Moarii *et al. Human Genomics* (2015) 9:26
DOI 10.1186/s40246-015-0048-9

Human Genomics

PRIMARY RESEARCH Open Access

Integrative DNA methylation and gene expression analysis to assess the universality of the CpG island methylator phenotype

Matahi Moarii[1,2,3], Fabien Reyal[4,5,6] and Jean-Philippe Vert[1,2,3*]

Abstract

Background: The CpG island methylator phenotype (CIMP) was first characterized in colorectal cancer but since has been extensively studied in several other tumor types such as breast, bladder, lung, and gastric. CIMP is of clinical importance as it has been reported to be associated with prognosis or response to treatment. However, the identification of a universal molecular basis to define CIMP across tumors has remained elusive.

Results: We perform a genome-wide methylation analysis of over 2000 tumor samples from 5 cancer sites to assess the existence of a CIMP with common molecular basis across cancers. We then show that the CIMP phenotype is associated with specific gene expression variations. However, we do not find a common genetic signature in all tissues associated with CIMP.

Conclusion: Our results suggest the existence of a universal epigenetic and transcriptomic signature that defines the CIMP across several tumor types but does not indicate the existence of a common genetic signature of CIMP.

Background

Epigenetic modifications have been recognized as important players in cancer etiology and development and constitute promising therapeutic targets for diagnosis or treatment due to their possible reversibility [1–3]. In particular, aberrant methylation of CpG islands (CGIs) located in promoter regions of tumor suppressor and DNA repair genes, leading to their silencing, is now considered a hallmark of cancer playing an important role in neoplasia [1–6].

The CpG island methylator phenotype (CIMP) was first defined and observed by [7] in a subset of colorectal cancers as the joint methylation of several promoter regions, leading to the inactivation of the corresponding genes. The stratification of patients based on CIMP was shown to be clinically relevant, as CIMP-positive patients had better prognosis than CIMP-negative ones, and could lead to personalized treatments. Since the identification of CIMP in colorectal cancers, many studies have tried to replicate the analysis to find CIMP in different types of cancers including but not limited to colon [8–12], breast [13, 14], lung [15], stomach [16], and glioblastoma [17–19]. While most of these works concluded in the existence of a CIMP in different cancers, other studies did not yield the same conclusions [20, 21], and the genes whose promoter CGI methylation are considered to define the CIMP differ between studies. This raises the question of whether the CIMP is tissue specific or is a universal phenomenon with common biological causes affecting common genes across cancers. A recent review of CIMP-related studies across different cancers pointed out the diversity of methods and measurement technologies used to define CIMP, which hinders the establishment of a molecular basis for CIMP in spite of growing evidence linking mutations in specific genes and CIMP in several cancers [22].

In the present study, we investigate the existence and universality of CIMP by performing a systematic genome-wide methylation analysis on several large datasets of different cancer types simultaneously. We propose a simple methodology to assess the existence of a CIMP phenotype

*Correspondence: jean-philippe.vert@mines-paristech.fr
[1] CBIO-Centre for Computational Biology, Mines Paristech, PSL-Research University, 35 Rue Saint-Honore, F-77300 Fontainebleau, France
[2] Department of Bioinformatics, Biostatistics and System Biology, Institut Curie, 11-13 Rue Pierre et Marie Curie, F-75248 Paris, France
Full list of author information is available at the end of the article

in each cancer and to identify a set of genes whose promoter methylation is a marker for the CIMP. This allows us to compare the different cancer types in search for a cross-cancer CIMP signature and to analyze the link between CIMP and gene expression in different cancers. Finally, we assess the clinical relevance of CIMP on the overall survival.

Results

A cross-cancer CIMP signature

We first assess with a common methodology whether a CIMP can be detected on different cancers and whether CIMP in different cancers share a common signature in terms of which gene promoters are hypermethylated in CIMP-positive patients. For that purpose, we collected high-density methylation datasets from the cancer genome atlas (TCGA) data portal providing more than 485,000 CpG methylation levels for more than 2000 samples from five tissues of origin: bladder, breast, colon, lung, and stomach (Table 1). For each sample, we aggregate the methylation levels of CpG probes by CGI, including the CGI itself and its shores and shelves, resulting in a single methylation level for each of 21,176 CGIs in each sample.

A CIMP corresponds to the joint hypermethylation of a subset of CGIs in a subset of samples [7]. To characterize from whole-genome methylation data whether a CIMP exists for a cancer and which CGIs characterize it, we follow a standard methodology: (i) select the 5 % most variant CGIs in the set of samples, which we call the *CIMP signature* and (ii) check by unsupervised classification whether the samples cluster into two main clusters (CIMP-positive and -negative clusters) when we restrict them to the methylation values they take on the CGIs in the CIMP signature.

We apply this methodology to each of the five families of tumors, cutting the tree obtained by hierarchical clustering to two clusters in order to enforce a classification of all samples into two subgroups based on the methylation of CGIs in the CIMP signature. Interestingly, in all five cases, one of the two clusters is clearly characterized by an overall hypermethylation of most CGIs in the signature compared to the second cluster, allowing us to characterize it as the CIMP-positive cluster, the second one being the CIMP-negative cluster (Additional file 1). The proportion of CIMP-positive samples according to this definition varies from about 20 % for breast and colon cancers to 30 % for bladder and about 60 and 70 % for stomach and lung cancers (Table 2). Proportion of the CIMP-positive group in each tissue is similar to previously reported studies [22]. Varying the size of the CIMP signature from 1 to 10 % of all CGIs had a small impact on the clustering stability (Additional file 2).

Comparing the epigenetic signatures that define CIMP for each tissue, we find a common set of 89 CGIs associated with 51 genes (Fig. 1a). If the signatures were random subsets of 5 % of all CGIs independent from each other, the overlap would contain on average $(5\ \%)^5 \simeq 3.10^{-5}$ % of all CGIs, namely 0.006 CGI. This provides a strong evidence that a common set of genes is involved in CIMP in different cancers. We call these 89 CGIs the *cross-cancer CIMP signature* (Table 3). A hierarchical clustering on all samples restricted to this cross-cancer CIMP signature is able to cluster CIMP-positive and CIMP-negative patients independently of the tissue of origin (Fig. 1b), suggesting that CIMP observed in each individual cancer share in common a significant proportion of genes whose promoter CGIs are hypermethylated in all CIMP-positive cancers. A functional enrichment analysis of the cross-cancer CIMP signature reveals that it is significantly enriched in genes involved in cell differentiation and neuronal developmental and immune response processes (Fig. 1c).

Are there 2 or 3 CIMP classes?

