Predicting gene expression using DNA methylation in two human populations

Huan Zhong 1, Soyeon Kim 2, Degui Zhi 2, Xiangqin Cui Corresp. 3

1 Department of Biology, Hong Kong Baptist University, Hong Kong, China
2 School of Biomedical Informatics, University of Texas Health Center at Houston, Houston, Texas, United States
3 Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, United States

Corresponding Author: Xiangqin Cui
Email address: xiangqin.cui@emory.edu

Background. DNA methylation, an important epigenetic mark, is well known for its regulatory role in gene expression, especially the negative regulation in the promoter region. However, its correlation with gene expression at population level has not been well studied. In particular, it is unclear if genome-wide DNA methylation profile of an individual can predict her/his gene expression profile. Previous studies were mostly limited to association analyses between single CpG site methylation and gene expression. It is not known whether DNA methylation of a gene has enough prediction power to serve as a surrogate for gene expression in existing human study cohorts with DNA samples but not RNA samples.

Results. We studied two human population datasets, Multiple Tissue Human Expression Resource Projects (MuTHER)’s Adipose tissue as well as asthma and normal peoples’ peripheral blood mononuclear cell (PBMC), for predicting gene expression using methylation of all CpG sites from the gene region. Three prediction models were investigated; single linear regression, multiple linear regression, and least absolute shrinkage and selection operator (LASSO) penalized regression. Our results showed that LASSO regression has superior performance among these methods. However, even with LASSO regression, very small prediction R2 was obtained for the majority of genes and only about one thousand genes had prediction R2 greater than 0.1. GO term and pathway analyses of these more predictable genes showed that they are enriched for immune and defense genes.

Conclusion. In human populations, DNA methylation of CpG sites at gene region have weak prediction power for gene expression. The relatively more predictable genes tend to be defense and immune genes.
Predicting Gene Expression Using DNA methylation in Two Human Populations

Huan Zhong\textsuperscript{1}, Soyeon Kim\textsuperscript{2}, Degui Zhi\textsuperscript{2}, Xiangqin Cui\textsuperscript{3}

\textsuperscript{1} Department of Biology, Hong Kong Baptist University, HONG KONG
\textsuperscript{2} School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA
\textsuperscript{3} Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, USA

Corresponding author: Xiangqin Cui
Email: xiangqin.cui@emory.edu

Keywords

DNA methylation, Methylation Microarray, transcriptome, LASSO
Abstract

Background

DNA methylation, an important epigenetic mark, is well known for its regulatory role in gene expression, especially the negative regulation in the promoter region. However, its correlation with gene expression at population level has not been well studied. In particular, it is unclear if genome-wide DNA methylation profile of an individual can predict her/his gene expression profile. Previous studies were mostly limited to association analyses between single CpG site methylation and gene expression. It is not known whether DNA methylation of a gene has enough prediction power to serve as a surrogate for gene expression in existing human study cohorts with DNA samples but not RNA samples.

Results

We studied two human population datasets, Multiple Tissue Human Expression Resource Projects (MuTHER)’s Adipose tissue and Asthma and normal peoples’ peripheral blood mononuclear cell (PBMC), for predicting gene expression using methylation of all CpG sites from the gene region. Three prediction models were investigated; single linear regression, multiple linear regression, and least absolute shrinkage and selection operator (LASSO) penalized regression. Our results showed that LASSO regression has superior performance among these methods. However, even with LASSO regression, very small prediction $R^2$ was obtained for the majority of genes and only about one thousand genes had prediction $R^2$ greater than 0.1. GO term and pathway analyses of these more predictable genes showed that they are enriched for immune and defense genes.
Conclusion

In human populations, DNA methylation of CpG sites at gene region have weak prediction power for gene expression. The relatively more predictable genes tend to be defense and immune genes.

Background

DNA methylation has long been recognized as an important epigenetic modification in regulating gene expression (Razin & Riggs, 1980). This process often occurs at CG dinucleotides sites (CpG site), adding a methyl group to the cytosine residue (You & Jones, 2012; Wagner et al., 2015). In mammals, more than 70% of CpG sites are methylated.

