Derek Tam
April 10, 2017

Predicting the CpG island methylator phenotype using the transcriptome

Abstract

The CpG island methylator phenotype is described as a global increase in methylation of promoter-associated CGIs, and has been consistently observed in different types of cancer. CpG methylation is classically correlated with decreased gene expression and promoter silencing. In this study, an attempt is made to use this correlation in the prediction of this unique CIMP+ methylome profile using only gene expression data. The correlation between gene expression and CGI methylation is found to not be strictly negative. A subset of protein coding genes (identified to be ideal markers of CIMP status due to their measurable change in expression) are used to design a predictor. In the CIMP classification of 444 colorectal cancers, the predictor achieved an overall accuracy of 66.3%. Further optimization may be possible through the use of a more succinct list of indicators, in combination with adjustments to the thresholds used for each.

1. Introduction

1.1 Cancer and its progression

Cancer is a disease of the genome, identified by unregulated cell division and non-specific differentiation. Normal cells are programmed to serve a defined role through the expression of genes specific to its tissue type, and undergo apoptosis at the end of their functional life cycle. Cancerous cells, however, acquire independence from apoptotic signals and divide unchecked through evasion of cell cycle blockades (1). In most cancers, this rapid cell division results in the formation of a primary solid-tumor. Mutations continue to accumulate as cells divide, resulting in heterogeneity within individual tumors (2). As cancer progresses, some cells may become invasive and lose adhesion to adjacent cells, metastasizing to other sites in the body and forming secondary tumors.

Unrestricted division and failure to differentiate is often the result of dis-regulation in a combination of genes associated with cell cycle maintenance. While a small number of these changes are attributed to mutations in the primary genetic sequence, most cancer phenotypes are the result of changes in the overall expression of key genes (3). How a specific gene should be expressed is not directly encoded in the primary sequence, but rather is the cumulative effect of various marks on the DNA itself. These marks are left, maintained, and written as the action of various epigenetic writers, readers, and erasers, which together can modulate the expression of large swathes of proteins. Through this network of regulation, it is possible to understand how changes in the expression of a small number of genes critical to the maintenance of many others could rapidly disseminate changes to expression. The combined effect is dis-regulation of these key oncogenes that we now know to be responsible for cancer development.

1.2 Epigenetics and DNA methylation

Epigenetic modifications can generally be separated into activating and repressive marks, covalently modifying the DNA molecule directly or its associated histone proteins. The addition of various molecular groups, such as methyl or acetyl moieties, to specific amino acids on histone subunits can directly or indirectly affect the packaging of chromatin (4). This in turn changes the accessibility of the bound DNA to various transcription factors and polymerases, which regulates how often that specific gene becomes transcribed and therefore translated. For example, trimethylation at Lys27 of histone subunit 3 (H3K27me3) is a well-studied deactivating mark, and its presence is strongly associated with the silencing of its bound promoter. Left by methyltransferase EZH2, a subunit of the polycomb repressive complex 2 (PRC2), this epigenetic mark is propagated by PRC2 recruitment to other trimethylated lysine residues. Conversely, the addition of an acetyl group to this same position (H3K27ac)

is activating, neutralizing the positive charge of the lysine that results in reduced adhesion of DNA to the histone (5). In combination with trimethylation at Lys4 (H3K4me3), chromatin remodeling factors are recruited that ultimately result in the formation of nucleosome depleted regions and activation of the underlying promoter. Through the action of various epigenetic writers and erasers, some histone marks are dynamically changed throughout the lifetime of even a single cell, delicately balancing the expression of proteins important to that cells function (6). Histone modifications can be considered in stark contrast to cytosine base methylation, a heritable epigenetic mark that has more permanent effects beginning in embryonic development (7).

DNA methylation is the covalent addition of a methyl group to the 5th carbon of a cytosine base (5mC). These marks are left at CpG dinucleotides by de novo methyltransferases DNMT3A and DNMT3B early in development, then later maintained and propagated throughout cell division by DNMT1 (8). As methylated cytosines are readily converted to thymidine by deamination, 5mCs are inherently mutagenic, resulting in their gradual genomic depletion which has been tracked throughout vertebrate evolution (9). Groups of preserved CpGs are generally found in dense clusters termed CpG islands (CGIs), and are often associated with gene promoters (10). However, individual CpGs can also be found scattered throughout the genome, both in intronic and exonic regions (11). Curiously, these individual CpGs are consistently methylated whilst CGIs acquire methylation throughout embryonic development and as cells differentiate. This implies that methylation is a default state, and that de-methylation is an active function directed to specific regions of the DNA. It is therefore likely that molecular machinery exists that is responsible for the maintenance of these regions, and their absence or malfunction will result in the "creeping" of 5mC marks into CGIs by natural DNMT1 propagation.

Clasically, methylation of DNA has been correlated with silencing of the affected genic region (12). An early example of this observation was in the study of X-chromosome inactivation, where differential methylation was experimentally observed to be an important difference between the active and inactive X-chromosomes (13). It was later found that these differences were primarily focused to CGIs, giving rise to the idea of CGI methylation being a sign of gene inactivation (14). With the advent of modern technologies such as whole-genome bisulfite sequencing (WGBS), the role of the methylome as a whole can be better studied. It is now clear that the methylation state of a CpG is not significant in isolation, but requires spatial context to be fully understood (15). Whilst methylated CGIs have been found proximal to suppressed promoters, isolated CpGs in gene bodies may potentially promote transcript elongation and affect how the transcript is ultimately spliced. Methylation may also play a protective role in repeat regions such as centromeres and telomeres, preventing shortening of chromosome ends and ensuring proper separation in mitosis. Furthermore, the palindromic structure of CpGs mean that methylation information is retained in one of the two daughter strands during DNA replication (16). As a result, this information can be recovered by DNMT1 post-division, supporting the idea of determined cell fates and strictly differentiated tissue types. These differences in the role of CpG methylation based on their density is an active area of research, and may shed light on the effects of their changes on diseases such as cancer.