Several studies suggest the existence of a third class in CIMP phenotype that corresponds to an intermediate level of methylation [12, 23, 24]. While we enforced an

Table 1 Patients' dataset. Number of samples available for the different cancer types (first column) for methylation (Meth) and gene expression (GE). The "Meth/GE" column summarizes the number of samples with both methylation and gene expression, while the "Meth/Mutations" column shows the number of samples with both methylation and DNA mutation data

Tissue	Meth	GE	Meth/GE	Meth/Mutations
Bladder	373	56	*43*	*28*
Breast	626	778	*478*	*468*
Colon	291	193	*34*	*219*
Lung	452	125	*82*	*411*
Stomach	338	373	*309*	*199*
Overall	*2090*	*1525*	*941*	*1325*

Table 2 CIMP proportion. For each cancer type, this table shows the number of samples clustered in the CIMP-negative and CIMP-positive clusters and the percentage of CIMP-positive samples

Tissue	Negative	Positive	Ratio (%)
Bladder	262	111	30
Breast	509	117	19
Colon	232	59	20
Lung	136	316	70
Stomach	144	194	57
Overall	*1283*	*797*	*38*

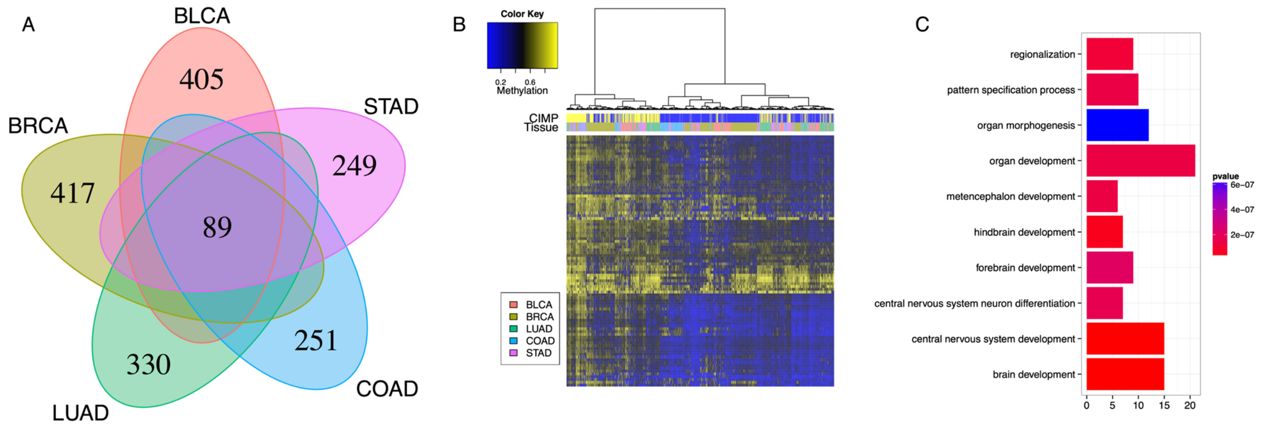

Fig. 1 Pan-cancer clustering on common epigenetic signature clusters CIMP-positive and CIMP-negative tumors independently of tissue type. **a** Venn diagram of the CIMP signatures for each tissue. **b** Hierarchical clustering on the common epigenetic signature for all tissues. **c** Gene ontology analysis of the genes associated with the promoters of the common epigenetic signature

analysis with 2 classes to define the CIMP of each sample as positive or negative in the previous section, we now examine whether the data call for a third class. Following [25], we assess the existence of an intermediate CIMP phenotype for each tissue by comparing the increase in empirical cumulative distributive distribution $\Delta(K)$ for different values of $K = 2, \ldots, 5$ where K is the number of clusters considered for CIMP.

Figure 2 shows how $\Delta(K)$ varies as a function of K for each cancer, suggesting how many clusters exist in each case. We observe that the existence of a third class is not clear-cut. While colon and breast tissues show a significant increase in $\Delta(K)$ for $K = 3$ suggesting a possible third cluster in CIMP, the bladder is flat between 2 and 3 clusters, while lung and gastric cancers do not support the presence of 3 classes. In addition, we assess the stability of 3 clusters by varying the number of CGIs that define CIMP and observed that while CIMP clusters are highly robust for $K = 2$, there is some high variability in the cluster definitions for $K = 3$ (Additional file 2). In summary, the presence of 2 clusters is well supported by the data in all cancers, while the third cluster is much more debatable.

Table 3 List of genes associated with the common set of CGIs that define CIMP in each tissue

Epigenetic Signature	LOC339524, GSTM1, CD1D, LMX1A
	CACNA1E ,NR5A2, WNT3A, GNG4
	EMX1, CTNNA2 ,LRRTM1, DLX1
	EVX2, HOXD13, GBX2, SYN2
	HAND2, NBLA00301, EBF1, HIST1H2BB
	HIST1H3C, HLA-DRB1, C6orf186, IKZF1
	CDKN2A, HMX3, KNDC1, KLHL35
	HOTAIR, SLC6A15, ALX1, RFX4
	CLDN10, ADCY4, RIPK3, NID2
	OTX2, OTX2OS1, GSC, KIF26A
	GREM1, SEC14L5, HS3ST3B1, IGF2BP1
	HOOK2, NFIX, ZNF577, ZNF649
	CPXM1, CDH22, CHRNA4

Similar gene expression variations are predictive of CIMP

To shed light on the relationship between methylation and transcription, we now assess to what extent a transcriptomic signature can classify the samples as CIMP positive or negative. For that purpose, we collected for each family of cancer samples with both methylation and gene expression data available, leading to a subset of samples with an overall proportion of CIMP-positive samples comparable to that of the original dataset (Table 4). We measure by cross-validation how well expression data alone can recover the two CIMP classes.

We first perform a multivariate regression analysis using the lasso technique to assess whether gene expression of a few genes can be predictive of the CIMP status for each tissue separately. The cross-validation accuracies for each family of cancer are shown in Table 5. We observe that while a classifier based on gene expression performs significantly better than random to recover CIMP-positive samples in breast, lung, and stomach cancers, the performance on the bladder and colon is not different from a random classifier. Moreover, we compare the lists of genes selected in the transcriptomic signature after bootstrap resampling of the samples in order to assess their robustness and potential biological significance (Fig. 3a). We observe that very few genes are robustly selected in the signatures, and in particular that no gene is associated with BLCA-CIMP and COAD-CIMP prediction in

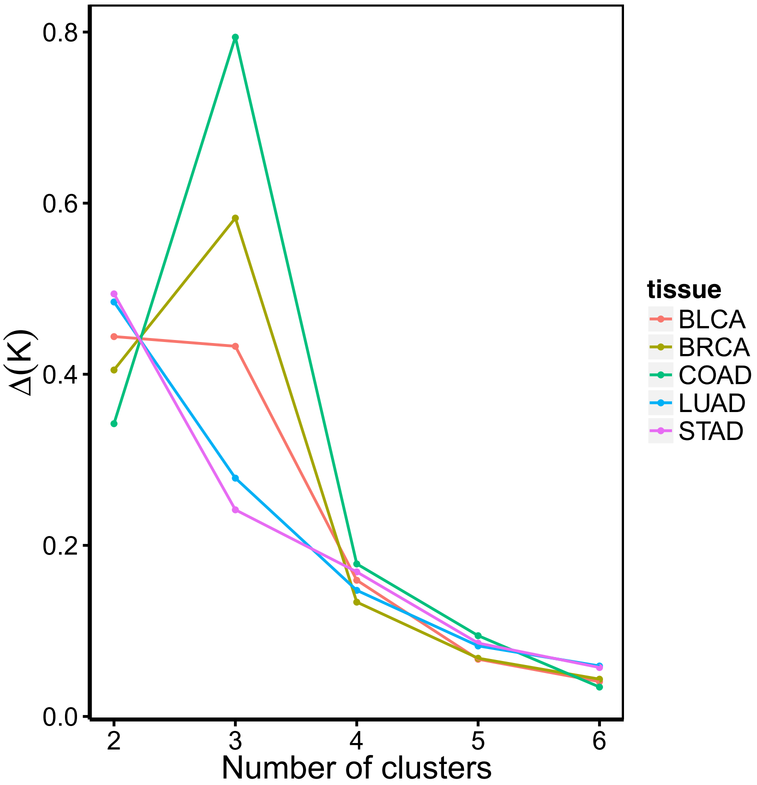

Fig. 2 Variation in the empirical cumulative distributive function $\Delta(K)$ for each tissue. The empirical cumulative distributive function $\Delta(K)$ is a data-driven criterion which can indicate the number of clusters K in the data when it reaches its maximum [25]. This plots shows how $\Delta(K)$ varies as a function of K, for the different tissues

more than 15 % of the bootstrap resampling . In addition, the transcriptomic signatures of different cancers are very diverse, and no gene is present in all of them (Fig. 3b). Overall, these results suggest that there is information in the transcriptome related to the CIMP status, but that a robust signature across cancers is difficult to obtain.

