SCIENTIFIC REPORTS

OPEN

Received: 06 September 2015
Accepted: 19 November 2015
Published: 22 December 2015

# Developmental genes significantly afflicted by aberrant promoter methylation and somatic mutation predict overall survival of late-stage colorectal cancer

Ning An[1,*], Xue Yang[1,*], Shujun Cheng[1], Guiqi Wang[2] & Kaitai Zhang[1]

**Carcinogenesis is an exceedingly complicated process, which involves multi-level dysregulations, including genomics (majorly caused by somatic mutation and copy number variation), DNA methylomics, and transcriptomics. Therefore, only looking into one molecular level of cancer is not sufficient to uncover the intricate underlying mechanisms. With the abundant resources of public available data in the Cancer Genome Atlas (TCGA) database, an integrative strategy was conducted to systematically analyze the aberrant patterns of colorectal cancer on the basis of DNA copy number, promoter methylation, somatic mutation and gene expression. In this study, paired samples in each genomic level were retrieved to identify differentially expressed genes with corresponding genetic or epigenetic dysregulations. Notably, the result of gene ontology enrichment analysis indicated that the differentially expressed genes with corresponding aberrant promoter methylation or somatic mutation were both functionally concentrated upon developmental process, suggesting the intimate association between development and carcinogenesis. Thus, by means of random walk with restart, 37 significant development-related genes were retrieved from a priori-knowledge based biological network. In five independent microarray datasets, Kaplan–Meier survival and Cox regression analyses both confirmed that the expression of these genes was significantly associated with overall survival of Stage III/IV colorectal cancer patients.**

Colorectal cancer (CRC) is the third most common cancer in men (746,000 cases, 10.0% of the total) and the second in women (614,000 cases, 9.2% of the total) worldwide, accounting for roughly 694,000 deaths per year[1]. The initiation of CRC is an incredibly complicated biological process, involving multiple genomic and epigenomic alterations, occurring over an extended time period of usually a decade[2]. Patient survival is limitedly dependent on the tumor stage at the time of diagnosis, and reduced sensitivity to chemotherapy is still a major obstacle in effective treatment of advanced disease. Therefore, the discovery of novel molecules promoting CRC progression and indicating prognostic status, is still urgently needed[3].

It is putatively accredited that carcinogenesis is caused by multi-level dysregulations, including genomics [majorly caused by somatic mutation and copy number variation (CNV)][4,5], DNA methylomics[6,7], and transcriptomics[8,9]. CNV plays a significant role in tumorigenesis in many cancers[10–14], whose accumulation during oncogenesis might be a result of preferential selection by which transforming cells gain evolutionary advantages[15]. Somatic mutation, together with CNV, could contribute to genomic instability[4]. It could also activate additional downstream pathways in many types of cancer to acquire proliferative advantages[16–18]. DNA methylation is substantially important in promoting embryonic development[19], aging[20], and nearly all types of cancer[21–24], by influencing DNA and chromatin structures[25]. Numerous investigations indicated that the dysregulation of promoter

[1]State Key Laboratory of Molecular Oncology, Department of Etiology and Carcinogenesis, Peking Union Medical College & Cancer Institute (Hospital), Chinese Academy of Medical Sciences, Beijing, 100021, China. [2]Department of Endoscopy, Cancer Hospital, Chinese Academy of Medical Sciences, Beijing, 100021, China. [*]These authors contributed equally to this work. Correspondence and requests for materials should be addressed to S.C. (email: chengshj@263.net.cn) or G.W. (email: wangguiq@126.com) or K.Z. (email: zhangkt_bingyin@sina.cn)

region, especially promoter hypermethylation of tumor suppressor genes, was the essential epigenetic events in carcinogenesis, prognostic marker discovery, and therapeutic utilities[26–29].

CNV, aberrant promoter methylation and somatic mutation could all influence gene activation or suppression, thereby influencing the process of carcinogenesis. CNVs may alter gene dosage by changing the number of copies of a gene that is present in the genome[30–33], explaining in most circumstances, CNV and corresponding gene expression are positively correlated in CRC[34]. Promoter hypomethylation might lead to gene activation, and promoter hypermethylation might cause gene suppression[35]. Genes with somatic mutation could probably lead to the activation or suppression of downstream signaling pathways[36]. For example, in thyroid cancer, somatic mutation of *BRAF* could activate *MAPK* pathway, thus influencing the massive dysregulation of gene activity[37].

The multi-level genomic dysregulations during carcinogenesis indicated that while looking into the dysregulation of gene expression in cancer, the aberrant patterns of multi-level events should also be paid considerable attention to shed light on the underlying intricate mechanisms of cancer initiation and deterioration. Therefore, the integrative analysis of cancer genomics, methylomics and transcriptomics is urgently needed to comprehensively dissect cancer etiology and provide clinical guidance.

The Cancer Genome Atlas (TCGA) database is an immeasurable source of knowledge launched in 2005, which provides publicly available cancer genomic datasets[38]. Based on abundant resources of RNA sequencing (RNAseq), DNA sequencing (DNAseq), single nucleotide polymorphism (SNP) based platforms and DNA methylation, integrative analysis of cancer genomics was exuberantly emerging, for instance, in breast cancer[39], ovarian cancer[40], glioma[41], lung cancer[42], renal cancer[43] and many other types of cancers. Multi-dimensional analyses (MDA) of the genome, epigenome, and transcriptome was proven to be greatly beneficial in facilitating the rational deduction of aberrant genes and pathways, delineating subtypes of cancer, and promoting derivation of diagnostic and prognostic signatures, which otherwise would be overlooked in single genomic dimension investigations[44]. Thus, the molecular abnormalities of multiple levels should be altogether taken into consideration and systematically identify genes or pathways critically important in carcinogenesis.

In this study, we first collected genes with significant dysregulations with regard to DNA copy number, DNA promoter methylation, gene expression, and somatic mutation from TCGA paired samples. Differentially expressed genes (DEGs) with consistent aberrant promoter methylation or somatic mutation were found both exhibiting remarkable functional unity in developmental process. Gene to gene regulatory network was constructed by means of merging Human Protein Reference Database (HPRD), and Kyoto Encyclopedia of Genes and Genomes (KEGG) networks. By combining multi-dimensional genomic data of CRC and priori knowledge network, we applied a computational strategy, i.e. random walk with restart, to obtain the genes which were affected considerably by aberrant promoter methylation or somatic mutation. The most of these significant genes were connected in the network, and proven to hold profound prognostic information in late stage (Stage III/IV) patients, which might be helpful for constructing prognosis prediction models and providing novel tools to guide clinical implementations for this deadly disease.

## Material and Methods

A schematic for the study is depicted in Fig. 1.

**Data retrieval.** The multi-dimensional data of CRC associated datasets were retrieved from The Cancer Genome Atlas (TCGA) database (https://tcga-data.nci.nih.gov/tcga/). Four levels of paired data (cancer and normal adjacent tissues from CRC patients) were downloaded, including 32 paired RNA sequencing level 3 data [raw counts and RNASeq by Expectation Maximization (RSEM) normalized read counts], 500 paired DNA copy number level 3 data [conducted with Affymetrix SNP 6.0 platform, and segmented by circular binary segmentation (CBS) method[45]], 45 paired DNA methylation level 3 data [using Illumina HumanMethylation450 chips, and the methylation level of each CpG site was calculated as the ratio ($\beta$ value) of the signal of methylated probes relative to the sum of methylated and unmethylated probes, which ranged continuously from 0 (unmethylated) to 1 (fully methylated)], and somatic mutation level 2 data of 300 patients (mutation information of 17,427 genes).

The raw data for five human CRC mRNA microarray studies with overall survival (OS) information (sample size $>60$, referred to as Clinicinfo superset; Table 1) were downloaded from the National Center for Biotechnology Information Gene Expression Omnibus (GEO). The flowchart of Clinicinfo dataset retrieval is presented in Supplementary Figure S1. The combined data set contained a total of 940 samples (936 samples with clear OS information) hybridized to probe sets present on both the Affymetrix HG-U133A (with GEO accession number GPL96) and the HG-U133A Plus2 (GPL570) platforms, composed of data sets with accession numbers GSE39582, GSE17536, GSE29621, GSE39084, and GSE12945. In total, 22,277 probes were common in all data sets, and of which the 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 effects, and the expression levels of 12,500 genes were obtained as the median value of all the probes which could be mapped to this gene. All clinical information was extracted from the original publications.

