Predicting gene expression using DNA methylation in three human populations

Huan Zhong[1], Soyeon Kim[4], Degui Zhi[2] and Xiangqin Cui[3]

[1] Department of Biology, Hong Kong Baptist University, Hong Kong, China

[2] School of Biomendical Informatics, University of Texas Health Center at Houston, Houston, TX, United States of America

[3] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, United States of America

[4] School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States of America

ABSTRACT

Background. DNA methylation, an important epigenetic mark, is well known for its regulatory role in gene expression, especially the negative correlation in the promoter region. However, its correlation with gene expression across genome at human 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 other than RNA samples.

Results. We examined DNA methylation in the gene region for predicting gene expression across individuals in non-cancer tissues of three human population datasets, adipose tissue of the Multiple Tissue Human Expression Resource Projects (MuTHER), peripheral blood mononuclear cell (PBMC) from Asthma and normal control study participates, and lymphoblastoid cell lines (LCL) from healthy individuals. 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, the prediction power is generally low and varies across datasets. Only 30 and 42 genes were found to have cross-validation R^2 greater than 0.3 in the PBMC and Adipose datasets, respectively. A substantially larger number of genes (258) were identified in the LCL dataset, which was generated from a more homogeneous cell line sample source. We also demonstrated that it gives better prediction power not to exclude any CpG probe due to cross hybridization or SNP effect.

Conclusion. In our three population analyses DNA methylation of CpG sites at gene region have limited prediction power for gene expression across individuals with linear regression models. The prediction power potentially varies depending on tissue, cell type, and data sources. In our analyses, the combination of LASSO regression and all probes not excluding any probe on the methylation array provides the best prediction for gene expression.

Subjects Bioinformatics, Computational Biology, Genomics, Epidemiology, Statistics

Keywords DNA methylation, Methylation microarray, Transcriptome, LASSO

Submitted 19 July 2018
Accepted 10 March 2019
Published 1 May 2019

Corresponding author
Xiangqin Cui,
xiangqin.cui@emory.edu

Academic editor
Jun Chen

Additional Information and Declarations can be found on page 15

DOI 10.7717/peerj.6757

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 sites), adding a methyl group to the cytosine residue (*You & Jones, 2012*). In mammalian genomes, more than 70% of CpG sites are methylated (*Jabbari & Bernardi, 2004*). Many CpGs are clustered into CpG islands and more than 30,000 CpG islands have been identified in the human genome, most of which are located in promoter region and are hypo-methylated (*Jeziorska et al., 2017*). The level of DNA methylation at a CpG site is often correlated with that of neighbouring CpG sites and influenced by other genome features, such as genome position and regulatory elements. When combined, these genome features can effectively predict methylation level of CpG sites in the genome (*Zheng et al., 2017*).

The regulatory role of DNA methylation on gene expression has traditionally been 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 human genome-wide methylation and expression studies in non-cancer tissues have small sample sizes for comparing controlled groups. Only a limited number of studies profiled both genome-wide DNA methylation and gene expression in larger human populations and examined their relationship across individuals. *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. They found negative correlations between methylation and gene expression across individuals in a large proportion of differentially expressed and differentially methylated genes, but they also uncovered substantial positive correlations. 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. (2014)* profiled the genome wide DNA methylation and gene expression in forearm skin fibroblast among 62 unrelated individuals. They observed that the association between gene expression and methylation is not always negative in promoter region or positive in gene body (*Yang et al., 2014*).

The complex relationship among DNA methylation, gene expression, and genetic variants in human populations has also attracted substantial research attention. *Bell et al. (2011)* investigated the genetic controls for both methylation QTL (mQTL) and expression QTL (eQTL) using 77 human lymphoblastoid cell lines (LCLs) from the HapMap collection. They identified hundreds of mQTLs and eQTLs and showed that these two types of QTLs overlap significantly. *Gutierrez-Arcelus et al. (2013)* further examined the relationship among genetic variants, DNA methylation, and gene expression in three cell types of umbilical cord samples from 204 newborn babies and found that the relationship between DNA methylation and gene expression across individuals has a different process from that across genes with in a genome. The inter-individual relationship is much less clear in terms of negative regulation. Both active and passive roles are played by DNA methylation in regulating gene expression.