The regulatory role of DNA methylation has been mostly studied with a small number of CpG sites in a limited number of genes. The more recent application of microarrays and next generation sequencing enables large-scale analysis of DNA methylation and gene expression across the whole genome (Krueger et al., 2012). However, most genome-wide methylation and expression studies have small sample sizes for comparing controlled groups. Only a few studies profiled both genome-wide DNA methylation and gene expression in larger human populations and examined their relationship. Del Rey et al., (Del Rey et al., 2013) studied the genome-wide DNA methylation and gene expression in 83 low-risk subtypes of Myelodysplastic syndrome (MDS) patients and 36 controls using microarrays. While they found negative correlations...
between methylation and gene expression in a large proportion of differentially expressed and
differentially methylated genes, they also found substantial positive correlations in them. In
another study of 648 twins, overall negative correlations were found in the adipose tissue,
promoter region (-0.018), gene body (-0.013) and 3-prime UTR (-0.007)(Grundberg et al., 2013).
More recently, Wagner et al.(Wagner et al., 2014) profiled the genome wide DNA methylation
and gene expression in forearm skin fibroblast among 62 unrelated individuals. They found that
the association between gene expression and methylation is not always negative in promoter
region or positive in gene body.

In this study, we examine the DNA methylation and gene expression relationship in two large
human datasets. One is from the MuTHER project that includes 856 female European individuals
from TwinsUK Adult Twin Registry(Spector & Williams, 2006). Another dataset is the
Childhood Asthma study that has 194 inner city children(Yang et al., 2015), which we call
PBMC dataset for convenience throughout the paper. We determine the overall relationship
between DNA methylation and gene expression for each gene and evaluate the predictive
potential of DNA methylation for gene expression. We demonstrate that a penalized regression
improves the overall prediction. However, the prediction power is still low for most genes. For
only a small set of genes, DNA methylation has substantial prediction power, especially the
immune genes. Similar to PrediXcan’s prediction of gene expression using genotype
variations(Gamazon et al., 2015), we wrapped our predictions based on DNA methylation into a
package named “MethylXcan”.

Methods

1) Datasets:
Adipose Dataset

This dataset is from the MuTHER study, consisting of 856 female European-descent individuals enrolled in the TwinsUK Adult Twin Registry. The quartile normalized gene expression and DNA methylation data from subcutaneous fat were downloaded from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/). The gene expression data (accession number E-TABM-1140) were generated for 25160 genes using Illumina HumanHT-12 v3.0 on 825 individuals. The log2-transformed signals were quartile normalized for each tissue followed by quartile normalized across the whole population (Grundberg et al., 2012). The DNA methylation data (accession number E-MTAB-1866) were generated using Illumina Infinium HumanMethylation450 from 649 female twins. The methylation beta values were already quantile normalized for each type of probe, ranging from 0 (unmethylated) to 1 (totally methylated) [6].

PBMC Dataset

This dataset was downloaded from Gene Expression Ominbus (GSE40736). It includes 194 inner-city children with 97 cases of atopy and persistent asthma and 97 healthy controls. All the study subjects were 6 to 12 years old from African American, Dominican-Hispanic and Haitian-Hispanic background [9]. DNA methylation data were generated using Illumina’s Infinium Human Methylation450k BeadChip. The normalized data matrix was downloaded from ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE40nnn/GSE40576/matrix/ and the methylation M-values were converted to beta values ranging from 0 to 1. Gene expression data were generated for 23612 genes using Nimblegen Human Gene Expression arrays (12x135k). The normalized data matrix was downloaded from
According to the publication, one outlier sample has been removed after principle component analysis, SWAN normalization was used for methylation data. Log2 transformation and RMA normalization were used for gene expression data. For each gene, expression level was standardized across samples.