However, the poor accuracy as well as the non-robustness of genetic signatures to predict CIMP may be due to the small size of some datasets ($n_{\text{BLCA}} = 43$, $n_{\text{COAD}} = 34$). To overcome the lack of statistical power due to small sample size, we combine in a second analysis the different datasets into a single multivariate regression analysis, based on the assumption that the CIMP signatures of different cancers may share the same genes. We train classifiers to predict CIMP status from gene expression data jointly across cancers using two methods, based on two different assumptions: (i) assuming that all tissues share the same gene signature and coefficients for the prediction task, we run a single lasso classification on the combined datasets ("Combined-Lasso" prediction) or (ii) assuming that all tissues share the same gene signature but with different coefficients, we jointly train several models with a group lasso approach to constrain the selected genes to be the same across cancers without imposing their coefficients to coincide ("Group-Lasso" prediction) (see supplementary methods in Additional file 3). The rationale for the group lasso approach is that while CIMP may be caused by a common subset of genes, the specific contribution of each gene may vary between tissues. Our results show that both methods significantly outperform the tissue-specific predictions ($P \leq 2.10^{-16}$,

Table 4 CIMP Proportion in samples with both methylation and gene expression data. This table shows the number of CIMP-positive and CIMP-negative samples characterized by both methylation and gene expression data, for each cancer type, as well as the proportion of CIMP-positive samples

Tissue	Negative	Positive	Ratio (%)
Bladder	27	16	37
Breast	385	93	20
Colon	27	7	20
Lung	22	60	75
Stomach	131	178	58
Overall	*592*	*354*	*37*

Table 5 Accuracy of CIMP prediction using gene expression profiles

	Accuracy			
	Random	Lasso	Combined lasso	Group lasso
Bladder	62.8	62.9 ($p = 1$)	74.2 $\left(p \leq 2.10^{-16}\right)$	72.1 $\left(p \leq 2.10^{-16}\right)$
Breast	80.5	83.9 $\left(p \leq 2.10^{-16}\right)$	84.7 $\left(p \leq 2.10^{-16}\right)$	85.5 $\left(p \leq 2.10^{-16}\right)$
Colon	79.4	79.5 ($p = 1$)	95.0 $\left(p \leq 2.10^{-16}\right)$	94.2 $\left(p \leq 2.10^{-16}\right)$
Lung	73.2	84.2 $\left(p \leq 2.10^{-16}\right)$	76.2 $\left(p \leq 2.10^{-16}\right)$	86.6 $\left(p \leq 2.10^{-16}\right)$
Stomach	57.6	81.2 $\left(p \leq 2.10^{-16}\right)$	83.0 $\left(p \leq 2.10^{-16}\right)$	84.8 $\left(p \leq 2.10^{-16}\right)$
Overall	*71.9*	*82.4*	*82.6*	*85.0*

This table shows the accuracy, assessed by threefold cross-validation repeated 100 times over each tissue (first column), of sample classification in CIMP-positive and CIMP-negative classes from gene expression data using random classification (second column), lasso logistic regression (third column), combined lasso (fourth column), or group lasso logistic regression (fifth column)

Fig. 3c, Table 5) in particular for the bladder and colon where the size of the initial datasets could not give sufficient statistical power to predict CIMP accurately. There is overall little difference between both methods, with the notable exception of lung cancer where the combined lasso approach is significantly worse than the group lasso (and even the single lasso) model, suggesting that in that case, the weights of the genes in the CIMP signature may differ from other cancers. More importantly, each method allows to identify a common genetic signature (51 genes for the "Combined" prediction and 58 genes for the "Group-Lasso" prediction) that distinguishes CIMP-positive and CIMP-negative class for each tumors which is more robust than all the tissue-specific signatures (Fig. 3d). In addition, these signatures share a large common set of genes (25 common genes, Table 6). We represented the gene expression distribution for this common set of genes on the different datasets and observe a clear separation between CIMP-positive and CIMP-negative classes for all tissues (Additional file 4). Gene ontology analysis on the intersection of the two predictive gene signatures showed specific enrichment only for genetic regulatory processes.

A genetic signature is associated to CIMP only for colon and gastric cancers

Several somatic mutations have been found to be tightly associated with epigenetic aberrations in CIMP. Recent studies have pointed out the causal role of IDH1 mutations in Glioblastoma-CIMP [17, 19] and tight associations between IDH2 and TET2 mutations with other CIMPs (leukemia [26], enchondroma, and spindle cell hemangioma [27, 28]). In the colon, BRAF and KRAS mutations are associated with microsatellite instability and COAD-CIMP [9].

We re-assess the association between mutations in these genes and CIMP in the different types of cancers (Fig. 4a). We recover a strong association between BRAF mutation and CIMP-positive colon tumors but no specific association with other tumor types. We also find no coordinated association between *IDH1*, *IDH2*, *KRAS*, *BRAF*, or *TET2* mutations and CIMP phenotypes for all tissues. In addition, we perform genome-wide mutation analysis to assess whether specific gene mutations are associated with CIMP. We find no significant gene mutation association for bladder, breast nor lung CIMPs. For colon and gastric cancer, we find respectively 459 and 1070 gene mutations associated with CIMP with a common intersection of 195 genes (Additional file 5 panel A). Gene ontology analysis of this set of genes shows significant enrichment for extracellular matrix organization and cell adhesion but also neuronal developmental processes (Additional file 5 panel B).

Finally, we also look at the rate of mutations in each tissue given the CIMP phenotype. We observe a significant association between the number of mutations and the CIMP status for colon and gastric cancer (Fig. 4b), in accordance with the tight association between CIMP and microsatellite instability for these two tissues [9, 29–31]. However, the same observation could not be made for the bladder, breast, and lung.

Clinical impact of CIMP

Survival analysis in several CIMP studies has often shown distinct outcome between CIMP-positive and CIMP-negative tumors. However, there is no consensus in the general survival associated with CIMP: while CIMP has been associated with improved survival and lower risk of metastasis in breast [14], colorectal [9], leukemia [32–35], or gliomas [17], it has also been reportedly associated with poor survival for bladder [36], lung [15, 37], or prostate cancers [38], and prognosis even remains unclear for gastric cancers [39–43].

We perform a systematic survival analysis on the different tissues to assess the clinical impact of CIMP. However, we observe no significant association between CIMP and survival, in any of the tissues (Table 7 and Additional file 6).

