Molecular Genetics and Genomics
https://doi.org/10.1007/s00438-020-01664-y

ORIGINAL ARTICLE

Integrative analysis of DNA methylation and gene expression in papillary renal cell carcinoma

Noor Pratap Singh[1] · **P. K. Vinod**[1]

Received: 1 May 2019 / Accepted: 3 March 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

Patterns of DNA methylation are significantly altered in cancers. Interpreting the functional consequences of DNA methylation requires the integration of multiple forms of data. The recent advancement in the next-generation sequencing can help to decode this relationship and in biomarker discovery. In this study, we investigated the methylation patterns of papillary renal cell carcinoma (PRCC) and its relationship with the gene expression using The Cancer Genome Atlas (TCGA) multi-omics data. We found that the promoter and body of tumor suppressor genes, microRNAs and gene clusters and families, including cadherins, protocadherins, claudins and collagens, are hypermethylated in PRCC. Hypomethylated genes in PRCC are associated with the immune function. The gene expression of several novel candidate genes, including interleukin receptor IL17RE and immune checkpoint genes HHLA2, SIRPA and HAVCR2, shows a significant correlation with DNA methylation. We also developed machine learning models using features extracted from single and multi-omics data to distinguish early and late stages of PRCC. A comparative study of different feature selection algorithms, predictive models, data integration techniques and representations of methylation data was performed. Integration of both gene expression and DNA methylation features improved the performance of models in distinguishing tumor stages. In summary, our study identifies PRCC driver genes and proposes predictive models based on both DNA methylation and gene expression. These results on PRCC will aid in targeted experiments and provide a strategy to improve the classification accuracy of tumor stages.

Keywords Renal cell carcinoma · Multi-omics · Epigenetic regulation · RNASeq · Data integration · Multiple kernel learning · Tumor stage prediction

Abbreviations

BEMKL	Bayesian efficient multiple kernel learning
ccRCC	Clear cell renal cell carcinoma
DEGs	Differentially expressed genes
DMCs	Differentially methylated CpG sites
GL	Group lasso
KNN	k-Nearest neighbors
MKL	Multiple kernel learning
NB	Naive Bayes
PRCC	Papillary renal cell carcinoma
RCC	Renal cell carcinoma
RF	Random forest
SC	Shrunken centroids
SVM	Support vector machine
TSGs	Tumor suppressor genes

Communicated by Stefan Hohmann.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00438-020-01664-y) contains supplementary material, which is available to authorized users.

✉ P. K. Vinod
vinod.pk@iiit.ac.in

[1] Center for Computational Natural Sciences and Bioinformatics, IIIT Hyderabad, Hyderabad 500032, India

Introduction

Cancer cell reprogramming involves aberrations of cancer genome at multiple levels. Besides genetic aberrations, epigenetic modifications such as DNA methylation are also prominent features associated with cancer onset and progression. Hypermethylation (gains) of promoter 5′-C-phosphate-G-3′ (CpG) rich regions known as CpG islands (CGIs) is linked to transcriptional silencing of tumor suppressor genes (TSGs) in various cancers (Baylin 2006). Hypomethylation (losses) is associated with genomic instability in cancer (Eden et al. 2003). Genome-wide studies

Published online: 17 March 2020

have revealed methylation of gene bodies and intergenic regions, the significance of these is not well understood. Gene body methylation is associated with gene expression in some cancers (Maunakea et al. 2010; Kulis et al. 2012; Jones 2012; Yang et al. 2014). Identifying the functional consequences of DNA methylation requires the integration of multiple forms of data. In this study, we integrated The Cancer Genome Atlas (TCGA) multi-omics data of papillary renal cell carcinoma (PRCC) to understand the influence of DNA methylation on the transcriptome and to identify biomarkers for accurate classification of tumor. Renal Cell Carcinoma (RCC) is a heterogeneous group of cancers arising from different regions of nephron (Chen et al. 2016). RCC is divided into multiple histological subtypes clear cell, papillary, chromophobe and collecting duct (Moch et al. 2016). PRCC is the second most common subtype accounting for 10–15% cases (Jonasch et al. 2014). It is an epithelial tumor with papillary or tubulopapillary architecture and is divided into type 1 and type 2 tumors based on histology and molecular features (Hsieh et al. 2018). There are still no effective treatments available for patients with an advanced stage of PRCC (Durinck et al. 2015; Modi and Singer 2016).

Epigenetic regulation by DNA methylation plays an important role in the carcinogenesis of RCC (Shenoy et al. 2015; Lasseigne and Brooks 2018; Morris and Latif 2017). Most studies on RCC have focused on epigenetic regulation in clear cell renal cell carcinoma (ccRCC) (Arai et al. 2012; Kluzek et al. 2015; Wei et al. 2015; Chen et al. 2017). Aberrant DNA methylation and mutation of genes involved in histone modifications and chromatin remodeling have been reported for ccRCC. Somatic inactivation of TSG VHL is common in ccRCC but is also shown to be epigenetically inactivated in ccRCC and PRCC. TSGs including BNC1, WIF1, FBN2 and SLIT2 are frequently inactivated by promoter methylation in ccRCC, while CDH1, IGFBP1, SFRP1, SPINT2 and RASSF1A are hypermethylated in both ccRCC and PRCC (Morris et al. 2005; de Caceres et al. 2006; Morris and Maher 2010; Ellinger et al. 2011; Klacz et al. 2016). A high percentage of CpG islands are hypermethylated in a subset of ccRCC and PRCC. The CpG Island methylator phenotype (CIMP) tumors are aggressive and associated with poor survival (The Cancer Genome Atlas Research Network 2016; Chen et al. 2016; Ricketts et al. 2018). In PRCC, CIMP is related with CDKN2A hypermethylation. These results suggest that DNA methylation patterns can serve as biomarkers and have a role in dysregulation of gene expression (Baylin and Jones 2011; McMahon et al. 2017).

However, a detailed analysis of DNA methylation patterns of PRCC and its relationship with the gene expression is yet to be performed. Further, most studies on RCC focused only on changes in the methylation profiles at promoters. These prompted us to examine the methylation pattern of PRCC using TCGA multi-omics data (The Cancer Genome Atlas Research Network 2016; Chen et al. 2016; Ricketts et al. 2018). This can help in detection of aberrantly methylated loci with respect to normal tissue and different stages of tumor, extracting causal relationships between DNA methylation and gene expression and identifying genes and pathways that are affected by cancer.

Although these cancer datasets are available, they pose a challenge for accurate cancer detection since there is a mismatch between a number of measurements and sample size. Machine learning methods can be used to extract biomarkers and built predictive models from complex single/multi-omics data to classify patients. Previously, we have developed machine learning models using RNASeq data of PRCC to predict the stages of PRCC (Singh et al. 2018). Studies have shown that integration of multi-omics data can yield superior performance compared to single omics data in different cancers (Mankoo et al. 2011; Kim et al. 2015, 2017; Taskesen et al. 2015; Yan et al. 2017; Jiang et al. 2016; Zhu et al. 2017). Different approaches exist for integrating multiple datasets in a supervised manner (Ritchie et al. 2015; Lin and Lane 2017; Huang et al. 2017; Wu et al. 2019). Multiple kernel learning (MKL) is one such approach of data integration which has been used for prediction of survival outcome across different cancers and drug sensitivity (Seoane et al. 2014; Thomas and Sael 2017; Ali et al. 2018; Costello et al. 2014; Zhu et al. 2017). Kernel-based methods can transform input data into a higher-dimensional space in which data points become linearly separable and can be used to integrate different types of data. Further, to develop predictive individual and integrative models based on DNA methylation, different realizations of methylation data have to be explored. Most studies have used a gene-based representation by aggregating only promoter CpG probes for developing integrative models (Kim et al. 2015; Thomas and Sael 2017). On the other hand, few studies have used individual CpG probes for building predictive models (Hao et al. 2017; Shen et al. 2017; Wang et al. 2018). It would be interesting to perform a comparative study using different representations of DNA methylation.

We investigated the methylation profiles of PRCC and its relationship with gene expression. We identified differential methylated CpGs (DMCs) between normal and tumor samples and mapped their location with respect to gene and CpG islands. DMCs are mostly hypermethylated at gene body and are distributed in open sea and Islands. Hypermethylated CpGs map to tumor suppressor genes, microRNAs and gene clusters and families, while hypomethylated CpGs map to immune response genes. We identified several novel candidate genes that show significant correlation between DNA methylation and gene expression. We also extracted methylation features to develop predictive models for distinguishing between early and late stages of PRCC. Further, we integrated features from DNA methylation, RNASeq and

clinical data using multiple strategies to improve the performance of the models. A comparison of model performance using three different representations of methylation data was also performed. Integration of multi-omics data improved the performance of models in distinguishing tumor stages across different feature sets.

Materials and methods

Dataset and preprocessing

DNA Methylation, RNASeq and clinical data of PRCC were downloaded from Genomics Data Commons (GDC) data portal (https://portal.gdc.cancer.gov/) using TCGABiolinks package in R (Colaprico et al. 2016). The beta values obtained from Illumina Infinium HumanMethylation450 (HM450) BeadChiparrays and HTSeq counts obtained from lluminaHiSeq RNASeqV2 were used for our analyses. A total of 321 samples are available for methylation data with 45 matched normal and tumor samples. There are 297 samples including normal that are common across both platforms.

IlluminaHumanMethylation450kanno.ilmn12.hg19 package was used for annotating the probes and mapping to their respective locations (Hansen 2016). For methylation data, probes with missing values in any sample, present on X and Y chromosomes, and overlapping with SNPs were removed. After preprocessing and filtering, there are around 375 K CpGs, whose distribution with respect to CpG Islands and genes is shown in Fig. S1. If a CpG site occurs at a distance situated within 2 kb from CGI, the location is referred to as shore; if it occurs at a location within 2–4 kb then it is referred to as shelf and any probe located further then it is referred to Open Sea (Sandoval et al. 2011).

Identification of DMCs and their association with the gene expression

We used the minfi package (Aryee et al. 2014) for finding DMCs between matched normal and tumor samples (45 each). DMCs with different mean beta value differences and adjusted p value < 0.05 were considered. EnrichR was used for performing functional enrichment analysis of the genes associated with DMCs (Kuleshov et al. 2016). We obtained Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways with Benjamini–Hochberg adjusted p-values < 0.05. We also computed Pearson correlation for every CpG-gene pair using 297 samples that are common across the two platforms.

Predictive models for stage predictions

We next build predictive models to distinguish between tumor stages of PRCC using RNASeq and DNA methylation. The pathological stages are known for 250 samples (common across both the platforms) with the following distributions: Stage I—167, Stage II—19, Stage III—50, and Stage IV—14. We divided the dataset containing these 250 samples into training (80%) and test (20%) datasets and combined Stage I and II into early and Stage III and IV into late stage to develop our predictive models. The beta values (β) were transformed into M values $\left(\log_2\left(\frac{\beta}{1-\beta}\right)\right)$ for the analysis. RNASeq raw count data was normalised using variance stabilizing transformation (VST) (Anders and Huber 2010). The machine learning pipeline used to predict the tumor stages is shown in Fig. 1.