1.3 Epigenetics in understanding cancer pathology

In 1999, Toyota et al. described two distinct methylation phenotypes that were found in the primary colorectal cancer tissues they were studying (17). Assessing methylation at a genome-wide scale, they observed that a subset of the tumors had acquired increased methylation independently of age. Identifying this pattern at a number of genic regions, in particular those encoding p16, THBS1, and MLH1, they hypothesized that a "CpG island methylator phenotype" (CIMP) could be an alternative driver of neoplasia. Due to its vague definition, the idea remained controversial until its study by the Weisenberger group (18). Using unsupervised hierarchical clustering, they found that CIMP-positive (CIMP+) status was almost completely correlated with mutations in the DNA mismatch-repair protein, BRAF, resulting in increased microsatellite instability (MSI). Additionally, they found that a different group with intermediate methylation was associated with KRAS mutations, separate from the CIMP+ cluster. While their findings confirmed the possibility for a phenotypic classification by methylation state, it also suggested that the separation may not be binary, but rather a continuum of increasing methylation that correlates naturally with the progression of cancer. As a result, some research groups have turned to classifying CIMP status as high, intermediate, or zero, though the existence of a distinct CIMP-intermediate group remains a topic of debate.

CIMP classification has since been extended to many other cancer types, and has been reported in (but is not limited to) breast, lung, and pancreatic cancers, as well as in gliomas and leukemias (19). Due to its vague definition, however, different gene panels have been used in determining CIMP status with inconsistent thresholds, making their study difficult to compare between research groups. Moarii et al. (20) proposed a general set of 89 hypermethylated CGIs, associated with 51 genes, that they found to be simultaneously involved in the development of CIMP, though they failed to identify a common genetic signature. While this implies that the accumulation of methylation is not stochastic, it is likely the result of different disease states converging to present a similar, but nonuniform epigenetic landscape. However, CIMP status has been observed to have an effect on patient outcome and response to various treatments, and may potentially be clinically significant. As a

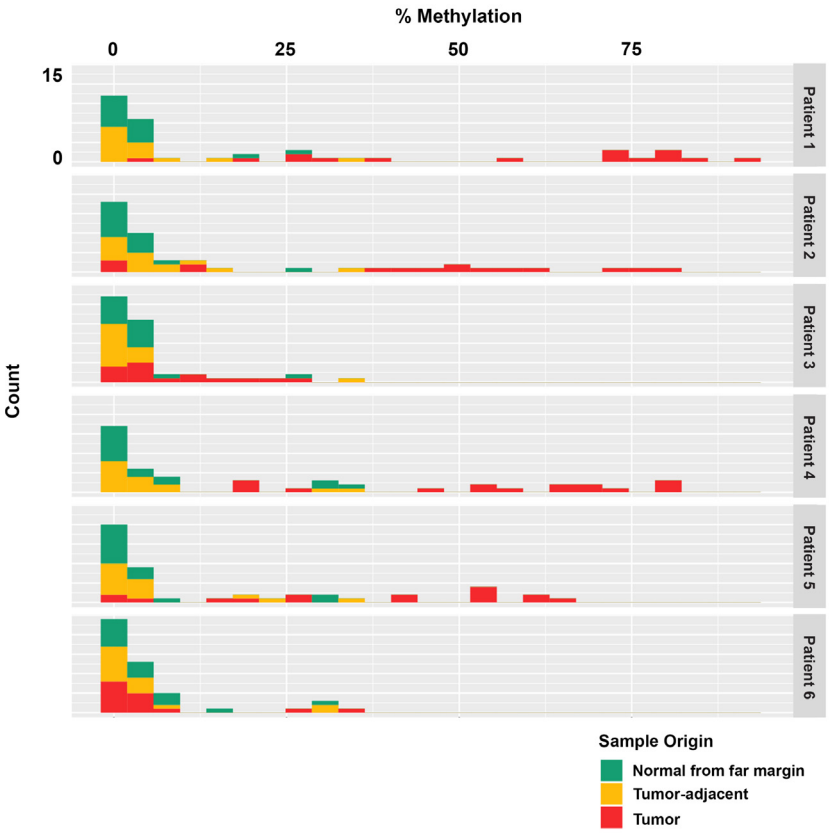

Figure 1 | **Distribution of methylation for the identification of CEMT patient CIMP status.** n = 11 for each sample; 33 for each patient. Percent methylation was determined by averaging methylation for all CpGs in a given CGI. 11 indicators are plotted for every sample (listed in Methods). The 4 true-normal reference samples are excluded on this plot. Increased methylation in tumor samples relative to normal samples are observed in patients 1, 2, 4, and 5; the colorectal cancers of these 4 patients are classified as CIMP+.

result, the ability to rapidly diagnose CIMP status in definite terms may become important in future cancer treatment.

Independent of the controversy, it is generally agreed that CIMP can be classified by a global increase in promoter-associated CGI methylation. Knowing that methylation at these sites has a repressive effect on gene expression, one would expect to see CIMP positivity reflected in the transcriptome. Based on this correlation, it should therefore be possible to predict CIMP status from gene expression data alone.