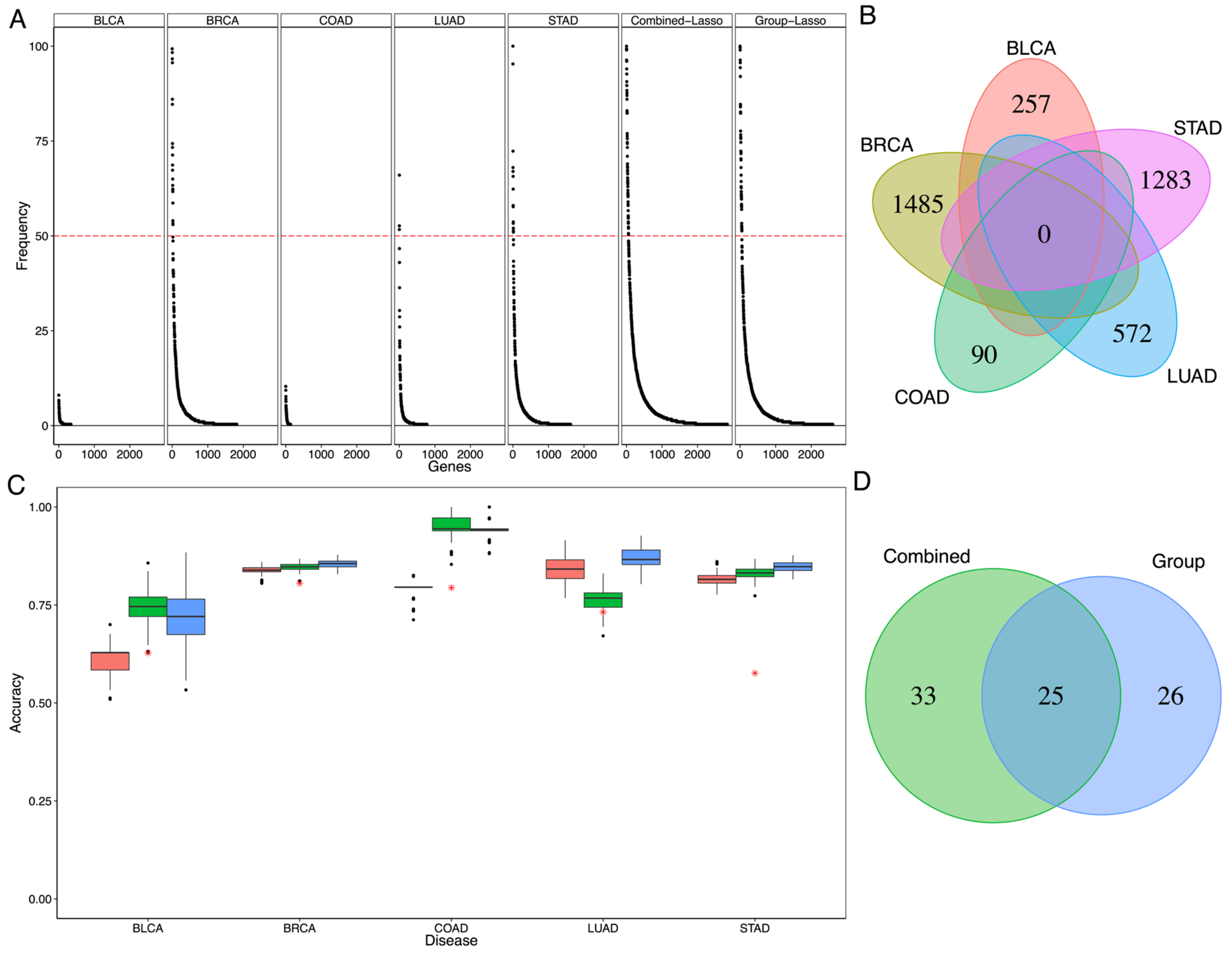

Fig. 3 Gene expression variations predictive of CIMP. **a** Stability of each gene signature for each tissue-specific CIMP prediction as well as the "Combined-Lasso" and the "Group-Lasso" CIMP prediction task obtained and ranked by frequency of appearance using bootstrap ($n = 100$ repeats). For bladder and colon CIMP prediction task, the signature was non-robust (frequency of the most redundant gene inferior to 10 %). The combined prediction task signature outperforms the tissue-specific signatures in robustness. **b** Venn diagram of the tissue-specific gene signatures using lasso for each tissue separately. **c** Distribution of the accuracy of the CIMP-phenotype prediction task given the patient gene expression profile using $n = 100$ bootstrap and threefold cross-validation for several methods (*pink* = "tissue-specific" lasso, *green* = "Combined-Lasso," *blue* = "Group-Lasso," *red star* = random prediction). **d** Venn diagram representing the intersection between the "Combined" and "Group" lasso gene signatures

Other clinical parameters have been associated with CIMP such as microsatellite instability (MSI) in the colon [9] and hormone receptor statuses in the breast [14]. We therefore assess the association between the CIMP status and eight clinical annotations provided in the TCGA, namely, age, MSI, ER status, PR status, HER2 status, tumor size, lymph node invasion, and presence of metastasis. We first observe that CIMP is significantly associated with a higher age in the breast, colon, and stomach ($P_{\text{breast}} = 2.10^{-4}$, $P_{\text{colon}} = 2.10^{-3}$, $P_{\text{stomach}} = 0.036$, student test, Additional file 7 panel A) but not in the bladder and lung. In the colon, we recover a significant association between CIMP and MSI ($P = 5.10^{-6}$, chi-squared test, Additional file 7 panel B). We also recover a significant association between CIMP and ER, PR, and HER2 statuses in breast ($P_{\text{ER}} = 2.10^{-5}$, $P_{\text{PR}} = 0.03$, $P_{\text{HER2}} = 5.10^{-8}$, chi-squared test, Additional file 7 panel C). However, we observed no significant association between CIMP and either tumor size, lymph node invasion, or metastasis in any tissue.

Discussion

CIMP has been thoroughly studied over the past few years in several tissue types but the heterogeneity of the methods and measurement technologies has hindered the assessment of a common epigenetic and genetic signature predictive of CIMP across all cancer sites [22]. In the present study, we analyze a large dataset of over