Representations of methylation data and feature selection

We considered three different methylation representations for our analysis. The first representation is CpgM, which is obtained by transforming beta values of each CpG into M values. The second representation is GeneM, which is obtained using the approach proposed by (Jiao et al. 2014). Here, the average beta value of all the probes mapping within 200 bp of the transcription start site (TSS) of a gene is calculated. If no such probes exist, then the average beta value of probes mapping to the 1st exon of the gene

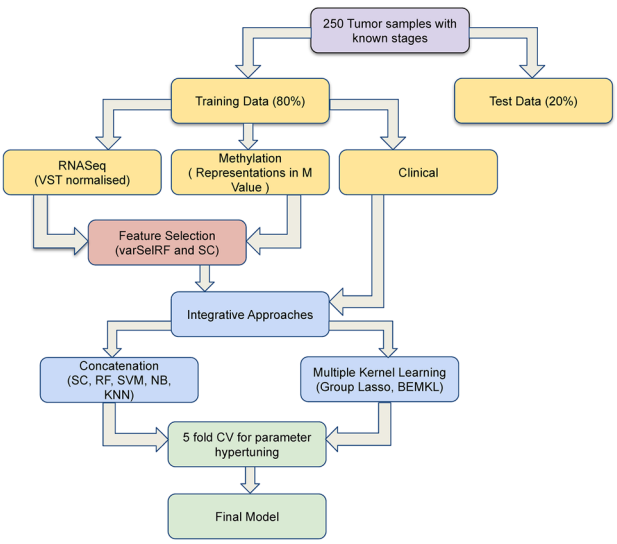

Fig. 1 The workflow used to develop integrated models for PRCC tumor stage prediction. This includes different models, feature selection algorithms and integrative approaches for RNASeq and DNA methylation

or probes mapping within 1500 bp of the TSS is used. The third representation is BmpM, which is obtained using bumphunter function (Jaffe et al. 2012). This function groups nearby genomic locations into regions and finds differentially methylated regions. The average beta values of probes located within significant regions are calculated. The mean beta values are transformed into M Values in the GeneM and BmpM representations.

Two different feature selection methods: Shrunken Centroids (SC) and varSelRF were applied on the training dataset for extracting features between the different stages of PRCC (Fig. 1) (Tibshirani et al. 2002; Díaz-Uriarte and De Andres 2006). SC computes a *t*-statistic for each feature for each class by comparing the overall centroid to class-specific centroid. It then shrinks the *t*-statistic by a threshold, such that if the threshold exceeds the *t*-statistic, then the *t*-statistic is set to zero, making the class-specific centroid to coincide with the overall centroid for that feature. After shrinking the centroid, a sample is classified by nearest centroid rule. Different thresholds are tried such that the one that yields the smallest misclassification error is chosen. The features left with a non-zero *t*-statistic at that threshold are the features selected by the algorithm. varSelRF is a Random Forest-based recursive feature selection algorithm where feature importance is computed first and then the features are removed at each iteration. The iteration that yields the least number of features with an out-of-bag (OOB) error comparable to the iteration yielding the lowest OOB error is chosen. We modified the above algorithms to replace the smallest misclassification error or OOB error with largest overall MCC (Matthews Correlation Coefficient) or OOB MCC.

Classification models

Different supervised machine learning algorithms: RF, NB, Linear-SVM, KNN, SC, GL and BEMKL classifiers were used to predict the tumor stages of PRCC. The models were developed based on the features extracted from the different representations of DNA methylation. Further, we integrated features extracted from DNA methylation, 104 features obtained from our previous study on RNASeq (Singh et al. 2018) and the clinical information including bmi, age, sex and race. One-hot encoding is used for binarizing categorical variables while the missing continuous variables were imputed by taking the median within the class.

We have explored three different approaches for integrating features from each platform. The first approach is a simple concatenation-based approach where we concatenate the scaled features from the individual platforms and then train the models (RF, NB, SC, KNN, Linear SVM) on the concatenated features. The other two approaches are multiple kernel learning (MKL): Group lasso (Xu et al. 2010; Rahimi and Gönen 2018) and BEMKL (Gonen 2012) which are described below.

In SVM, to learn the appropriate hyperplanes, we often solve the dual optimization problem which is (Cortes and Vapnik 1995):

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ & \text{with respect to } \alpha \in R^N \\ & \text{subject to} \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & 0 \le \alpha_i \le C \quad \forall i \end{aligned} \tag{1}$$

where N is the total number of training samples, α is the weight vector corresponding to samples, y represents the class labels, k represents the kernel function and C represents the cost.

The class label for a sample x_* is determined by:

$$\text{sign}\left(f\left(x_*\right)\right),$$

$$f\left(x_*\right) = \alpha^{\top} Y_D k_* + b \tag{2}$$

where Y_D is the $N \times N$ diagonal matrix containing class labels, $k_* = [k(x_1, x_*) \ldots k(x_N, x_*)]^{\top}$ and b is the bias.

MKL instead of using a single kernel substitutes it with a combined kernel computed as a function of input kernels $k(x_i, x_j) = f(\{k_m(x_i, x_j)\}_{m=1}^{P})$ where P is the number of kernel functions used (Gönen and Alpaydın 2011). One way of combining input kernels is to use a weighted sum such that:

$$f(\{k_m(x_i, x_j)\}_{m=1}^{P}) = \sum_{m=1}^{P} e_m k_m(x_i, x_j), \tag{3}$$

where e_m represents the learnt weight of a kernel function k_m.

Group Lasso modifies the optimization problem highlighted in (1) by replacing the final kernel term $k(x_i, x_j)$ with (3) and imposing an additional constraint of l_1 norm on the kernel weights i.e.:

$$e \in R^P, \ \sum_{m=1}^{P} e_m = 1 \text{ and } e_m \ge 0 \quad \forall m \tag{4}$$

For solving the above optimization problem, an iterative strategy is used. The algorithm begins by setting the kernel weights to uniform values such that at the first iteration, $e_m^1 = 1/P$. Then at each iteration t, kernel weights are used to solve the SVM optimization problem giving us support vector coefficients α^t. These are then used to update the weights at iteration $t+1$ as follows and the cycle continues:

$$e_m^{(t+1)} = \frac{e_m^{(t)}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i^{(t)}\alpha_j^{(t)}y_iy_jk_m(x_i,x_j)}}{\sum_{k=1}^{P}e_k^{(t)}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i^{(t)}\alpha_j^{(t)}y_iy_jk_k(x_i,x_j)}}\forall m$$

BEMKL on the other hand uses a Bayesian approach where the parameters such as e_m, α_i, b are assumed to follow a normal distribution with their priors following a gamma distribution. The intermediate output is computed from each kernel and then combined with bias and kernel weights to predict f. It uses deterministic variational Bayesian formulation to efficiently infer the posterior mean and covariance for the above distributions. The final formulations can be found in (Gonen 2012).

We created a list of gaussian and polynomial kernels with varying sigma and degrees for integrating using the MKL approaches. These were applied to each of the feature sets extracted from DNA methylation and RNASeq datasets. Further, the kernel matrix obtained for each kernel was unit normalized. Different parameters such as cost for linear-SVM, k for KNN, number of trees for Random Forest and the threshold for Shrunken Centroids were optimized using fivefold cross validation. For MKL approaches, we observed that the final performance was also dependent on factors, such as scaling of the data from the individual platform and on the range of degrees/sigma that was used for creating the multiple kernels, in addition to cost parameter for GL and gamma-prior for BEMKL. Therefore, we performed a grid search and selected the best combination using fivefold cross validation.

The performance of the models was evaluated on the 20% test dataset (Fig. 1). The metrics such as PR AUC, MCC, Accuracy, Sensitivity and Specificity were used to quantify the performance (Matthews 1975; Sokolova and Lapalme 2009; Saito and Rehmsmeier 2015). We specifically used PR AUC and MCC to compare the performance of the models due to the class imbalance in our dataset (Saito and Rehmsmeier 2015; Chicco 2017). The code for building the models is provided in the Github repository.

Results

Methylation pattern of PRCC

First, we identified the differentially methylated CpGs (DMCs) between 45 normal and tumor-matched samples. Table 1 shows the number of CpGs obtained based on different beta values cut-off and q-value < 0.001. A significant number of probes are hypermethylated compared to hypomethylated across different thresholds. Figure 2a shows the principal component analysis (PCA) plot using DMCs (|mean beta difference| ≥ 0.4). A clear separation between normal and matched tumor samples is observed barring a couple of tumor samples. This separation is also observed using DMCs obtained via other thresholds or by either using hyper- or hypo-methylated DMCs (Fig. S2). Further, these CpGs are also sufficient to separate most tumor samples available for PRCC in TCGA (321 samples) from normal samples (Fig. 2b). The genomic distribution of hyper- and hypo-methylated DMCs (|mean beta difference| ≥ 0.2) with respect to gene and CpG islands is shown in Fig. 3. Both hyper- and hypo-methylated probes are mostly located at gene body followed by promoter regions (TSS200, 1st exon, TSS1500). The distribution based on CpG Islands shows that hypermethylated probes are mostly located at both OpenSea and Islands, while hypomethylated probes are predominantly located at OpenSea.

Table 1 Total number of CpGs that are differentially methylated between matched normal and tumor samples

Beta value	Hypermethylated	Hypomethylated	Total
0.2	12,034	2343	14,387
0.3	1942	440	2382
0.4	256	63	319
0.5	30	8	38

Of the total 14,387 DMCs (|mean beta difference| ≥ 0.2), 11,972 CpGs map to 7160 genes and non-coding RNAs. We performed the enrichment analysis to identify biological processes that are associated with the differentially methylated genes and non-coding RNAs. The hypermethylated genes are associated with different cancer signalling pathways (Table 2). This includes Hippo Signalling, Sonic Hedgehog (Shh), Wnt, Notch and Ras signaling pathways. We found HHIP, ZIC1 and ZIC4 that are antagonists of Shh signaling to be hypermethylated. HHIP is an antagonist of Shh ligand while ZIC1 and ZIC4 are antagonists of transcriptional factor GLI (Llinàs-Arias and Esteller 2017). In the Wnt pathway, several Wnt ligands notably WNT5A, members of 'frizzled' gene family (FZD4, FZD5, FZD7, FZD9), transcriptional repressors (HIC1, HIC2) and pathway inhibitors FRZB, SFRP5 and TMEM88 are methylated. SOX1, KCNQ1 and KCTD1 that are known to interfere with Wnt signaling by modulating β-catenin are also hypermethylated (Guan et al. 2014; Li et al. 2014; Rapetti-Mauss et al. 2017). Most of these genes are methylated either at promoter or at both promoter and gene body. However, FZD4 and FZD5 are hypermethylated only at gene body. Further, Notch receptors (NOTCH1, NOTCH4) and ligand DLL1, and its downstream transcriptional co-repressor NCOR2 are mostly hypermethylated within gene body.