2. Results

2.1 Creating a predictive model

2.1.1 Analysis of CEMT datasets

To generate a predictor, differences between the epigenetic landscapes of CIMP+ and CIMP– cancers first had to be identified. In this project, colorectal cancer was chosen as a point of study as it was where CIMP was originally described, and multiple WGBS and gene expression datasets were available for both normal and tumor tissues. 22 colorectal samples were made available through the Centre for Epigenome Mapping Technologies (CEMT) project. These samples originated in 10 patients, of which 6 had colorectal cancer and 4 provided normal tissue references. For each of the 6 cancer patients, samples were obtained from the tumor tissue, tumor-adjacent tissue, and normal tissue from a far margin. All samples were analyzed for methylation by WGBS, and for gene expression by RNA sequencing.

2.1.2 Determination of CIMP status

CIMP status of the 18 CEMT colorectal cancer samples were first determined by assessing methylation at promoter-associated CGIs of specific indicator genes. Four genes used by the Weisenberger group became the pseudo-standard following their publication in 2006 (17). In 2007, a meta-analysis performed by Ogino et al (21) identified 8 markers that were both sensitive and specific for CIMP. The analysis performed in this thesis used a gene list (see Methods) compiled from these and other publications (22-24). Average methylation over the promoter-associated CGI of each indicator gene was determined for each of the 22 samples (Figure 1). Methylation of the normal and normal-adjacent samples were found on average to fall below 0.4 for all six patients. Above normal methylation was observed in four of the six tumor samples, however, whilst near-normal levels of methylation were observed in the remaining two. By the prescribed criteria, PT-1, -2, -4, and -5 were deemed CIMP positive.

2.1.3 Identifying differences in methylation and gene expression

In order to find trends in global methylation between the positive and negative phenotype, WGBS data was analyzed for the two groups. Average methylation was determined over the range of each promoter-associated CGI for all protein coding genes, and were visualized using a heatmap (Figure 2A).

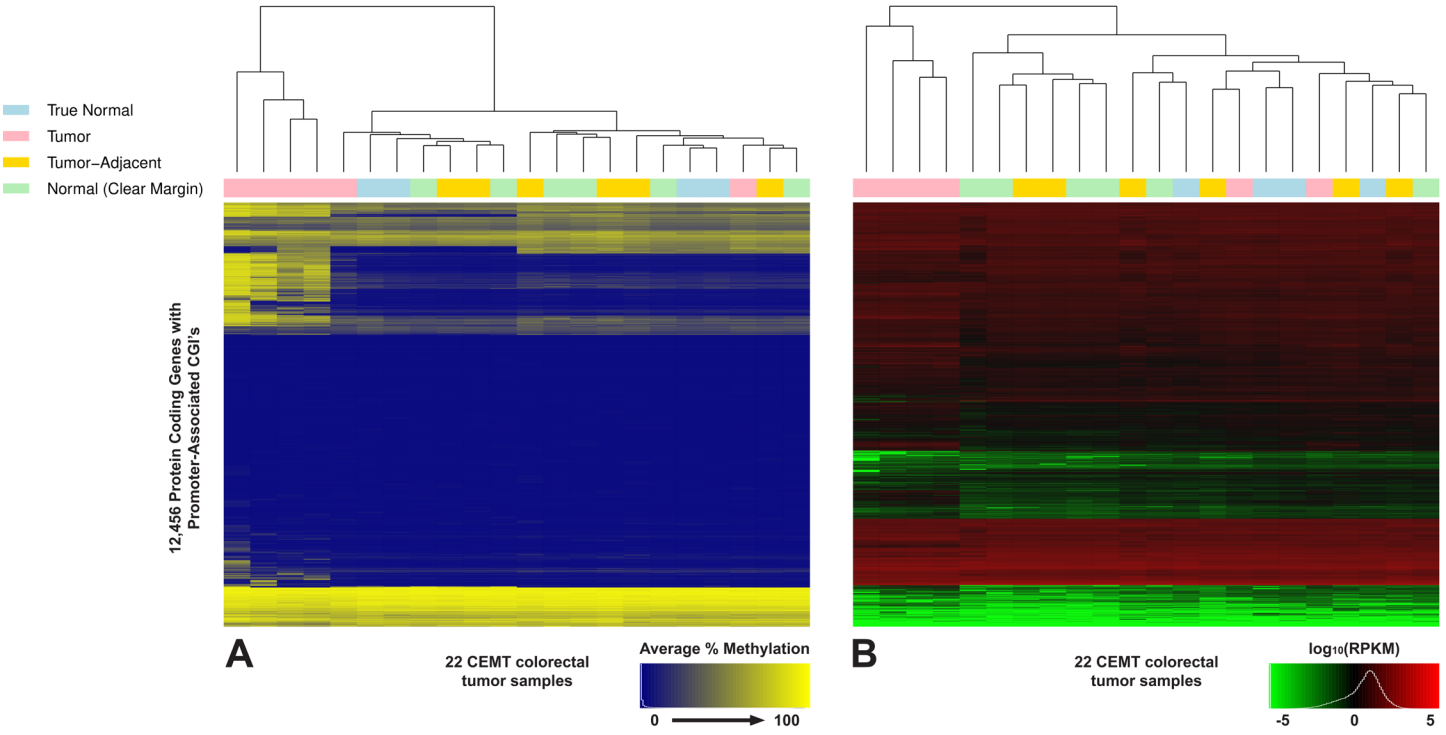

Figure 2A | **Heatmap of average CGI methylation for all protein-coding genes.** Methylation was measured by averaing all CpGs in a given CGI. Samples are color labeled by tissue origin. Distance was calculated using Spearman rank correlation. Distinct clustering is observed separating 4 of the 6 tumor samples, which was in agreement with CIMP status determination using the 11 indicator genes. A subset of genes are consistently methylated across all CIMP+ samples.