Unlike genome-wide DNA methylation, the inter-individual relationship between genetic variants and gene expression in human populations has been well-studied in both eQTL identification (*Deelen et al., 2015*) and gene expression prediction (*Xie et al., 2017*; *Zeng, Zhou & Huang, 2017*). Predicted gene expression is also used as an instrument in genome wide association studies to reduce multiple testing and identify associated genes (*Gamazon et al., 2015*). Similar studies in DNA methylation is lacking since 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 when all CpGs are considered together to serve as a surrogate for gene expression or enable gene expression to be an instrument in genome wide methylation studies in human populations. In this study, we examine the DNA methylation and gene expression relationship in three large human datasets. We determine the overall relationship between DNA methylation and gene expression across individuals for each gene and evaluate the predictive potential of DNA methylation data for gene expression. We also demonstrate that a penalized regression improves the overall prediction.

METHODS

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 25,160 genes using Illumina HumanHT-12 v3.0 on 825 individuals. The log2-transformed signals were quantile normalized for each tissue followed by quantile normalized across the whole population (*Grundberg et al., 2012*). The DNA methylation data (accession number E-MTAB-1866) were generated using Illumina Infinium Human Methylation 450 from 649 female twins. The methylation beta values were already quantile normalized for each type of probe, ranging from 0 (unmethylated) to 1 (total-methylated).

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 participants were 6 to 12 years old from African American, Dominican-Hispanic and Haitian-Hispanic background (*Yang et al., 2015*). DNA methylation data were generated using Illumina's Infinium Human Methylation450k BeadChip. The normalized methylation M value matrix was downloaded from ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE40nnn/GSE40576/matrix/. Gene expression data were generated for 23,612 genes using Nimblegen Human Gene Expression arrays (12 × 135 k). The normalized data matrix was downloaded from ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE40nnn/GSE40732/matrix/.

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.

LCL dataset

The LCL dataset was generated from Lymphoblastoid cell lines (LCL) of 280 healthy individuals (96 Han Chinese-American, 96 Caucasian-American and 95 African-American). Data were downloaded from GSE23120 and GSE36369. Gene expression microarray data were generated using Affymetrix Human Genome U133 Plus 2.0 Array, which contains 38,500 well-characterized human genes covered by 54,000 probe sets (https://www.affymetrix.com/support/technical/datasheets/human_datasheet.pdf) DNA methylation data were generated using Infinium HumanMethylation450 BeadChip platform. Quantile normalized M values were used in the analyses.

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 Grundberg's (*2012*) paper. The probes on the methylation array but excluded from 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 of the respective gene. Among the 485,679 probes in the dataset, 344,201 probes remained in Adipose dataset after filtering. For the PBMC dataset, 344,180 out of the 485,461 probes in the dataset remained after filtering and 344,202 out of 485,578 probes remained for the LCL dataset. In order to make the method comparable and the analysis consistent, only genes that have the LASSO models were used, 8040 genes and 149,152 CpG sites in Adipose dataset, 4,252 genes and 73,553 CpG sites in PBMC dataset and 7514 genes and 143,599 CpG sites in LCL dataset (Table S1).

Modelling 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 and LCL 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 with default parameters, 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.

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.

Model comparisons on significant genes

We first identified genes that showed overall model prediction *p* values less than 0.0001 in multiple regressions and then compared the three regression models on these genes.

Gene Ontology (GO) and pathway enrichment analysis

For top 2,000 genes with highest R^2, 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.

Gene expression prediction using different type of probes on the methylation microarray

The probes excluded by Table S1 of Grundberg's (*2012*) 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.

