Deng *et al. BMC Genomics* (2018) 19:138
https://doi.org/10.1186/s12864-018-4525-0

BMC Genomics

DATABASE Open Access

MethCNA: a database for integrating genomic and epigenomic data in human cancer

Gaofeng Deng[1], Jian Yang[1], Qing Zhang[2,3], Zhi-Xiong Xiao[1] and Haoyang Cai[1*]

Abstract

Background: The integration of DNA methylation and copy number alteration data promises to provide valuable insight into the underlying molecular mechanisms responsible for cancer initiation and progression. However, the generation and processing of these datasets are costly and time-consuming if carried out separately. The Illumina Infinium HumanMethylation450 BeadChip, initially designed for the evaluation of DNA methylation levels, allows copy number variant calling using bioinformatics tools.

Results: A substantial amount of Infinium HumanMethylation450 data across various cancer types has been accumulated in recent years and is a valuable resource for large-scale data analysis. Here we present MethCNA, a comprehensive database for genomic and epigenomic data integration in human cancer. In the current release, MethCNA contains about 10,000 tumor samples representing 37 cancer types. All raw array data were collected from The Cancer Genome Atlas and NCBI Gene Expression Omnibus database and analyzed using a pipeline that integrated multiple computational resources and tools. The normalized copy number aberration data and DNA methylation alterations were obtained. We provide a user-friendly web-interface for data mining and visualization.

Conclusions: The Illumina Infinium HumanMethylation450 BeadChip enables the interrogation and integration of both genomic and epigenomic data from exactly the same DNA specimen, and thus can aid in distinguishing driver from passenger mutations in cancer. We expect MethCNA will enable researchers to explore DNA methylation and copy number alteration patterns, identify key oncogenic drivers in cancer, and assist in the development of targeted therapies. MethCNA is publicly available online at http://cgma.scu.edu.cn/MethCNA.

Keywords: Copy number aberration, DNA methylation, Data integration, Cancer, Infinium HumanMethylation450 BeadChip, Genomic data, Epigenomic data

Background

Genomic instability is a hallmark of malignant tumors, causing DNA copy number changes in most cancer types [1, 2]. These copy number aberrations (CNAs) are important influential factors for altered gene expression levels in cancer [3]. Genomic alterations may confer growth advantage of cancer cells and are usually associated with poor prognosis. Recurrent CNA occurs in multiple tumor samples across the same genomic region and plays crucial roles in tumorigenesis [4]. The characterization of CNAs in various cancer types has lead to identification of a large number of genes that contribute to cancer initiation and progression [5–7].

In addition to copy number aberrations, DNA methylation is an important regulator of gene transcription, and is one of the most studied epigenetic modifications [8]. The methylated cytosines are almost exclusively located in CpG dinucleotide sequences [9]. CpGs are uniformly distributed across the genome, and some of them are concentrated in short regions named CpG islands. Methylation in CpG islands within gene promoters usually leads to gene silencing. Association of altered DNA methylation patterns of the promoter CpG islands with the expression profile of cancer genes has

* Correspondence: haoyang.cai@scu.edu.cn
[1]Center of Growth, Metabolism, and Aging, Key Laboratory of Bio-Resources and Eco-Environment, College of Life Sciences, Sichuan University, Chengdu, Sichuan 610064, China
Full list of author information is available at the end of the article

been found in many tumor types [10–12]. Aberrant hypomethylation may induce genome instability and overexpression of oncogenes, while hypermethylation in promoter regions of tumor suppressor genes may perturb cell cycle regulation, apoptosis and DNA repair, and result in malignant cellular transformation [13]. Therefore, DNA methylation status may serve as an epigenetic biomarker for cancer diagnosis and prognosis, and has been studied extensively in human cancer.

The integration of different 'omics' data types is an increasingly important approach for understanding the fundamental mechanisms of cancer development. Genomic and epigenomic alterations are key regulators of gene expression, and may act in concert to drive tumorigenesis and promote progression towards a malignant phenotype [14]. Genes that are both amplified and hypomethylated or deleted and hypermethylated are likely to play a key role in cancer development. DNA methylation and copy number alteration data from the same tumor specimen may facilitate to elucidate the synergistic mechanisms for the inactivation of tumor suppressor genes or the activation of oncogenic pathways [6, 15]. Copy number aberration profiles and DNA methylation patterns can be measured genome-wide with microarrays. Although at present arrays offer the most cost-effective solutions for producing genomic and epigenomic data, the costs remains a major concern for the large-scale assessment of multiple datasets. Moreover, the computational burden and storage requirements are increased substantially if these data are generated on separate array platforms.

The Illumina Infinium HumanMethylation450 BeadChip (450 k) is based on similar biochemical reaction principle and technology as the single nucleotide polymorphism (SNP) arrays, and is able to robustly detect CNAs as a no-cost byproduct of methylation studies [16, 17]. A variety of bioinformatics tools have been developed for copy number calling from the 450 k methylation array. The detection of copy number and methylation alterations in a single experiment is particularly important when considering the potential effects of tumor heterogeneity on cancer development [17]. The subtle areas of a tumor may be genetically and epigenetically different, and thus confer a different phenotypic trait, such as differing metastatic potential. In recent years, a growing number of studies have generated genome-wide DNA methylation profiles of thousands of cancer samples using Infinium HumanMethylation450 platform, including The Cancer Genome Atlas (TCGA) project, which represents one of the largest efforts to systematically characterize the molecular profiles of cancers [18]. These data are valuable resources for meta-analysis and may provide insight into molecular mechanisms underlying tumorigenesis. However, the processing and integration of these datasets are laborious and time-consuming.

Here we present MethCNA, a comprehensive database for the integrated analysis of DNA methylation and copy number alterations in human cancer. Currently, the database contains about 10,000 publicly available tumor samples interrogated by Infinium HumanMethylation450 BeadChip. The raw array data were collected from TCGA and NCBI Gene Expression Omnibus (GEO) database [19] and processed through a pipeline that integrates several widely used Illumina Infinium 450 k array analysis tools. We developed a user-friendly web-interface and online tools for data access, analysis and visualization. 'Omics' data integration and exploration hold great promise for the identification of novel cancer genes, and will enable the development and selection of optimal therapies targeting driver aberrations. MethCNA is designed to meet the growing interest in integrating 'omics' data, and is a resource for large-scale integration analysis of genomic and epigenomic data in human cancer.