**Circos plot of TCGA colorectal data in terms of DNA copy number, DNA methylation and somatic mutation.** Colorectal primary tumor datasets in TCGA database, including 617 DNA copy number data, 393 DNA methylation data, and 300 somatic mutation data, were enrolled for integrative Circos plot construction via Perl software "Circos plot" (Fig. 2). Bioconductor package "cghMCR" was used to compute the segment gain or loss (SGOL) scores to quantify chromosome regions showing common gains/losses by summation of the score in each patient. For DNA methylation, the whole genome was segmented into contiguous 500,000 base pair (bp) bins, and the median and 75th percentile of methylation levels of CpGs which could be

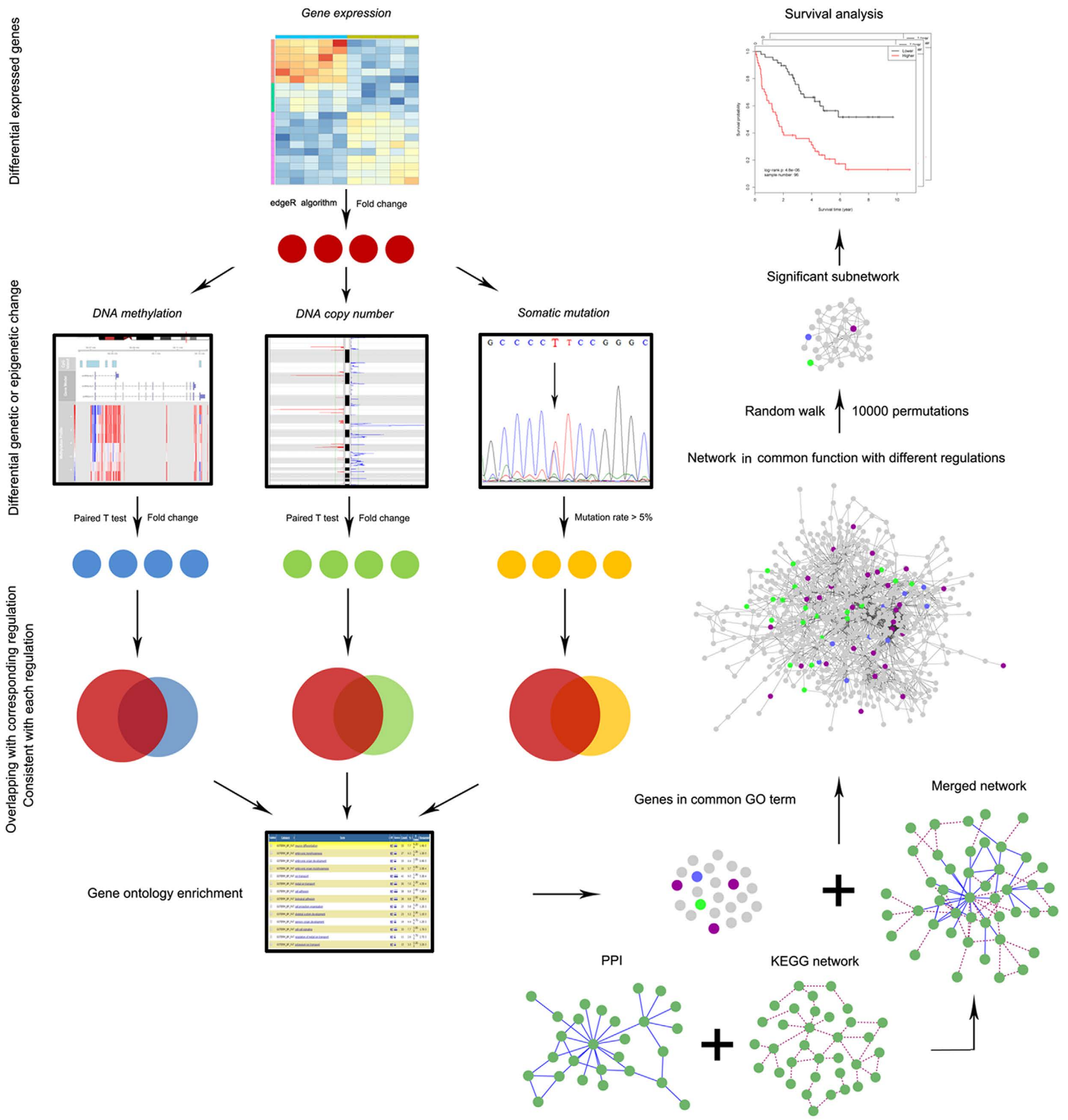

**Figure 1. Schematic of methodology applied in this study.** Step I: Integration of genomic, DNA methylomic, and transcriptomic data to identify three candidate gene groups; Step II: Identification of gene ontology (GO) function term and corresponding gene groups of interest based on GO enrichment analysis; Step III: Identification of genes within the identified functional groups significantly afflicted by genetic or epigenetic dysregulation, by applying random walk with restart algorithm in the merged network. Step IV: Survival analysis of identified significant genes to evaluate their prognostic value.

mapped onto each bin were plotted. As for somatic mutation data, genes with mutation rate >5% were shown in scatter plot.

## Identification of candidate genes with significant alteration at multi-level.

DEGs were identified using edgeR algorithm[46] with RNA sequencing raw counts (FDR < 0.01, fold change >2). As for DNA copy number data, Bioconductor package "CNTools" was used to process segmentation data and format the data into a gene-level matrix based on corresponding genomic location of 26,863 genes. Genes with genomic amplification and deletion were identified with paired $t$ statistic test (FDR < 0.001, fold change >1.2). In methylation analysis, promoter region was defined as the region between 1,000 bp upstream transcription start site (TSS) and 300 bp downstream TSS. The $\beta$ value of the probe which could be mapped to the CpG site located in the promoter region of a given gene was used to quantify the methylation level of this gene. If more than one probe could be mapped to the promoter region of a given gene, the mean value was adopted. In this manner, the methylation level of 16,996 genes were obtained with DNA methylation data, and significant hypermethylated and hypomethylated genes were identified with paired $t$ statistic test (FDR < 0.001, fold change >1.5).

By virtue of dysregulation pattern at different levels, three groups of candidate genes of interest were collected: (i) genes with differential expression and corresponding copy number alteration (i.e. genes with overexpression and amplification, and genes with underexpression and deletion); (ii) genes with differential expression and

| Characteristics | Samples | | | | |
|---|---|---|---|---|---|
| | GSE12945 | GSE17536 | GSE39582 | GSE29621 | GSE39084 |
| ***Number*** | 62 | 177 | 566 | 65 | 70 |
| ***Year*** | 2009 | 2009 | 2013 | 2014 | 2014 |
| ***Country*** | Germany | American | France | American | France |
| ***Gender*** | | | | | |
| Male | 34 | 96 | 310 | 40 | 35 |
| Female | 28 | 81 | 256 | 25 | 35 |
| ***Age*** | | | | | |
| Mean ± SD (years) | 64.4 ± 11.8 | 65.5 ± 13.1 | 63.0 ± 19.0 | NR | 59.2 ± 18.3 |
| ***T status*** | | | | | |
| T1 + T2 | 16 | NR | 57 | 8 | 13 |
| T3 + T4 | 46 | NR | 486 | 57 | 57 |
| ***N status*** | | | | | |
| N0 | 36 | NR | 302 | 32 | 35 |
| N1 | 14 | NR | 134 | 25 | 20 |
| N2 | 12 | NR | 104 | 7 | 15 |
| ***M status*** | | | | | |
| M0 | 56 | NR | 482 | 46 | 48 |
| M1 | 5 | NR | 61 | 18 | 22 |
| ***AJCC stage*** | | | | | |
| Stage I + II | 36 | 81 | 297 | 29 | 31 |
| Stage III + IV | 26 | 96 | 265 | 36 | 38 |
| ***Pathologic grade*** | | | | | |
| G I | 0 | 16 | NR | 4 | NR |
| G II | 31 | 134 | NR | 51 | NR |
| G III | 31 | 27 | NR | 10 | NR |
| ***AdjCTX*** | | | | | |
| Yes | NR | NR | 233 | 38 | NR |
| No | NR | NR | 316 | 27 | NR |

**Table 1. Colorectal cancer microarray datasets included in survival analysis.** Abbreviations: *SD* = standard deviation; *AdjCTX* = whether chemotherapy was used; *NR* = not reported.

corresponding promoter methylation (i.e. genes with overexpression and promoter hypomethylation, and genes with underexpression and promoter hypermethylation); and (iii) genes with differential expression and somatic mutation.

**Identification of significant genes through random walk.** Gene ontology (GO) enrichment analysis was conducted using Bioconductor package "clusterProfiler". The protein–protein interaction network was downloaded from HPRD database, and KEGG network was constructed with Bioconductor package "KEGGgraph". Therefore, gene regulatory network was established by merging HPRD and KEGG network, including 10,479 nodes and 60,689 edges after eliminating self-loops and duplicated edges.