Analysis codes

We wrapped up our major analysis codes into a package at https://github.com/dorothyzh/MethylXcan. It includes all three regressions and calculates the squared correlation for each model. The package is written in R and Perl, and has been tested under linux or MACSOX system. Users can use this package on the datasets described here or on their own data after formatting their methylation data, expression profiling data, and annotation files as specified by the package.

RESULTS

Association between single CpG methylation and gene expression is often conducted in human populations when both transcriptome and methylome are profiled. 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 three human datasets, an Adipose dataset generated from subcutaneous fat tissue, a PBMC dataset from Childhood Asthma study, and a lymphoblastoid cell line (LCL) dataset. 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. Squared correlation (R^2) was 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 (S&C probes). In addition, since some genes fail to establish a LASSO model due to the lack of predictive information in DNA methylation, we only focus on genes with valid LASSO models for comparing different regression methods. In the three datasets, the total number of genes varies from 26,736 to 32,946 after quality control and normalization. About 1/6 to 1/3 of these genes have valid LASSO models with slightly bigger numbers when S&C probes are included (Table 1). In general, a large fraction of the genes with LASSO models have prediction R^2 greater than 0.1, but the number of genes quick reduces to hundreds and tens when R^2 increases to 0.2 and 0.3 (Table 1).

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

As a reference, we first conducted association analysis on each methylation probe in predicting gene expression using single regression. Most of the genes with valid LASSO models have at least one significant CpG at nominal significance level of 0.05. For example, in the Adipose dataset, 7,326 out of 8,040 genes have at least one CpG site significant at 0.05 level and 3460 out of 8040 genes have at least one CpG site significant at 0.0001. However, the prediction power represented by the largest R^2 in each gene is generally low. Only 19 genes have R^2 greater than 0.3 and 486 genes have R^2 larger than 0.1 when the most predictive CpG site is considered (Table 1). Similar results were obtained from the PBMC and LCL datasets, except that the PBMC dataset has a substantially smaller number of genes with a CpG significant at 0.0001 level (582 out of 4,252 genes) although the distribution of the estimated R^2 is similar to that from the Adipose dataset. This could be due to the smaller sample size or the nature of the PBMC tissue source. On the other hand, the LCL

Table 1 The number of genes with prediction R^2 larger than thresholds in single, multiple and LASSO regressions.

	Dataset	Regress model	Model fitting R^2			Cross validation R^2			Genes w/ LASSO model	All genes
			>0.1	>0.2	>0.3	>0.1	>0.2	>0.3		
Methylation probes	Adipose	Single	486	87	19	106	16	2	8,040	26,736
		Multiple	2,178	476	116	722	166	38		
		LASSO	1,702	360	113	827	179	42		
	PBMC	Single	851	108	14	381	33	4	4,252	31,030
		Multiple	3,358	1,163	382	746	109	30		
		LASSO	2,382	561	142	1,022	165	30		
	LCL	Single	1,753	465	126	419	82	21	7,514	32,946
		Multiple	5,138	2,170	975	1,663	575	185		
		LASSO	4,246	1,740	805	2,030	751	258		
All probes	Adipose	Single	591	115	33	103	21	5	8,864	26,736
		Multiple	3,037	760	211	898	226	64		
		LASSO	2,283	536	178	1,008	259	76		
	PBMC	Single	1,330	212	58	666	90	34	5,064	31,030
		Multiple	4,455	1,902	694	994	197	64		
		LASSO	3,207	870	235	1,465	289	66		
	LCL	Single	1,888	533	155	425	88	32	7,498	32,946
		Multiple	5,870	2,573	1,267	1,757	627	221		
		LASSO	4,646	2,029	999	2,155	840	335		

Notes.