In Ras pathway, we found promoter of RASSF1, RASAL2, RIN1 and PAK6 and body of RASA3 are hypermethylated. Members of Ras oncogene family RAB1B and

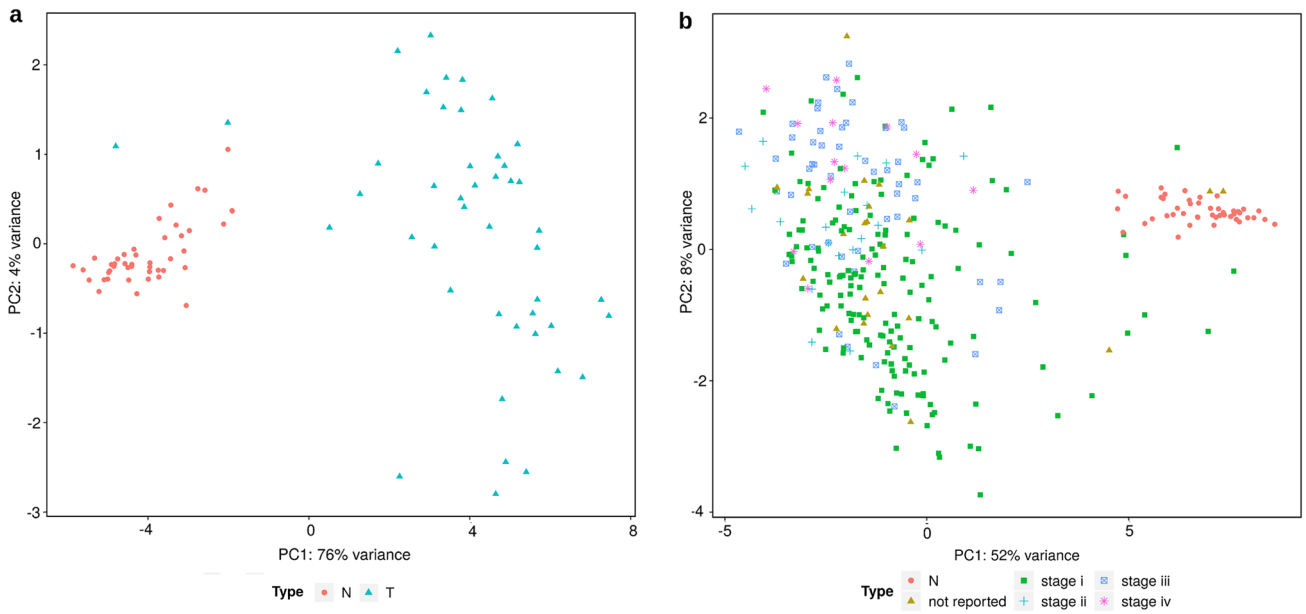

Fig. 2 Principal component analysis (PCA) plot using differentially methylated CpGs (DMCs) between matched normal and tumor samples having |mean beta difference| ≥ 0.4. **a** Matched normal and tumor samples, **b** all normal and tumor samples with their respective tumor stages

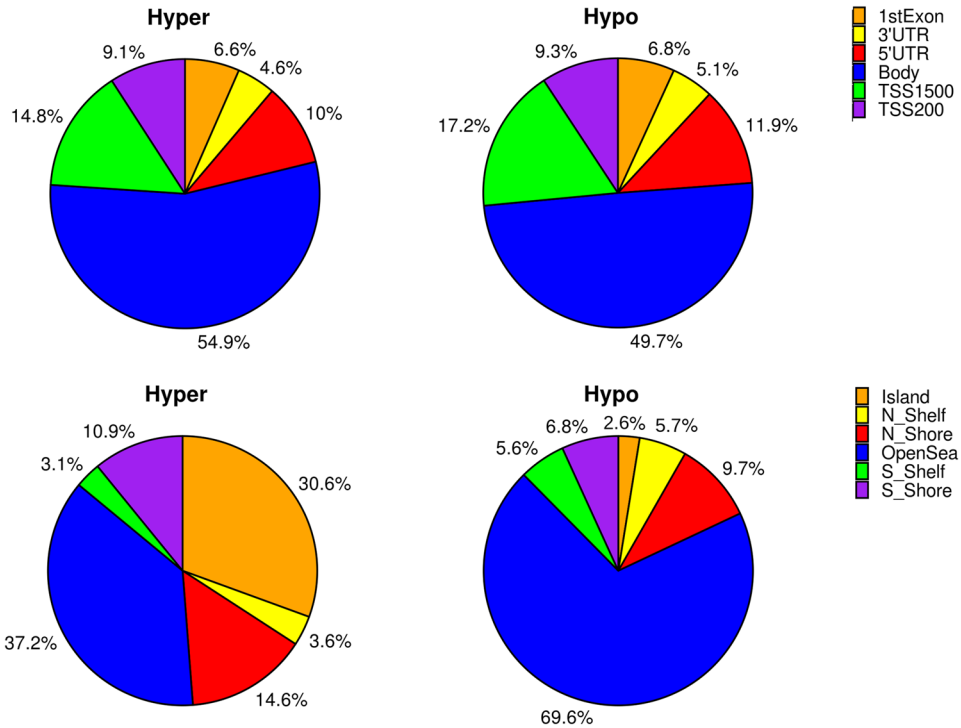

Fig. 3 The relative distribution of hyper and hypomethylated DMCs obtained between matched normal and tumor samples having |mean beta difference| ≥ 0.2 with respect to their location from gene (upper panel) and CpG Islands (lower panel)

RAB25, and a gene encoding RAB GTPase activating protein, RABGAP1L are hypermethylated at several CpGs in the promoter region. Genes of signaling pathways regulating the pluripotency of stem cells are also significantly hypermethylated at both promoter and gene body regions. These include TGFβ ligands BMP2 and BMP4 and its downstream target SMAD3, and JARID2 that plays an important role in gene silencing by binding to Polycomb repressive complex 2 (PRC2).

Further, genes involved in cell adhesion and extracellular matrix are also hypermethylated (Table 2). This includes cadherins, protocadherins, claudins and collagens gene

Table 2 Biological processes and KEGG pathways associated with hypermethylated Cpgs having a mean beta difference greater than 0.2

	adj *p*-value
Biological process	
Transmembrane receptor protein tyrosine kinase signaling pathway	0.000037
Regulation of BMP signaling pathway	0.000991
Positive regulation of transcription of Notch receptor target	0.001398
Regulation of cell migration	0.006823
Positive regulation of pathway-restricted SMAD protein phosphorylation	0.008569
Regulation of cell proliferation	0.024873
DNA damage response	0.032839
Regulation of cellular response to transforming growth factor beta stimulus	0.033076
Cell–cell adhesion via plasma-membrane adhesion molecules	0.034136
Kidney epithelium development	0.034713
Cellular response to growth factor stimulus	0.038848
Mesenchymal to epithelial transition	0.041761
KEGG pathway	
Rap1 signaling pathway	0.00006
Hippo signaling pathway	0.000206
AMPK signaling pathway	0.001358
Signaling pathways regulating pluripotency of stem cells	0.001528
TGF-beta signaling pathway	0.006286
cGMP-PKG signaling pathway	0.006663
cAMP signaling pathway	0.008218
MAPK signaling pathway	0.016953
Ras signaling pathway	0.026518
Wnt signaling pathway	0.026518
Hedgehog signaling pathway	0.026691
Calcium signaling pathway	0.026877

families. Genes belonging to claudins family (CLDN14, CLDN10, CLDN19, CLDN22, CLDN8, CLDN9) are hypermethylated at promoter while genes belonging to collagen family are hypermethylated at promoter (COL18A1-AS1, COL9A2, COL26A1, COL11A2, COL5A2, COL11A1) and gene body (COL18A1, COL23A1, COL4A2, COL4A1, COL9A3). Most genes belonging to cadherin (CDH17, CDH5, CDH23, CDH22, CDH15) and protocadherin (PCDHGA1, PCDHGA2, PCDHGA3, PCDHGA4, PCDHGA5, PCDHGB1, PCDHGB2, PCDHGB3, PCDHGA6, PCDHGA7, PCDHGB4) families are hypermethylated at gene body. These protocadherins are hypermethylated at several CpGs (> 30). On the other hand, CDH9 is hypermethylated at promoter and PCDHGC4, PCDHGB8P, PCDHA13 and PCDHB3 are hypermethylated at both promoter and gene body.

Genes with promoter hypermethylation also include mucin family (MUC12, MUC13, MUC15, MUC20, MUC17), histone cluster family (HIST1H1A, HIST1H2AL, HIST1H3E, HIST1H3I, HIST1H4L), keratins (KRT81, KRT86, KRT9, KRT72, KRTAP17-1), E3 ubiquitin ligases (SIAH3, NEDD4L), neuroactive ligand-receptor interaction (ADRA1A, HRH2, DRD4, GABBR1), solute carrier family (SLC6A3, SLC25A2, SLC16A5, SLC4A11) and potassium voltage-gated channel subfamily (KCNAB3, KCNH2), zinc-finger proteins (ZNF106, ZNF154, ZNF177, ZNF217, ZNF233, ZNF577, ZNF750), inhibitors of EMT(OVOL1, GRHL2) and transmembrane proteins (TMPRSS2, TMPRSS12, TMPRSS13, TMEM178A, TMEM263, TMEM30B). ZNF154 and ZNF577 are tumor suppressors whereas TMPRSS2 is a transmembrane protease that is known to be hypermethylated in ccRCC (Revill et al. 2013). HOX family (HOXA5, HOXA3, HOXB3, HOXB-AS3, HOXA-AS2, HOXA-AS3, HOXA4, HOXB5, HOXB6, HOXB7, HOXA10, HOXA11, HOXA2, HOXC4) and forkhead-box (FOXD2, FOXJ1, FOXS1, FOXP2, FOXG1, FOXK1, FOXO1, FOXO3, FOXD2-AS1, FOXA1) genes are hypermethylated at both promoter and gene body. The promoter of HOXA5 is hypermethylated at 28 CpGs. We also found the promoter of genes encoding different microRNAs in cancer and ECM (MIR25, MIR26B, MIR10B, MIR429, MIR30C1, MIRLET7E, MIR219A1, MIR125A, MIR125B1), and long intergenic non-protein coding RNA (LINC00028, LINC00461, LINC00638, LINC00887, LINC00925, LINC01101 and LINC01366) to be hypermethylated.

On the other hand, genes with promoter hypomethylation are associated with immune (adj p-value = 0.01) and inflammation response (adj p-value = 0.02). These include genes belonging to cytokine production (HHLA2, PDE4D, IL18, C3, IFI16, IL1B, TRIM15, HAVCR2), defense response (FNDC4, TNFAIP6, NR1H4), regulation of viral life cycle (ISG20, SLPI) and regulation of phagocytosis (AZU1, SIRPB1). Genes encoding MHC class II protein complex (HLA-DMA, HLA-DRA, HLA-DPA1) are also hypomethylated at the promoter region. Further, the protein tyrosine phosphatase receptor PTPRN2 and the transcription factor CUX1 are hypomethylated at multiple CpGs (19 and 9) located at gene body.

Relationship between gene expression and methylation

We studied the link between methylation patterns and gene expression across normal and tumor samples. The correlation between DMCs with |mean beta difference| ≥ 0.2 and their corresponding genes was computed. The relative distribution of correlation (q-value ≤ 0.01) with respect to DMCs' location is shown in Fig. 4. A larger proportion of DMCs located at promoter region shows negative correlation while those located at gene body shows both positive and negative correlation with the gene expression. Further, DMCs located at promoter CpG Islands and shores also show large number of negative correlations (Fig. S3). On the other hand, DMCs located at gene body CpG Islands show large number of positive correlations (Fig. S4). Overall, a significant number of negative correlations between methylation and gene expression is observed. This might explain our previous observation that genes are predominantly downregulated in PRCC compared to normal samples (Singh et al. 2018).