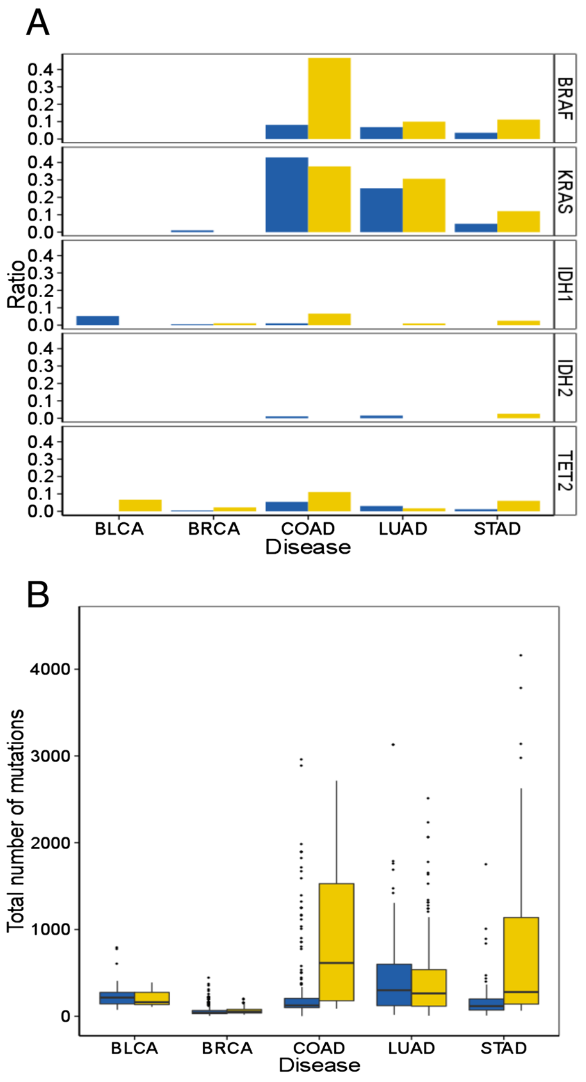

Fig. 4 Mutation analysis. **a** Association between specific mutations (*IDH1*, *IDH2*, *BRAF*, and *KRAS*) with the CIMP phenotype for all tissues (*yellow* = CIMP positive, *blue* = CIMP negative). **b** Significantly higher mutation rate for CIMP-positive (*yellow*) compared to CIMP-negative (*blue*) tumors is observed for colon and gastric cancers only and is concordant with CIMP association with microsatellite instability for these tissues

2000 tumor methylation profiles measured with a single technology from 5 different tissue types. We observe a universal epigenetic signature that defines CIMP independently from the tissue of origin, which might suggest a common molecular basis to CIMP across tissues. Genes associated with these CGIs are enriched in several biological pathways linked to organ development and include several interesting genes such as *CDKN2A* coding for p16, a well-characterized tumor suppressor protein [44], which is aberrantly hypermethylated in CIMP-positive tumors and might contribute to tumor development.

Table 6 Intersection of the genetic signatures for "Combined-Lasso" and "Group-Lasso" predictive of CIMP ranked by decreasing level of robustness

Over-expressed	*ZIC2*, *AMH*, *LHX1*, *ZIC3*, *XKR9*,*TNNT1*, *CAMK2N2*,*PCDHB9*, *RAET1K*, *HIST1H2AB*, *C2CD4C*, *FBXL20*, *TFCP2L1*
Under-expressed	*MAGEC2*, *ZNF300*,*SLC15A1*,*TSPYL5*, *MLF1*, *GATA2*, *MAGEA12*, *LOC441666*, *MAGEA2*, *LOC389493*, *H2AFY2*, *LDHC*

Other genes present in the cross-cancer CIMP signature such as *HOTAIR*, which is known to reprogram the chromatin state and is associated with breast cancer metastasis [45], might on the contrary be repressed in CIMP tumors and be linked with a better prognosis for breast cancer patients. *GREM1* is another gene present in the CIMP signature and is associated with tumor cell proliferation [46]. Less documented genes present in the CIMP signature could potentially be investigated for a biological validation of their role in tumor development.

Recent studies have pointed out that epigenetic aberrations could be derived from genetic aberrations [47]. By combining the different datasets into a single prediction task, we are able to identify a common set of genes whose expression levels can predict the CIMP status for each tissue. This gene list is enriched mostly in genetic regulatory pathways, suggesting that the epigenetic reprogramming and thus CIMP might be an intermediate step in the regulatory mechanism. Among the genes contained in the signature, *ZIC2*, which is robustly selected in each bootstrap of the CIMP prediction task and is significantly more expressed in CIMP-positive tumors for each tissue, has been known to act as a Wnt/β-catenin signalling inhibitor [48] which is

Table 7 Clinical impact of CIMP. Overall survival proportion given the CIMP phenotype and the *p* value associated with the survival analysis (logrank test)

Tissue	Event		*p* value
	CIMP−	CIMP+	
BLCA	47/214	21/96	0.74
BRCA	29/495	9/114	0.20
COAD	28/218	6/54	0.57
LUAD	24/127	67/295	0.49
STAD	26/141	20/193	0.29

usually upregulated in several cancers. Another interesting characteristic of this genetic predictive signature from a clinical point of view is the recurrence of cancer/testis antigens (CTAs) such as *MAGEC2* [49–51], *MAGEA12* [52, 53], *MAGEA2* [54], and *LDHC* [55], which are interesting targets for cancer immunotherapy [56] and are consistently under-expressed in CIMP-positive tumors. Recently, Gevaert et al. [57] also showed a strong association between *MAGEA4* hypomethylation and CIMP-positive tumors which further supports the link between CTAs and the absence of a methylator phenotype.

Mutation analyses are not very conclusive in defining a set of specific somatic mutations significantly associated with CIMP. In particular, lowly mutated cancer sites such as the bladder, breast, or even lung do not show any mutations significantly associated with CIMP. For highly mutated cancer sites such as colon or stomach, our results confirm a strong association between *BRAF* mutation and COAD-CIMP [9] but do not show any particular associations with *IDH1/2*, which have been reported to be causal in gliomas and leukemia [19, 26]. There is a strong association between COAD and STAD-CIMP and the specific mutations of genes related with extracellular matrix and cell adhesion, both reported to be strongly associated with metastasis [58–61]. Interestingly, neuronal developmental processes are highly enriched but affecting different genes from the universal epigenetic signature. Associations with neuronal development were already mentioned in [17].

Studies have often reported a clear distinct clinical prognosis associated with CIMP [9, 14, 17, 32]. This reiterates that a main reason for defining CIMP in each tissue site is its potential use as a prognosis marker. However, CIMP could be associated with a good or bad prognosis depending on the type of tumors. In the current study, we do not observe a significant association with any good nor bad prognosis linked with CIMP.

Conclusion

This meta-analysis of more than 2000 samples sheds new light on CIMP across cancers, its link with gene expression, and its clinical relevance. We found strong evidence that a panel of genes, which we call the pan-cancer CIMP signature, is involved simultaneously in the establishment of the CIMP in various cancer sites, which might be an indicator of a universal biological process behind CIMP. We found that differences in the CIMP status of a sample is associated to differences in the transcriptome, and also found a core set of genes whose expression levels differentiates CIMP-positive and CIMP-negative samples, in all cancers studied. Finally, we found little evidence of association between CIMP and mutations, except for the well-known BRAF mutation in colon cancer and also little association with patient survival.

Materials and methods

Patient selection

All data were retrieved from the TCGA data portal. We selected samples from bladder, breast, colon, lung and gastric adenocarcinomas because large matched datasets were available for methylation, gene expression, and mutation profiles. Moreover, all these tissues were previously reported to exhibit a methylator phenotype. The datasets are detailed in Table 1 and the different institutions that released the data are mentioned in the "Acknowledgements" section.

Methylation profiling

Methylation profiles were retrieved from level 2 TCGA data. They were obtained with the Illumina HumanMethylation450K DNA Analysis BeadChip assay, which is based on genotyping of bisulfite-converted genomic DNA at individual CpG sites to provide a quantitative measure of DNA methylation [62]. Following hybridization, the methylation value for a specific probe was calculated as the ratio $M/(M+U)$ where M is the methylated signal intensity and U is the unmethylated signal intensity. Across the genome, 485,577 CpG methylation levels, associated with 27,176 CGIs and 21,231 genes, were measured as such.