2) Dataset cleaning and filtering

To assess the DNA methylation effect in prediction gene expression, we defined the “methylation probes” as the 344,303 probes in Table S1 of the Grundberg’s paper. The probes on the methylation array but excluded from the Table S1, which have potential SNP effects or cross hybridization effects, are termed “S&C probes”. The combinations of these two types of probes are termed “all probes”.

The Adipose dataset has 32,478 missing values in the DNA methylation data. Samples with missing values were excluded from regression analysis. Among the 485679 probes in the dataset, 344201 probes remained in Adipose dataset after filtering. For the PBMC dataset, 344180 out of the 485461 probes in the dataset remained after filtering. For method comparisons and the consistency of the analysis, only genes that have the LASSO models were used, 8121 genes and 150349 CpG sites in Adipose dataset and 4251 genes and 73553 CpG sites in PBMC datasets.

(Supplementary Table 1).

3) Modeling the relationship between gene expression and DNA methylation.

CpG probes were mapped to genes using UCSC RefGene annotation. Gene expression and DNA methylation data for each gene were extracted using in-house perl script. Since there was no
missing value for methylation of the PBMC dataset, all samples were used in the regression analysis.

We used three types of regressions, single linear regression, multiple regression, and least absolute shrinkage and selection operator (LASSO) regression (Tibshirani, 1996), to model the linear relationship between gene expression and DNA methylation. Squared correlation ($R^2$) between predicted and observed data was used to compare the three types of regressions. In the single linear regression, each CpG site was modeled separately to predict gene expression level. The CpG site that provides maximum $R^2$ was used to represent each gene. In multiple regressions, all the CpG sites in each gene were used as predictors and the $R^2$ was calculated. In the LASSO regression, all CpG sites were used to predict the gene expression. We used the GLMNET package in R to fit the LASSO model in which penalized parameters were obtained using 10-fold cross-validation to minimize the mean squared error, while the predictors and response variables were all standardized.

4) Cross validation

In addition to calculating the $R^2$ from fitting the models ($fitting \ R^2$), we also conducted five-fold cross validation to compare the prediction power of the three regression models using the validation $R^2$ ($R^2.cv$). Specifically, the samples were randomly separated into training set (4/5 of data) and testing set (1/5 of data). The procedure was iterated 10 times and the mean $R^2$ of the 10 five-fold cross validations was used as our final cross validation $R^2$ for each model. For single regression, cross validation was conducted for each CpG site and the maximum $R^2$ was used for
each gene. For LASSO regression analysis, we first obtained the optimal penalty parameter using ten-fold cross validation and then used another five-fold cross validation to evaluate the predictive performance of the model.

Note that we calculated fitting $R^2$ in the LASSO cross validation models. We used the entire datasets as testing in the LASSO cross validation models in order to obtain the fitting $R^2$ in a fashion consistent with the multiple and single cross-validation models. In this case, all the $R^2$ values in the paper are squared correlation of the predicted and the true values in the training set.

5) Model comparisons on significant genes

We first identified genes that showed overall model prediction p values less than 0.05 in Lasso regressions and then compared the three regression models on these genes.

6) Gene Ontology (GO) and pathway enrichment analysis.

For genes with $R^2$ greater than 0.1, we use The Database for Annotation, Visualization and Integrated Discovery (DAVID) at https://david.ncifcrf.gov/ (Huang, Sherman & Lempicki, 2008) to conduct GO term enrichment analysis based on modified Fisher Exact Test. The background genes were set to be the genes on the expression array, HumanHT-12_V3_0_R2_11283641_A. The significantly overrepresented GO terms were selected based on the EASE Score, which is the geometric mean of p-values on logarithm scale for the member terms. We applied medium classification stringency in the DAVID website to our data. “GOTERM_BP_FAT” was used to obtain more information in biological processes of the Gene Ontology enrichment analysis. “KEGG_PATHWAY” was selected for pathway enrichment analysis in the same fashion. The most enriched GO terms and pathways with low p-value or FDR were shown in the results.
7) Gene expression prediction using different type of probes on the methylation microarray

The probes excluded by Table S1 of the Grundberg’s paper were treated as probes with SNP and/or hybridization effects (S&C probes). We compared these probes, the methylation probes, and the combination of these two types of probes in predicting gene expression.

