges and J. A. Timmons, *Genome Biol.*, 2015, 16, 185.
39. N. An, X. Shi, Y. Zhang, N. Lv, L. Feng, X. Di, N. Han, G. Wang, S. Cheng and K. Zhang, *PLoS ONE*, 2015, 10, e0137171.
40. S. B. Edge and American Joint Committee on Cancer., *AJCC cancer staging manual*, Springer, New York, 7th edn., 2010.
41. S. Stintzing, *F1000Prime Rep*, 2014, 6, 108.
42. S. Ahmed, K. Johnson, O. Ahmed and N. Iqbal, *Int. J. Colorectal Dis.*, 2014, 29, 1031-1042.
43. S. Benoist, K. Pautrat, E. Mitry, P. Rougier, C. Penna and B. Nordlinger, *Br. J. Surg.*, 2005, 92, 1155-1160.
44. G. Galizia, E. Lieto, M. Orditura, P. Castellano, V. Imperatore, M. Pinto and A. Zamboli, *Arch. Surg.*, 2008, 143, 352-358; discussion 358.
45. C. Zemmour, F. Bertucci, P. Finetti, B. Chetrit, D. Birnbaum, T. Filleron and J. M. Boher, *Cancer Inform*, 2015, 14, 129-138.
46. S. Y. Tian, C. Wang and M. W. An, *Biol. Direct*, 2015, 10.
47. D. Cangelosi, M. Muselli, S. Parodi, F. Blengio, P. Becherini, R. Versteeg, M. Conte and L. Varesio, *BMC Bioinformatics*, 2014, 15.