All genes, the total number of genes in a dataset after quality control and normalization; Genes w/LASSO model, the number of genes with valid LASSO models; All probes, the combination of methylation probes and probes with cross-hybridization/SNP effects.

dataset has a larger number of genes with higher R^2 from single CpG regression analysis, which could be related to the homogeneous nature of cell lines.

Since multiple CpG sites from each gene were assayed on the methylation microarray, we applied multiple linear regression to utilize all methylation CpG sites in the gene region as predictors simultaneously. The R^2 explained by the regression model did improve substantially for the majority of genes compared with that from the single linear regression (Fig. 1). As expected, the significant genes from multiple regression analyses tend to have larger R^2 compared with the non-significant genes. The improvement of R^2 from the multiple regression over single regress in the PBMC and LCL datasets is similar to that in the Adipose dataset (Fig. 1).

Compared with multiple regressions, LASSO regression did not generate R^2 quite as high in all datasets (Figs. 1D–1F), which is also indicated by smaller number of genes with R^2 exceeding each threshold (Table 1). For example, the number of genes with R^2 greater than 0.2 decreased from 476 to 360 for the Adipose dataset, from 1,163 to 561 for the PBMC datasets, and from 2,170 to 1,704 for the LCL dataset. Similar trend was observed at the other two thresholds for all three datasets.

It is widely recognized that gene expression is negatively correlated with DNA methylation level in the promoter region but often positively correlated with DNA methylation level in gene body (*Ball et al., 2009*; *Wu et al., 2010*; *Jones, 2012*). The different directions of correlation among CpG sites in the same gene may lead to perceptions that

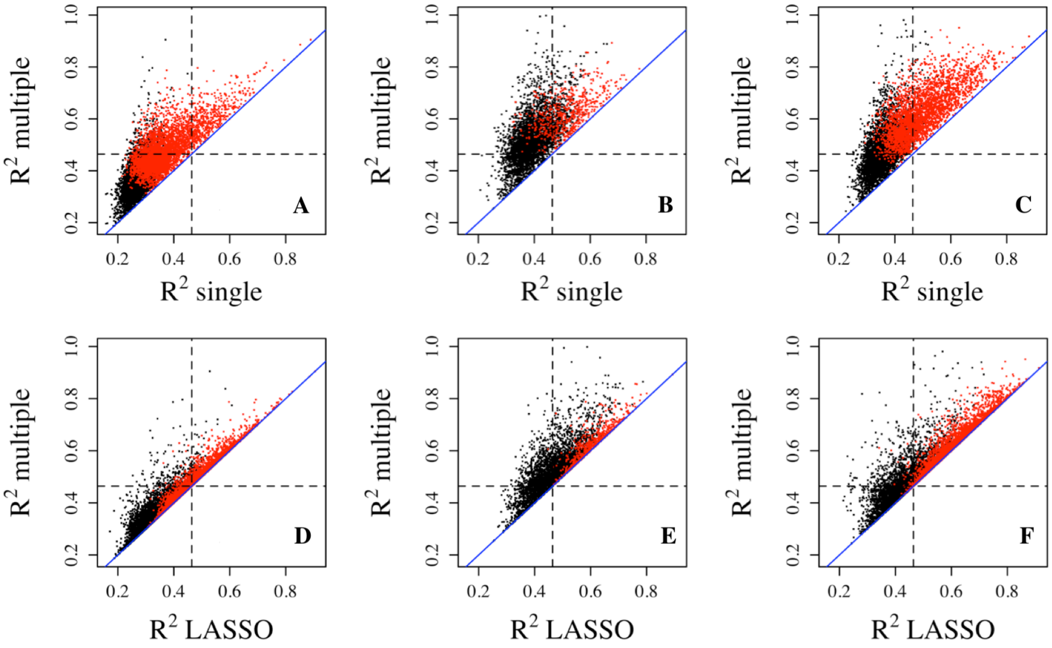

Figure 1 Goodness of fit R^2 comparison among three regression models. R^2 values from multiple regression are compared to those from single (top three panels) and lasso (bottom three panels) regressions in three datasets, Adipose (A, D), PBMC (B, E) and LCL (C, F). R^2 values shown here are on cubic 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 0.0001. Blue solid line is the identity line and the dashed lines represent R^2 of 0.1.