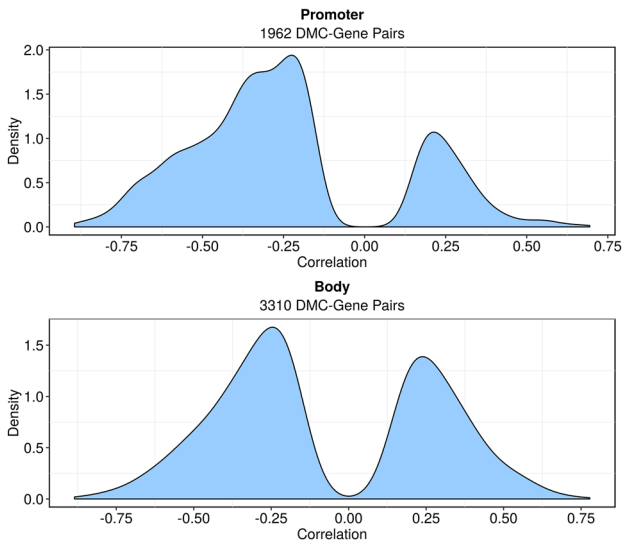

Fig. 4 Correlation distribution of gene-CpG pairs for all DMCs with |mean beta difference| ≥ 0.2 and q value ≤ 0.01

We identified candidate genes that show promoter hypermethylation and downregulated gene expression with respect to normal samples. 445 (1098) gene-CpG pairs show significant ($r \leq -0.5$, q-value ≤ 0.01) correlation between methylation and gene expression with mean beta difference ≥ 0.2 (0.1) and $\log_2$FC ≤ − 1. We found genes (ATP1A1, ATP6V0A4, GGT6, KCNQ1, PROM2, CYFIP2) that have been previously associated with ccRCC to be hypermethylated and downregulated in PRCC (Fig. S5). Further, we also found several novel candidate genes that have not been reported in PRCC (Fig. 5). FAM83F, CNKSR1 and IL17RE are hypermethylated at multiple promoter CpGs and their expression is downregulated in PRCC compared to normal samples. In some PRCC samples, these genes are upregulated and show similar pattern to normal samples, which suggests heterogeneity in gene expression and methylation within the tumor samples. PXDNL, FYB2, NECTIN4 and PYGM are hypermethylated at multiple promoter CpGs and their expression is downregulated in most cancer samples. LYNX1 is hypermethylated and downregulated in graded pattern in tumor samples. It is hypermethylated at 6 CpG islands probes in the promoter with this pattern also present in the more aggressive late-stage samples. Therefore, LYNX1 is a possible CpG island methylator phenotype (CIMP) gene.

We also identified candidates which show promoter hypomethylation and upregulation of gene expression with respect to normal samples. 137 (436) gene-CpG pairs show significant ($r \leq -0.5$, q-value ≤ 0.01) correlation between methylation and gene expression with mean beta difference ≤ − 0.2 (− 0.1) and $\log_2$FC ≥ 1. Genes (TNFAIP6, CHI3L2, C3, EHBP1L1, IFI16, CMTM3) associated with ccRCC are also hypomethylated and upregulated in PRCC (Fig. S6). Figure 6 shows several novel candidates that have not been reported previously in PRCC. Most tumor samples show hypomethylation and upregulated expression of CNTN6, SPATA12, HHLA2, SIRPA, APOL1 and HAVCR1. In the case of ARL4C, the expression and methylation showed a graded pattern within tumor samples. SPATA12 and APOL1 are hypomethylated at multiple CpGs.

Further, we studied the correlation between gene body hypomethylation and gene expression. Interestingly, we found MET, which is known to be mutated in PRCC, to be hypomethylated and upregulated in tumor samples compared to normal samples (Fig. 6). PVT1 and ABC33 are also hypomethylated in tumor samples compared to normal samples. Within tumor samples, we found that gene body of RRM2, NCAPG and SLC7A11 to be hypomethylated and upregulated in late-stage samples. Although we observed a large number of CpG-gene pairs to have a negative correlation, there also exist pairs with strong positive correlation. These

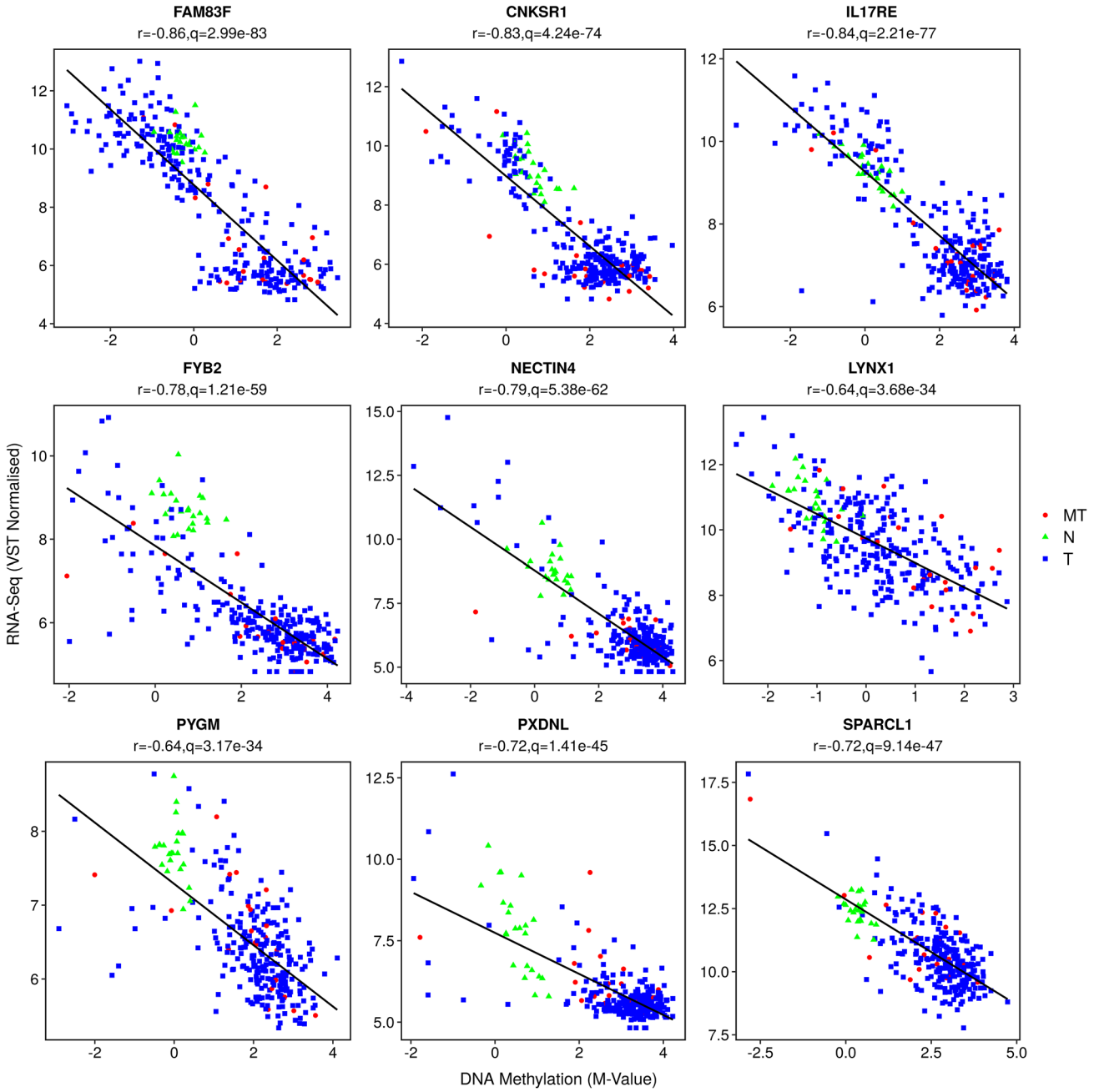

Fig. 5 Scatter plot showing the correlation between downregulated genes and their corresponding hypermethylated CpGs in PRCC. Green represents normal samples, red matched tumor samples and blue remaining tumor samples

include SPI1, SATB2, PLEKHN1, TNFSF9 and MNX1 mostly hypemethylated at gene body. The significance of such a relationship is yet to be explored in PRCC.

Machine learning models for tumor stage prediction

We observed that DMCs, which distinguish normal and tumor samples, failed to distinguish tumor samples based on tumor stages (Fig. 2). Therefore, we performed further analysis using machine learning techniques to distinguish early (stage I and II) and late (stage III and IV) stages of PRCC. We applied SC and varSelRF on each of the three representations of methylation training data to extract features between early and late stages of PRCC (see "Materials and methods"). The number of features obtained for each representation is shown in Table 3. Most of the features obtained across all the representations are hypermethylated in the late stage of tumor. Different models were trained

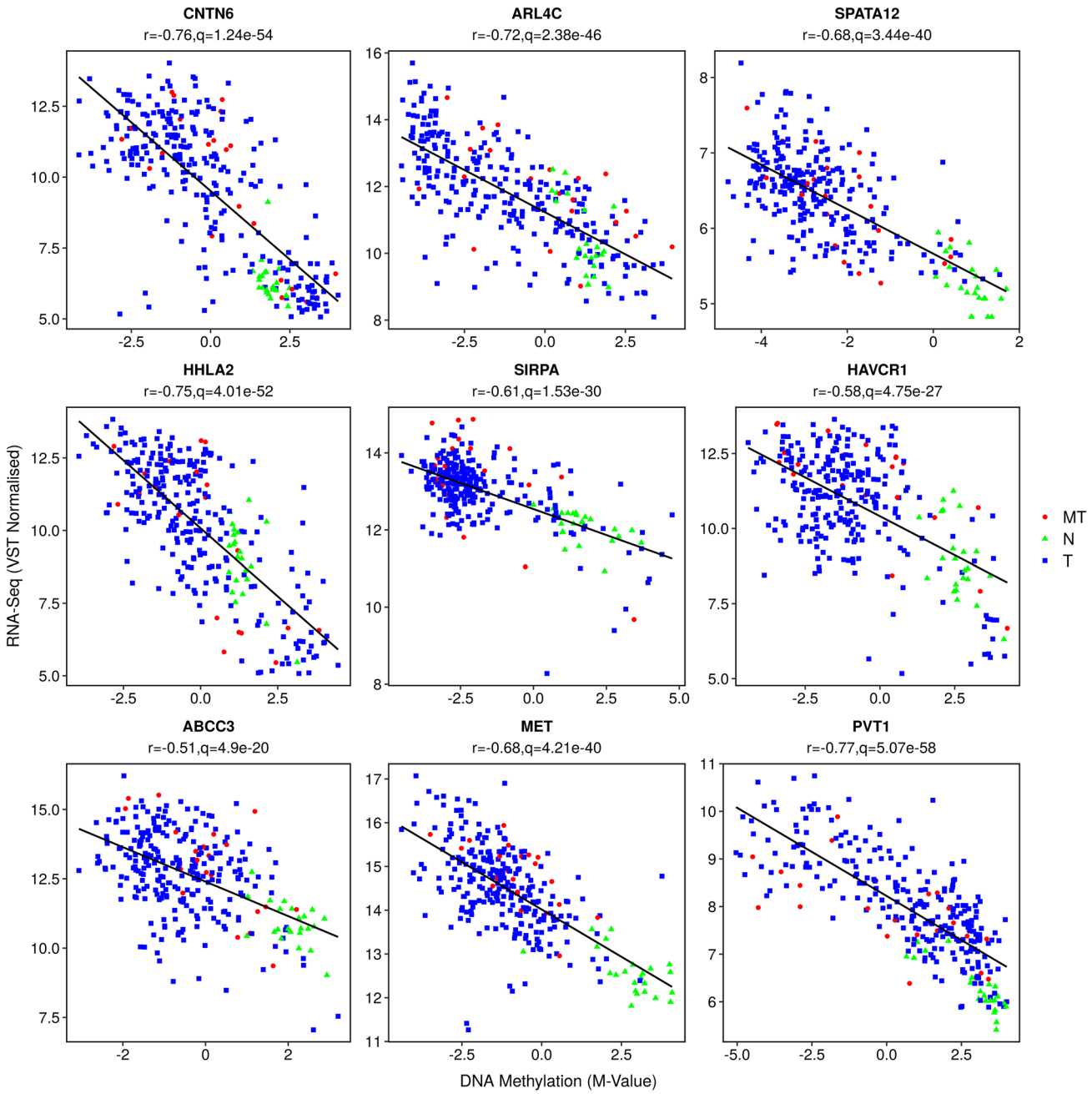

Fig. 6 Scatter plot showing the correlation between upregulated genes and their corresponding hypomethylated CpGs in PRCC. Green represents normal samples, red matched tumor samples and blue remaining tumor samples