Construction and content

Data source for MethCNA

MethCNA includes genome-wide copy number alterations and DNA methylation profiles, which were simultaneously interrogated by the Illumina Infinium HumanMethylation450 BeadChip. Over 10,000 publicly available tumor samples were collected from TCGA and GEO database. Our data selection criteria are that (1) the tumor samples must be assayed by Illumina Infinium HumanMethylation450 BeadChip; (2) the raw signal intensity data (.IDAT) files must be downloadable for re-analysis and (3) the tumor type matched tissue-specific normal samples are available as a reference for data analysis. In the current release of MethCNA, we collected 28 and 30 data series from TCGA and GEO, respectively. These data series consist of 9964 tumor samples across 37 cancer types. Cancers were defined by their histological types and sites of origin. To provide standardized information on cancer types, International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) morphology and topography terms and codes were assigned to each tumor sample (Additional file 1: Table S1 and Additional file 2: Table S2). For five tumor types in TCGA (Brain Lower Grade Glioma [LGG], Acute Myeloid Leukemia [LAML], Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC], Mesothelioma [MESO], Testicular Germ Cell Tumors [TGCT]), the normal samples could not be obtained, and thus are not included in our database. Detailed statistics of cancer types and samples are shown in Table 1.

Data processing pipeline

To reanalyze the raw array data, we downloaded signal intensity data (.IDAT) files from public microarray

Table 1 Summary of array data contained in MethCNA

Cancer type	TCGA		GEO	
	Sample	Series	Sample	Series
Acute myeloid leukemia	0	0	88	3
Adrenocortical carcinoma	80	1	0	0
Atypical teratoid rhabdoid tumor	0	0	150	1
Bladder urothelial carcinoma	419	1	0	0
Breast invasive carcinoma	796	1	316	3
Cervical cancer	309	1	0	0
Cholangiocarcinoma	36	1	0	0
Colon adenocarcinoma	316	1	104	2
Diffuse intrinsic pontine glioma	0	0	25	1
Esophageal carcinoma	186	1	0	0
Glioblastoma multiforme	153	1	72	1
Head and neck squamous cell carcinoma	530	1	0	0
Kidney Chromophobe	65	1	0	0
Kidney renal clear cell carcinoma	325	1	46	1
Kidney renal papillary cell carcinoma	276	1	0	0
Liver hepatocellular carcinoma	380	1	0	0
Lung adenocarcinoma	475	1	164	1
Lung squamous cell carcinoma	370	1	0	0
Medulloblastoma	0	0	119	2
Neuroblastoma	0	0	32	1
Oligodendroglioma	0	0	46	1
Ovarian adenocarcinoma	10	1	345	3
Pancreatic adenocarcinoma	185	1	167	1
Pheochromocytoma and Paraganglioma	184	1	0	0
Prostate adenocarcinoma	503	1	20	3
Rectum adenocarcinoma	99	1	6	1
Sarcoma	265	1	0	0
Schwannoma	0	0	40	1
Skin cutaneous melanoma	473	1	160	2
Small cell lung carcinoma	0	0	11	1
Squamous cell skin carcinoma	0	0	7	1
Stomach adenocarcinoma	396	1	0	0
Thymoma	124	1	0	0
Thyroid carcinoma	515	1	0	0
Uterine carcinosarcoma	57	1	0	0
Uterine corpus endometrial carcinoma	439	1	0	0
Uveal melanoma	80	1	0	0
Total	8046	28	1918	30

repositories. The data processing pipeline integrates a collection of widely used R packages, including minfi [20], limma [21], ChAMP [22] and CopyNumber450kCancer [23]. A schematic overview of the data processing pipeline is shown in Fig. 1. For all these packages, we used the default settings of the tools. The data files were processed at a per data series level. For each data series, the raw array files were run through the pipeline in batch mode, which integrates several analysis steps from preprocessing (e.g. normalization and batch effect analysis) to basic analysis (e.g. copy number aberrations calling and detection of differentially methylated regions) as well as tool for cancer-specific analysis (e.g. baseline correction for accurate copy number calling). The Bioconductor package minfi was used to perform quality control and generate normalized beta values [20]. The Subset-quantile within array normalization (SWAN) [24] and Beta-Mixture Quantile (BMIQ) [25] normalization methods were employed. The differentially methylated probes and regions were identified by limma package [21]. Benjamini-Hochberg method was used for *p*-value adjustment. For detecting copy number alterations from the Illumina 450 k array platform, we employed the R package ChAMP, which infers copy number changes using intensity of individual and surrounding probes [22]. We used the Beta-value method for differential methylation analysis, since the Beta-value has an intuitive biological interpretation [26]. However, for data series that less than 30 samples, we employed M-value method for the analysis, because in the limit of small sample size, M-values allow more reliable identification of true positives [27]. In order to optimize CNA calling, we integrated the CopyNumber450kCancer R package into the pipeline for baseline estimation and correction using the maximum density peak estimation (MDPE) method [23]. For TCGA datasets, we performed batch effect reduction by ComBat [28] method in the sva R package [29]. We mainly focused on two covariates, namely tissue source sites (TSS), which may introduce bias during sample preparation, and Slides, which may cause bias in data generation process. We treated Slide as batch variable and TSS as covariate. For cases that TSS are confounded with Slide, we only used Slide as batch variable for the correction. This correction was applied to both beta values and signal intensities, which were used for differentially methylated regions calling and copy number alterations calling, respectively. Batch correction was followed by manual inspection of quality control plots generated by ChAMP to evaluate the correction performance.