Taking advantage of knowledge-based network topology, random walk algorithm was utilized to identify genes algorithmically most affected by aberrant promoter methylation and somatic mutation[47]. In the network, genes of interest were designated as information source (i.e., source nodes) and the remaining genes in the network as the information target (i.e., target nodes). The information flow originates from source nodes iteratively and randomly transmits to their neighbors with a probability proportional to their topological features. At each step, the information can flow back to the source nodes with the same probability. The final steady-state probability assigned to each gene in the network reflects the integrated influence imposed by source nodes combining network topology. Formally, the random walk with restart is defined as:

$$p^{t+1} = (1 - r)\,Wp^{t} + rp^{0} \tag{1}$$

where $W$ is the column-normalized adjacency matrix of network, and $p^t$ is a vector in which the genes in the network holds probability in the iterative process up to step $t$. Source nodes were weighted with initial probability vector $p^0$ (the sum of its elements was equal to 1), and $r$ represents restart probability ($r = 0.7$ in this study). All the genes in the network were ranked according to the values in the steady-state probability vector $p^\infty$. This was obtained at query time by performing the iteration until the difference between $p^t$ and $p^{t+1}$ (measured by the L1 norm) was lower than $10^{-10}$. In order to obtain genes with significantly high steady-state probability, 10,000 permutations of node labels (with network topology remained the same) were conducted to calculate the null distribution of final probability for each gene. The $p$ value was termed as the ratio of random values that were greater than the

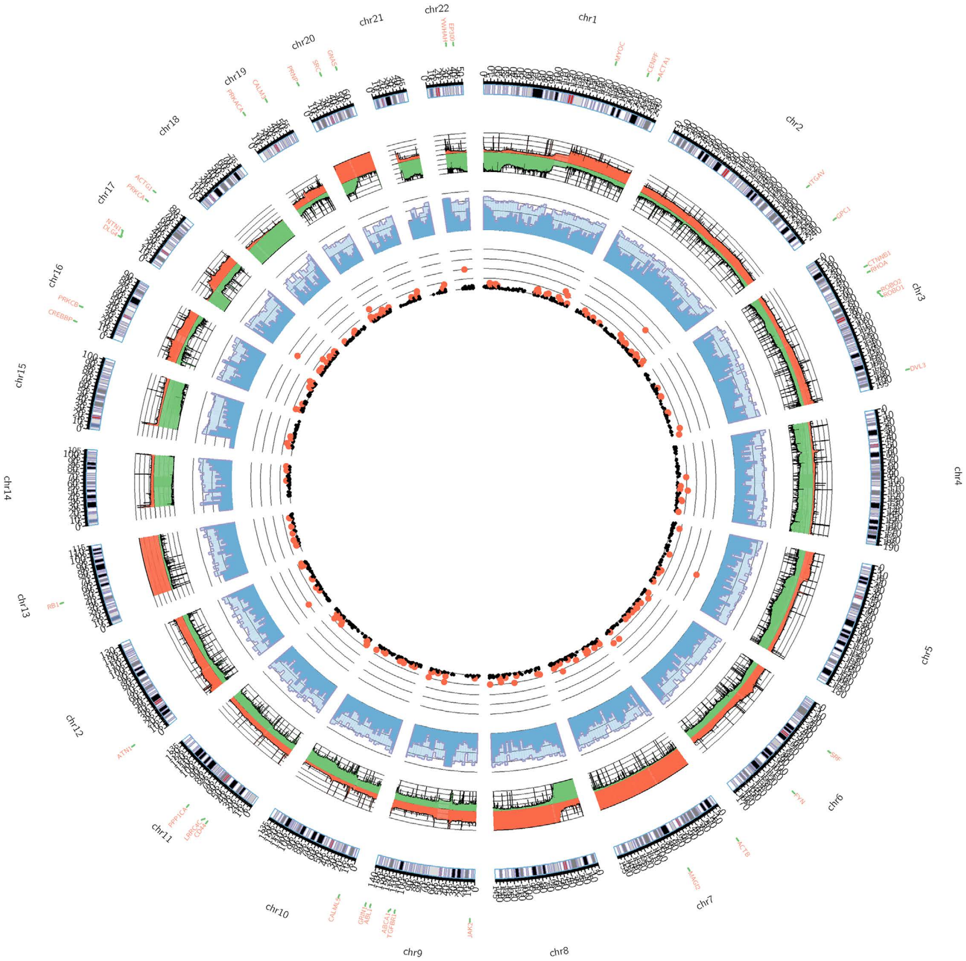

**Figure 2. Circos plot in terms of DNA copy number, DNA methylation and somatic mutation.** An ideogram of a normal karyotype is shown in the outermost ring. The next outermost ring represents DNA copy number at corresponding genomic coordinates, calculated by the segment gain or loss (SGOL) scores (red represents amplification and green represents deletion). The next ring represents the amount of DNA methylation. The whole genome was segmented into contiguous 500,000 base pair (bp) bins, and the median (dark blue) and 75th percentile (light blue) of methylation levels of CpGs which could be mapped onto each bin were plotted. The innermost ring is scatter plot illustrating somatic mutation data, genes with mutation rate >5% were shown in black and those >10% were shown in red.

observed final probability. Genes with $p < 0.01$ were regarded as the genes significantly afflicted by these genetic or epigenetic abnormalities.

**Validation of gene signature's prognostic value in Clinicinfo superset.** In order to assess the prognostic value of the significant genes 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 ($x_1, x_2, \ldots, x_n$) 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 four Clinicinfo data sets were treated as test cohorts. The regression coefficient $\beta$ 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:

$$rj = \sum_{i=1}^{n} \beta_i x_{ij} \tag{2}$$

Five-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.

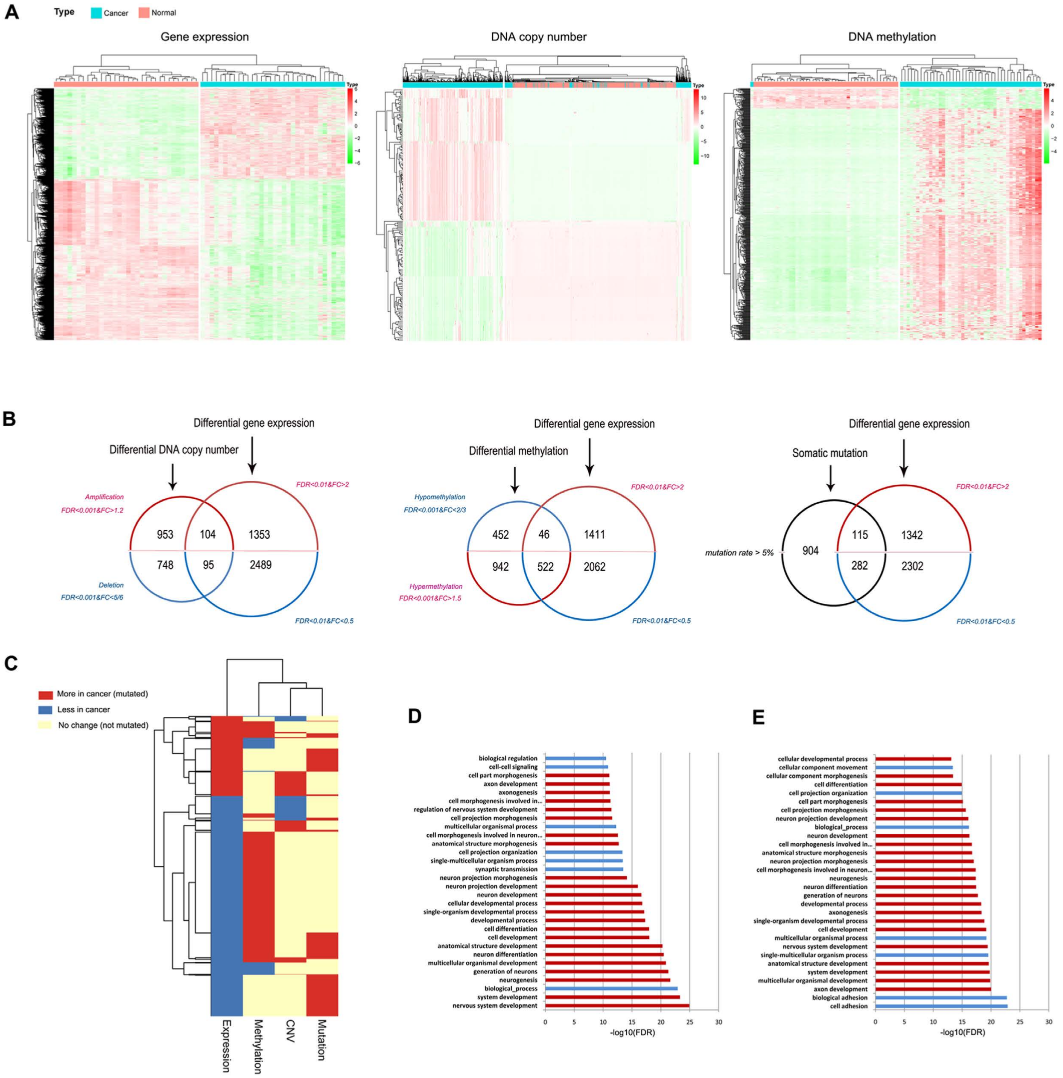

**Figure 3. Identification of three gene groups with multi-omic data in the Cancer Genome Atlas (TCGA) database.** (**a**) Heat maps of differentially expressed genes (DEGs) [log2 transformed RNASeq by Expectation Maximization (RSEM) normalized read counts], DEGs with copy number variation (CNV), and DEGs with aberrant promoter methylation in corresponding paired TCGA data, respectively. (**b**) Venn diagram illustrating three groups of candidate genes with differential expression and another altered molecular level, such as DNA copy number, promoter methylation and somatic mutation. (**c**) Integrated genetic and epigenetic alteration patterns of differentially expressed genes. Rows represent DEGs, and columns represent four dysregulation types. Red denotes the more in cancer (overexpression, promoter hypermethylation or DNA amplification) or mutated DEGs. Blue denotes less in cancer (underexpression, promoter hypomethylation or DNA deletion) or not mutated DEGs. (**d**) Gene ontology (GO) enrichment analysis of Group B genes [differentially expressed genes (DEGs) with abnormal promoter methylation]. Red bar represents enriched GO terms which are offspring of developmental process. (**e**) GO enrichment analysis of Group C genes (DEGs with somatic mutation).