Following [63], we considered not only the CGI methylation profile but also included in the analysis proximal regions in the near vicinity (up to 4 kb), namely the CGI Shores and Shelves regions in a general CGI+SS methylation profile.

Gene expression profiling

Gene expression profiles were retrieved from level 3 TCGA data. They were obtained from the Illumina HiSeq RNASeq technology and processed following [64]. We used the reads per kilobase per million mapped reads (RPKM) to quantify the gene expression level from RNA sequencing data.

Mutation profiling

Mutation profiles were retrieved from somatic mutation profiles from level 2 TCGA data obtained through whole exome sequencing. To compare the rate of mutation given the CIMP status, we performed a hypergeometric test and corrected for multiple testing using Benjamini-Hochberg correction.

CIMP analysis

To assess the existence of CIMP, we performed Ward hierarchical clustering using euclidean distance on the top 5 % most variant CGIs. Variations from 1 to 10 % of the most variant CGIs had a small impact on the clustering stability (Additional file 3). We then cut the hierarchical clustering tree in two classes namely CIMP-positive and CIMP-negative tumors given their average level of methylation (CIMP-positive = high level of methylation, CIMP-negative = low level of methylation). Robustness of the clustering was obtained through consensus clustering [25].

Predicting CIMP status from gene expression profiles

We performed logistic regression using a lasso penalty [65] to predict CIMP status from gene expression profiles for each tissue separately. Accuracy is calculated through threefold cross-validation averaged over 100 repeats. To combine the different datasets into a single prediction task, we performed group-lasso logistic regression (Additional file 1). Given the imbalanced proportion of CIMP in each datasets, we defined the "random" predictor as a predictor that always predicts the majority class. The statistical significance of a gene expression-based predictor over the "random" predictor was calculated using a Student *t* test.

To determine the genetic predictive signature, genes were ranked according to the frequency at which they appeared in the optimal lasso estimator signature averaged over the different folds and repeats. Genes with a frequency of at least 50 % were selected.

Survival analysis

Overall survival was estimated using the Kaplan-Meier method [66] to compare the survival between CIMP-positive and CIMP-negative tumors. A multivariate Cox proportional hazards regression model [67] was also fitted to assess the CIMP odd ratio.

Endnotes

[1] Bladder tissue
[2] Breast tissue
[3] Colon tissue
[4] Lung tissue
[5] Stomach tissue

Additional files

Additional file 1: Hierarchical clustering and CIMP status of samples in each tissue. Each sample is represented by the methylation levels of the 5 % of the probes that vary most in the tissue considered. Heatmaps range from hypomethylated (*blue*) to hypermethylated (*yellow*). The *column colorbar* represents the resulting assignment of each sample as CIMP positive (*yellow*) or CIMP negative (*blue*). Panel A. bladder; panel B. breast; panel C. colon; panel D. lung; panel E. stomach. (PDF 13722 kb)

Additional file 2: Stability of CIMP clusters with respect to the size of the CIMP signature. Robustness of cluster assignment for each sample (*columns*) as a function of the proportion of variant CGIs kept to define the CIMP signature, from 1 to 10 % (*rows*) and given the number of CIMP clusters considered (*left panels*: K = 2, *right panels*: K = 3, *yellow* = CIMP-positive, *blue* = CIMP-negative, *black* = CIMP-low) for bladder (panel A/B), breast (panel C/D), colon (panel E/F), lung (panel G/H), stomach (panel I/J). Panel K. Table summarizing the stability of the cluster assignments for each tissue and different number of CIMP clusters considered. (PDF 381 kb)

Additional file 3: Supplementary methods. (PDF 153 kb)

Additional file 4: Gene expression profiling on the common genetic predictive signature for each tissue. The *column color bar* represents the CIMP status (*yellow* = CIMP-positive, *blue* = CIMP-negative) while the *row color bar* represents the clustering of genes (*green* = under-expressed in CIMP, red = over-expressed in CIMP). Panel A. bladder; panel B. breast; panel C. colon; panel D. lung; panel E. stomach. (PDF 561 kb)

Additional file 5: Study of a genetic signature associated with CIMP. Panel A. Venn diagram representing the intersection of the mutations significantly associated with CIMP in colon and gastric cancers. Panel B. Gene ontology analysis of the common genes associated with CIMP. (PDF 82 kb)

Additional file 6: Clinical impact of CIMP on the patient survival. The plots show the Kaplan Meier survival curves based on CIMP status for different tissues. Panel A. bladder; panel B. breast; panel C. colon; panel D. lung; panel E. stomach. (PDF 89 kb)

Additional file 7: Association between CIMP and clinical annotations. Panel A. Association between CIMP and age: distribution of patients' age given their CIMP phenotype in each tissue. Panel B. Association between CIMP and MSI in colon: ratio of MSI-positive and MSI-negative patient given the CIMP phenotype in the colon. Panel C. Association between CIMP and ER status in breast: ratio of ER-positive patients given the CIMP phenotype in the breast. Panel D. Association between CIMP and PR status in the breast: ratio of PR-positive patients given the CIMP phenotype in the breast. Panel E. Association between CIMP and HER2 status in the breast: ratio of HER2-positive patients given the CIMP phenotype in the breast. (PDF 86 kb)

Abbreviations

CIMP: CpG island methylator phenotype; CGI: CpG island; BLCA: Bladder carcinoma; BRCA: Breast carcinoma; COAD: Colon adenocarcinoma; LUAD: Lung adenocarcinoma; STAD: Stomach adenocarcinoma.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MM participated in the statistical analyses and the writing of the manuscript. FR and JPV conceived the study and participated in its design and coordination. All authors read and approved the final manuscript.

Acknowledgements

This study was financially supported by "La Ligue Nationale Contre le Cancer" (to MM) and the European Research Council (SMAC-ERC-280032 to JPV and MM).

The authors would like to acknowledge the following: the Cancer Genome Atlas, IUPUI[1,2,4,5], Lahey Clinic[1,4], Research Metrics Pakistan[2], Asterand[1,2,3,4,5], Baylor[1,3,4], Cleveland Clinic[1], UT Southwestern Medical Center at Dallas[1], University of Chicago[1], University of Miami[1], Barretos Cancer Hospital[1,5], Penrose Colorado[1,4], Candler[1,3,4,5], Christiana Healthcare[1,2,3,4,5], Cornell Medical College[1], Cureline[2,3,4,5], Duke[2], Erasmus MC[1], Fox Chase[4,5], Global Bioclinical-Moldova[5], Greater Poland Cancer Center[2,3,5], Gundersen Lutheran Health System[1,4], Hartford Hospital[1], Harvard[3], ILSBio[1,2,3,4,5], Indivumed[1,2,3,4,5], International Genomics Consortium[1,2,3,4,5], Johns Hopkins[4], MD Anderson Cancer Center[1,2,5], Memorial Sloan Kettering Cancer Center[2,3,5], Mayo Clinic[1,2], Medical College of Georgia[1], National Cancer Center Korea[5], Ontario Institute for Cancer Research[1,2,3,4,5], Peter MacCallum Cancer Center[5], Prince Charles Hospital[4], Roswell Park[1,2,3,4], Spectrum Health[1], St Joseph Medical Center[3,4], Tayside Tissue Bank[5], Thoraxklinik at University Hospital Heidelberg[4], UCSF[2], UNC[1,2,3,4,5], University Health Network[5], University of Michigan[3], University of Colorado Denver[1], University of Miami[2,4], University of Minnesota[1], University of Oklahoma[1], University of Pittsburgh[1,2,3,4,5], University of Puerto Rico[1], University of Sheffield[1], University of Southern California[1], Walter Reed[2], Washington University[4] for the distribution of patients data (as specified below).