Table 3 Number of features extracted for different representations of methylation data

Representation/feature selection	CpgM	GeneM	BmpM
varSelRF	14 (14, 0)	74 (71, 3)	14 (14, 0)
Shrunken centroids	79 (79, 0)	109 (109, 0)	192 (192, 0)

The numbers in the bracket denotes hyper- and hypo-methylated features in late stage samples

using these features sets to distinguish the tumor stages of PRCC and their performance (MCC and PR AUC) on the test dataset were compared. The MCC and PR AUC values range from 0.41 to 0.61 and 0.57 to 0.73 for CpgM representation, from 0.38 to 0.50 and 0.57 to 0.72 for GeneM representation and from 0.32 to 0.55 and 0.44 to 0.70 for BmpM representation, respectively (Fig. S7).

The best performing models are shown in Table 4 (Tables S1, S2). The best MCC of 0.61 and PR AUC of 0.72, 0.68

Table 4 Best performing models on the test dataset for the CpgM representation of methylation data

Classifier	Feature set	MCC	PR AUC	Accuracy (%)
RF	varSelRF	0.606	0.720	84.6
KNN	varSelRF	0.606	0.676	84.6
GL	varSelRF, SC	0.606, 0.558	0.722, 0.699	84.6, 82.7
BEMKL	varSelRF	0.558	0.733	82.7
Linear SVM	varSelRF, SC	0.551, 0.502	0.708, 0.713	82.7, 80.8

and 0.72 are obtained for CpgM representation with RF, KNN and GL using varSelRF features, respectively. BEMKL has the highest PR AUC of 0.733. Most models have MCC greater than 0.55 and PR AUC greater than 0.65 for this representation with both varSelRF and SC feature sets. The performance with respect to GeneM and BmpM representations is lower with best MCC of 0.50 (multiple models) and 0.55 (KNN), and best PR AUC of 0.717 and 0.695 (both GL), respectively. These results indicate, irrespective of the feature sets, CpgM representation provides better performance than the other feature representations. Figure 7 shows the heat map of varSelRF and SC features obtained for CpgM representation on the entire data (Tables S3, S4). The features obtained are a subset of DMCs (q-value < 0.05 and |mean beta difference| ≥ 0.1) between the early and late stages of PRCC.

Multi-omics data integration for tumor stage prediction

In our previous study, we showed that 104 features extracted from RNASeq can predict the stages of tumor (Singh et al. 2018). The performance of various models using 104 features are shown in Table S5. We have extended our previous study to include GL and BEKML models. The best performing models are SC, GL and BEMKL with MCC of 0.71, 0.72 and 0.66 and with similar PR AUC of around 0.8, respectively. The performance of models using RNASeq data is superior compared to models using methylation data in terms of both MCC and PR AUC.

Further, we studied the effect of integrating features extracted from both RNASeq and each representation of methylation data on the model performance. The concatenation-based integration approach was used for the models KNN, NB, RF, SC and SVM, and respective MKL-based integration method was used for the models GL and BEMKL. Multi-omics data integration with CpgM representation improves the performance with increase in the MCC value of GL and RF using varSelRF features to 0.77 in comparison to RNASeq or DNA methylation data (Fig. 8 and Table 5). The PR AUC decreases for SVM, BEMKL and GL models while for other models it increases with the SC model having a higher PR AUC of 0.813. On the other hand, MCC value of most models using SC features decreases with only GL showing an increase (0.77) (Fig. 8 and Table 5). PR AUC follows a similar trend as MCC but a higher PR AUC of 0.82 is obtained for GL. Further, for both GeneM and BmpM representations, the best performance is observed with GL using SC features (Figs. S8, S9 and Tables S6, S7). The MCC value increases to 0.77 without much change in PR AUC value. Thus, the GL model shows the overall best performance across different representations of methylation data and feature sets.

We also integrated clinical features such as age, sex, weight and race and analyzed the performance of the models. There is no change in the performance of best models obtained using multi-omics data. However, we observed an improvement in the performance of KNN and BEMKL with the inclusion of clinical features. MCC and PR AUC values increase across representations for the different feature sets. An improvement in the performance of SVM and RF models is also observed for CpgM representation with SC features. Further, we performed the analysis on another three different training (80%) and test (20%) datasets using CpgM representation to study how the performance of models can be affected depending on the split of training and test datasets. Although a variation in performance is seen across partitions, we observed that the integration of multi-omics data mostly increases the performance of models (GL, BEMKL) in comparison to RNASeq or methylation features (Table S8). An overlapping set of features are obtained from at least two partitions of data. We obtained with varSelRF only 11 CpGs that are common since it selects only very few features (10 to 20) across different partitions. However, a significant number of common features (107) are obtained with shrunken centroid. The performance metrics for other models across different partitions are provided in the GitHub repository.

Discussion

In this study, we performed an integrative analysis of DNA methylation and gene expression to characterize the patterns of DNA methylation in PRCC. Our analysis showed that most probes are hypermethylated in PRCC, and both hyper- and hypo-methylated probes can distinguish normal from cancer samples. The differentially methylated probes map to genes of various cancer signaling pathways, immune response, cell adhesion, gene families and ECM and are located at gene promoter, body or both. Several novel candidate genes including immune checkpoint genes show significant correlation between DNA methylation and gene

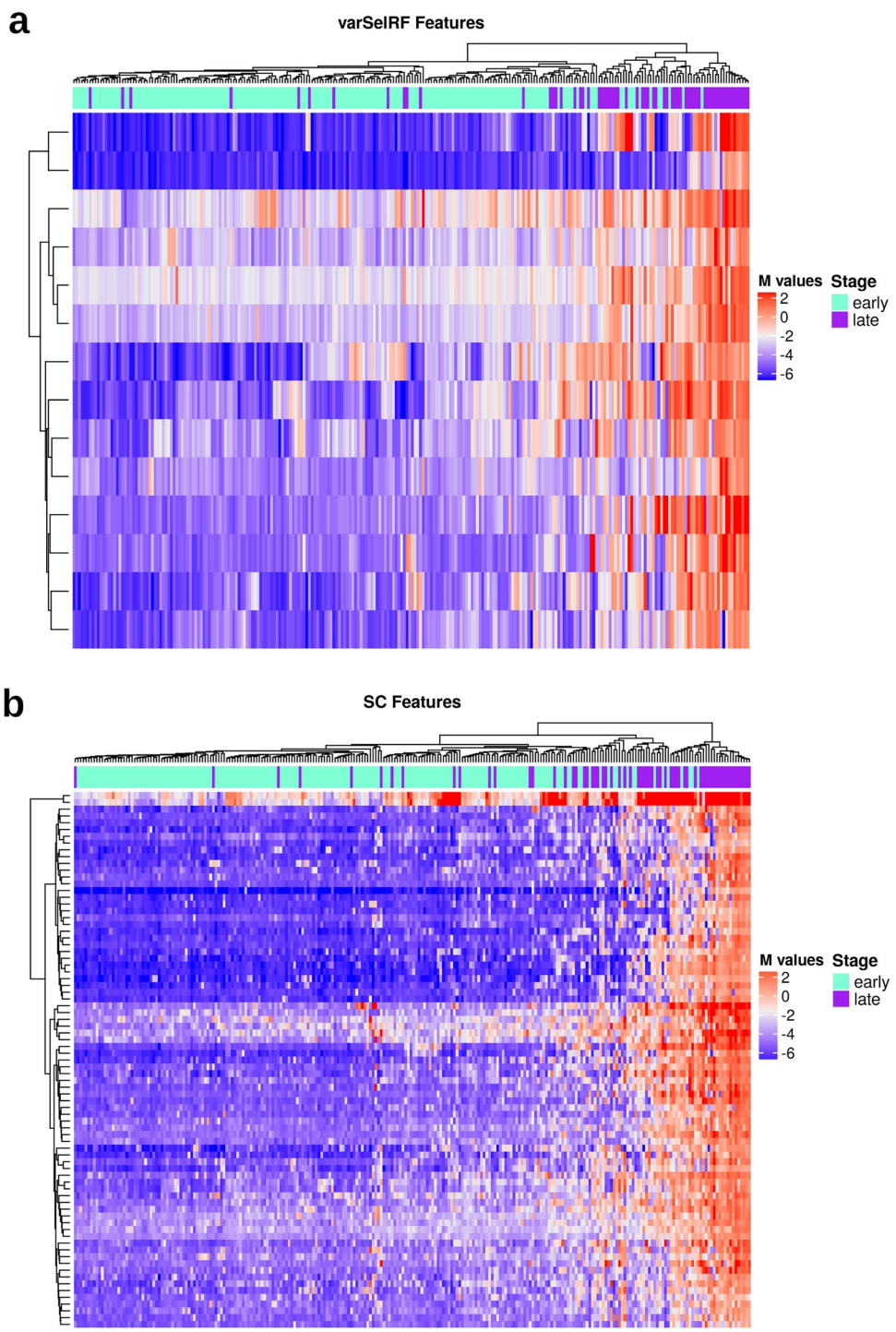

Fig. 7 Heatmap of CpGs obtained as features between early and late stages of PRCC for CpgM representation using **a** varSelRF and **b** Shrunken Centroids feature selection algorithm

expression. Further, we developed machine learning models that integrate features of DNA methylation and gene expression to distinguish early and late stages of PRCC. This study provides a comprehensive framework for data integration to improve the accuracy of classification of tumor stages, which can be extended to other cancer datasets as well.