## Results

**Collection of genes with somatic mutation, differential expression, DNA copy number and promoter methylation with paired TCGA samples.** Due to abundant resources of TCGA database, paired samples of CRC were used to obliterate individual difference. DEGs, calculated using edgeR algorithm, were composed of 1,457 up-regulated genes and 2,584 down-regulated genes (Fig. 3A). In addition, 1,057 genes were significantly amplified and 843 genes were found significantly deleted (Fig. 3A). Integrative Circos plot indicated there were severe copy number alteration in Chromosome 7, 8, 13, 17, 18 and 20, highly consistent with previous investigations[34,48–52] (Fig. 2). By means of paired *t* statistic test, 1,464 genes with promoter hypermethylation and 498 genes with promoter hypomethylation were also identified (Fig. 3A), and 1,301 genes with mutation rate >5% were regarded as mutated genes.

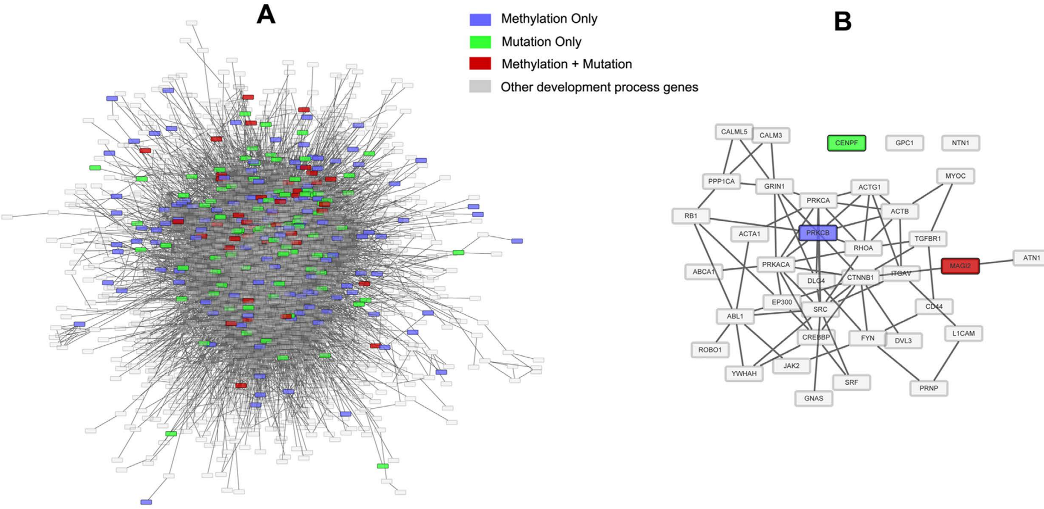

**Figure 4. Random walk of DEGs of Group B and C in developmental process related network (DPRN).** (**a**) The biggest connected component (BCC) of DPRN containing 3,271 developmental process related genes (DPRGs) and 20,652 edges. DEGs with abnormal promoter methylation and DEGs with somatic mutation in DPRN were regarded as source nodes, and the rest of DPRGs in the network were target nodes. (**b**) The subgraph of DPRN composed of 37 significant genes retrieved via random walk with restart.

**Identification of candidate gene groups associated with DNA copy number alterations, promoter methylation, and somatic mutation.** Three groups of DEGs with aberrant genetic or epigenetic dysregulations (Fig. 3B) were categorized as follows: (i) 104 genes with overexpression and copy number amplification, and 95 genes with underexpression and copy number loss (altogether 199 genes, termed as Group A); (ii) 46 genes with overexpression and promoter hypomethylation, and 522 genes with underexpression and promoter hypermethylation (altogether 568 genes, termed as Group B); (iii) 397 genes (termed as Group C) with somatic mutation and differential expression (115 overexpression and 282 underexpression). Genetic and epigenetic dysregulation of DEGs were shown in Fig. 3C. Consistent with classic knowledge of gene regulation, promoter methylation exerted trans-regulation, while DNA copy number exerted cis-regulation upon gene expression, and the promoter of DEGs tended to be hypermethylated in CRC (Fig. 3C).

The overlapping among these three gene groups was conducted, and hypergeometric distribution was used to assess the statistical significance. The formula of hypergeometric distribution is as follows:

$$p = 1 - F(x/N,\, K,\, M) = 1 - \sum_{i=0}^{x-1} \frac{\binom{K}{t}\binom{N-K}{M-t}}{\binom{N}{M}} \tag{3}$$

where $N$ is the number of all DEGs ($N = 4041$, the background gene number since all candidate genes were DEGs); $K$ is the gene number of one target gene groups; $M$ is the gene number of the other target gene group; $x$ is the number of common genes shared by the both gene groups. As shown in Supplementary Figure S2, the result of hypergeometric distribution test indicated that there was no significant overlapping between Group A and Group B ($p = 0.966$) or Group C ($p = 0.398$), while Group B significantly overlapped with Group C ($n = 107$, $p = 6.309\text{e-}13$).

**Random walk in developmental process related network.** GO analysis of aforementioned three gene groups indicated Group A was found no GO terms significantly enriched, whereas Group B (Fig. 3D, Supplementary Table S1) and Group C (Fig. 3E, Supplementary Table S2) were both significantly enriched with a variety of GO terms (Bonferroni adjusted $FDR < 1\text{e-}07$). The enriched GO terms were increasingly ordered with FDR value, and top 30 GO terms were shown in Fig. 3D,E. All the offspring GO terms of "developmental process" were highlighted in red. Among top 30 enriched GO terms, 76.67% (23/30) of these terms were the offspring of "developmental process" for both Group B and Group C. Moreover, 48.33% (232/480, Supplementary Table S1) of Group B genes and 52.39% (186/355, Supplementary Table S2) of Group C genes belonged to this GO term (Fig. 3D,E). Among the 107 overlapping genes between Group B and Group C, 54.2% (58/107) of these genes belonged to the GO term "developmental process.

Since DEGs with abnormal promoter methylation and somatic mutation were both functionally concentrated on developmental process, developmental process related genes (DPRG, $n = 5{,}161$) were extracted from GO term "GO: 0032502". Developmental process related network (DPRN) was established by extracting DPRGs and edges between DPRGs from the aforementioned merged network. The biggest connected component (BCC) of DPRN containing 3,271 DPRGs and 20,652 edges was established as walking graph for random walk (Fig. 4A). Genes in Group B or C and also present in the BCC were used as source nodes ($n = 249$). Genes only afflicted with dysregulated promoter methylation or somatic mutation were scored as 1, and genes afflicted with both abnormalities were scored as 2. The initial probability vector $p_0$ was obtained by normalizing the score vector ($n = 249$) so that