2B | **Heatmap of average gene expression for all protein-coding genes.** Gene expression is presented in RPKM, and is scaled using log base 10. Samples are color labeled by tissue origin. Distance was calculated using Spearman rank correlation. Although clustering correctly separates the 4 CIMP+ samples from the other 18, the difference in gene expression is subtle when compared to the difference observed in methylation.

Two distinct clusters were observed, with the same grouping as those identified by the indicator genes. It is interesting to note that CIMP– samples were clustered indistinguishably from the normal and normal-adjacent samples, rather than forming a separate clade. A group of genes were also identified to have increased methylation in the four CIMP+ samples. The heterogeneity in this group is consistent with previous findings of CIMP being characterized by a global increase in methylation, but lacking in a single reliable gene indicator.

As methylation at promoter-associated CpGs is classically correlated with expression of its associated gene, it is expected that this global increase in methylation will have an effect on the overall gene expression observed. RNA-seq data is presented in RPKMs and similarly visualized using a heatmap (Figure 2B). Clustering successfully identifies CIMP+ samples from CIMP– and normal samples, though differences are not as obvious as those observed in methylation. Furthermore, the cluster of genes identified in the methylation data are not represented in the expression data, alluding to a weaker correlation between methylation and expression than expected. Regardless, the successful separation of positive and negative samples suggests that it maybe possible to predict these states from gene expression alone.

2.1.4 Correlating methylation and gene expression

The exact predictive power of gene expression was quantified by determining its correlation with methylation. Separating the samples into positive and negative pools, the methylation and expression profiles of each gene were determined, and a correlation was calculated (Figure 3A). In the normal and CIMP– pools, gene expression and methylation had non-significant Spearman's correlation coefficients of ρ = -0.317 and -0.318, respectively. This same analysis in the CIMP+ pool, however, showed a significant correlation of ρ = -0.533. From the plot, this can be attributed to a cluster of genes that have acquired methylation at their promoter-associated CGIs in combination a decrease in overall gene expression. As this is the expected correlation between methylation and gene expression, its observation only in the CIMP+ group should be noted. From this analysis, the best indicators of CIMP status would be genes that have changed most significantly both in CGI methylation and gene expression, and should therefore be the subset used in the predictive model.

2.1.5 Isolating a predictor subset by euclidean distance

To identify this group of genes, their changes in methylation and gene expression profile were quantified. The euclidean distance was calculated between each gene's location in the CIMP– correlation and its corresponding location in the CIMP+ correlation (Figure 3B). By normalizing the location of each data point on the CIMP– plot to the origin, genes with the greatest distance from the center had the greatest combined difference in either metric between CIMP– and CIMP+ profiles. In general, changes in methylation were in the positive direction, while the few genes that decreased in methylation did not have significant changes in expression. It is interesting to note that increases in promoter-associated CGI methylation were not strictly correlated with decreases in expression, as suggested by the correlation. The subset of genes observed to increase significantly in both CGI methylation and expression conflict with the classic definition of methylation as a repressor of gene expression. From this analysis, a gene list sorted by absolute Euclidean distance was constructed, effectively ranking each gene by their utility as a predictor.

Extracting the average expression observed in the positive phenotype of each gene on this list, a rudimentary classifier of CIMP status was created. Changes in expression were separated into two groups: ones that had an increase in the positive phenotype, and ones that decreased. Using these expressions as a threshold, the number of genes exceeding their respective expressions in either direction could be determined for any given sample. After a predetermined number of these positive indications, the sample could then be identified to be CIMP positive with some degree of confidence. The optimal number of indicators to use, the expression thresholds, and the cutoff after which CIMP status could be determined would have to be found in-practice, however, and necessitated the use of another dataset for validation.

2.2 Validation of the predictor

2.2.1 Analysis of TCGA datasets

A testing set was obtained from the Cancer Genome Atlas. 460 colorectal cancer cases were published, of which 444 had paired DNA methylation and transcriptome data (25). 45 cases