Full-size DOI: 10.7717/peerj.6757/fig-1

combining all CpG sites is not advantageous in the prediction of gene expression. However, the multiple and lasso regressions can accommodate coefficients in different directions without affecting prediction power. Nevertheless, we tested CpG sites from promoter region and those from gene body for prediction separately in the LCL dataset. As expected, neither performs as well as combined (Fig. S1).

LASSO regression shows better 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 prediction R^2. The results showed that the LASSO regression produced much larger R^2 values than the single regression and less dramatic but discernible increases over multiple regressions (Fig. 2). These differences are also reflected in the number of genes with R^2 exceeding the three thresholds. For example, 827 genes (10.29%) from the Adipose dataset have R^2 greater than 0.1 from LASSO regression, while 722 genes (8.98%) and 106 genes (1.32%) have R^2 greater than 0.1 from multiple regression and single regression, respectively (Table 1). For genes with R^2 greater than 0.3, LASSO regression has 42 genes (0.52%) while multiple regression and single regression have 38 (0.47%) and 2 genes (0.02%), respectively. These results indicate that penalized regression has better prediction than multiple or single regressions in cross-validation. Cross validation tends to overcome bias and over-fitting

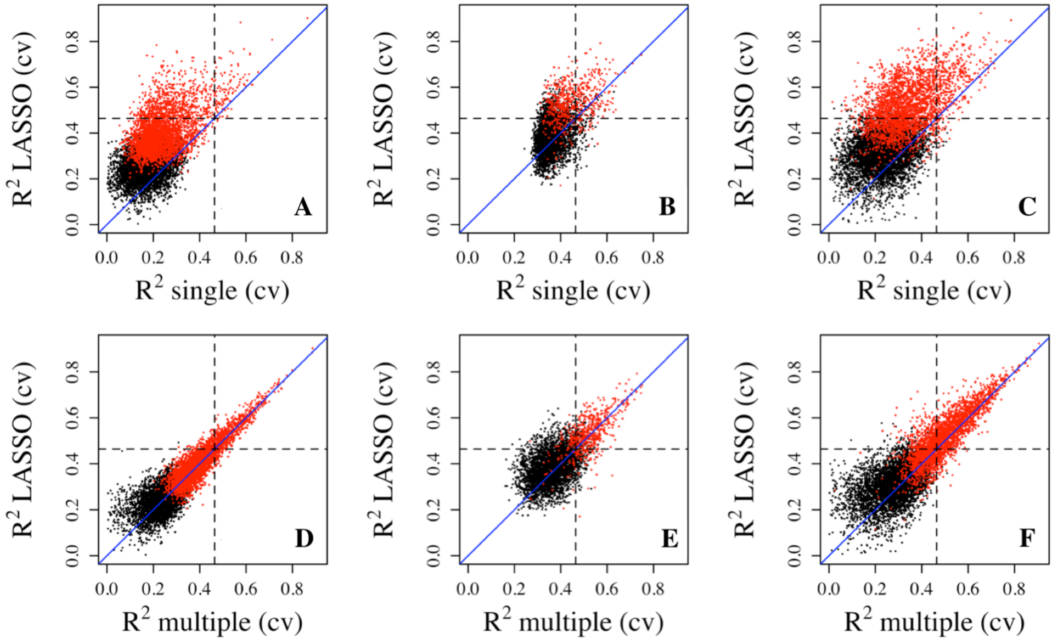

Figure 2 Prediction R^2 comparison among regression models in cross validation. Cross-validation R^2 from LASSO regression are compared to those from single regression (top three panels) and multiple regression (bottom three panels) for three datasets, Adipose (A, D), PBMC (B, E) and LCL (C, F). Five-fold validation was used for all regression models. R^2 shown here are on cubic root scale for visualization clarity. The red points represent the significant genes from multiple regressions ($p < 0.0001$). Blue solid line is the identity line and the dashed lines represent R^2 of 0.1. single.cv, cross-validation R^2 of single regression; multiple.cv, cross-validation R^2 of multiple regression; cross-validation R^2 of LASSO regression.