8) Software Package: MethylXcan

We created a software package, MethylXcan, for predicting gene expression using DNA methylation array data in the analyses of this paper. This package includes all three regressions and calculates the squared correlation for each model. The program was written in R and Perl, and has been tested under linux or MACSOX system. Users can use this package on their own data after formatting their methylation data, expression profiling data, and annotation files as specified by the package. URL: https://github.com/dorothyzh/MethylXcan

Results

Most DNA methylation studies analyze one CpG at a time in explaining gene expression. In this study we set out to find whether combining all CpG sites in a gene can better predict the gene expression in a human population. We obtained two human datasets, the Adipose dataset generated from subcutaneous fat tissue and the Childhood Asthma dataset generated from PBMC. To evaluate the predicting power of DNA methylation on gene expression, we conducted three types of linear regression analyses, single regression, multiple regression, and LASSO regression for each gene. The squared correlations (R²) were used for model comparisons. To focus on DNA methylation effect, we first left out CpG probes that overlap SNPs or cross-
hybridize to multiple locations. In addition, since some genes fail to establish a LASSO model, we only focus on genes with LASSO models for comparing different regression methods.

Multiple regressions using all CpGs from a gene predict gene expression the best in model fitting.

For the Adipose dataset, the single regression identified a large number of genes with significant CpG sites, about half (3509) with at least one CpG site significant at significance level of 0.0001 and most of them (7423) have at least one CpG site significant at significance level of 0.05. However, the prediction power represented by the maximum $R^2$ for these CpG sites is generally low. Only 20 genes have maximum $R^2$ greater than 0.3 and 491 (3.0%) genes have maximum $R^2$ larger than 0.1 (Table 1). Since multiple CpG sites from each gene were assayed on the methylation microarray, we applied multiple linear regression to utilize all CpG sites as predictors simultaneously. The $R^2$ explained by the regression model did improve substantially for the majority of genes compared with that of the single linear regression model (Figure 1A). The significant genes (red color) have relatively higher $R^2$ compared with the non-significant genes (Figure 1). For the PBMC data, a much smaller number of significant CpG sites were identifies at the same significance level for the 4251 genes analyzed (582 with p value less than 0.0001 and 3112 with p value less than at 0.05 for single regression) potentially due to the smaller sample size. However, the numbers of genes with larger $R^2$ values are comparable with those from the Adipose dataset (Table 1). The improvement of $R^2$ from the multiple regressions over single regress in the PBMC dataset is also similar to that in the Adipose dataset (Figure 1A).
Compared with multiple regressions, LASSO regression did not generate $R^2$ as high as that from multiple regressions in both datasets (Figure 1B and D). For example, the number of genes with $R^2$ greater than 0.2 decreased from 472 to 343 for the Adipose dataset and from 1163 to 561 for the childhood asthma PBMC datasets (Table 2). With LASSO regression, only about 19.7% of the genes (1608 out of 8123) from the Adipose dataset and about 56% of the genes (2382 out of 4251) from the PBMC dataset have prediction $R^2$ greater than 0.1.