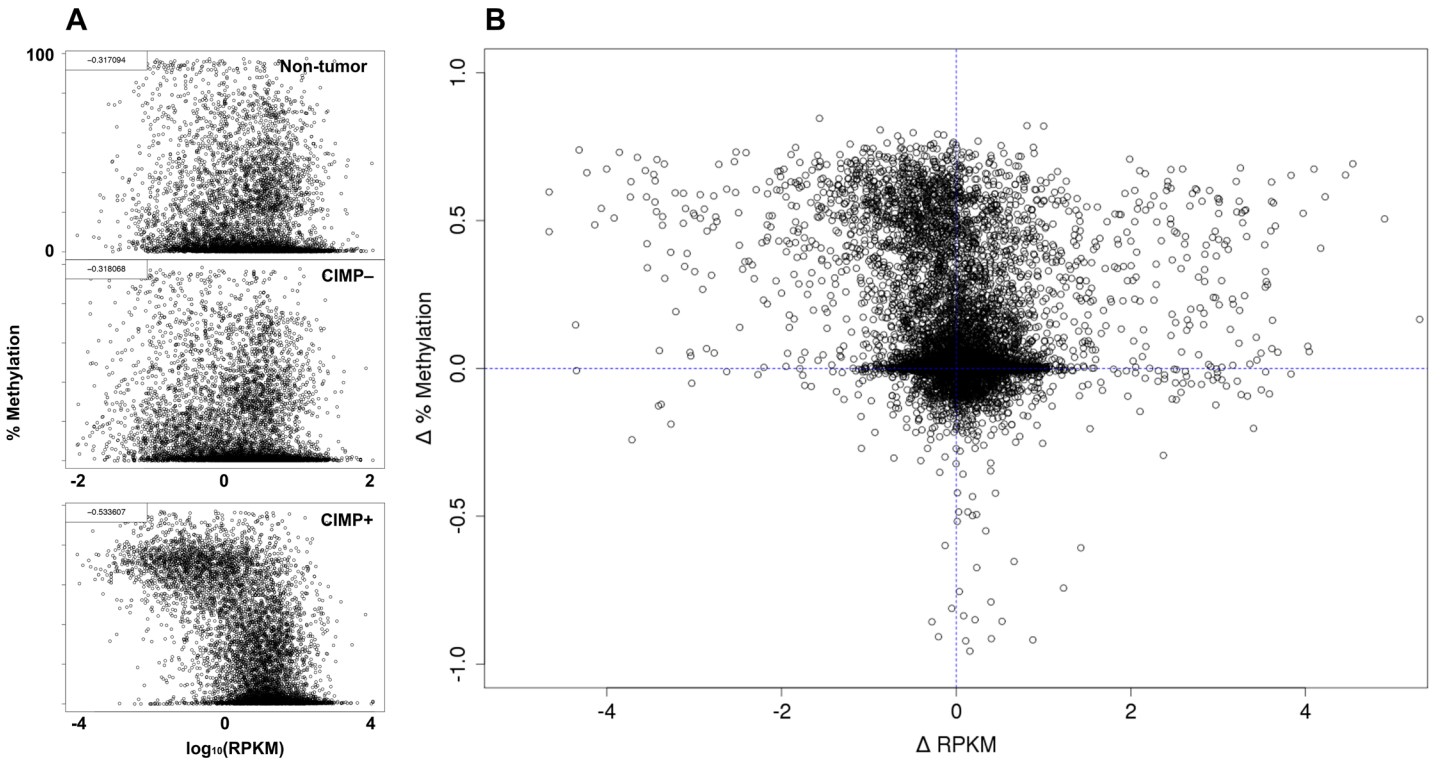

Figure 3A | **Correlation plots of percent methylation vs $\log_{10}$(RPKM).** n = 12, 456. Gene expression and methylation values plotted are averages of pooled samples withiin each of 3 groups. A Spearman's rank correlation of – 0.317, – 0.318 and – 0.534 were measured in the non-tumor, CIMP–, and CIMP+ plots respectively. While correlations were insignificant in the non-tumor and CIMP– groups, a distinct cluster of genes have increased methylation and decreased gene expression in the CIMP+ group.

3B | **Plot of change in methylation against change in gene expression between CIMP– and CIMP+ phenotypes.** n = 12, 456. Differences in methylation and gene expression were measured between the CIMP+ and CIMP– samples for all gene plotted. Quadrants are separated by dashed lines. Genes with increased methylation observed in the CIMP+ pool have a corresponding reduction in gene expression (top left quadrant). A smaller subet is also observed in the top right quadrant, representing an increase in both gene expression and methylation.

had a corresponding normal tissue sample analyzed by DNA methylation only. It is important to note that methylation data was measured using the Illumina HumanMethylation27 and HumanMethylation450 arrays, provided in β-values. This data was filtered for probes that measured CpGs in promoter-associated CGIs of protein-coding genes. β-values were averaged for all CpGs that were associated with one CGI. Gene expression was calculated by high-throughput RNA sequencing, and provided in FPKM values.

2.2.2 Difficulties in classifying CIMP status

An initial attempt to determine the CIMP status of each TCGA case was made using CIMP indicator genes. Fractional methylation at each of the indicators was determined for all samples. While the methylation at some indicators distinctly separate normal from increased methylation, others exhibit a gradual tailing into regions of higher methylation. For the indicators that had a clear separation, cutoffs were determined for high and normal degrees of methylation, and samples with increased methylation at two or more indicators were tentatively deemed positive.

To assess the accuracy of this strategy, correlation between CGI methylation and gene expression was determined for the pooled positive and negative samples as before. It would be expected that a similar change would be observed between the two, from which a subset of CIMP indicator genes could be extracted that has a significant intersection with the genes identified in the CEMT analysis. However, when the Euclidean distance of the combined changes in CGI methylation and gene expression were calculated, little to no significant movement of the data points was observed. As this was diagnostic of a failure to separate the samples in a meaningful way, classification was re-attempted with a different strategy.

CIMP status was re-determined using hierarchical clustering by CGI methylation. Unlike before, where an obvious subset of the genes had increased methylation in the positive phenotype, heterogeneity was observed for many CGIs and was inconsistent across samples. As a result, the clusters were grouped insignificantly due to an apparent gradual increase in CGI methylation from true-normal samples to samples that were presumably CIMP+. While the original problem arose from a dependence on too few genes, analyzing the whole methylome obfuscated any trends that could be used to differentiate the two phenotypes. To extract genes that would highlight differences in methylation, the standard deviation of the β-values for every gene across the samples was calculated. Clustering was performed on the subset that had a standard deviation greater than 0.35, resulting in four distinct groups (Figure 4). While CIMP+, CIMP–, and true normal groups were readily separated, a fourth group exhibited heterogeneous increases in methylation. For the purposes of the predictor, this group was considered CIMP+ due an inability to differentiate this CIMP "high" group from the CIMP "low" group. Of the 444 samples analyzed, 102 were