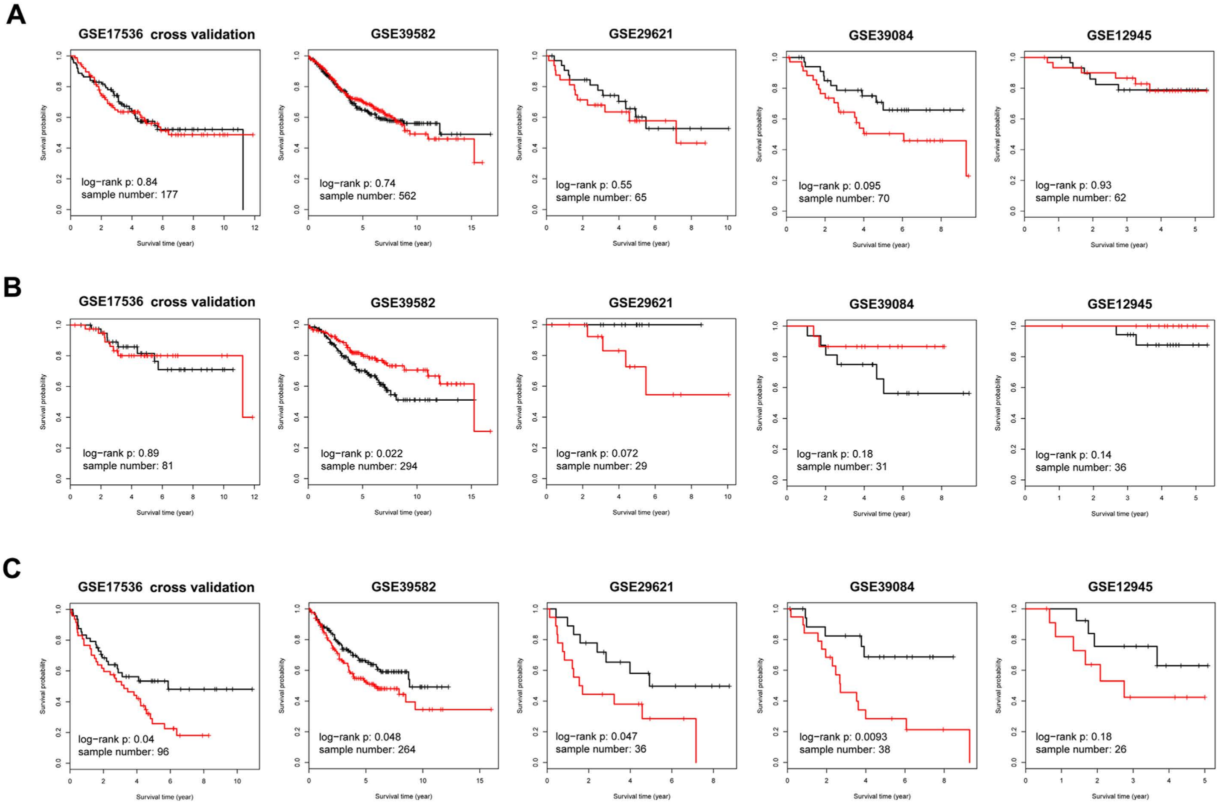

**Figure 5. Kaplan–Meier survival analysis of random-walk significant genes in five independent datasets of Clinicinfo.** (**a**) Kaplan–Meier survival analysis of 37 significant genes with all-stage samples in five independent data sets of Clinicinfo superset. GSE17536 was treated as training cohort, and five-fold cross validation was conducted to calculate risk score. Survival analysis was performed to discriminate OS between risk score assigned groups. (**b**) Kaplan–Meier survival analysis of 37 significant genes in five independent data sets with Stage I/II samples. (**c**) Kaplan–Meier survival analysis of 37 significant genes in five independent data sets with Stage III/IV samples. ***Note***: in Kaplan–Meier survival analysis, red curve represents the subgroup with higher risk score, and black curve represents lower risk score.

the sum of the vector is equal to 1 (the input of random walk algorithm). When the steady-state was finally reached, all the genes in the BCC (including 249 source nodes) were scored with $p^{\infty}$ (n = 3271, output of random walk algorithm), and thus the genes with significantly high score were mostly affected by both of these dysregulations. Therefore, 37 significant genes in respect to steady-stage probability were collected through 10,000 permutations (Fig. 4B), and algorithmically these genes received the most influence imposed by source genes with severe genetic and epigenetic dysregulations.

### Validation of significant genes' prognostic value via survival analysis.

We used GSE17536 in Clinicinfo superset as training cohort to train Cox regression model with 37 significant genes and then used the constructed model to evaluate the risk score of patients in test cohorts. Patients in each test data set were further divided into high risk and low risk subgroups based on the median of their risk score. Kaplan–Meier survival analysis was performed to evaluate the actual survival difference between the two risk score assigned groups in samples from all American Joint Committee on Cancer (AJCC) stages (Fig. 5A), Stage I/II (Fig. 5B), and Stage III/IV (Fig. 5C) in each data set, respectively. Risk score calculated in all stage and Stage I/II samples were not significantly or consistently associated with patient's OS in both self-cross validation and four individual test cohorts (Fig. 5A,B). However, patients with higher risk score in Stage III/IV patient groups tended to live significantly shorter than those with lower risk score. The ability of risk score to discriminate OS was quite satisfactory in Stage III/IV samples in each data set (Fig. 5C, GSE17536 cross validation, n = 96, $p = 0.04$; GSE39582, n = 264, $p = 0.048$; GSE29621, n = 36, $p = 0.047$; GSE39084, n = 38, $p = 0.0093$; GSE12945, n = 26, $p = 0.18$), suggesting the genes most influenced by promoter methylation dysregulation and somatic mutation probably hold great prognostic value in late stage CRC patients.

### Confirmation of the prognostic value of these 37 genes by means of meta-analysis and Cox regression analysis.

Meta-analysis of 37 significant genes and risk score in five Clinicinfo data sets also confirmed the result of survival analysis with both fixed-effect model (Fig. 6A) and random-effect model (Fig. 6B), corroborating the prognostic value of these significant genes in late stage (conducted with R package "meta"). Fixed-effect and random-effect model are the most commonly used methods in conducting meta-analysis. The

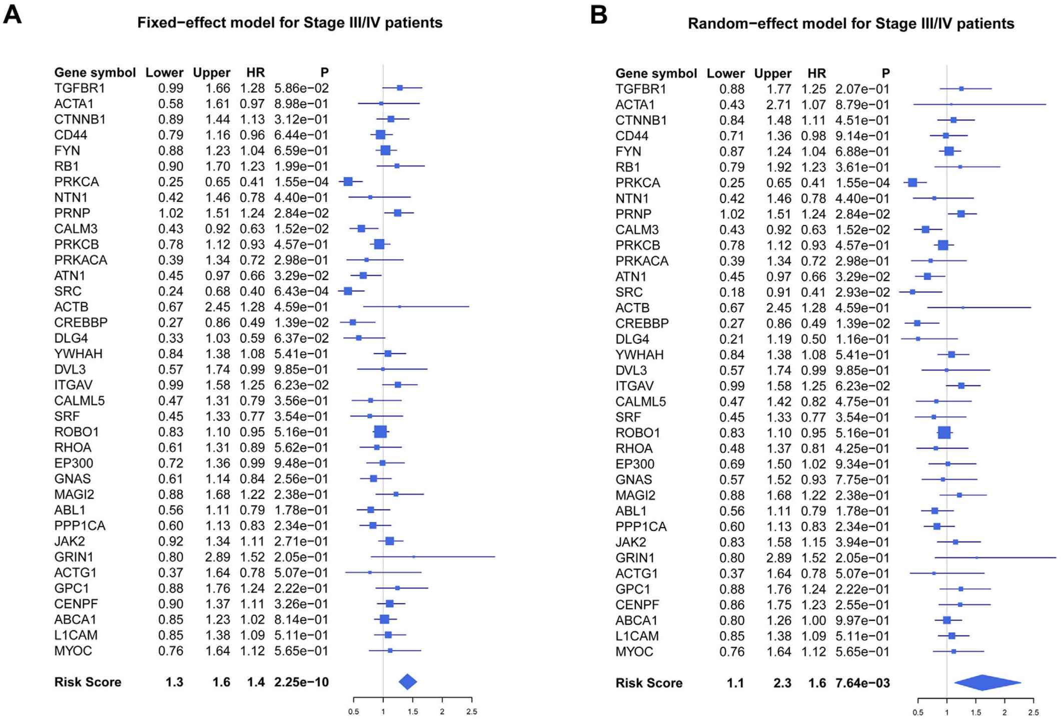

**Figure 6. Forest plots in Stage III/IV patients in terms of overall survival (OS).** (**a**) Forest plot of 37 significant genes with fixed-effect model with Stage III/IV patients in Clinicinfo superset. Meta-analysis of 37 significant genes in five independent data set of Clinicinfo superset was conducted, and hazard ratio (*HR*), 95% confidence interval (*CI*), and corresponding *p* value of each gene and risk score was calculated and plotted in the forest plot for Stage III/IV samples. (**b**) Forest plot of 37 significant genes with random-effect model.

| Factors | Univariate Cox regression | | Multivariate Cox regression | |
|---|---|---|---|---|
| | HR (95% CI) | *P* | HR (95% CI) | *P* |
| Age | 1.017 (0.998 ~ 1.037) | 0.076 | – | – |
| Gender (Male/ Female) | 1.245 (0.761 ~ 2.036) | 0.380 | – | – |
| Stage (IV/III) | 4.384 (2.671 ~ 7.194) | **7.894e-09** | 4.709 (2.839 ~ 7.812) | **1.960e-09** |
| Grade (III/I + II) | 1.620 (0.974 ~ 2.696) | 0.071 | – | – |
| Risk score | 2.225 (1.740 ~ 2.845) | **4.047e-10** | 2.223 (1.739 ~ 2.842) | **1.831e-10** |

**Table 2. Univariate and multivariate analyses of overall survival in late stage CRC patients.** Abbreviations: *HR* = hazard ratio; *CI* = confidence interval. Note: Significant *P* values were in bold ($P < 0.05$).

two models are different from the way of pooling the effect sizes obtained from the individual studies into an overall effect size. The fixed-effect model assumes that the differences between the studies are so important that during the effect-size pooling process, individual effect sizes should be retained; while random-effect model assumed that the individual trial effect sizes are "random" quantities[53,54]. Additionally, overall concordance index (C-index) analysis was also meta-analytically conducted to evaluate its OS predictive ability[55], and the result indicated that these 37 genes could significantly predict OS of late stage CRC patients (Supplementary Fig. S3). The Cox proportional hazards regression model was used to evaluate the independence of the prognostic factors in a stepwise manner (Table 2). We collected 122 Stage III/IV samples in Clinicinfo superset with definite information of OS, age, gender, stage and grade, and univariate Cox regression analysis indicated stage [hazard ratio (*HR*): 4.384; 95% confidence interval (*CI*): 2.671 ~ 7.194; *p* = 7.894e-09] and the risk score (*HR*: 2.225; 95% *CI*: 1.740 ~ 2.845; *p* = 4.047e-10) generated by these 37 significant genes were significantly associated with patient's OS. Multivariate Cox analysis indicated the risk score was an independent prognostic factor (*HR*: 2.223; 95% *CI*: 1.739 ~ 2.842; *p* = 1.831e-10).