Author details
[1]CBIO-Centre for Computational Biology, Mines Paristech, PSL-Research University, 35 Rue Saint-Honore, F-77300 Fontainebleau, France. [2]Department of Bioinformatics, Biostatistics and System Biology, Institut Curie, 11-13 Rue Pierre et Marie Curie, F-75248 Paris, France. [3]U900, INSERM, 11-13 Rue Pierre et Marie Curie, F-75248 Paris, France. [4]UMR932, Immunity and Cancer Team, Institut Curie, 26 Rue d'Ulm, 75006 Paris, France. [5]Department of Translational Research, Residual Tumor and Response to Treatment Team, Institut Curie, 26 Rue d'Ulm, 75006 Paris, France. [6]Department of Surgery, Institut Curie, 26 Rue d'Ulm, 75006 Paris, France.

Received: 20 July 2015 Accepted: 1 October 2015
Published online: 13 October 2015

References

1. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007;128(4):683–92.
2. Esteller M. Epigenetics in cancer. New Eng J Med. 2008;358(11):1148–59.
3. Rodriguez-Paredes M, Esteller M. Cancer epigenetics reaches mainstream oncology. Nat Med. 2011;17(3):330–339.
4. Jones P. DNA methylation and cancer. Cancer Res. 1986;46(2):461–6.
5. Baylin SB, Herman JG. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. Trends Genet. 2000;16(4):168–74.
6. Esteller M, Corn PG, Baylin SB, Herman JG. A gene hypermethylation profile of human cancer. Cancer Res. 2001;61(8):3225–9.
7. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa J-PJ. CpG island methylator phenotype in colorectal cancer. Proc Nat Acad Sci. 1999;96(July):8681–6.
8. Issa J-PJ, Shen L, Toyota M. CIMP, at last. Gastroenterology. 2005;129(3): 1121–4.
9. Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse Ma, et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat Genet. 2006;38(7):787–93.
10. Estécio MRH, Yan PS, Ibrahim AEK, Tellez CS, Shen L, Huang TH-M, et al. High-throughput methylation profiling by MCA coupled to CpG island microarray. Genome Res. 2007;17(10):1529–36.
11. Curtin K, Slattery ML, Samowitz WS. CpG island methylation in colorectal cancer: past, present and future. Pathol Res Int. 2011;2011:902674.
12. Hinoue T, Weinsenberger D, Lange C, Shen H, Byun H, Van Den Berg D, et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res. 2012;22(2):271–82.
13. Van der Auwera I, Yu W, Suo L, Van Neste L, van Dam P, Van Marck EA, et al. Array-based DNA methylation profiling for breast cancer subtype discrimination. PloS One. 2010;5(9):e12616. doi:10.1371/journal.pone.0012616.
14. Fang F, Turcan S, Rimner A, Kaufman A, Giri D, Morris LGT, et al. Breast cancer methylomes establish an epigenomic foundation for metastasis. Sci Trans Med. 2011;3(75):75–25.
15. Suzuki M, Shigematsu H, Lizasa T, Hiroshima K, Nakatani Y, Minna J, et al, Cancer. Exclusive mutation in epidermal growth factor receptor gene, HER-2, and KRAS, and synchronous methylation of nonsmall cell lung cancer. 2006;106(10):2200–7.
16. Chen HY, Zhu BH, Zhang CH, Yang DJ, Peng JJ, Chen JH, et al. High CpG island methylator phenotype is associated with lymph node metastasis and prognosis in gastric cancer. Cancer Sci. 2012;103(1):73–9.
17. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010;17(5):510–22.
18. Baysan M, Bozdag S, Cam MC, Kotliarova S, Ahn S, Walling J, et al. G-CIMP status prediction of glioblastoma samples using mRNA expression data. PloS One. 2012;7(11):47839.
19. Yilmaz E, Campos C, Fabius AWM, Lu C, Ward PS, Viale A, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature. 2012;483(7390):479–83.
20. Bae YK, Brown A, Garrett E, Bornman D, Fackler MJ, Sukumar S, et al. Hypermethylation in histologically distinct classes of breast cancer. Clinical Cancer Res. 2004;10(18):5998–6005.
21. Anacleto C, Leopoldino A, Rossi B, Soares FA, Lopes A, Rocha JC, et al. Colorectal cancer "methylator phenotype": fact or artifact? Neoplasia. 2005;7(4):331–5.
22. Hughes LAE, Melotte V, de Schrijver J, de Maat M, Smit VTHBM, Bovee JVMG, et al. The CpG island methylator phenotype: what's in a name? Cancer research. 2013;73(19):5858–68.
23. Ogino S, Kawasaki T, Kirkner GJ, Loda M, Fuchs CS. CpG island methylator phenotype-low (CIMP-low) in colorectal cancer: possible associations with male sex and KRAS mutations. J Mol Diagn. 2006;8(5):582–8.
24. Shen L, Toyota M, Kondo Y, Lin E, Zhang L, Guo Y, et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. Proc Natl Acad Sci USA. 2007;104(47):18654–9.
25. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering : A Resampling-Based Method for Class Discovery and Visualization of Gene. Machine Learning. 2003;52(1):91–118.
26. Figueroa M, Abdel-Wahab O, Lu C, Ward P, Patel J, Shih A, et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell. 2010;18(6):553–67.
27. Amary M, Damato S, Halai D, Eskandarpour M, Berisha F, Bonar F. Ollier disease and Maffucci syndrome are caused by somatic mosaic mutations of IDH1 and IDH2. Nat Genet. 2011;43(12):1262–5.
28. Pansuriya T, van Eijk R, d'Adamo P, van Ruler M, Kuijjer M, Oosting J, et al. Somatic mosaic IDH1 and IDH2 mutations are associated with enchondroma and spindle cell hemangioma in Ollier disease and Maffucci syndrome. Nat Genet. 2011;43(12):1256–61.
29. Herman J, Umar A, Polyak K, Graff J, Ahuja N, Issa J, et al. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. Proc Natl Acad Sci U S A. 1998;95(12):6870–5.
30. Jones S, Li M, Parsons D, Zhang X, Wesseling J, Kristel P, et al. Somatic mutations in the chromatin remodeling gene ARID1A occur in several tumor types. Hum Mutat. 2012;33(1):100–3.
31. Zang Z, Cutcutache I, Poon S, Zhang S, McPherson J, Tao J, et al. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nat Genet. 2012;44(5):570–4.
32. Toyota M, Kopecky K, Toyota M, Jair K, Willman C, Issa J. Methylation profiling in acute myeloid leukemia. Blood. 2001;97(9):2823–9.
33. Garcia-Manero G, Daniel J, Smith T, Kornblau S, Lee M, Kantarjian H, et al. DNA methylation of multiple promoter-associated CpG islands in adult acute lymphocytic leukemia. Clinical Cancer Res. 2002;8(7):2217–24.
34. Roman-Gomez J, Jimenez-Velasco A, Agirre X, Prosper F, Heiniger A, Torres A. Lack of CpG island methylator phenotype defines a clinical subtype of T-cell acute lymphoblastic leukemia associated with good prognosis. J Clin Oncol. 2005;23(28):7043–9.
35. Roman-Gomez J, Jimenez-Velasco A, Agirre X, Castillejo J, Navarro G, Calasanz M, et al. CpG island methylator phenotype redefines the prognostic effect of t(12;21) in childhood acute lymphoblastic leukemia. Clinical Cancer Res. 2006;12(16):4845–50.
36. Maruyama R, Toyooka S, Toyooka K, Harada K, Virmani A, Zochbauer-Muller S, et al. Aberrant promoter methylation profile of bladder cancer and its relationship to clinicopathological features. Cancer Res. 2001;61(24):8659–63.
37. Liu Z, Zhao J, Chen X, Li W, Liu R, Lei Z, et al. CpG island methylator phenotype involving tumor suppressor genes located o chromosome 3p in non-small cell lung cancer. Lung Cancer. 2008;62(1):15–22.
38. Maruyama R, Toyooka S, Toyooka K, Virmani A, Zochbauer-Muller S, Farinas A, et al. Aberrant promoter methylation profile of prostate cancers and its relationship to clinicopathological features. Clinical Cancer Res. 2002;8(2):514–9.
39. Toyota M, Ahuja N, Suzuki H, Itoh F, Ohe-Toyota M, Imai K, et al. Aberrant methylation in gastric cancer associated with the CpG island methylator phenotype. Cancer Res. 1999;59:5438–42.
40. Oue N, Oshimo Y, Nakayama H, Ito R, Yoshida K, Matsusaki K, et al. DNA methylation of multiple genes in gastric carcinoma: association with histological type and CpG island methylator phenotype. Cancer Sci. 2003;94(10):901–5.
41. Kim H, Kim Y, Kim S, Kim N, Noh S. Concerted promoter hypermethylation of hMLH1, p16INK4A, and E-cadherin in gastric carcinomas with microsatellite instability. J Pathol. 2003;200(1):23–31.
42. Etoh T, Kanai Y, Ushijima S, Nakagawa T, Nakanishi Y, Sasako M, et al. Increased DNA methyltransferase 1 (DNMT1) protein expression correlates significantly with poorer tumor differentiation and frequent DNA hypermethylation of multiple CpG islands in gastric cancers. Am J Pathol. 2004;164(2):689–99.
43. Kusano M, Toyota M, Suzuki H, Akino K, Aoki F, Fujita M, et al. Genetic, epigenetic, and clinicopathologic features of gastric carcinomas with the