**LASSO regression shows best prediction in cross-validation.**

To better assess the accuracy of the predictive models, we performed 5-fold cross validation on single, multiple regressions, and LASSO regressions to estimate the $R^2$. For comparison purpose, we present $R^2$ with genes for which the LASSO models are available. The results showed that the LASSO regression produced much larger $R^2$ values than the single regression and less dramatic but discernable increases over multiple regressions (Figure 2). These differences are also reflected in the number of genes with $R^2$ exceeding certain thresholds. For example, 893 genes (11%) from the Adipose dataset have $R^2$ greater than 0.1 from LASSO regression, while 791 genes (9.7%) and 125 genes (1.5%) have $R^2$ greater than 0.1 from multiple regression and single regression, respectively (Table1). For genes with $R^2$ greater than 0.3, LASSO regression has 44 genes (0.54%) while multiple regression and single regression have 40 (0.49%) and 2 genes (0.03%), respectively. These results indicate that penalized regression has better prediction than multiple or single regressions for cross-validation test datasets. Cross validation tends to overcome bias and over-fitting issues. As expected, cross-validation $R^2$ values are generally
lower than those from the model fittings reflected by the smaller number of genes with $R^2$ values greater than certain cut-offs (Table 1). In addition, the significant genes from multiple regression fitting have relatively higher cross-validation $R^2$ than non-significant genes (red dots in Figure 2) as expected. Similar results were obtained from the PBMC dataset (Table 1 and Figure 2 C and D).

To make sure that the prediction $R^2$ is larger than those from random chance, we compared the cumulative $R^2$ from the two real datasets with those from the the null distribution of correlations based on Fisher z-transformation in quantile-quantile plots (Figure 3). Both datasets show that the observed $R^2$ values are much larger than the expected $R^2$ values from random chance (Figure 3A and 3C). In addition, the Adipose dataset has a larger departure than the PBMC dataset, indicating that the LASSO model could capture more proportion of the transcriptome variability in the Adipose dataset. This is potentially due to the larger sample size or the regulative nature of the different tissues.

Using all probes improves prediction power for gene expression

In order to evaluate DNA methylation power in predicting gene expression, we first left out a large proportion of probes potentially affected by genetic or cross-hybridization effects (Supplementary Table 1). However, using all probes on the array is preferred if our goal is to achieve better prediction accuracy of gene expression. To evaluate the prediction power from all probes, we included all available probes in LASSO regression and found that the overall prediction power did increase compared to the models using only the methylation probes (Figure...
We observed more genes with $R^2$ values exceeding the thresholds (Table 1). In addition, the largest $R^2$ value is much larger when all probes are used compared to only the methylation probes. For example, the largest $R^2$ is 0.92 from all probes compared to 0.54 from only methylation probes in the Adipose dataset (Figure 3B and Figure 3A). Likewise, the largest $R^2$ is 0.88 from all probes compared to 0.46 from methylation probes alone in the PBMC dataset (Figure 3C and 3D). Furthermore, LASSO models are available for more genes when all probes are used in both datasets (Supplementary Figure 2).

The increase of prediction power on gene expression from using all probes on the methylation microarray suggests that there is contribution from the probes with potential SNP effects or cross hybridization effects (S&C probes) we initially excluded in estimating methylation prediction. To further examine the size and nature of their contribution, we separately estimated the prediction power of the methylation probes, S&C probes, and the combination of them (all probes). The results showed that the S&C probes have independent prediction power from the methylation probes and the combination of both has increased prediction power over the methylation probes alone (red points vs black line in the left two panels in Figure 4). The prediction power from the S&C probes was estimated for genes with enough SNP probes to form a LASSO model and their prediction power are mostly above zero (blue points in Figure 4). The fact that the blue points are randomly distributed instead of following the black line suggests that the two sources of $R^2$ are not correlated; therefore, the larger genetic effect and larger epigenetic effect do not seem to coexist in the same genes. Figure 5 shows some examples of genes with large prediction powers from either methylation probes or S&C probes. As expected, the methylation probes show continuous methylation values while the S&C probes show categorical values due to limited genotypes of the samples.
Methylation probes tend to better predict the expression of defense and immune gene expressions.