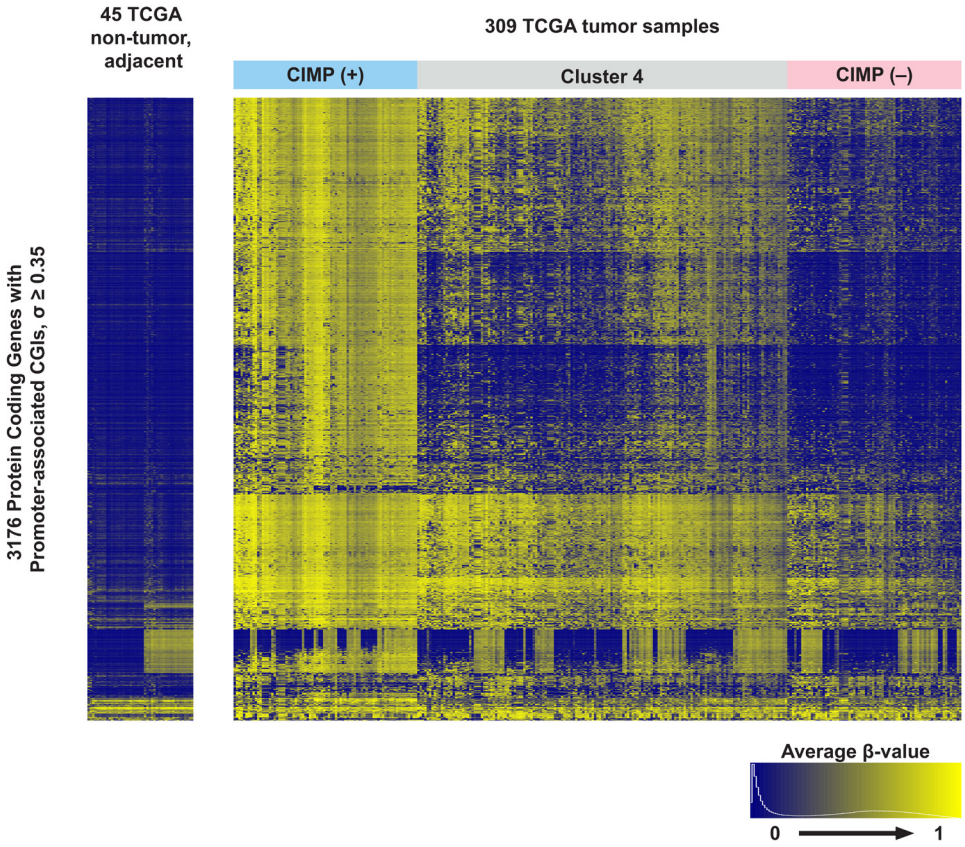

Figure 4 | **Heatmap of CGI methylation for 354 samples with HumanMethylation450 array data.** Average β values are measured for all probes that associated with the same CGI. Genes with methylation calls that had standard deviations greater than 0.35 across the 354 samples were used in clustering. Removing the cluster containing normal tissue, 3 distinct groups are observed. While CIMP+ and CIMP– clusters have similar methylation patterns, the fourth cluster has significant heterogeneity. Data from HumanMethylation27 array are excluded from this plot.

found to be CIMP "high", 157 were found to be CIMP–, and 185 were found to be CIMP "low".

2.2.3 Assessing the predictor

Using the prepared testing set, the predictor was assessed for accuracy. An initial adjustment had to be made due to the difference in scale between FPKM and RPKM values. On average, for any given gene's expression, the RPKM measured in the CEMT dataset was 2 to 4x larger than its corresponding FPKM in the TCGA data. As a result, the thresholds were scaled down to 1/3 of their original value. Without this correction, all thresholds checking for increased expression were consistently negative, whilst thresholds checking for decreased expression were consistently positive, and did not result in any meaningful separation. To determine the number of thresholds met before a sample is predicted to be positive, as well as the number of thresholds to be used, the predictor was ran iteratively. Using the top 100 genes that had the greatest increase and decrease in expression, and using a cut-off of 36, the predictor achieved an accuracy of 66.3% in the analysis of 445 samples. Of the incorrectly predicted samples, 72 were false positives and 78 were false negatives.

3. Discussion

While the accuracy of the predictor is underwhelming, its shortcomings highlight some points of discussion regarding CIMP status and the role of methylation in general as a repressor of gene expression. The lack of correlation between the two in normal and CIMP– tissues suggest that a resting methylation state, whether quantitatively high or low, has a more complicated relationship with gene expression than a simple negative correlation. However, aberrant increases in methylation, such as those observed in the CIMP+ disease state did correlate with a reduction in expression of the associated gene. By measuring the magnitude of change in methylation, it was clear that increasingly methylated CGIs were negatively correlated with the expression of some genes. A positive correlation is also observed in a small subset, however, suggesting diverging roles of methylation depending on their context. Curiously, of the genes that decreased in methylation in the positive phenotype, none of them changed significantly in expression. Taken together, this data implies that the current model for the relationship between CGI methylation and gene expression holds in cases where methylation is increased above the norm, but is not strictly observed at all loci.

To account for this disparity, the predictor was designed to check for both increased and decreased expression in different indicator genes. Although not the original intention, it is likely that monitoring for an increase in expression is a more robust test than checking for a decrease in expression. In practice, a negative result may be due to a failure to detect a signal and is often more difficult to confirm as correct. Alternatively, a positive result generally only occurs when a process has succeeded and is more readily accepted as the true outcome. To assess any bias that may have been introduced into the predictor, the expression thresholds exceeded for every sample was separated into indicators that were positive due to a decrease in expression, and indicators that were positive due to an increase in expression. While no significant difference was observed within the samples that were predicted to be CIMP+, the samples predicted to be CIMP– largely appeared to be the result of a decrease in the number of indicators that returned a positive result due to a decrease in expression. As the same number of indicators measuring increases and decreases in expression were used, the result suggests that the latter was the determining metric in the classifier. Although counterintuitive, this is in agreement with increased methylation being the primary bisector in identifying CIMP, assuming methylation negatively correlates with expression.