## Discussion

The booming amount of high-throughput and multi-dimensional genomic data usher us into a new era, when the tremendously complicated molecular mechanism of carcinogenesis were perceived and dissected in a more integrative perspective. In this study, we systematically analyzed CRC genomic data, including CNV, somatic mutation, DNA promoter methylation and gene expression, to discover novel and important molecules and genomic dysregulations in a more comprehensive manner. Paired samples in TCGA database were used to identify differential gene expression and genetic or epigenetic abnormalities, respectively, and collected three groups of candidate genes with differential gene expression pattern and upstream corresponding dysregulations. The result of GO analysis indicated the functions of DEGs with abnormal promoter methylation (Group B) and somatic mutation (Group C) both majorly concentrated on developmental process, of which the outcome is an anatomical structure (which may be a subcellular structure, cell, tissue, or organ), or organism over time from an initial condition to a later condition[56]. Additionally, the DEGs with CNV didn't significantly overlap with the other DEG groups, while the majority of the significantly overlapping DEGs between Group B and Group C belonged to the GO term "developmental process" (Supplementary Fig. S2). These common DEGs shared by Group B and Group C play a pivotal role in both development and carcinogenesis. For instance, the germline gain-of-function mutation of *ALK* could disrupt the development of central nervous system[57], of which the same anomaly was also identified in sporadic and familial neuroblastoma cases[58–61]. *TIAM1*, expressed in the base of intestinal crypts, established a fundamental role for *Wnt*-signaling pathway in the development and maintenance of normal intestinal physiology[62]. Its expression was greatly elevated in mouse intestinal tumors and human colon adenomas, and the cross-talk between *TIAM1* and canonical *Wnt*-signaling pathways could significantly influence intestinal tumor formation and progression[63]. Based on GO and overlapping analyses, it is quite plausible that DEGs with aberrant promoter methylation and somatic mutation intimately cooperated together to facilitate the dysregulation of developmental process. DEGs with CNV, however, were not found functionally specific in terms of influencing certain biological process.

It has been more than 150 years since Rudolf Virchow first advocated that neoplasms arise "in accordance with the same law, which regulates development" in 1858. Emerging evidences supported the cellular behavioral similarity between ontogenesis and oncogenesis, for instance, in the process of epithelial-to-mesenchymal transition (EMT)[64], mesenchymal-to-epithelial transition (MET)[65] and immune-surveillance evasion[66]. The molecular resemblances have been documented between certain malignant tumors and developing tissues on the basis of transcription factor activity[67], regulation of chromatin structure[68] and cellular signaling[69]. Important molecules were reported to play substantial role in both development and carcinogenesis. For example, *PTCH1* is a key regulator of development, whose overexpression could drive skin carcinogenesis[70]. 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[71–73]. Many important signaling pathways, including *Notch1* signaling pathway, activated during development, are proven to be reactivated in the process of carcinogenesis[74,75]. In addition, there were some pioneering works discovering that mRNA and microRNA expression profile of cancer could recapitulate the expression pattern of development[72,76–79]. The intimate association between developmental process and carcinogenesis, together with astounding synchronization of promoter methylation dysregulation and somatic mutation in developmental process related genes (DPRGs), compelled us to propose the hypothesis that DPRGs affected most by the aberrance of promoter methylation and somatic mutation, probably hold meaningful explanation for the underlying mechanism of carcinogenesis, and might be intimately associated with clinicopathological characteristics, for instance, OS.

In our study, we adopted a simple and effective computational strategy to randomly walk DPRGs with aberrant promoter methylation or somatic mutation in HPRD and KEGG merged biological network. Random walk with restart was adopted to decipher gene to disease association in priori-knowledge based network, whose performance was proven to be much more superior to other methods, such as neighborhood approaches[80–82]. The advantage of this strategy is that it subtly combines observed multi-omic data with knowledge based regulatory network, tracing the information flow which would be greatly accumulated in significant genes.

The majority of these significant genes were connected to form a relatively compact biological module (Fig. 4B), implying enormous biological association existing among these genes. Many of these significant genes obtained through random walk algorithm were closely related to the initiation and progression of CRC. *TGFBR1* is a central molecule in *TGF-$\beta$* pathway, whose alteration could strikingly enhance the susceptibility to CRC[83]. The high microsatellite instability and expressional loss of *EP300* may be a feature of gastric and colorectal cancers[84]. *PRKCA* and *PRKCB* are both member of Protein kinase C (*PKC*) family, which have a role in cell proliferation, differentiation, angiogenesis, and apoptosis[85]. *PRKCB* inhibition by enzastaurin could lead to mitotic missegregation and preferential cytotoxicity toward colorectal cancer cells with chromosomal instability; loss of *PRKCA* signaling is a general characteristic of colorectal tumors regardless of other underlying genetic defects, pointing to the importance of this pathway[86].

Since candidate genes were collected based on aberrant patterns in multi-omic level of TCGA genomic data, we used microarray data sets with OS information from GEO database instead of TCGA to test the prognostic value of these significant genes. Recent expression profiling datasets lack of consistent results between the studies due to different technological platforms and lab protocols[87,88], and the microarray expression value of a particular genes could only be calculated based on different type of probes, 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 could cause the inconsistency of biological conclusions. To address these challenges, we collected five Affymetrix microarray data sets ($n = 940$, each sample number $>60$) with 22,277 common probes to get robust result of their significant clinical relevance. The expression value of 37 significant genes was retrieved and the prognostic value was evaluated with Cox regression model. The result indicated these 37 genes were significantly associated with OS in late stage (Stage III/IV) patients, rather than early stage (Stage I/II). According to AJCC staging system (7th edition)[89], the lesion of early stage CRC (Stage I/II) is relatively contained with neither lymph node invasion nor distant metastasis; when tumor advances 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 and Stage II patients only need to receive radical treatment to defuse the peril caused by molecularly chaotic tumors. However, with the deterioration of the disease, Stage III patients principally should be treated with neoadjuvant chemoradiation 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[90]. Surgical resection of the primary tumor is not beneficial for most of Stage IV patients[91,92]. Prognostic genes have the ability to predict patient's OS status, probably by means of exerting influence on or reflecting tumor encroachment in the patient. Suppose the tumor is completely removed from the patient, and then the expression of this gene signature would probably not 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 indicative function of prognostic genes continues monitoring the interaction between the residual neoplasms and CRC patients, probably explaining the question why these genes were only significantly associated with the OS of late stage CRC patients.

In summary, with the increasing availability of multidimensional genomic data, we collected genes with high rate of somatic mutation, differential expression, promoter methylation dysregulation and significant CNV, using paired samples in TCGA database. Three groups of DEGs with corresponding genetic or epigenetic abnormalities were obtained; the GO enrichment and overlapping analysis suggested DEGs with aberrant promoter methylation or somatic mutation were both functionally centering on developmental process. Random walk with restart was used to extract significant developmental genes most affected by aberrant promoter methylation and somatic mutation in merged regulatory network. In addition, the significant genes were closely related to OS of late stage patient. It is also very tempting that the identification of the functional regulators of these genes might be profusely beneficial to the discovery of new drug targets for CRC treatment. It is our hope that our preliminary exploration would be helpful for the further study upon cancer etiology and treatment guidance.