For TCGA datasets, the tumor type abbreviations were assigned as data series IDs, and the TCGA barcodes were extracted from the annotations file and used as sample IDs. A full list of sample and series IDs, cancer types and related publications is maintained on the MethCNA website. These information can also be found in Additional file 1: Table S1 and Additional file 2: Table S2. For gene level analysis, the standard gene names and locations were downloaded from the UCSC Genome Browser FTP server [30]. All genome

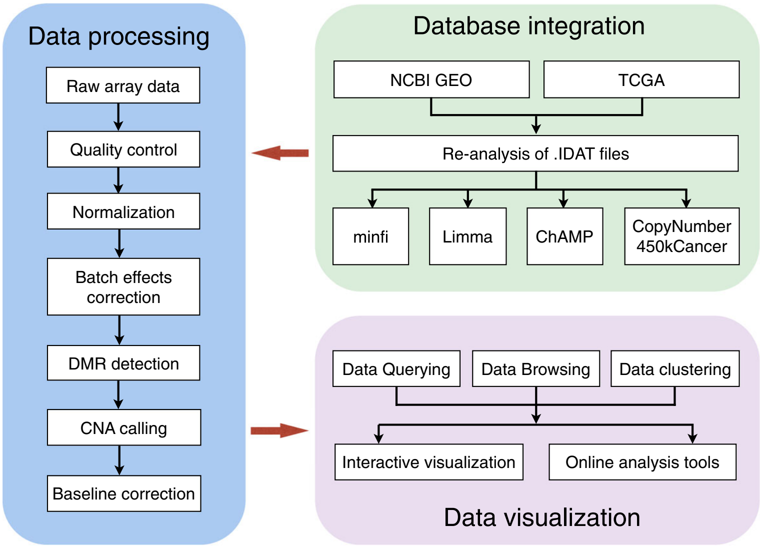


Fig. 1 The schematic overview of the data processing pipeline. It consists of three major modules. The raw array data are processed with a series of bioinformatics tools, and the normalized data are used for visualization and analysis

coordinates were based on human genome assembly NCBI Build 37/hg19. Since cancer-related DNA methylation studies concentrate on different biologically relevant genome regions, we provide several gene and CpG island regions for data analysis, including promoter, transcriptional start site (TSS), untranslated region (UTR), exon, gene body, CpG island, shore and shelf. The genomic annotations of these functional regions were obtained from the UCSC Genome Browser [30].

Database architecture and implementation details

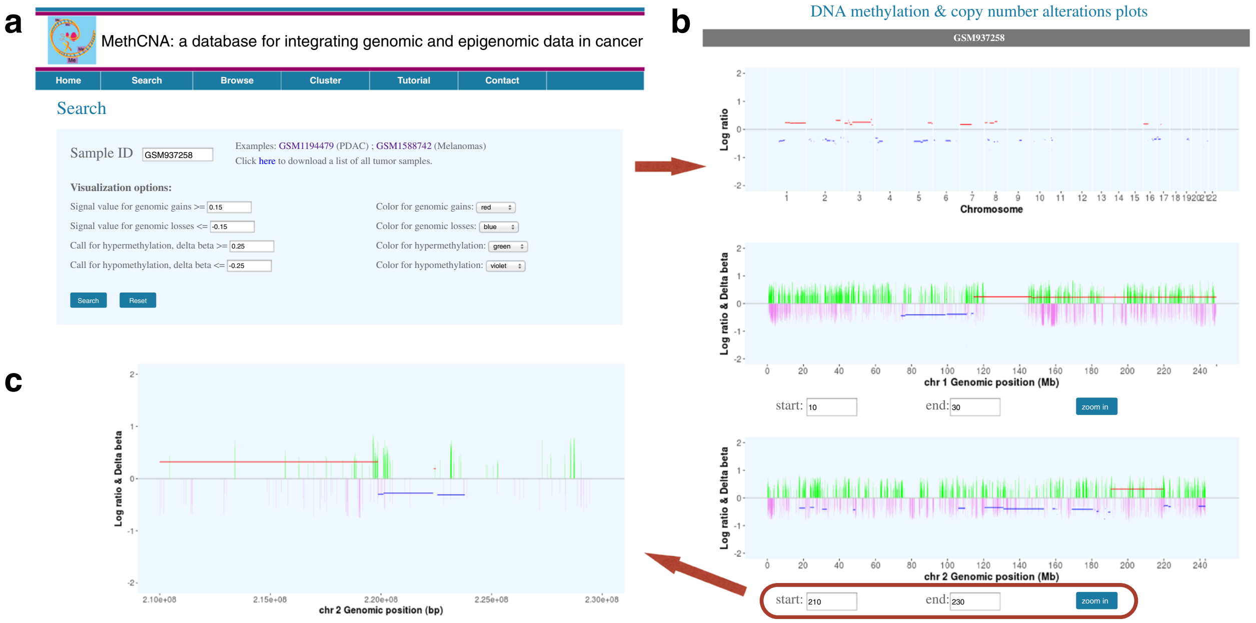
The normalized DNA methylation data at both the probe and gene level, and the called copy number states (gain, loss or neutral) of the segmentation data were stored in a MySQL database (version 5.5.49). We developed a user-friendly web interface for users to query and visualize the processed array data. The main functions of MethCNA include search, browse and clustering analysis. The web server runs on a dedicated Linux machine with the Apache HTTP server version 2.4.7. The web application used PHP (version 5.5.9) and HTML to serve web pages. The client-side interactive user interface was designed using JavaScript libraries and jQuery plugins, and the ggplot2 R package [31] and self-written R scripts were used for data visualization. The website has been tested to ensure functionality across different operating systems and browsers, including Internet Explorer, Safari, Firefox, Opera and Chrome.

Utility and discussion

Data querying

MethCNA contains 58 data series consisting of 9964 tumor samples across 37 cancer types. The 'Search' page provides two options for users to explore data of interest: query by array ID and query by gene. For query by ID, GEO array accession number or TCGA barcode can be input directly into the query box to retrieve copy number and methylation data of the corresponding array. The entire list of array IDs and related information is provided on MethCNA web page. The search interface offers a set of parameters or thresholds for customized data visualization (Fig. 2a). The thresholds for signal value represent cut-offs from which a segment is considered a genomic gain or loss. The delta beta values are used to calculate the percentage methylation difference between cancer and normal tissue. The positive and negative values of delta beta correspond to relative hyper- and hypomethylation, respectively. The empirical optimal thresholds for the Infinium 450 K platform are displayed as defaults. A set of colors can be used to intuitively display both types of data in the same figure. The results page shows CNAs and differentially methylated positions (DMP) for each chromosome (Fig. 2b). The interactive visualization interface allows users to zoom-in to an area of interest by inputting the start and stop genomic locations (Fig. 2c). This will facilitate to identify potential genes that are affected by both CNAs and DNA methylation.

The second option is query by genes, which allows users access to gene-specific aberration information in selected datasets (Fig. 3a). Cancer types and data series can be chosen in the list box. The data series field supports multiple selections, and the field content changes dynamically according to the selected cancer type. Gene names can be input in the text field, and multiple gene names may be specified separated with a semicolon. The



MethCNA: a database for integrating genomic and epigenomic data in cancer

Search

Sample ID GSM937258

Examples: GSM1194479 (PDAC) ; GSM1588742 (Melanomas)
Click here to download a list of all tumor samples.