Full-size DOI: 10.7717/peerj.6757/fig-2

issues. As expected, cross-validation R^2 values are generally lower than those from the model fittings, which is reflected by the smaller number of genes with R^2 values greater than the R^2 thresholds (Table 1). Similar results were obtained from the PBMC dataset and the LCL dataset.

To make sure that the prediction R^2 is larger than thoses from random chance, we compared the cumulative R^2 from the three datasets with those from the the null distribution of correlations based on Fisher z-transfromation in quantile–quantile plots (Fig. 3). All datasets showed that the observed R^2 values are much larger than the expected R^2 values from random chance. In addition, the departure is the largest in the LCL dataset followed by the Adipose dataset and the PBMC dataset when methylation probes were considered, indicating that the LASSO models capture a larger proportion of the transcriptome variability in the LCL dataset than in the other two datasets. This is potentially due to the combination of sample size and nature of different tissues.

To rule out the possibility that prediction R^2 is mainly driven by the variability of gene expression and the variability of DNA methylation across individuals in the study population, we first examined the correlation between the variability of gene expression with R^2 from LASSO regression. No obvious correlation was observed (Fig. S2). For assessing the correlation between DNA methylation variability and prediction R^2, we took

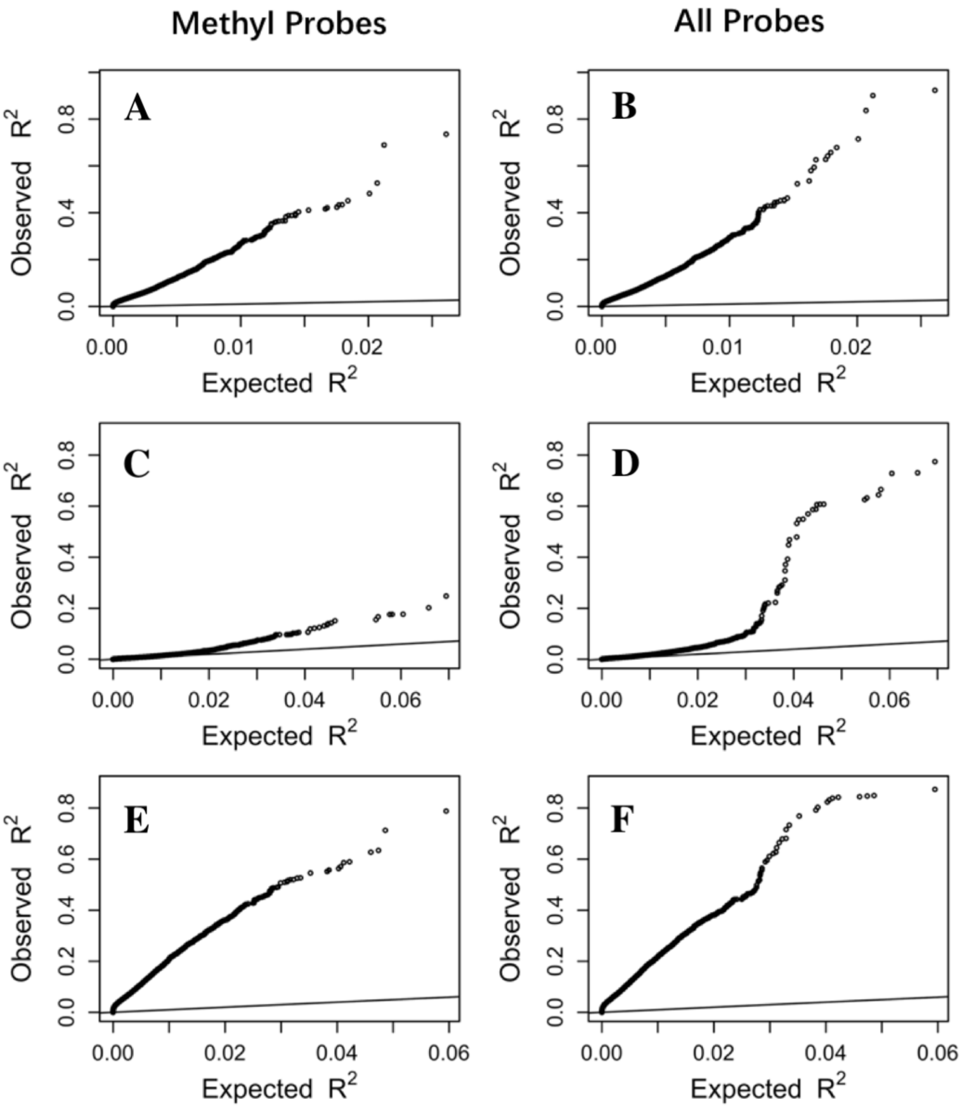

Figure 3 The prediction R^2 is beyond random chance. Sorted R^2 values from three datasets, Adipose (A, B), PBMC (C, D) and LCL (E, F), are compared with those from the null distribution of R^2 based on Fisher z-transformation (straight line). (A, C, E) are from methylation probes, after excluding probes that have cross-hybridization or SNP effects. (B, D, F) are from all probes, the combination of methylation probes and probes with cross-hybridization/SNP effects. Five-fold cross validation was used for LASSO regression models.

Full-size DOI: 10.7717/peerj.6757/fig-3

the CpG with maximum R^2 from single regression and examined the correlation between its variability with the prediction R^2 from LASSO regression. We only observed a potential positive correlation in the Adipose dataset when the R^2 is greater than 0.5, where there are a small number of genes (Fig. S3).