We found several candidate genes that function as tumor suppressors to be hypermethylated suggesting that epigenetic inactivation might play a role in PRCC. Negative regulators of Wnt, Sonic Hedgehog and Ras signaling pathway are hypermethylated (HHIP, ZIC1, ZIC4, RASAL2, RASSF1, FRZB, SFRP5 and TMEM88). This might lead to the enhanced transcriptional activity of β-catenin and Gli1, which are linked to cell proliferation, migration and EMT (Wils and Bijlsma 2018). Hypermethylation of HHIP, ZIC1 and ZIC4 of sonic hedgehog pathway have been reported

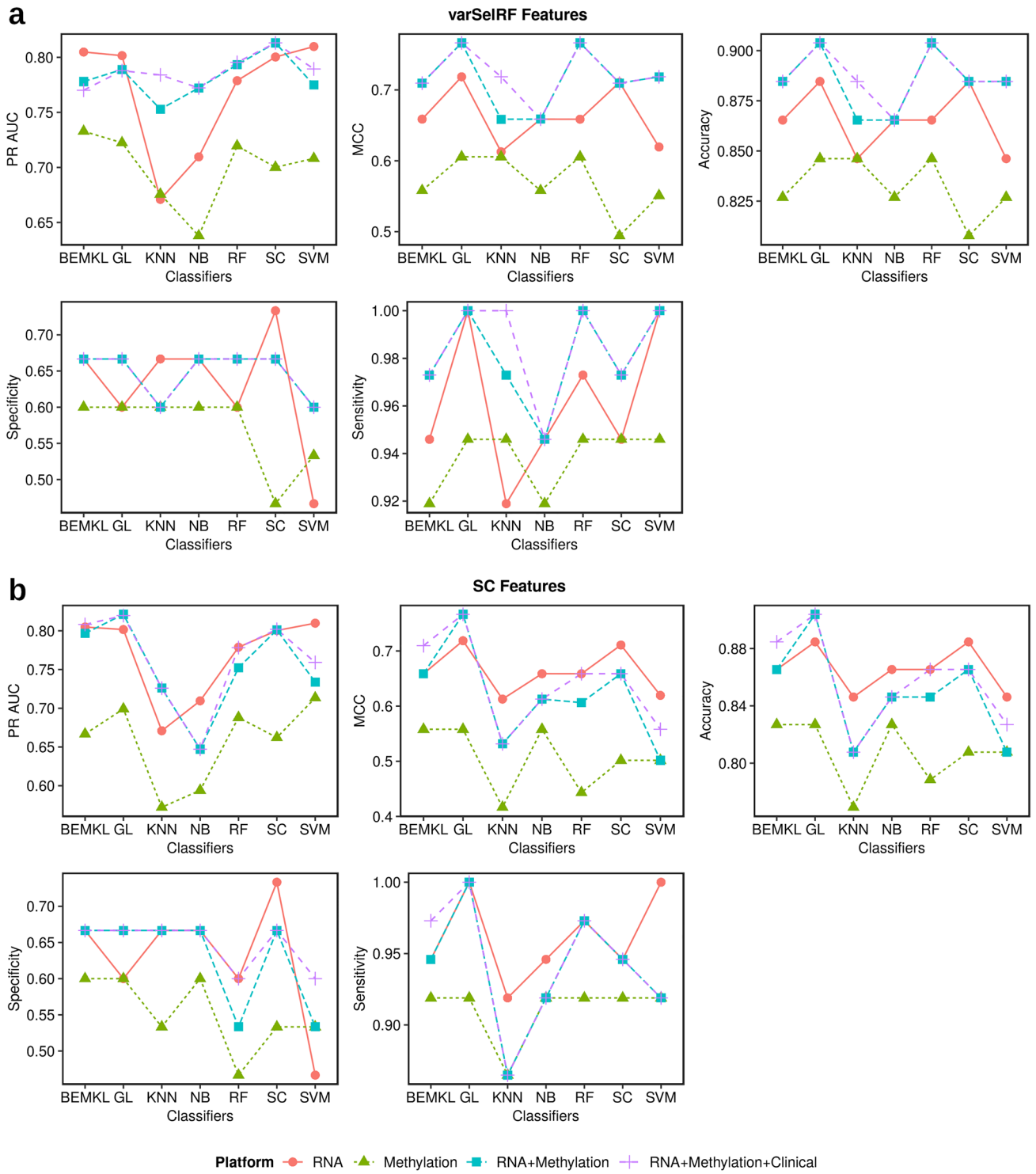

Fig. 8 Performance of different models on the test data with features extracted from CpgM representation using **a** varSelRF and **b** Shrunken Centroids. The metrics PR AUC, MCC, Accuracy, Sensitivity and Specificity are shown for individual and integrated models based on gene expression (RNA), DNA methylation and clinical data. For integration of data, concatenation-based approach is used for classifiers KNN, NB, RF, SC, SVM, while MKL-based approaches are used for BEMKL and GL

Table 5 Best performing models on the test dataset obtained by integrating features from RNASeq and CpgM representation of methylation data and clinical data

Classifier	Feature set	MCC	PR AUC	Accuracy (%)
RF	varSelRF	0.77	0.79	90.4
GL	varSelRF, SC	0.77, 0.77	0.79, 0.82	90.4, 90.4
Linear SVM	varSelRF	0.72	0.78	88.5
SC	varSelRF, SC	0.71, 0.66	0.81, 0.80	88.5, 86.5
BEMKL[a]	varSelRF	0.71	0.78	88.5
BEMKL[b]	SC	0.71	0.81	88.5
KNN[b]	varSelRF	0.72	0.78	88.5

[a]Implies the performance is observed by integrating RNASeq and methylation data

[b]Implies the performance is observed by integrating RNASeq, methylation and Clinical data, otherwise similar performance is observed with or without clinical data

for gastric cancer and hepatocellular carcinoma, but not in renal cancer (Wils and Bijlsma 2018; Paluszczak et al. 2017). RASAL2 and RASSF1 are tumor suppressors, known to be epigenetically silenced in many cancers including RCC (Morrissey et al. 2001; McLaughlin et al. 2013). RAS effectors RIN1 and PAK6 are also hypermethylated in PRCC. They have been implicated in the prognosis of ccRCC patients (Feng et al. 2017; Liu et al. 2014). Other ccRCC associated genes that are also hypermethylated in PRCC include DLEC1, RUNX3 and TNFRSF10C (Ricketts et al. 2012; Chen et al. 2013). RUNX3 is involved in suppression of migration, invasion and angiogenesis.

Our study revealed that hypermethylation of gene families as another feature of PRCC. This includes clusters of cadherins and protocadherins. Interestingly, most genes of PCDHGs cluster showed hypermethylation of many CpGs at gene body. We also found homeobox genes clustered on different chromosomes (7 and 17) to be hypermethylated at promoter and gene body. These genes are shown to be hypermethylated in lung and pancreatic cancers (Vincent et al. 2011). Polycomb group of proteins promote hypermethylation of homeobox proteins and can facilitate the efficient coordination of chromatin modifications during carcinogenesis (Vincent et al. 2011; Soshnikova and Duboule 2009). In this context, we found JARID2 that controls Polycomb machinery (Sanulli et al. 2015) to be hypermethylated at gene body. We also found microRNAs (MIR10B, MIR125B1 and MIR-429) associated with cancer to be hypermethylated in PRCC. MIR10B is a tumor suppressor in ccRCC while MIR125B1 has been known to modulate PI3K/AKT and MAPK/ERK signaling pathways (He et al. 2015; Wang et al. 2017). MIR-429 suppresses tumor migration and invasion (Guo et al. 2018). Hypomethylated genes in PRCC are associated with immune function, which is consistent with previous studies on ccRCC (Lasseigne et al. 2014; Wozniak et al. 2013). PTPRN2 is overexpressed in various cancers while another candidate CUX1 is involved in cell cycle progression, differentiation and regulates tumor invasiveness (Ripka et al. 2010; Sorokin et al. 2015).

Although we found several genes to be differentially hypermethylated, only a small proportion showed evidence of downregulation in gene expression. Therefore, aberrant methylation might not be always linked to selection-driven gene silencing. However, we found novel candidates including FAM83F, CNKSR1, IL17RE and NECTIN4 to be hypermethylated and downregulated in PRCC. Several studies have found FAM83 family to be associated with poor prognosis (Snijders et al. 2017; Bartel et al. 2016). Recently, FAM83F downregulation is shown to decrease DNA-damage induced response and increase cell proliferation (Salama et al. 2019). CNKSR1, a kinase suppressor of Ras1, and IL17RE, a receptor for interleukin-17C, are linked to poor survival in pancreatic cancer and hepatocellular carcinoma, respectively (Quadri et al. 2017; Liao et al. 2013). NECTIN4 is a prognostic biomarker for breast cancer while PYGM is downregulated in colon, breast, bladder and head and neck cancers (Nishiwada et al. 2015; Smutna et al. 2014).

We found genes including HHLA2, HAVCR2, SIRPA, APOL1 and CMTM3 to be hypomethylated at the promoter and upregulated in PRCC. HHLA2 is an immune checkpoint gene that is known to be upregulated in ccRCC (Chen et al. 2019). However, we found that HHLA2 is both hyper- and hypo-methylated in tumor samples compared to normal samples suggesting immune suppression only in hypomethylated samples. HAVCR2 is a T-cell immunoglobulin and like HHLA2 is an immune checkpoint gene (Anderson 2014). SIRPA encodes proteins that are expressed on the surface of macrophages and serves as an immune checkpoint (Weiskopf 2017). These genes encoding proteins are targeted by cancer immunotherapy. APOL1 is linked to autophagy induction and postulated to protect against RCC (Hu et al. 2012). CMTM3 is upregulated in PRCC but is known to inhibit migration and invasion in gastric cancer (Su et al. 2014). Further, we found a correlation between gene-body hypomethylation and gene expression for MET, PVT1 and ABCC3. The upregulation of long noncoding RNA PVT1 is known to inhibit apoptosis and is associated with poor prognosis in many cancers including ccRCC while upregulation of ABCC3 is related to cell proliferation, drug resistance and aerobic glycolysis (He et al. 2018; Liu et al. 2016a). This gene body correlation also exists for RRM2, NCAPG and SLC7A11. RRM2 is upregulated in various cancers and is linked to chemoresistance (Lu et al. 2012). We have also shown that RRM2, NCAPG and SLC7A11 are biomarkers for distinguishing the early and late stages of PRCC in our previous study (Singh et al. 2018).

Further, predictive models were developed to distinguish between early and late stages of PRCC. A comparative

study of different feature selection algorithms, predictive models, data integration techniques and representations of methylation data was performed. The feature selection using varSelRF yielded least number of features that are mostly hypermethylated in the late stage of PRCC compared to SC features (Fig. 7). The CpgM representation of DNA methylation yielded the best performance with varSelRF features across multiple models (Fig. S7). The extracted features from CpgM representation also include probes from the gene body, which perhaps explain the drop in performance using GeneM representation. This also suggests that BmpM representation can yield better performance with appropriate choice of bin size and quantifying summary such as median for aggregating the probes. Both varSelRF and SC extracted features from at least two partitions of data map to the promoter (TSS200) of CDO1, BHLHE23 and CLDN6 and body of GDF6. CDO1 is part of taurine biosynthesis pathway and its methylation is associated with poor survival of ccRCC patients (Deckers et al. 2015). Hypermethylation of CLDN6 is linked to breast cancer cell migration and invasion (Liu et al. 2016b). Further, varSelRF features also map to the promoter (TSS200) of GPR150 and body of SCRT2, while SC features map to the promoter of PROM1, SLC6A3, MIR375, TTBK1, RGS22, C2CD4B, PTGDR and ESRRG, 1st exon of TRH, BARHL2, TOX2, ASCL2, ASCL4 and EPHX3, and body of TNFRSF10C, B4GALNT1 and CYP26C1. We found both PROM1 and PROM2 (Figure S5) to be hypermethylated. Their expression levels are associated with clinical prognosis of cancer (Saha et al. 2019). SLC6A3 encodes dopamine transporter and is known to be overexpressed in ccRCC (Hansson et al. 2017). ESRRG, RGS22 and MIR375 are known to act as a tumor suppressor (Yan et al. 2014; Hu et al. 2011; Kang et al. 2018). EPHX3 and PTGDR (prostaglandin D2 receptor) are hypermethylated in the advanced stage of cancers (Sugino et al. 2007; Stott-Miller et al. 2014). ACSL2 is associated with the aggressive CIMP phenotype (Arai et al. 2012).