To examine the potential biological function of the genes showing good predictability of gene expression by DNA methylation, we conducted gene ontology (GO) and pathway enrichment analysis using DAVID. For the Adipose dataset, we examined the 1608 genes with $R^2$ greater than 0.1 from LASSO regression. With false discovery rate (FDR) of 0.1, the significantly enriched biological process (GOTERM_BP_FAT) GO terms are response to inflammatory, defense activities and wounding, as well as antigen (Table 2), which seem to be consistent with the previous findings for subcutaneous fat cells (Berg and Scherer 2005). For the PBMC dataset, we conducted the same gene set enrichment analysis on the 2382 genes with $R^2$ greater than 0.1 and found that the significantly enriched terms are mostly related to defense and immune functions, lymphocyte activation, leukocyte activation, and immune response (Supplementary Table 2). These results seem to be reasonable for atopy and persistent asthma blood cells. Similar GO terms and pathways were enriched in results obtained from all probes (Supplementary Tables 3 and 4).

Discussion

We examined the relationship between gene expression and DNA methylation across the genome using data from two large human studies. We explored three linear regression models for predicting gene expression and found that shrinkage based LASSO multiple regression provides.
the best prediction. However, even with LASSO regression, the methylation probes can predict
expression in only a small proportion of genes with moderate prediction power. These genes are
mostly immune and defense genes. Using all probes on the methylation array does improve
prediction power to some degree.

We used squared correlation ($R^2$) to assess predictive power in regression models. The single
linear regression is based only on the best predictive CpG in each gene while the multiple
regression models uses all CpG sites in each gene. However, the multiple regression model has
substantial over-fitting problem for genes with large number of CpGs. The shrinkage based
LASSO regression model overcomes the over-fitting problem without losing predictability.
LASSO imposes sparsity among the coefficients and puts constraint on the overall absolute
values of the regression coefficients, which forces certain coefficients to be zero. This property is
beneficial in variable selection and improves model interpretability. LASSO is not the only
shrinkage-based regression method. There are other penalty regression models, such as the Ridge
(Hoerl & Kennard, 1970) and elastic net (Zou & Hastie, 2005). However, ridge regression scales
all coefficients by a constant without setting some of them to zero, which is less ideal for
variable selection. The elastic net method is a hybrid of the ridge and LASSO. There are also
other advanced shrinkage methods, such as elastic net with rescaled-coefficients and grouped
lasso (Yuan & Lin, 2006; Meier, Van De Geer & Bühlmann, 2008), as well as other variants of
LASSO (Zou et al., 2013). Further evaluation is needed for their merits in improving prediction
of gene expression in this setting.
One potential reason for low prediction power from DNA methylation on gene expression is the complex mechanisms of gene expression regulation. In addition to DNA methylation, transcription factors, histone modification (Verdin & Ott, 2015), and non-coding RNAs (Janowski et al., 2005; Ting et al., 2005; Ting, McGarvey & Baylin, 2006; Kaikkonen, Lam & Glass, 2011) all play critical roles in gene transcription regulation (Jones, 2015). Another potential reason for low prediction power of methylation is that the landscape of DNA methylation differs dramatically across cell types, tissues (Lokk et al., 2014), ages (Teschendorff et al., 2010), and races (Song et al., 2015). The relationship between gene expression and DNA methylation could also vary substantially across these factors. The correlation between gene expression and DNA methylation from bulk studies at population level encompasses all these variability; therefore, it is not surprising to see low prediction power based on DNA methylation alone. It is likely to be limited in general that the potential of DNA methylation alone as surrogate for gene expression. The combination of DNA methylation and other predictors, such as genotype, may be more powerful for this purpose.

Although the overall prediction power of DNA methylation on gene expression was low in our analyses of both datasets, much better prediction power was found with the immune and defense genes. This result is consistent with the extensive evidence linking aberrant DNA methylation patterns with genes involved in immune deficiencies, autoimmune disorders, and inflammatory disease (Bayarsaihan, 2011). Some key immune genes, such as mb1 (CD79a) (Gao et al., 2009) and MHCII, (HLA-DR, DQ, DP) (Seguín-Estévez et al., 2009; Suárez-Álvarez et al., 2010; Majumder & Boss, 2011) have been shown to have strong association between DNA methylation and their key immune functions. On cellular level, hypomethylation in T cell promoter regions plays an important role in lupus (Deng et al., 1998; Gorelik & Richardson, 2010). In addition,
DNA methylation remodeling is critical in the differentiation of naïve immune cells into antigen-specific effector cells in the response to the pathogen or virus (Scharer et al., 2013). By altering DNA methylation pattern, cells can adapt to external stimuli via changes in gene expression. Given these evidences and the results in our study, DNA methylation could be more useful in predicting the gene activities in the immune and defense system.