A significant challenge was encountered in trying to classify the TCGA datasets into CIMP+ and CIMP– groups. Using specific genes as indicators for increases in methylation, the group classified as CIMP+ did not have a significant difference in either gene expression or CGI methylation from the CIMP– group. As distinct clustering was observed when a much larger gene set was used, it is likely that inconsistency in methylation of this small set is to blame, which would have a much smaller effect when averaged across 3000 genes. As WGBS analysis provides fractional methylation information on all CpGs, it is possible to determine the percent methylation, on average, observed within any given CGI. Conversely, methylation assessed at a single probe in the HumanMethylation450 microarray is extrapolated to be the methylation state of all CpGs in its associated CGI. As a result, the importance of a single methylation call in the microarray is significantly greater than that of each converted 5mC in WGBS. As our current understanding is limited on how consistently methylation is propagated within a CGI, it is entirely possible that the degree of heterogeneity observed in the CIMP-intermediate group is an artifact of the assay used. While the existence of this transitional phenotype is generally accepted, the actual deviation in methylation for any given gene may be much smaller than what would be seen in Figure 4. As a result, it is possible that the use of a small number of indicators confounded the classifier due to an inconsistency in the methylation observed, whilst using a larger number of genes hid low-frequency errors when extrapolating methylation state from a single CpG.

Although the predictor was unable to reliably identify CIMP status, a number of optimizations may be made that could increase its accuracy. In the TCGA data analysis, the CIMP-intermediate samples that were identified were clustered into the CIMP+ pool for the purposes of the predictor, which was designed to classify in a binary fashion. Dissecting the incorrectly categorized samples, it was found that 61 of the 78 false negatives, but only 3 of the 72

false positives, originated from this CIMP-intermediate cluster. This suggests, firstly, that heterogeneity in CIMP-intermediate methylation is potentially being reflected in gene expression, but also that the predictor may have a positive bias. A solution to this problem would be to adjust the classifier to separately identify this third transitional group, which would indirectly increase the accuracy of predicting the CIMP+ and CIMP– groups. A problem with this approach would be the apparent gradient in phenotype between the two. This observation would suggest that it is only possible to accurately identify the CIMP+ phenotype, and that heterogeneity in the CIMP-intermediate group makes it too phenotypically similar to the CIMP– expression profile to confidently make a separation.

Improvements could also be made to the thresholds used in the predictor. Scaling of the RPKM values calculated in the CEMT analysis to fit the FPKM values from the TCGA dataset is non-ideal. By performing a conversion of the FPKM values into RPKMs, or vice versa, the accuracy of the predictor could be better assessed. Furthermore, adjustments to the thresholds themselves could be made, as average expression for a single indicator erroneously identifies half of the positive samples as negative, by definition. The specific method of making this adjustment would be fairly complex, however, as a large range of RPKM values may be observed that do not scale in a linear fashion. For example, a reduction in expression of 100 RPKM from 150 to 50 would be fairly obvious, and may be easily encompassed by a 10% adjustment in the threshold to 55. However, the same reduction in 100 RPKM from 2100 to 2000 is much more subtle, and a 10% adjustment to a threshold of 2200 would incorrectly identify all the samples to be positive. A potential solution to this could be to measure differences in expression proportional to the original value, and tailor adjustments on this scale. In the same example, the first reduction in expression would represent a 66% change, whilst the second represents only a 5% change. As a result, if the thresholds were to be adjusted by 10%, the former could be increased by 6.6%, whilst the later would be increased by 0.5%, resulting in new thresholds of 53.3 and 2010 respectively.

4. Conclusion

A predictor was developed that could classify gene expression data from colorectal cancer samples into CIMP+ and CIMP– groups with 66.3% accuracy. A statistically significant correlation was observed between gene expression and changes in promoter-associated CGI methylation, though a more granular analysis reveals that the trend is not a simple inverse relationship. While it may be possible to optimize the parameters used, a better classification of the dataset used in validation may more appropriately train the predictor to separate the two phenotypes. Additionally, the use of more sophisticated classification techniques, such as machine-learning, could potentially identify an ideal set of genes that best separate the two groups with more specificity than the techniques used here.

5. Methods

5.1 CIMP Classification

Genes used in identification of CIMP status include CACNA1G, CDKN2A, CRABP1, IGF2, MLH1, MGMT, RBP1, NEUROG1, RUNX3, SOCS1, and THBS1. Thresholds for separating high from normal methylation were tailored to each gene. As average methylation was distributed bimodally in each indicator, the means of the two subpopulations were determined using Gaussian Mixture Modeling, and a threshold was determined that sufficiently separated the two. If more than 3 indicators exceeded their respective thresholds, the sample was deemed CIMP+.

5.2 Hierarchical clustering

All calculations were performed using R. Heatmaps were generated using heatmap.2, and clustering was performed using the hclust function from the R stats package. Distance between samples were calculated using Spearman rank correlation. Complete linkage hierarchical clustering was used to separate sample groups. RPKM values were scaled using log base 10.

5.3 Creation of the predictor

The predictor is written in shell script, and utilizes AWK for streaming data processing. Thresholds were produced based on the average expressions observed in the CIMP+ CEMT samples, and are expressed in RPKMs. For analysis of TCGA data, thresholds were adjusted to 1/3 of their original values to accommodate the difference in scale. ENSGs used in the key are from human genome build hg38. Code for the predictor can be found at https://github.com/ddtam/cimp-predictor.