This is the first attempt to build integrative models for PRCC. Integration of methylation and gene expression data improved the performance for multiple models across different feature selection algorithms. The increase in MCC with data integration is due to the increase in either specificity, sensitivity or both. Although we observed a higher performance with integrated models, there is still scope for improvement, especially with respect to the late-stage classification. It is a complex problem since the sample size is small which in turn is compounded by class imbalance and heterogeneity within PRCC. The availability of more samples will aid in further testing and improving the performance of the models. Integrating the other available data such as CNVs, SNPs and miRNAs might also help towards improving the performance.

Funding P.K.V acknowledges financial support from the Early Career Research Award Scheme, Science and Engineering Research Board, DST, India (ECR/2016/000488).

Data availability The code for building the models and additional data can be found at https://github.com/NPSDC/IGAMS.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ali M, Khan SA, Wennerberg K, Aittokallio T (2018) Global proteomics profiling improves drug sensitivity prediction: results from a multi-omics, pan-cancer modeling approach. Bioinformatics 34(8):1353–1362

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11(10):R106

Anderson AC (2014) Tim-3: an emerging target in the cancer immunotherapy landscape. Cancer Immunol Res 2(5):393–398

Arai E, Chiku S, Mori T, Gotoh M, Nakagawa T, Fujimoto H, Kanai Y (2012) Single-CpG-resolution methylome analysis identifies clinicopathologically aggressive CpG island methylator phenotype clear cell renal cell carcinomas. Carcinogenesis 33(8):1487–1493

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA (2014) Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. Bioinformatics 30(10):1363–1369

Bartel CA, Parameswaran N, Cipriano R, Jackson MW (2016) FAM83 proteins: fostering new interactions to drive oncogenic signaling and therapeutic resistance. Oncotarget 7(32):52597–52612

Baylin SB (2006) DNA methylation and gene silencing in cancer. ChemInform. https://doi.org/10.1002/chin.200622296

Baylin SB, Jones PA (2011) A decade of exploring the cancer epigenome—biological and translational implications. Nat Rev Cancer 11(10):726–734

Chen F, Bai J, Li W, Mei P, Liu H, Li L, Pan Z, Wu Y, Zheng J (2013) RUNX3 suppresses migration, invasion and angiogenesis of human renal cell carcinoma. PLoS ONE 8(2):e56241

Chen F, Zhang Y, Şenbabaoğlu Y, Ciriello G, Yang L, Reznik E, Shuch B et al (2016) Multilevel genomics-based taxonomy of renal cell carcinoma. Cell Rep 14(10):2476–2489

Chen G, Wang Y, Wang L, Xu W (2017) Identifying prognostic biomarkers based on aberrant DNA methylation in kidney renal clear cell carcinoma. Oncotarget 8(3):5268–5280

Chen D, Chen W, Xu Y, Zhu M, Xiao Y, Shen Y, Zhu S, Cao C, Xu X (2019) Upregulated immune checkpoint HHLA2 in clear cell renal cell carcinoma: a novel prognostic biomarker and potential therapeutic target. J Med Genet 56(1):43–49

Chicco D (2017) Ten quick tips for machine learning in computational biology. BioData Mining 10:35

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS et al (2016) TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 44(8):e71

Cortes C, Vapnik V (1995) Support-Vector networks. Mach Learn 20(3):273–297

Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, Bansal M et al (2014) A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol 32(12):1202–1212

De Caceres II, Dulaimi E, Hoffman AM, Al-Saleem T, Uzzo RG, Cairns P (2006) Identification of novel target genes by an epigenetic reactivation screen of renal cancer. Cancer Res 66(10):5021–5028

Deckers IA, Schouten LJ, Van Neste L, Van Vlodrop IJ, Soetekouw PM, Baldewijns MM, Jeschke J et al (2015) Promoter methylation of CDO1 identifies clear-cell renal cell cancer patients with poor survival outcome. Clin Cancer Res 21(15):3492–3500

Díaz-Uriarte R, De Andres SA (2006) Gene selection and classification of microarray data using random forest. BMC Bioinform 7:3

Durinck S, Stawiski EW, Pavía-Jiménez A, Modrusan Z, Kapur P, Jaiswal BS, Zhang N et al (2015) Spectrum of diverse genomic alterations define non-clear cell renal carcinoma subtypes. Nat Genet. https://doi.org/10.1038/ng.3146

Eden A, Gaudet F, Waghmare A, Jaenisch R (2003) Chromosomal instability and tumors promoted by DNA hypomethylation. Science 300(5618):455

Ellinger J, Holl D, Nuhn P, Kahl P, Haseke N, Staehler M, Siegert S et al (2011) DNA hypermethylation in papillary renal cell carcinoma. BJU Int 107(4):664–669

Feng ZH, Fang Y, Zhao LY, Lu J, Wang YQ, Chen ZH, Huang Y et al (2017) RIN1 promotes renal cell carcinoma malignancy by activating EGFR signaling through Rab25. Cancer Sci 108(8):1620–1627

Gonen M (2012) bayesian efficient multiple kernel learning. arXiv [cs. LG]. arXiv. http://arxiv.org/abs/1206.6465

Gönen M, Alpaydın E (2011) Multiple kernel learning algorithms. J Mach Learn Res 12:2211–2268

Guan Z, Zhang J, Wang J, Wang H, Zheng F, Peng J, Xu Y et al (2014) SOX1 down-regulates β-catenin and reverses malignant phenotype in nasopharyngeal carcinoma. Mol Cancer 13:257

Guo C, Zhao D, Zhang Q, Liu S, Sun MZ (2018) miR-429 suppresses tumor migration and invasion by targeting CRKL in hepatocellular carcinoma via inhibiting Raf/MEK/ERK pathway and epithelial–mesenchymal transition. Sci Rep 8(1):2375

Hansson J, Lindgren D, Nilsson H, Johansson E, Johansson M, Gustavsson L, Axelson H (2017) Overexpression of functional SLC6A3 in clear cell renal cell carcinoma. Clin Cancer Res 23(8):2105–2115

Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J, Flagg K et al (2017) DNA methylation markers for diagnosis and prognosis of common cancers. Proc Natl Acad Sci USA 114(28):7414–7419

He C, Zhao X, Jiang H, Zhong Z, Xu R (2015) Demethylation of miR-10b plays a suppressive role in ccRCC cells. Int J Clin Exp Pathol 8(9):10595–10604

He RQ, Qin MJ, Lin P, Luo YH, Ma J, Yang H, Hu XH, Chen G (2018) Prognostic significance of LncRNA PVT1 and its potential target gene network in human cancers: a comprehensive inquiry based upon 21 cancer types and 9972 Cases. Cell Physiol Biochem 46(2):591–608

Hsieh JJ, Le V, Cao D, Cheng EH, Creighton CJ (2018) Genomic classifications of renal cell carcinoma: a critical step towards the future application of personalized kidney cancer care with pan-omics precision. J Pathol 244(5):525–537

Hu Y, Xing J, Wang L, Huang M, Guo X, Chen L, Lin M et al (2011) RGS22, a novel cancer/testis antigen, inhibits epithelial cell invasion and metastasis. Clin Exp Metas 28(6):541–549

Hu CA, Klopfer EI, Ray PE (2012) Human apolipoprotein L1 (ApoL1) in cancer and chronic kidney disease. FEBS Lett. https://doi.org/10.1016/j.febslet.2012.03.002

Huang S, Chaudhary K, Garmire LX (2017) More is better: recent progress in multi-omics data integration methods. Front Genet 8:84

IlluminaHumanMethylation450kanno, Hansen K (2014) ilmn12. hg19: annotation for illumina's 450k methylation arrays. R Package Version 0.2. 1

Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol 41(1):200–209

Jiang Y, Shi X, Zhao Q, Krauthammer M, Rothberg BE, Ma S (2016) Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. Genomics 107(6):223–230

Jiao Y, Widschwendter M, Teschendorff AE (2014) A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. Bioinformatics 30(16):2360–2366

Jonasch E, Gao J, Rathmell WK (2014) Renal cell carcinoma. BMJ 349:g4797

Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13(7):484–492

Kang MH, Choi H, Oshima M, Cheong JH, Kim S, Lee JH, Park YS et al (2018) Estrogen-related receptor gamma functions as a tumor suppressor in gastric cancer. Nat Commun 9(1):1920

Kim D, Joung JG, Sohn KA, Shin H, Park YR, Ritchie MD, Kim JH (2015) Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. J Am Med Inform Assoc JAMIA 22(1):109–120

Kim D, Li R, Lucas A, Verma SS, Dudek SM, Ritchie MD (2017) Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. J Am Med Inform Assoc JAMIA 24(3):577–587

Klacz J, Wierzbicki PM, Wronska A, Rybarczyk A, Stanislawowski M, Slebioda T, Olejniczak A, Matuszewski M, Kmiec Z (2016) Decreased expression of rassf1a tumor suppressor gene is associated with worse prognosis in clear cell renal cell carcinoma. Int J Oncol 48(1):55–66

Kluzek K, Bluyssen HA, Wesoly J (2015) The epigenetic landscape of clear-cell renal cell carcinoma. J Kidney Cancer VHL 2(3):90–104

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw377

Kulis M, Heath S, Bibikova M, Queirós AC, Navarro A, Clot G, Martínez-Trillos A et al (2012) Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet 44(11):1236–1242

Lasseigne BN, Brooks JD (2018) The role of DNA methylation in renal cell carcinoma. Mol Diagn Ther. https://doi.org/10.1007/s40291-018-0337-9

Lasseigne BN, Burwell TC, Patil MA, Absher DM, Brooks JD, Myers RM (2014) DNA methylation profiling reveals novel diagnostic biomarkers in renal cell carcinoma. BMC Med 12:235

Li X, Chen C, Wang F, Huang W, Liang Z, Xiao Y, Wei K et al (2014) KCTD1 suppresses canonical Wnt signaling pathway by enhancing β-catenin degradation. PLoS ONE 9(4):e94343

Liao R, Sun J, Wu H, Yi Y, Wang JX, He HW, Cai XY et al (2013) High expression of IL-17 and IL-17RE associate with poor prognosis of hepatocellular carcinoma. J Exp Clin Cancer Res 32:3

Lin E, Lane HY (2017) Machine learning and systems genomics approaches for multi-omics data. Biomark Res 5:2