Visualization options:

Signal value for genomic gains >= 0.15
Signal value for genomic losses <= -0.15
Call for hypermethylation, delta beta >= 0.25
Call for hypomethylation, delta beta <= -0.25

Color for genomic gains: red
Color for genomic losses: blue
Color for hypermethylation: green
Color for hypomethylation: violet

Search Reset

start: 10 end: 30

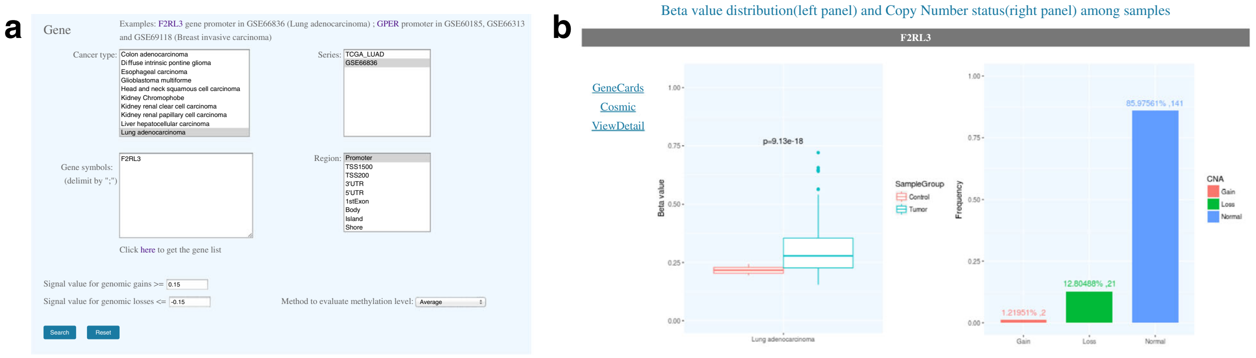
start: 210 end: 230

Fig. 2 The web interface for query by sample. **a** A set of parameters and thresholds can be used for customized data visualization. **b** The result page displays both copy number aberrations and differentially methylated positions. **c** Areas of interest can be zoomed in for more detail

standard gene names and symbols can be downloaded from MethCNA web page. The thresholds for defining genomic gains and losses are also provided. There are seven gene regions and four CpG island (CGI) regions that can be selected for analysis. The gene promoter region is assigned as default. If genetic and epigenetic variant data is available for the inquired genes, the results page shows the differential methylation status of each gene in selected tumors by boxplot, and the frequency of genomic gains and losses by histogram (Fig. 3b). The 'View Detail' link opens a new page to show detailed alteration data about the gene of interest in each sample. For further information of gene annotations, the result page provides links to the corresponding entries of GeneCards [32] and Catalogue of Somatic Mutations in Cancer (COSMIC) databases [33]. This feature will allow researchers to identify promising genetic and epigenetic biomarkers in human cancer.

Data browsing

MethCNA also provides data browsing interface to allow researchers to investigate mutation profiles of each study. The 'Browse' page contains all MethCNA data organized by studies. Clicking on the article title leads to a page that includes a table with detailed information of the dataset and genome-wide frequency plots of copy number alterations across all samples in the study (Fig. 4a). The frequency plots for each chromosome provide an intuitive view of regional chromosomal rearrangement hotspots. The user can also access



Gene

Examples: F2RL3 gene promoter in GSE66836 (Lung adenocarcinoma) ; GPER promoter in GSE60185, GSE66313 and GSE69118 (Breast invasive carcinoma)

Cancer type: Colon adenocarcinoma; Diffuse intrinsic pontine glioma; Esophageal carcinoma; Glioblastoma multiforme; Head and neck squamous cell carcinoma; Kidney Chromophobe; Kidney renal clear cell carcinoma; Kidney renal papillary cell carcinoma; Liver hepatocellular carcinoma; Lung adenocarcinoma

Series: TCGA_LUAD; GSE66836

Gene symbols: (delimit by ";") F2RL3

Click here to get the gene list

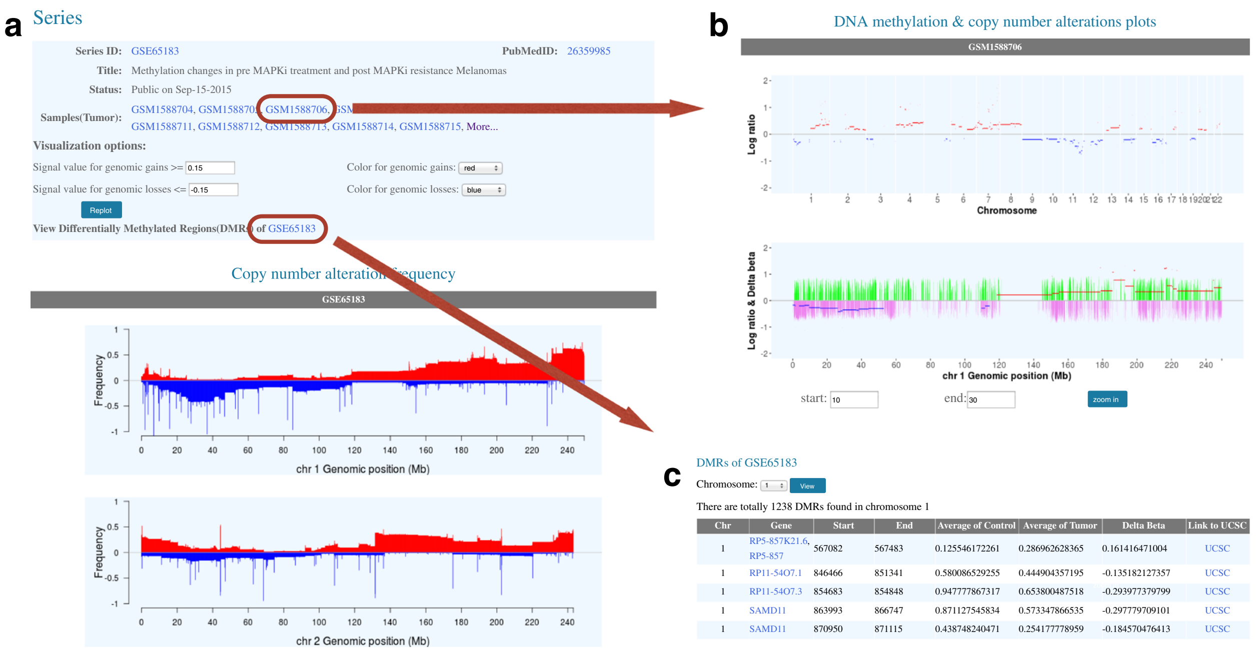
Region: Promoter; TSS1500; TSS200; 3'UTR; 5'UTR; 1stExon; Body; Island; Shore