Abbrevations used:

CGI – CpG island
CIMP – CpG island methylator phenotype
FPKM – fragments per kilobase of transcript per million mapped reads
RPKM – reads per kilobase of transcript per million mapped reads
TSS – transcription start site
WGBS – whole genome bisulfite sequencing

Acknowledgements

I would like to thank Dr. Martin Hirst for taking me on as a directed studies student for the past eight months. Your continuous guidance and support throughout this project have been invaluable, especially at times when I haphazardly decided to complicate the project for myself. I would also like to thank Annaick Carles and Thomas Sierocinski for their advice and company as I struggled through shell scripting, AWK, and even the intricacies of machine-learning at one point; I have learned an immense amount from the both of you in the past 8 months, and I will be perpetually grateful. And to all members of the Hirst lab, your counsel in and out of lab meetings have brought this project much farther than I ever could have hoped to bring it alone.

Finally, I would like to thank my family and friends, who have put up with my late nights and missed calls due to my nose being glued to a terminal; your unwavering support means the world to me as I continue to struggle through the vast world of academia.

References

1. **Wong RS. 2011**. Apoptosis in cancer: from pathogenesis to treatment. J Exp Clin Cancer Res **30**:87.
2. **Easwaran H, Tsai H-C, Baylin SB**. 2014. Cancer Epigenetics: Tumor Heterogeneity, Plasticity of Stem-like States, and Drug Resistance. Mol Cell **54**:716–727.
3. **You JS, Jones PA. 2012**. Cancer Genetics and Epigenetics: Two Sides of the Same Coin? Cancer Cell **22**:9–20.
4. **Lee JY, Lee T-H**. 2012. Effects of histone acetylation and CpG methylation on the structure of nucleosomes. Biochim Biophys Acta **1824**:974–982.
5. **Kim J, Lee J, Lee T-H**. 2015. Lysine Acetylation Facilitates Spontaneous DNA Dynamics in the Nucleosome. J Phys Chem B **119**:15001–15005.
6. **Barth TK, Imhof A**. 2010. Fast signals and slow marks: the dynamics of histone modifications. Trends in Biochemical Sciences **35**:618–626.
7. **Cedar H, Bergman Y**. 2009. Linking DNA methylation and histone modification: patterns and paradigms. Nature Reviews Genetics **10**:295–304.
8. **Hackett JA, Surani MA**. 2013. DNA methylation dynamics during the mammalian life cycle. Philos Trans R Soc Lond, B, Biol Sci **368**:20110328–20110328.
9. **Deaton AM, Bird A**. 2011. CpG islands and the regulation of transcription. Genes Dev **25**:1010–1022.
10. **Panchin AY, Makeev VJ, Medvedeva YA**. 2016. Preservation of methylated CpG dinucleotides in human CpG islands. Biol Direct **11**:11.
11. **Jones PA**. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics **13**:484–492.
12. **Newell-Price J, Clark AJL, King P**. 2000. DNA Methylation and Silencing of Gene Expression. Trends in Endocrinology & Metabolism **11**:142–148.
13. **Razin A, Riggs AD**. 1980. DNA methylation and gene function. Science **210**:604–610.
14. **Lock LF, Takagi N, Martin GR**. 1987. Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. Cell **48**:39–46.
15. **Messerschmidt DM, Knowles BB, Solter D**. 2014. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. Genes Dev **28**:812–828.
16. **Bird A**. 2002. DNA methylation patterns and epigenetic memory. Genes Dev **16**:6–21.
17. **Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP**. 1999. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci USA 96:8681–8686.
18. **Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse MA, Kang GH, Widschwendter M, Weener D, Buchanan D, Koh H, Simms L, Barker M, Leggett B, Levine J, Kim M, French AJ, Thibodeau SN, Jass J, Haile R, Laird PW**. 2006. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat Genet **38**:787–793.
19. **Miller B, Sánchez-Vega F, Elnitski L**. 2016. The Emergence of Pan-Cancer CIMP and Its Elusive Interpretation. Biomolecules 2016, Vol 6, Page 45 **6**:45.
20. **Moarii M, Boeva V, Vert J-P, Reyal F**. 2015. Changes in correlation between promoter methylation and gene expression in cancer. BMC Genomics **16**:873.
21. **Ogino S, Kawasaki T, Kirkner GJ, Kraft P, Loda M, Fuchs CS**. 2007. Evaluation of markers for CpG island methylator phenotype (CIMP) in colorectal cancer by a large population-based sample. Journal of Molecular Diagnostics, The **9**:305–314.
22. **van Rijnsoever M, Elsaleh H, Joseph D, McCaul K, Iacopetta B**. 2003. CpG Island Methylator Phenotype Is an Independent Predictor of Survival Benefit from 5-Fluorouracil in Stage III Colorectal Cancer. Clin Cancer Res **9**:2898–2903.
23. **Galamb O, Kalmar A, Peterfia B, Csabai I, Bodor A, Ribli D, Krenacs T, Patai AV, Wichmann B, Bartak BK, Toth K, Valcz G, Spisak S, Tulassay Z, Molnar B**. 2016. Aberrant DNA methylation of WNT pathway genes in the development and progression of CIMP-negative colorectal cancer. Epigenetics **11**:588–602.
24. **Lee S, Cho N-Y, Yoo EJ, Kim JH, Kang GH**. 2008. CpG island methylator phenotype in colorectal cancers: comparison of the new and classic CpG island methylator phenotype marker panels. Arch Pathol Lab Med **132**:1657–1665.
25. **Cancer Genome Atlas Network**. 2012. Comprehensive molecular characterization of human colon and rectal cancer. Nature **487**:330–337.