**Conclusions**

We explored three regression methods to predict gene expression using DNA methylation, single regressions, multiple regressions, and LASSO penalized regression. LASSO regression reduces over-fitting and improved the prediction power. Both datasets we analyzed shows overall low prediction power. The better predictive genes (with squared correlations larger than 0.1 in LASSO model) are mostly involved in immune, inflammatory, and wound responses. Overall, we will recommend caution for using one’s methylation profile to predict one’s transcriptome.

**List of Abbreviations**

- **SNP**: single nucleotide polymorphisms
- **LASSO**: least absolute shrinkage and selection operator
- **R2**: squared correlation
- **S&C**: the probes have potential SNP effects or cross hybridization effects
- **MuTHER**: Multiple Tissue Human Expression Resource Project
- **PBMC**: peripheral blood mononuclear cell
- **MDS**: myelodysplastic syndrome
Availability of data and material
For Adipose Dataset, both gene expression and DNA methylation data can be found at http://www.ebi.ac.uk/arrayexpress/
Software package, ‘MethylXcan’ are at https://github.com/dorothyzh/MethylXcan.

References


Majumder P., Boss JM. 2011. DNA methylation dysregulates and silences the HLA-DQ locus by


Teschendorff AE., Menon U., Gentry-Maharaj A., Ramus SJ., Weisenberger DJ., Shen H., Campan M.,
Noushmehr H., Bell CG., Maxwell AP. 2010. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res 20:440–446.


Figure 1

$R^2$ comparison among three regression models

Multiple regression is compared to single (left two panels) and lasso (right two panels) regressions in two datasets, Adipose and PBMC. $R^2$ shown here are on cubic root scale for visualization clarity. “single”, single linear regression with the most significant CpG site as predictor; “multiple”, multiple regression with all methylation CpG sites in a gene as predictors. Red points represent significant genes from multiple regressions at significance level of 0.05. Blue solid line is the identity line and the dashed lines represent $R^2$ of 0.1.
$R^2$ comparison among regression models in cross validation.

LASSO regression is compared to single (left two panels) and multiple (right two panels) regressions for two datasets, Adipose and PBMC. Five-fold validation was used for all regression models. The red points represent the significant genes from multiple regressions ($p < 0.05$). single.cv, single regression with cross validation; multiple.cv, multiple regression with cross validation; lasso.cv, LASSO regression with cross validation.
Figure 3

Figure 3. The prediction $R^2$ is beyond random noise

Sorted $R^2$ values are compared with those from the null distribution of $R^2$ based on Fisher z-transformation (straight line). Left two panels are from the methylation probes alone. Right two panels are from all probes. Five-fold cross validation was used for LASSO regression models.
Figure 4

Comparison of \( R^2 \) from methylation probes (black line), S&C probes (blue points), and all probes (red points).

The 95% confidence intervals of \( R^2 \) from methylation probes are shown as gray shadow.
Figure 5

Example genes with high prediction power.

$R^2$ is from LASSO regression models. Adipose and PBMC are the two datasets. methyl, methylation probes; all, all probes. X-axis indicates the predicted expression levels from LASSO regression models, y-axis were observed expression levels for each dataset.
Table 1 (on next page)

The number of genes with prediction R2 larger than thresholds (0.1, 0.2, and 0.3) using single, multiple and LASSO regressions.