Liu W, Liu H, Liu Y, Xu L, Zhang W, Zhu Y, Xu J, Gu J (2014) Prognostic significance of p21-activated kinase 6 expression in

patients with clear cell renal cell carcinoma. Ann Surg Oncol 21(Suppl 4):S575–S583
Liu X, Yao D, Liu C, Cao Y, Yang Q, Sun Z, Liu D (2016a) Overexpression of ABCC3 promotes cell proliferation, drug resistance, and aerobic glycolysis and is associated with poor prognosis in urinary bladder cancer patients. Tumour Biol 37(6):8367–8374
Liu Y, Jin X, Li Y, Ruan Y, Lu Y, Yang M, Lin D et al (2016b) DNA methylation of claudin-6 promotes breast cancer cell migration and invasion by recruiting MeCP2 and deacetylating H3Ac and H4Ac. J Exp Clin Cancer Res 35(1):120
Llinàs-Arias P, Esteller M (2017) Epigenetic inactivation of tumour suppressor coding and non-coding genes in human cancer: an update. Open Biol. https://doi.org/10.1098/rsob.170152
Lu AG, Feng H, Pu-Xiong-Zhi Wang DP, Han XH, Zheng MH (2012) Emerging roles of the ribonucleotide reductase M2 in colorectal cancer and ultraviolet-induced DNA damage repair. World J Gastroenterol 18(34):4704–4713
Mankoo PK, Shen R, Schultz N, Levine DA, Sander C (2011) Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. PLoS ONE 6(11):e24709
Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochem Biophys Acta 405(2):442–451
Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE et al (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466(7303):253–257
McLaughlin SK, Olsen SN, Dake B, De Raedt T, Lim E, Bronson RT, Beroukhim R et al (2013) The RasGAP gene, RASAL2, is a tumor and metastasis suppressor. Cancer Cell. https://doi.org/10.1016/j.ccr.2013.08.004
McMahon KW, Karunasena E, Ahuja N (2017) The roles of DNA methylation in the stages of cancer. Cancer J. https://doi.org/10.1097/ppo.0000000000000279
Moch H, Cubilla AL, Humphrey PA, Reuter VE, Ulbright TM (2016) The 2016 WHO classification of tumours of the urinary system and male genital organs-part A: renal, penile, and testicular tumours. Eur Urol 70(1):93–105
Modi PK, Singer EA (2016) Improving our understanding of papillary renal cell carcinoma with integrative genomic analysis. Ann Transl Med 4(7):143
Morris MR, Latif F (2017) The epigenetic landscape of renal cancer. Nat Rev Nephrol 13(1):47–60
Morris MR, Maher ER (2010) Epigenetics of renal cell carcinoma: the path towards new diagnostics and therapeutics. Genome Med 2(9):59
Morris MR, Gentle D, Abdulrahman M, Maina EN, Gupta K, Banks RE, Wiesener MS et al (2005) Tumor suppressor activity and epigenetic inactivation of hepatocyte growth factor activator inhibitor type 2/SPINT2 in papillary and clear cell renal cell carcinoma. Cancer Res 65(11):4598–4606
Morrissey C, Martinez A, Zatyka M, Agathanggelou A, Honorio S, Astuti D, Morgan NV et al (2001) Epigenetic inactivation of the RASSF1A 3p21.3 tumor suppressor gene in both clear cell and papillary renal cell carcinoma. Cancer Res 61(19):7277–7281
Nishiwada S, Sho M, Yasuda S, Shimada K, Yamato I, Akahori T, Kinoshita S, Nagai M, Konishi N, Nakajima Y (2015) Nectin-4 expression contributes to tumor proliferation, angiogenesis and patient prognosis in human pancreatic cancer. J Exp Clin Cancer Res 34:30
Paluszczak J, Wiśniewska D, Kostrzewska-Poczekaj M, Kiwerska K, Grénman R, Mielcarek-Kuchta D, Jarmuż-Szymczak M (2017) Prognostic significance of the methylation of wnt pathway antagonists-CXXC4, DACT2, and the inhibitors of sonic hedgehog signaling-ZIC1, ZIC4, and HHIP in head and neck squamous cell carcinomas. Clin Oral Investig 21(5):1777–1788
Quadri HS, Aiken TJ, Allgaeuer M, Moravec R, Altekruse S, Hussain SP, Miettinen MM, Hewitt SM, Rudloff U (2017) Expression of the scaffold connector enhancer of kinase suppressor of Ras 1 (CNKSR1) is correlated with clinical outcome in pancreatic cancer. BMC Cancer 17(1):495
Rahimi A, Gönen M (2018) Discriminating early- and late-stage cancers using multiple kernel learning on gene sets. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty239
Rapetti-Mauss R, Bustos V, Thomas W, McBryan J, Harvey H, Lajczak N, Madden SF et al (2017) Bidirectional KCNQ1:β-catenin interaction drives colorectal cancer cell differentiation. Proc Natl Acad Sci USA 114(16):4159–4164
Revill K, Wang T, Lachenmayer A, Kojima K, Harrington A, Li J, Hoshida Y, Llovet JM, Powers S (2013) Genome-wide methylation analysis and epigenetic unmasking identify tumor suppressor genes in hepatocellular carcinoma. Gastroenterology 145(6):1424–1435.e1–25
Ricketts CJ, Morris MR, Gentle D, Brown M, Wake N, Woodward ER, Clarke N, Latif F, Maher ER (2012) Genome-wide CpG island methylation analysis implicates novel genes in the pathogenesis of renal cell carcinoma. Epigenetics 7(3):278–290
Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, Bowlby R et al (2018) The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. Cell Rep 23(12):3698
Ripka S, Neesse A, Riedel J, Bug E, Aigner A, Poulsom R, Fulda S et al (2010) CUX1: target of Akt signalling and mediator of resistance to apoptosis in pancreatic cancer. Gut 59(8):1101–1110
Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015) Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet 16(2):85–97
Saha SK, Islam SR, Kwak KS, Rahman MS, Cho SG (2019) PROM1 and PROM2 expression differentially modulates clinical prognosis of cancer: a multiomics analysis. Cancer Gene Ther. https://doi.org/10.1038/s41417-019-0109-7
Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. https://doi.org/10.1371/journal.pone.0118432
Salama M, Benitez-Riquelme D, Elabd S, Munoz L, Zhang P, Glanemann M, Mione MC et al (2019) Fam83F induces p53 stabilisation and promotes its activity. Cell Death Differ. https://doi.org/10.1038/s41418-019-0281-1
Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, Esteller M (2011) Validation of a dna methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. https://doi.org/10.4161/epi.6.6.16196
Sanulli S, Justin N, Teissandier A, Ancelin K, Portoso M, Caron M, Michaud A et al (2015) Jarid2 methylation via the PRC2 complex regulates H3K27me3 deposition during cell differentiation. Mol Cell 57(5):769–783
Seoane JA, Day IN, Gaunt TR, Campbell C (2014) A pathway-based data integration framework for prediction of disease progression. Bioinformatics 30(6):838–845
Shen S, Wang G, Shi Q, Zhang R, Zhao Y, Wei Y, Chen F, Christiani DC (2017) Seven-CpG-based prognostic signature coupled with gene expression predicts survival of oral squamous cell carcinoma. Clin Epigenet 9:88
Shenoy N, Vallumsetla N, Zou Y, Galeas JN, Shrivastava M, Hu C, Susztak K, Verma A (2015) Role of DNA methylation in renal cell carcinoma. J Hematol Oncol 8:88
Singh NP, Bapi RS, Vinod PK (2018) Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma. Comput Biol Med 100:92–99
Smutna V, Dieci MV, Lefebvre C, Scott V, Andre F, Fromigue O (2014) Abstract 435: is PYGM dysregulation involved in breast cancer

cell metabolism. Cancer Res. https://doi.org/10.1158/1538-7445.am2014-435
Snijders AM, Lee SY, Hang B, Hao W, Bissell MJ, Mao JH (2017) FAM83 family oncogenes are broadly involved in human cancers: an integrative multi-omics approach. Mol Oncol 11(2):167–179
Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45(4):427–437
Sorokin AV, Nair BC, Wei Y, Aziz KE, Evdokimova V, Hung MC, Chen J (2015) Aberrant expression of proPTPRN2 in cancer cells confers resistance to apoptosis. Cancer Res 75(9):1846–1858
Soshnikova N, Duboule D (2009) Epigenetic regulation of vertebrate hox genes: a dynamic equilibrium. Epigenetics 4(8):537–540
Stott-Miller M, Zhao S, Wright JL, Kolb S, Bibikova M, Klotzle B, Ostrander EA, Fan JB, Feng Z, Stanford JL (2014) Validation study of genes with hypermethylated promoter regions associated with prostate cancer recurrence. Cancer Epidemiol Biomark Prev 23(7):1331–1339
Su Y, Lin Y, Zhang L, Liu B, Yuan W, Mo X, Wang X et al (2014) CMTM3 inhibits cell migration and invasion and correlates with favorable prognosis in gastric cancer. Cancer Sci 105(1):26–34
Sugino Y, Misawa A, Inoue J, Kitagawa M, Hosoi H, Sugimoto T, Imoto I, Inazawa J (2007) Epigenetic silencing of prostaglandin E receptor 2 (PTGER2) is associated with progression of neuroblastomas. Oncogene. https://doi.org/10.1038/sj.onc.1210550
Taskesen E, Babaei S, Reinders MM, de Ridder J (2015) Integration of gene expression and DNA-methylation profiles improves molecular subtype classification in acute myeloid leukemia. BMC Bioinform 16(Suppl 4):S5
The Cancer Genome Atlas Research Network (2016) Comprehensive molecular characterization of papillary renal-cell carcinoma. N Engl J Med. https://doi.org/10.1056/nejmoa1505917
Thomas J, Sael L (2017) Multi-kernel LS-SVM based integration bioclinical data analysis and application to ovarian cancer. Int J Data Mining Bioinform. https://doi.org/10.1504/ijdmb.2017.089281
Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.082099299
Vincent A, Omura N, Hong SM, Jaffe A, Eshleman J, Goggins M (2011) Genome-wide analysis of promoter methylation associated with gene expression profile in pancreatic adenocarcinoma. Clin Cancer Res 17(13):4341–4354
Wang Y, Zhao M, Liu J, Sun Z, Ni J, Liu H (2017) miRNA-125b regulates apoptosis of human non-small cell lung cancer via the PI3K/Akt/GSK3β signaling pathway. Oncol Rep 38(3):1715–1723
Wang Z, Teng D, Li Y, Hu Z, Liu L, Zheng H (2018) A six-gene-based prognostic signature for hepatocellular carcinoma overall survival prediction. Life Sci 203:83–91
Wei JH, Haddad A, Wu KJ, Zhao HW, Kapur P, Zhang ZL, Zhao LY et al (2015) A CpG-methylation-based assay to predict survival in clear cell renal cell carcinoma. Nat Commun 6:8699
Weiskopf K (2017) Cancer immunotherapy targeting the CD47/SIRPα axis. Eur J Cancer 76:100–109
Wils LJ, Bijlsma MF (2018) Epigenetic regulation of the hedgehog and Wnt pathways in cancer. Crit Rev Oncol Hematol 121:23–44
Wozniak MB, Le Calvez-Kelm F, Abedi-Ardekani B, Byrnes G, Durand G, Carreira C, Michelon J et al (2013) Integrative genome-wide gene expression profiling of clear cell renal cell carcinoma in Czech Republic and in the United States. PLoS ONE 8(3):e57886
Wu C, Zhou F, Ren J, Li X, Jiang Y, Ma S (2019) A selective review of multi-level omics data integration using variable selection. High-Throughput. https://doi.org/10.3390/ht8010004
Xu Z, Jin R, Yang H, King I, Lyu MR (2010) Simple and efficient multiple kernel learning by group lasso. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 1175–1182. Citeseer
Yan JW, Lin JS, He XX (2014) The emerging role of miR-375 in cancer. Int J Cancer 135(5):1011–1018
Yan KK, Zhao H, Pang H (2017) A comparison of graph- and kernel-based –omics data integration algorithms for classifying complex traits. BMC Bioinform. https://doi.org/10.1186/s12859-017-1982-4
Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G (2014) Gene body methylation can alter gene expression and is a therapeutic target in cancer. Cancer Cell 26(4):577–590
Zhu B, Song N, Shen R, Arora A, Machiela MJ, Song L, Landi MT et al (2017) Integrating clinical and multiple omics data for prognostic assessment across human cancers. Sci Rep. https://doi.org/10.1038/s41598-017-17031-8

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