Signal value for genomic gains >= 0.15
Signal value for genomic losses <= -0.15

Method to evaluate methylation level: Average

Search Reset

Fig. 3 The web interface for query by gene. **a** The interface to select cancer type, data series and gene regions. **b** The differential methylation status and the frequency of genomic imbalances of a gene are illustrated by boxplot and histogram, respectively


a
b
c
DNA methylation & copy number alterations plots
GSM1588706
Log ratio
Chromosome
Log ratio & Delta beta
chr 1 Genomic position (Mb)
Copy number alteration frequency
GSE65183
Frequency
chr 1 Genomic position (Mb)
chr 2 Genomic position (Mb)
Replot
zoom in
View

Series

Series ID: GSE65183 PubMedID: 26359985

Title: Methylation changes in pre MAPKi treatment and post MAPKi resistance Melanomas

Status: Public on Sep-15-2015

Samples(Tumor): GSM1588704, GSM158870, GSM1588706, GSM
GSM1588711, GSM1588712, GSM1588713, GSM1588714, GSM1588715, More...

Visualization options:

Signal value for genomic gains >= 0.15 Color for genomic gains: red

Signal value for genomic losses <= -0.15 Color for genomic losses: blue

View Differentially Methylated Regions(DMRs) of GSE65183

start: 10 end: 30

DMRs of GSE65183

Chromosome: 1

There are totally 1238 DMRs found in chromosome 1

Chr	Gene	Start	End	Average of Control	Average of Tumor	Delta Beta	Link to UCSC
1	RP5-857K21.6, RP5-857	567082	567483	0.125546172261	0.286962628365	0.161416471004	UCSC
1	RP11-54O7.1	846466	851341	0.580086529255	0.444904357195	-0.135182127357	UCSC
1	RP11-54O7.3	854683	854848	0.947777867317	0.653800487518	-0.293977379799	UCSC
1	SAMD11	863993	866747	0.871127545834	0.573347866535	-0.297779709101	UCSC
1	SAMD11	870950	871115	0.438748240471	0.254177778959	-0.184570476413	UCSC

Fig. 4 An example of GEO data series browsing. **a** The copy number gain and loss frequencies for each chromosome across the data series (only chromosome 1 and 2 are shown here). **b** Illustration of sample-level data. **c** The annotations of the differentially methylated regions and the associated genes

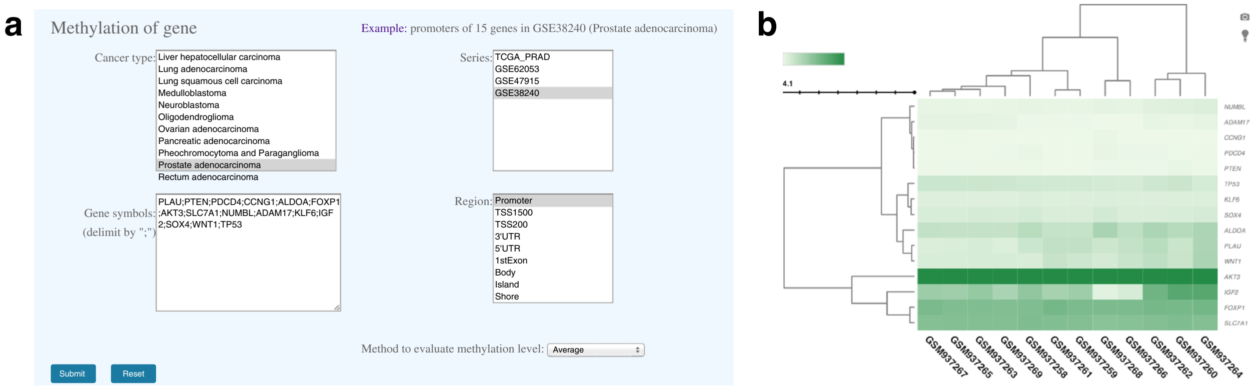
sample-level data by clicking on the link for each array ID (Fig. 4b). The result page contains a link to view the annotations of the differentially methylated regions (DMR), as well as detailed information about the genes that overlap with these regions (Fig. 4c). The DMR page also generates links to UCSC Genome Browser [30] and COSMIC database [33] for further annotation.

Data clustering

Most tumors display genome-wide abnormal genetic and epigenetic events, which provide their genomes with complex mutational patterns. Hierarchical clustering is a widely used method for large-scale genomic data analysis. The 'Cluster' page of MethCNA provides clustering of copy number alterations and DNA methylation data for users to characterize different aberration patterns in data series. The clustering of CNA and methylation data will be investigated separately. For CNA data, the users are able to investigate tumor type specific datasets. The data series can be chosen in the list box according to cancer types. In the clustering graph, tumor samples with similar CNA profiles are grouped together. The regional hotspots can be seen intuitively, and it is particularly helpful when comparing CNA profiles across multiple datasets. Investigation of these hotspots has proven to be an effective methodology to identify novel cancer genes. Furthermore, the identified clusters may represent distinct cancer subtypes. For DNA methylation profiles, we developed gene-level clustering analysis. The user can input a list of gene symbols and specify cancer type and gene region to run the analysis (Fig. 5a). The result page will return a clustering diagram to show the methylation profile of queried genes among the selected samples (Fig. 5b). The interactive diagram can be zoomed in for more detailed information. It enables researchers to investigate the synergistic effect of aberrant methylation of multiple genes in tumorigenesis.