## References

1. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136,** E359–86 doi: 10.1002/ijc.29210 (2015).
2. Han, D. *et al.* Long noncoding RNAs: novel players in colorectal cancer. *Cancer Lett.* **361,** 13–21 doi: 10.1016/j.canlet.2015.03.002 (2015).
3. Gonzalez-Pons, M. & Cruz-Correa, M. Colorectal Cancer Biomarkers: Where Are We Now? *Biomed Res Int* doi:Artn 149014 doi: 10.1155/2015/149014 (2015).
4. Ferguson, L. R. *et al.* Genomic instability in human cancer: Molecular insights and opportunities for therapeutic attack and prevention through diet and nutrition. *Semin. Cancer Biol.* doi: 10.1016/j.semcancer.2015.03.005 (2015).
5. Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18,** 11–22, doi: 10.1016/j.ccr.2010.05.026 (2010).
6. Shenker, N. & Flanagan, J. M. Intragenic DNA methylation: implications of this epigenetic mechanism for cancer research. *Br. J. Cancer* **106,** 248–53, doi: 10.1038/bjc.2011.550 (2012).
7. Akhavan-Niaki, H. & Samadani, A. A. DNA methylation and cancer development: molecular mechanism. *Cell Biochem. Biophys.* **67,** 501–13, doi: 10.1007/s12013-013-9555-2 (2013).
8. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–52, doi: 10.1038/nature10983 (2012).
9. Domany, E. Using High-Throughput Transcriptomic Data for Prognosis: A Critical Overview and Perspectives. *Cancer Res.* **74,** 4612–4621, doi: 10.1158/0008-5472.Can-13-3338 (2014).
10. Leary, R. J. *et al.* Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl. Acad. Sci. USA* **105,** 16224–16229, doi: 10.1073/pnas.0808041105 (2008).
11. Despierre, E. *et al.* Somatic copy number alterations predict response to platinum therapy in epithelial ovarian cancer. *Gynecol. Oncol.* **135,** 415–422, doi: 10.1016/j.ygyno.2014.09.014 (2014).
12. Xu, H. T. *et al.* Non-invasive Analysis of Genomic Copy Number Variation in Patients with Hepatocellular Carcinoma by Next Generation DNA Sequencing. *Journal of Cancer* **6,** 247–253, doi: 10.7150/Jca.10747 (2015).
13. Silveira, S. M. *et al.* Genomic screening of testicular germ cell tumors from monozygotic twins. *Orphanet J. Rare Dis.* **9,** doi: Artn 181, doi: 10.1186/S13023-014-0181-X (2014).
14. Horpaopan, S. *et al.* Genome-wide CNV analysis in 221 unrelated patients and targeted high-throughput sequencing reveal novel causative candidate genes for colorectal adenomatous polyposis. *Int. J. Cancer* **136,** E578–E589, doi: 10.1002/Ijc.29215 (2015).
15. Liang, L., Fang, J. Y. & Xu, J. Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy. *Oncogene*, doi: 10.1038/onc.2015.209 (2015).
16. Davies, M. A. & Samuels, Y. Analysis of the genome to personalize therapy for melanoma. *Oncogene* **29,** 5545–5555, doi: 10.1038/Onc.2010.323 (2010).
17. Jiang, B. H. & Liu, L. Z. PI3K/PTEN signaling in tumorigenesis and angiogenesis. *Biochim. Biophys. Acta* **1784,** 150–8, doi: 10.1016/j.bbapap.2007.09.008 (2008).
18. Yuan, T. L. & Cantley, L. C. PI3K pathway alterations in cancer: variations on a theme. *Oncogene* **27,** 5497–5510, doi: 10.1038/Onc.2008.245 (2008).
19. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11,** 204–220, doi: 10.1038/Nrg2719 (2010).
20. Yuan, T. *et al.* An integrative multi-scale analysis of the dynamic DNA methylation landscape in aging. *PLoS Genet.* **11,** e1004996, doi: 10.1371/journal.pgen.1004996 (2015).
21. Docherty, S. J., Davis, O. S. P., Haworth, C. M. A., Plomin, R. & Mill, J. DNA methylation profiling using bisulfite-based epityping of pooled genomic DNA. *Methods* **52,** 255–258, doi: 10.1016/j.ymeth.2010.06.017 (2010).
22. Laird, P. W. The power and the promise of DNA methylation markers. *Nat Rev Cancer* **3,** 253–266, doi: 10.1038/Nrc1045 (2003).
23. Costello, J. F. & Plass, C. Methylation matters. *J. Med. Genet.* **38,** 285–303, doi: 10.1136/Jmg.38.5.285 (2001).
24. Baylin, S. B. Tying it all together: Epigenetics, genetics, cell cycle, and cancer. *Science* **277**, 1948–1949, doi: 10.1126/science.277.5334.1948 (1997).
25. Akhavan-Niaki, H. & Samadani, A. A. DNA Methylation and Cancer Development: Molecular Mechanism. *Cell Biochem. Biophys.* **67,** 501–513, doi: 10.1007/s12013-013-9555-2 (2013).
26. De Carvalho, D. D. *et al.* DNA Methylation Screening Identifies Driver Epigenetic Events of Cancer Cell Survival. *Cancer Cell* **21,** 655–667, doi: 0.1016/j.ccr.2012.03.045 (2012).
27. Deckers, I. A. *et al.* Promoter Methylation of CDO1 Identifies Clear-Cell Renal Cell Cancer Patients with Poor Survival Outcome. *Clin. Cancer Res.*, doi: 10.1158/1078-0432.CCR-14-2049 (2015).
28. Busche, S. *et al.* Integration of High-Resolution Methylome and Transcriptome Analyses to Dissect Epigenomic Changes in Childhood Acute Lymphoblastic Leukemia. *Cancer Res.* **73,** 4323–4336, doi: 10.1158/0008-5472.Can-12-4367 (2013).

29. Choudhury, J. H. & Ghosh, S. K. Promoter Hypermethylation Profiling Identifies Subtypes of Head and Neck Cancer with Distinct Viral, Environmental, Genetic and Survival Characteristics. *PLoS ONE* **10,** e0129808, doi: 10.1371/journal.pone.0129808 (2015).
30. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39,** 1256–60, doi: 10.1038/ng2123 (2007).
31. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470,** 59–65, doi: 10.1038/nature09708 (2011).
32. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–12, doi: 10.1038/nature08516 (2010).
33. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–54, doi: 10.1038/nature05329 (2006).
34. Ali Hassan, N. Z. *et al.* Integrated analysis of copy number variation and genome-wide expression profiling in colorectal cancer tissues. *PLoS ONE* **9,** e92553, doi: 10.1371/journal.pone.0092553 (2014).
35. Ouadid-Ahidouch, H., Rodat-Despoix, L., Matifat, F., Morin, G. & Ahidouch, A. DNA methylation of channel-related genes in cancers. *Biochim. Biophys. Acta*, doi: 10.1016/j.bbamem.2015.02.015 (2015).
36. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674, doi: 10.1016/j.cell.2011.02.013 (2011).
37. Xing, M. Molecular pathogenesis and mechanisms of thyroid cancer. *Nat Rev Cancer* **13,** 184–99, doi: 10.1038/nrc3431 (2013).
38. Tomczak, K., Czerwinska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **19,** A68–77, doi: 10.5114/wo.2014.47136 (2015).
39. Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70, doi: 10.1038/nature11412 (2012).
40. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–15, doi: 10.1038/nature10166 (2011).
41. Cancer Genome Atlas Research Network. Corrigendum: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **494,** 506, doi: 10.1038/nature11903 (2013).
42. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–25, doi: 10.1038/nature11404 (2012).
43. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499,** 43–9, doi: 10.1038/nature12222 (2013).
44. Chari, R., Coe, B. P., Vucic, E. A., Lockwood, W. W. & Lam, W. L. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Syst. Biol.* **4,** 67, doi: 10.1186/1752-0509-4-67 (2010).
45. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5,** 557–72, doi: 10.1093/biostatistics/kxh008 (2004).
46. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140, doi: 10.1093/bioinformatics/btp616 (2010).
47. Kohler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82,** 949–958, doi: 10.1016/j.ajhg.2008.02.013 (2008).
48. Nakao, M. *et al.* Identification of DNA copy number aberrations associated with metastases of colorectal cancer using array CGH profiles. *Cancer Genet. Cytogenet.* **188,** 70–6, doi: 10.1016/j.cancergencyto.2008.09.013 (2009).
49. Lassmann, S. *et al.* Array CGH identifies distinct DNA copy number profiles of oncogenes and tumor suppressor genes in chromosomal- and microsatellite-unstable sporadic colorectal carcinomas. *J Mol Med (Berl)* **85,** 293–304, doi: 10.1007/s00109-006-0126-5 (2007).
50. Jones, A. M. *et al.* Array-CGH analysis of microsatellite-stable, near-diploid bowel cancers and comparison with other types of colorectal carcinoma. *Oncogene* **24,** 118–29, doi: 10.1038/sj.onc.1208194 (2005).
51. Alcock, H. E., Stephenson, T. J., Royds, J. A. & Hammond, D. W. Analysis of colorectal tumor progression by microdissection and comparative genomic hybridization. *Gene Chromosome Canc* **37,** 369–80, doi: 10.1002/gcc.10201 (2003).
52. Lipska, L. *et al.* Tumor markers in patients with relapse of colorectal carcinoma. *Anticancer Res.* **27,** 1901–5 (2007).
53. Helfenstein, U. Data and models determine treatment proposals–an illustration from meta-analysis. *Postgrad. Med. J.* **78,** 131–4 (2002).
54. Senn, S. Trying to be precise about vagueness. *Stat. Med.* **26,** 1417–1430, doi: 10.1002/sim.2639 (2007).
55. Pencina, M. J. & D'Agostino, R. B. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat. Med.* **23,** 2109–2123, doi: 10.1002/sim.1802 (2004).
56. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).
57. de Pontual, L. *et al.* Germline gain-of-function mutations of ALK disrupt central nervous system development. *Hum. Mutat.* **32,** 272–6, doi: 10.1002/humu.21442 (2011).
58. Chen, Y. *et al.* Oncogenic mutations of ALK kinase in neuroblastoma. *Nature* **455,** 971–4, doi: 10.1038/nature07399 (2008).
59. Mosse, Y. P. *et al.* Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature* **455,** 930–5, doi: 10.1038/nature07261 (2008).
60. Janoueix-Lerosey, I. *et al.* Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. *Nature* **455,** 967–70, doi: 10.1038/nature07398 (2008).
61. George, R. E. *et al.* Activating mutations in ALK provide a therapeutic target in neuroblastoma. *Nature* **455,** 975–8, doi: 10.1038/nature07397 (2008).
62. Clarke, A. R. Wnt signalling in the mouse intestine. *Oncogene* **25,** 7512–21, doi: 10.1038/sj.onc.1210065 (2006).
63. Malliri, A. *et al.* The Rac activator Tiam1 is a Wnt-responsive gene that modifies intestinal tumor development. *J. Biol. Chem.* **281,** 543–548, doi: 10.1074/jbc.M507582200 (2006).
64. Nieto, M. A. Epithelial Plasticity: A Common Theme in Embryonic and Cancer Cells. *Science* **342,** 708-+, doi: 10.1126/science.1234850 (2013).
65. Eastham, A. M. *et al.* Epithelial-mesenchymal transition events during human embryonic stem cell differentiation. *Cancer Res.* **67,** 11254–11262, doi: 10.1158/0008-5472.Can-07-2253 (2007).
66. Ridolfi, L., Petrini, M., Fiammenghi, L., Riccobon, A. & Ridolfi, R. Human embryo immune escape mechanisms rediscovered by the tumor. *Immunobiology* **214,** 61–76, doi: 10.1016/j.imbio.2008.03.003 (2009).
67. Hartwell, K. A. *et al.* The Spemann organizer gene, Goosecoid, promotes tumor metastasis. *Proc. Natl. Acad. Sci. USA* **103,** 18969–18974, doi: 10.1073/pnas.0608636103 (2006).
68. Sparmann, A. & van Lohuizen, M. Polycomb silencers control cell fate, development and cancer. *Nat Rev Cancer* **6,** 846–856, doi: 10.1038/Nrcd1991 (2006).
69. Liu, S. L. *et al.* Hedgehog signaling and Bmi-1 regulate self-renewal of normal and malignant human mammary stem cells. *Cancer Res.* **66,** 6063–6071, doi: 10.1158/0008-5472.Can-06-0054 (2006).
70. Kang, H. C. *et al.* Ptch1 overexpression drives skin carcinogenesis and developmental defects in K14Ptch(FVB) mice. *J. Invest. Dermatol.* **133,** 1311–20, doi: 10.1038/jid.2012.419 (2013).
71. Kho, A. T. *et al.* Conserved mechanisms across development and tumorigenesis revealed by a mouse development perspective of human cancers. *Genes Dev.* **18,** 629–640, doi: 10.1101/Gad.1182504 (2004).
72. Liu, H. Y., Kho, A. T., Kohane, I. S. & Sun, Y. Predicting survival within the lung cancer histopathological hierarchy using a multi-scale genomic model of development. *PLoS Med.* **3,** 1090–1102, doi: Artn E232 doi: 10.1371/Journal.Pmed.0030232 (2006).