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

(Table S1). 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 (Fig. 3). 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 that from only the methylation probes. For example, the largest R^2 is 0.92 from all probes compared to 0.74 from only methylation probes in the Adipose dataset. Similarly, the largest R^2 increases from 0.71 to 0.88 in the PBMC dataset and from 0.76 to 0.87 in the LCL dataset. Furthermore, valid LASSO models are available for more genes when all probes are used (Fig. S4).

The increase of prediction power on gene expression from all types of probes on the methylation microarray suggests that there is contribution from the probes with potential SNP effects or cross hybridization effects. To further assess 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 Figs. 4A, 4C and 4E). The prediction power from the S&C probes was also estimated for genes with enough SNP probes to form a LASSO model and their prediction power are mostly above zero (blue points in Fig. 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 genetic effect and 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 tend to show continuous methylation values while the S&C probes tend to show categorical values due to limited genotypes of the samples.

GO term analysis of better predicted genes

To examine the potential biological function of the genes showing relatively higher predictability, we conducted gene ontology (GO) enrichment analysis using DAVID on genes with R^2 larger than 0.2 from LASSO regression of methylation CpGs. At false discovery rate (FDR) of 0.01, cell adhesion, lipids metabolism, and regulation of immune system are among the most significantly enriched terms in the Adipose dataset (Table S2), which seem to be consistent with the previous findings for subcutaneous fat cells (*Berg & Scherer, 2005*). For the PBMC dataset, the most significantly enriched terms are mostly related to defense and immune functions, lymphocyte aggregation, T cell activation, inflammatory response, as well as cell adhesion. These results appear to be reasonable for atopy and persistent asthma blood cells. For the LCL data, some terms related to cell adhesion, migration, communication, and morphogenesis are highly enriched. Same GO term analyses were also conducted for R^2 from all CpG probes and similar results were obtained.