Case study

As an illustration of MethCNA functionalities, we analyzed two glioblastoma (GBM) datasets that were integrated into our database: GSE60274 and TCGA-GBM. There are 72 and 153 tumor samples in GEO and TCGA datasets, respectively. In the 'Browse' page, clicking on the series title of GSE60274 opens a result page with basic information of the dataset. It also contains histograms of each chromosome representing the CNA frequencies. The genome-wide profile of chromosomal alteration hotspots may point to genomic loci harbouring cancer related genes. We found that the whole chromosome 7 gain and chromosome 10 loss were the most prevalent genomic imbalances in this dataset (Fig. 6a). According to recent studies, these alterations are the evolutionary first driver events in the development of primary GBM [34, 35]. Furthermore, we observed a recurrent focal deletion of the short arm of chromosome 9 (Fig. 6a). The peak region of focal deletion on chromosome 9p encompassing the tumor suppressor gene CDKN2A, which may



Methylation of gene

Example: promoters of 15 genes in GSE38240 (Prostate adenocarcinoma)

Cancer type: Liver hepatocellular carcinoma
Lung adenocarcinoma
Lung squamous cell carcinoma
Medulloblastoma
Neuroblastoma
Oligodendroglioma
Ovarian adenocarcinoma
Pancreatic adenocarcinoma
Pheochromocytoma and Paraganglioma
Prostate adenocarcinoma
Rectum adenocarcinoma

Series: TCGA_PRAD
GSE62053
GSE47915
GSE38240

Gene symbols: (delimit by ";") PLAU;PTEN;PDCD4;CCNG1;ALDOA;FOXP1;AKT3;SLC7A1;NUMBL;ADAM17;KLF6;IGF2;SOX4;WNT1;TP53

Region: Promoter
TSS1500
TSS200
3'UTR
5'UTR
1stExon
Body
Island
Shore

Method to evaluate methylation level: Average

Fig. 5 An example of hierarchical clustering of DNA methylation data. **a** The interface for user to input gene symbols and select cancer type, data series and gene regions. **b** The clustering diagram shows the methylation profile of queried genes among the selected samples. The specific region can be zoomed in for an enlarged view

play a crucial role to promote GBM progression [36, 37]. CDKN2A has been reported as an indicator of poor prognosis and is hypothesized to be a cancer driver gene [38]. The similar CNA frequency profile was observed in the TCGA-GBM dataset (Additional file 3: Figure S1).

The heatmaps of both datasets can be inspected on the 'Cluster' page. For example, we selected 'Glioblastoma' and 'GSE60274' in the 'Cancer type' and 'Series' list box, respectively. The overall CNA profile of the 72 tumor samples in GSE60274 was shown. The high prevalence of whole chromosome 7 and 10 abnormalities can be clearly seen (Fig. 6b). The distribution of focal aberrations can also be explored. For the TCGA-GBM dataset, we got the similar overall pattern of results (Additional file 4: Figure S2). The specific pattern of genomic copy number profiles indicates the non-random occurrence of CNAs, and may provide important information for understanding the molecular mechanisms involved in tumorigenesis.

We also investigated the synergistic effects of genomic aberrations and epigenetic changes in GSE60274. In the 'query by genes' interface of 'Search' page, we searched for three genes that are implicated in cancer initiation or progression: EGFR, PTEN and HOXA10. We selected 'Glioblastoma' and 'GSE60274' in 'Cancer type' and 'Series' list box, respectively. In the 'Gene symbols' text field, we input the above gene names separated by semicolon. For the gene region, we selected 'Promoter' in the drop-down list for analysis. In the results page, box plot and histogram were generated to present DNA methylation and copy number status for each gene across glioblastoma samples. According to the results, EGFR promoter region was significantly hypomethylated ($P < 0.01$, two-sided t test) in glioblastoma, as compared with normal brain specimens (Fig. 6c). Simultaneously, the copy number of EGFR was increased in most samples (63 out of 72 samples, 87.5%). EGFR is a transmembrane tyrosine kinase receptor that plays a central role in regulating cell proliferation and differentiation. Overexpression of EGFR has been reported and implicated in the pathogenesis of many cancer types [39, 40]. On the contrary, the tumor suppressor gene PTEN was inactivated by hypermethylation ($P < 0.001$, two-sided t test) and genomic loss (54 out of 72 samples, 75%) throughout the dataset (Fig. 6d). The genomic and epigenomic alterations affecting the expression of PTEN may indicate Knudson's 'two-hit' hypothesis in tumorigenesis [41], and were reported to be crucial driver events in glioblastoma initiation and progression [42, 43]. Interestingly, the HOXA10 gene was found to be hypermethylated ($P < 10–16$, two-sided t test) and showed copy number gain (55 out of 72 samples, 76%) in the dataset (Fig. 6e). It revealed the theory that gene dosage may contribute to the aberrant gene expression, as reported recently by Kurscheid et al. [44]. Our results demonstrate the utility of MethCNA in integrating genomic and epigenomic alteration data and the ability to identify cancer driver genes.

Conclusion

MethCNA is an effort to further our understanding of the relationship between copy number alterations and DNA methylation status, both of which are known to be hallmarks of human cancer. There are several publicly available resources similar to our database. The cBioPortal [45] is a resource for interactive exploration of multidimensional cancer genomics data sets, including CNA and DNA methylation data. CNAmet [46] is an R package for integrative analysis of high-throughput copy number, DNA methylation and gene expression data. However, these tools were designed to integrate data derived from different patients or platforms. The most important advantage of our database is that the methylation and