The total number of predicted genes is 8122 excluding SNP-probes and 8929 including SNP-probes in the Adipose data. The total number of predicted genes is 4252 excluding SNP-probes in Adipose data and 5064 including SNP-probes in the PBMC data.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>SNP-probe Inclusion</th>
<th>R² &gt;=0.1</th>
<th>R² &gt;=0.1 cv</th>
<th>R² &gt;=0.2</th>
<th>R² &gt;=0.2 cv</th>
<th>R² &gt;=0.3</th>
<th>R² &gt;=0.3 cv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>NO</td>
<td>491</td>
<td>125</td>
<td>85</td>
<td>18</td>
<td>20</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>591</td>
<td>137</td>
<td>114</td>
<td>25</td>
<td>34</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Adipose</td>
<td>NO</td>
<td>2211</td>
<td>791</td>
<td>472</td>
<td>177</td>
<td>114</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>3105</td>
<td>966</td>
<td>765</td>
<td>232</td>
<td>217</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>LASSO</td>
<td>NO</td>
<td>1608</td>
<td>893</td>
<td>343</td>
<td>197</td>
<td>110</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>2189</td>
<td>1094</td>
<td>538</td>
<td>272</td>
<td>195</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>NO</td>
<td>851</td>
<td>381</td>
<td>108</td>
<td>33</td>
<td>14</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>1330</td>
<td>666</td>
<td>212</td>
<td>90</td>
<td>58</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>PBMC</td>
<td>NO</td>
<td>3358</td>
<td>746</td>
<td>1163</td>
<td>109</td>
<td>382</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>4455</td>
<td>994</td>
<td>1902</td>
<td>197</td>
<td>694</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>LASSO</td>
<td>NO</td>
<td>2382</td>
<td>1022</td>
<td>561</td>
<td>165</td>
<td>142</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>3207</td>
<td>1465</td>
<td>870</td>
<td>289</td>
<td>235</td>
<td>66</td>
<td></td>
</tr>
</tbody>
</table>
Gene Ontology and KEGG pathway enrichment analysis for more predictive genes in the Adipose dataset.

Genes with R2 greater than 0.1 (1680) were selected for conducting the GO and Pathway enrichment analyses. Enriched and representative terms with FDR less than 0.01 are included in the table.
<table>
<thead>
<tr>
<th>Term</th>
<th>Fold Enrich</th>
<th>P-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006952~defense response</td>
<td>2.28</td>
<td>1.50E-08</td>
<td>2.70E-05</td>
</tr>
<tr>
<td>GO:0006954~inflammatory response</td>
<td>2.69</td>
<td>2.72E-07</td>
<td>4.80E-04</td>
</tr>
<tr>
<td>GO:0009611~response to wounding</td>
<td>2.22</td>
<td>5.79E-07</td>
<td>0.001</td>
</tr>
<tr>
<td>GO:0048002~antigen processing and presentation of peptide antigen</td>
<td>8.93</td>
<td>8.43E-07</td>
<td>0.0015</td>
</tr>
<tr>
<td>GO:0002504~antigen processing and presentation of peptide or polysaccharide antigen via MHC class II</td>
<td>7.58</td>
<td>3.94E-06</td>
<td>0.007</td>
</tr>
<tr>
<td>KEGG_PATHWAY:hsa05330~Allograft rejection</td>
<td>5.91</td>
<td>8.58E-06</td>
<td>0.01</td>
</tr>
<tr>
<td>KEGG_PATHWAY:hsa05322~Systemic lupus erythematosus</td>
<td>3.52</td>
<td>9.17E-06</td>
<td>0.011</td>
</tr>
<tr>
<td>KEGG_PATHWAY:hsa04612~Antigen processing and presentation</td>
<td>3.73</td>
<td>1.65E-05</td>
<td>0.02</td>
</tr>
<tr>
<td>KEGG_PATHWAY:hsa05332~Graft-versus-host disease</td>
<td>5.45</td>
<td>1.87E-05</td>
<td>0.02</td>
</tr>
<tr>
<td>KEGG_PATHWAY:hsa04940~Type I diabetes mellitus</td>
<td>5.06</td>
<td>3.77E-05</td>
<td>0.046</td>
</tr>
</tbody>
</table>