73. Kaiser, S. *et al.* Transcriptional recapitulation and subversion of embryonic colon development by mouse colon tumor models and human colon cancer. *Genome Biol.* **8,** doi: Artn R131, doi: 10.1186/Gb-2007-8-7-R131 (2007).
74. Rhim, A. D. & Stanger, B. Z. Molecular Biology of Pancreatic Ductal Adenocarcinoma Progression Aberrant Activation of Developmental Pathways. *Development, Differentiation and Disease of the Para-Alimentary Tract* **97,** 41–78 (2010).
75. Hu, H., Zhou, L., Awadallah, A. & Xin, W. Significance of Notch1-signaling pathway in human pancreatic development and carcinogenesis. *Appl. Immunohistochem. Mol. Morphol*. **21,** 242–7, doi: 10.1097/PAI.0b013e3182655ab7 (2013).
76. Hu, M. & Shivdasani, R. A. Overlapping gene expression in fetal mouse intestine development and human colorectal cancer. *Cancer Res.* **65,** 8715–22, doi: 10.1158/0008-5472.CAN-05-0700 (2005).
77. Borczuk, A. C. *et al.* Non-small-cell lung cancer molecular signatures recapitulate lung developmental pathways. *Am. J. Pathol.* **163,** 1949–1960, doi: 10.1016/S0002-9440(10)63553-5 (2003).
78. Kho, A. T. *et al.* Conserved mechanisms across development and tumorigenesis revealed by a mouse development perspective of human cancers. *Genes Dev.* **18,** 629–40, doi: 10.1101/gad.1182504 (2004).
79. Monzo, M. *et al.* Overlapping expression of microRNAs in human embryonic colon and colorectal cancer. *Cell Res.* **18,** 823–833, doi: 10.1038/Cr.2008.81 (2008).
80. Navlakha, S. & Kingsford, C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics* **26,** 1057–1063, doi: 10.1093/bioinformatics/btq076 (2010).
81. Wang, X. J., Gulbahce, N. & Yu, H. Y. Network-based methods for human disease gene prediction. *Brief Funct Genomics* **10,** 280–293, doi: 10.1093/Bfgp/Elr024 (2011).
82. Zhang, C. L. *et al.* Identification of miRNA-Mediated Core Gene Module for Glioma Patient Prediction by Integrating High-Throughput miRNA, mRNA Expression and Pathway Structure. *PLoS ONE* **9,** doi:ARTN e96908, doi: 10.1371/journal.pone.0096908 (2014).
83. Xu, Y. & Pasche, B. TGF-beta signaling alterations and susceptibility to colorectal cancer. *Hum. Mol. Genet.* **16 Spec No 1**, R14–20, doi: 10.1093/hmg/ddl486 (2007).
84. Kim, M. S., Lee, S. H. & Yoo, N. J. Frameshift mutations of tumor suppressor gene EP300 in gastric and colorectal cancers with high microsatellite instability. *Hum. Pathol.* **44,** 2064–70, doi: 10.1016/j.humpath.2012.11.027 (2013).
85. Ali, A. S., Ali, S., El-Rayes, B. F., Philip, P. A. & Sarkar, F. H. Exploitation of protein kinase C: a useful target for cancer therapy. *Cancer Treat. Rev.* **35,** 1–8, doi: 10.1016/j.ctrv.2008.07.006 (2009).
86. Hao, F. *et al.* Protein kinase Calpha signaling regulates inhibitor of DNA binding 1 in the intestinal epithelium. *J. Biol. Chem.* **286,** 18104–17, doi: 10.1074/jbc.M110.208488 (2011).
87. Chen, R. *et al.* A Meta-analysis of Lung Cancer Gene Expression Identifies PTK7 as a Survival Gene in Lung Adenocarcinoma. *Cancer Res.* **74,** 2892–2902, doi: 10.1158/0008-5472.Can-13-2775 (2014).
88. Goonesekere, N. C. W., Wang, X. S., Ludwig, L. & Guda, C. A Meta Analysis of Pancreatic Microarray Datasets Yields New Targets as Cancer Genes and Biomarkers. *PLoS ONE* **9,** doi: ARTN e93046, doi: 10.1371/journal.pone.0093046 (2014).
89. Edge, S. B. & American Joint Committee on Cancer. *AJCC cancer staging manual*, xiv, 648 p. (Springer, New York, 2010).
90. Ahmed, S., Johnson, K., Ahmed, O. & Iqbal, N. Advances in the management of colorectal cancer: from biology to treatment. *Int. J. Colorectal Dis.* **29,** 1031–42, doi: 10.1007/s00384-014-1928-5 (2014).
91. Benoist, S. *et al.* Treatment strategy for patients with colorectal cancer and synchronous irresectable liver metastases. *Br. J. Surg.* **92,** 1155–60, doi: 10.1002/bjs.5060 (2005).
92. Galizia, G. *et al.* First-line chemotherapy vs bowel tumor resection plus chemotherapy for patients with unresectable synchronous colorectal hepatic metastases. *Arch. Surg.* **143,** 352–8, discussion 358; doi: 10.1001/archsurg.143.4.352 (2008).

## 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) received by K.Z.

## Author Contributions

All authors have made substantial contributions to the conception and design of the study. N.A. and X.Y contributed to protocol design, search, data extraction, quality assessment, statistical analysis and writing the article. G.W., S.C. and K.Z. contributed to study design, interpretation of data and revision of the article. All authors have seen and approved the final version.

## Additional Information

**Supplementary information** accompanies this paper at http://www.nature.com/srep

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article**: An, N. *et al.* Developmental genes significantly afflicted by aberrant promoter methylation and somatic mutation predict overall survival of late-stage colorectal cancer. *Sci. Rep.* **5**, 18616; doi: 10.1038/srep18616 (2015).