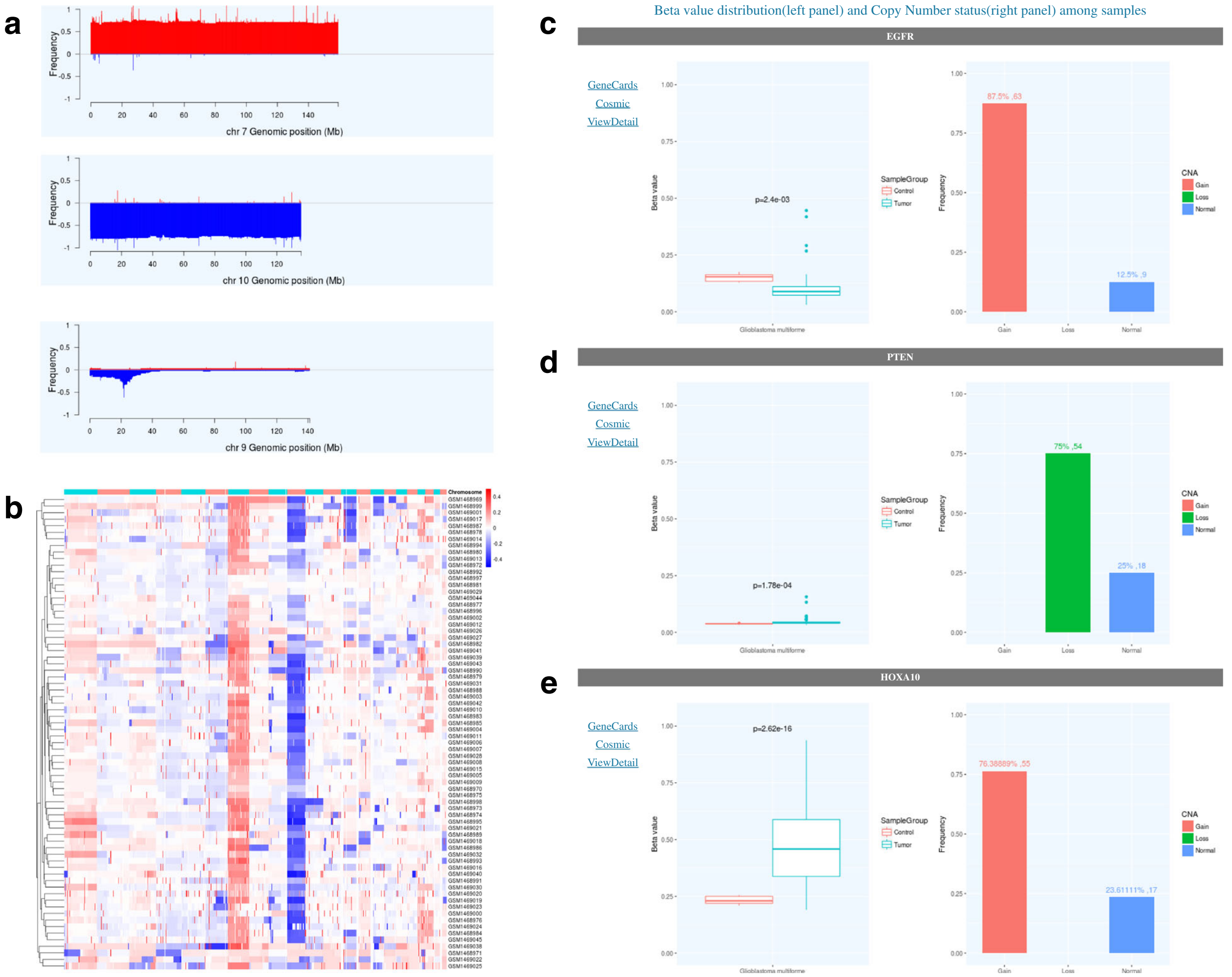


Fig. 6 An example of glioblastoma data analysis. **a** The copy number frequency plots exhibit the whole chromosome 7 gain, whole chromosome 10 loss and a focal deletion on chromosome 9p. **b** Heatmap of copy number profiles across 72 glioblastoma samples. **c** EGFR promoter hypomethylation and copy number gain are shown by boxplot and histogram, respectively. **d** PTEN promoter hypermethylation and copy number loss. **e** HOXA10 promoter hypermethylation and copy number gain

copy number data are derived from exactly the same genomic loci. As previous studies have indicated, genetic and epigenetic alterations may act coordinately to fulfil the two-hit paradigm at a gene-specific level, leading to cancer initiation. For example, tumor suppressor genes that undergo a 'double hit,' such as heterozygous loss and hypermethylation, or oncogenes in an amplified region that are hypomethylated are most likely to be drivers of tumorigenic processes. Thus, the efficient mining of this large-scale dataset can provide valuable insight into the underlying molecular bases of oncogenesis, and facilitate to distinguish driver from passenger alterations. We believe that our database is a powerful tool not only for bioinformaticians but also for experimental researchers. Concerning the future development of our database, we plan to control variations in the tumor cell content between samples. There are several tools developed for the correction of cell type heterogeneity and control for false positives in large-scale epigenome data analysis, such as ReFACTor [47] and MeDeCom [48]. In the next release of our database, we will utilize these tools to remove confounding variation and to provide a better framework for data interpretation.

Additional files

Additional file 1: Table S1. List of GEO tumor samples that are integrated in MethCNA. (XLS 290 kb)

Additional file 2: Table S2. List of TCGA samples that are integrated in MethCNA. (XLS 1181 kb)

Additional file 3: Figure S1. The genome-wide copy number aberration frequency profile of the TCGA GBM dataset. (PDF 1661 kb)

Additional file 4: Figure S2. The copy number heatmap for 153 glioblastoma samples of the TCGA GBM dataset. (PDF 438 kb)

Abbreviations

450 k: Illumina Infinium HumanMethylation450 BeadChip; CGI: CpG island; CNAs: Copy number aberrations; COSMIC: Catalogue of Somatic Mutations in Cancer; DMP: Differentially methylated positions; DMR: Differentially methylated regions; GBM: Glioblastoma; GEO: Gene Expression Omnibus; MDPE: Maximum density peak estimation; SNP: Single nucleotide polymorphism; TCGA: The Cancer Genome Atlas; TSS: Transcriptional start site; UTR: Untranslated region

Acknowledgements

We would like to thank Yang Cao for his valuable comments and help in website construction.

Funding

This work was supported by the National Natural Science Foundation of China (grant numbers 31571314 and U1603120).

Availability and requirements

MethCNA is freely available at http://cgma.scu.edu.cn/MethCNA. The datasets generated and/or analyzed during the current study are available in TCGA (https://gdc-portal.nci.nih.gov) and GEO (https://www.ncbi.nlm.nih.gov/geo/). A full list of dataset IDs is available at http://47.88.3.162/MethCNA/download/CancerSamples.xls. In the case study section, the related dataset GSE60274 can be found at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60274, and TCGA-GBM can be found at https://portal.gdc.cancer.gov/legacy-archive/search/f.

Authors' contributions

HC conceived the study. GD and JY created the work-flow and processed the raw array data. GD, JY and QZ developed the web interface. HC, ZXX and GD drafted the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publications

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Center of Growth, Metabolism, and Aging, Key Laboratory of Bio-Resources and Eco-Environment, College of Life Sciences, Sichuan University, Chengdu, Sichuan 610064, China. [2]Jiangsu Center for the Collaboration and Innovation of Cancer Biotherapy, Xuzhou, Jiangsu 221002, China. [3]Cancer Institute, Xuzhou Medical University, Xuzhou, Jiangsu 221002, China.