CpG island methylator phenotype and an association with Epstein-Barr virus. Cancer. 2006;106(7):1467–79.

44. Nobori T, Miura K, Wu DJ, Lois A, Takabayashi K, Carson DA. Deletions of the cyclin-dependent kinase-4 inhibitor gene in multiple human cancers. Nature. 1994;368(April):753–6.
45. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, et al. Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464(7291):1071–6.
46. Sneddon J, Zhen H, Montgomery K, van de Rijn M, Tward A, West R, et al. Bone morphogenetic protein antagonist gremlin 1 is widely expressed by cancer-associated stromal cells and can promote tumor cell proliferation. Proc Natl Acad Sci USA. 2006;103(40):14842–7.
47. Reddington JP, Sproul D, Meehan RR. DNA methylation reprogramming in cancer: does it act by re-configuring the binding landscape of Polycomb repressive complexes? BioEssays: News Rev Mol Cell Dev Biol. 2014;36(2):134–40.
48. Pourebrahim R, Houtmeyers R, Ghogomu S, Janssens S, Thelie A, Tran H, et al. Transcription factor Zic2 inhibits Wnt/beta-catenin protein signaling. J Biol Chem. 2011;286(43):37732–40.
49. von Boehmer L, Keller L, Mortezavi A, Provenzano M, Sais G, Hermanns T, et al. MAGE-C2/CT10 protein expression is an independent predictor of recurrence in prostate cancer. PLoS ONE. 2011;6(7):1–7.
50. Yang F, Zhou X, Miao X, Zhang T, Hang X, Tie R, et al. MAGEC2, an epithelial-mesenchymal transition inducer, is associated with breast cancer metastasis. Breast Cancer Res Treatment. 2014;145(1):23–32.
51. Reinhard H, Yousef S, Luetkens T, Fehse B, Berdien B, Kröger N, et al. Cancer-testis antigen MAGE-C2/CT10 induces spontaneous CD4+ and CD8+ T-cell responses in multiple myeloma patients. Blood Cancer J. 2014;4:e212. doi:10.1038/bcj.2014.31.
52. Heidecker L, Brasseur F, Probst-Kepper M, Guéguen M, Boon T, Van den Eynde BJ. Cytolytic T lymphocytes raised against a human bladder carcinoma recognize an antigen encoded by gene MAGE-A12. J Immunol (Baltimore, Md. : 1950). 2000;164(11):6041–5.
53. Mollaoglu N, Vairaktaris E, Nkenke E, Neukam FW, Ries J. Expression of MAGE-A12 in oral squamous cell carcinoma. Disease Markers. 2008;24(1): 27–32.
54. Peche LY, Scolz M, Ladelfa MF, Monte M, Schneider C. MageA2 restrains cellular senescence by targeting the function of PMLIV/p53 axis at the PML-NBs. Cell Death Differentiation. 2012;19(6):926–36.
55. Tang H, Goldberg E. Homo sapiens lactate dehydrogenase c (Ldhc) gene expression in cancer cells is regulated by transcription factor Sp1, CREB, and CpG island methylation. J Androl. 2009;30(2):157–67.
56. Scanlan MJ, Gure AO, Jungbluth AA, Old LJ, Chen YT. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. Immunol Rev. 2002;188(1):22–32.
57. Gevaert O, Tibshirani R, Plevritis SK. Pancancer analysis of DNA methylation-driven genees using MethylMix. Genome Biol. 2015;16(1):17.
58. Gilkes DM, Semenza GL, Wirtz D. Hypoxia and the extracellular matrix: drivers of tumour metastasis. Nat Rev Cancer. 2014;141(6):430–439.
59. Lu P, Weaver VM, Werb Z. The extracellular matrix: a dynamic niche in cancer progression. J Cell Biol. 2012;196(4):395–406.
60. Bendas G, Borsig L. Cancer cell adhesion and metastasis: selectins, integrins, and the inhibitory potential of heparins. Int J Cell Biol. 2012;ID 676731:10. doi:10.1155/2012/676731.
61. Okegawa T, Pong R, Hsieh J. The role of cell adhesion molecule in cancer progression and its application in cancer therapy. Acta Biochim Pol. 2004;51(2):445–57.
62. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288–95.
63. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. Genome-wide methylation analysis of human colon cancer reveals similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41(2):178–86.
64. Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):1–8.
65. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc. 1996;58(1):267–88.
66. Kaplan EL, Meier D. Nonparametric estimation from incomplete observation. J Am Statist. 1958;58:457–81.
67. Cox DR, Oakes D. Analysis of Survival Data. London: Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis; 1984.