Received: 26 October 2017 Accepted: 5 February 2018
Published online: 13 February 2018

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
2. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458(7239):719–24.
3. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315(5813):848–53.
4. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-Cancer patterns of somatic copy number alteration. Nat Genet. 2013;45(10):1134–40.
5. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009;462(7276):1005–10.
6. Network CGAR, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519–25.
7. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The life history of 21 breast cancers. Cell. 2012;149(5):994–1007.
8. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462(7271):315–22.
9. Illingworth RS, Bird AP. CpG Islands–'a rough guide'. FEBS Lett. 2009;583(11):1713–20.
10. Esteller M. Epigenetics in cancer. N Engl J Med. 2008;358(11):1148–59.
11. Hitchins MP, Rapkins RW, Kwok CT, Srivastava S, Wong JJ, Khachigian LM, et al. Dominantly inherited constitutional epigenetic silencing of MLH1 in a cancer-affected family is linked to a single nucleotide variant within the 5' UTR. Cancer Cell. 2011;20(2):200–13.
12. Network CGAR, et al. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609–15.
13. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41(2):178–86.
14. Aryee MJ, Liu W, Engelmann JC, Nuhn P, Gurel M, Haffner MC, et al. DNA methylation alterations exhibit intraindividual stability and interindividual heterogeneity in prostate cancer metastases. Sci Transl Med. 2013;5(169):169ra10.
15. Liu Q, Su PF, Zhao S, Shyr Y. Transcriptome-wide signatures of tumor stage in kidney renal clear cell carcinoma: connecting copy number variation, methylation and transcription factor activity. Genome Med. 2014;6(12):117.
16. Kwee I, Rinaldi A, Rancoita P, Rossi D, Capello D, Forconi F, et al. Integrated DNA copy number and methylation profiling of lymphoid neoplasms using a single array. Br J Haematol. 2012;156(3):354–7.
17. Feber A, Guilhamon P, Lechner M, Fenton T, Wilson GA, Thirlwell C, et al. Using high-density DNA methylation arrays to profile copy number alterations. Genome Biol. 2014;15(2):R30.
18. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20.
19. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res. 2013;41(D1):D991–5.
20. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363–9.
21. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;7:e47.
22. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, et al. ChAMP: 450k chip analysis methylation pipeline. Bioinformatics. 2014;30(3):428–30.
23. Nordlund J, Bäcklin CL, Lönnerholm G, Syvänen AC, Carlsson Almlof J, et al. CopyNumber450kCancer: baseline correction for accurate copy number calling from the 450k methylation array. Bioinformatics. 2016;32(7):1080–2.
24. Maksimovic J, Gordon L, Oshlack A. SWAN: subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol. 2012;13(6):R44.
25. Teschendorff A, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013;29(2):189–96.
26. Du P, Zhang X, Huang C, Jafari N, Kibbe W, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.
27. Zhuang J, Widschwendter M, Teschendorff A. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. BMC Bioinformatics. 2012;13:59.
28. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35.
29. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–3.
30. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, et al. The UCSC genome browser database: 2017 update. Nucleic Acids Res. 2017;45(D1):D626–34.

31. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2016.
32. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards version 3: the human gene integrator. Database. 2010;2010:baq020.
33. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45(D1):D777–83.
34. Ozawa T, Riester M, Cheng YK, Huse JT, Squatrito M, Helmy K, et al. Most human non-GCIMP glioblastoma subtypes evolve from a common proneural-like precursor glioma. Cancer Cell. 2014;26(2):288–300.
35. Sintupisut N, Liu PL, Yeang CH. An integrative characterization of recurrent molecular aberrations in glioblastoma genomes. Nucleic Acids Research. 2013;19:8803–21.
36. Reyes-Botero G, Dehais C, Idbaih A, Martin-Duverneuil N, Lahutte M, Carpentier C, et al. Contrast enhancement in 1p/19q-codeleted anaplastic oligodendrogliomas is associated with 9p loss, genomic instability, and angiogenic gene expression. Neuro-Oncology. 2014;16(5):662–70.
37. Crespo I, Vital AL, Nieto AB, Rebelo O, Tão H, Lopes MC, et al. Detailed characterization of alterations of chromosomes 7, 9, and 10 in glioblastomas as assessed by single-nucleotide polymorphism arrays. J Mol Diagn. 2011;13(6):634–47.
38. Zhao R, Choi BY, Lee MH, Bode AM, Dong Z. Implications Of genetic and epigenetic alterations of CDKN2A (p16 INK4a) in cancer. EBioMedicine. 2016;8:30–9.
39. Day KC, Hiles GL, Kozminsky M, Dawsey SJ, Paul A, Broses LJ, et al. HER2 and EGFR overexpression support metastatic progression of prostate cancer to bone. Cancer Res. 2017;77(1):74–85.
40. Song S, Honjo S, Jin J, Chang SS, Scott AW, Chen Q, et al. The hippo coactivator YAP1 mediates EGFR overexpression and confers chemoresistance in esophageal cancer. Clin Cancer Res. 2015;21(11):2580–90.
41. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci. 1971;68(4):820–3.
42. Benitez JA, Ma J, D'Antonio M, Boyer A, Camargo MF, Zanca C, et al. PTEN regulates glioblastoma oncogenesis through chromatin-associated complexes of DAXX and histone H3. 3. Nat Commun. 2017;8:15223.
43. Koul D. PTEN signaling pathways in glioblastoma. Cancer Biol Ther. 2008;7(9):1321–5.
44. Kurscheid S, Bady P, Sciuscio D, Samarzija I, Shay T, Vassallo I, et al. Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma. Genome Biol. 2015;16(1):16.
45. Gao J, Aksoy B, Dogrusoz U, Dresdner G, Gross B, Sumer S, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1.
46. Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics. 2011;27(6):887–8.
47. Rahmani E, Zaitlen N, Baran Y, Eng C, Hu D, Galanter J, et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. Nat Methods. 2016;13(5):443–5.
48. Lutsik P, Slawski M, Gasparoni G, Vedeneev N, Hein M, Walter J. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. Genome Biol. 2017;18(1):55.