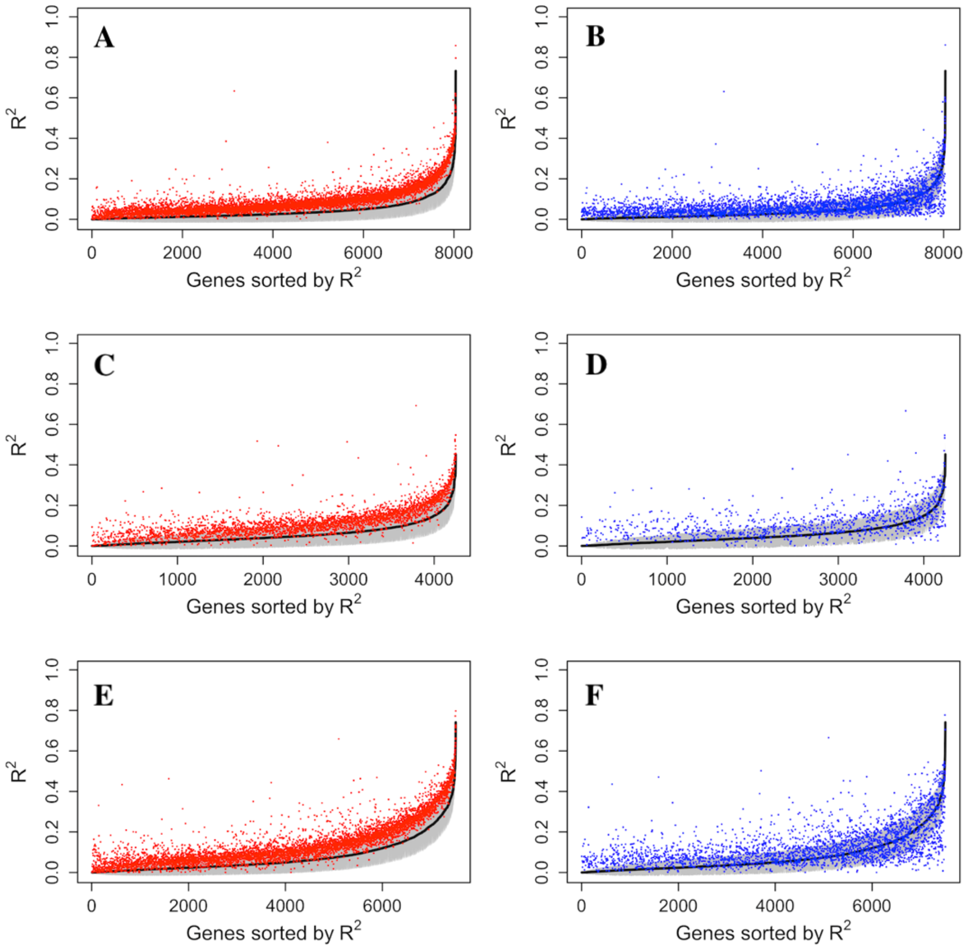

Figure 4 **Comparison of R^2 from using methylation probes, S&C probes and all probes.** LASSO regression R^2 values from three datasets, Adipose (A, B), PBMC (C, D) and LCL (E, F), were generated from methylation probes (black line), S&C probes (probes with cross-hybridization/SNP effects) (blue points), and all probes (the combination of methylation probes and probes with cross-hybridization/SNP effects) (red points). The 95% confidence interval of R^2 from methylation probes is shown as a grey shadow.

Full-size DOI: 10.7717/peerj.6757/fig-4

DISCUSSION

We examined the relationship between gene expression and DNA methylation across the genome using data from three 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. We also demonstrated that using all probes on the methylation array does improve prediction power to some degree.

Three types of regression models were examined in our study for their prediction power evaluated by squared correlation (R^2). The single linear regression is based only on the best predictive CpG in each gene, therefore, has least prediction power. The

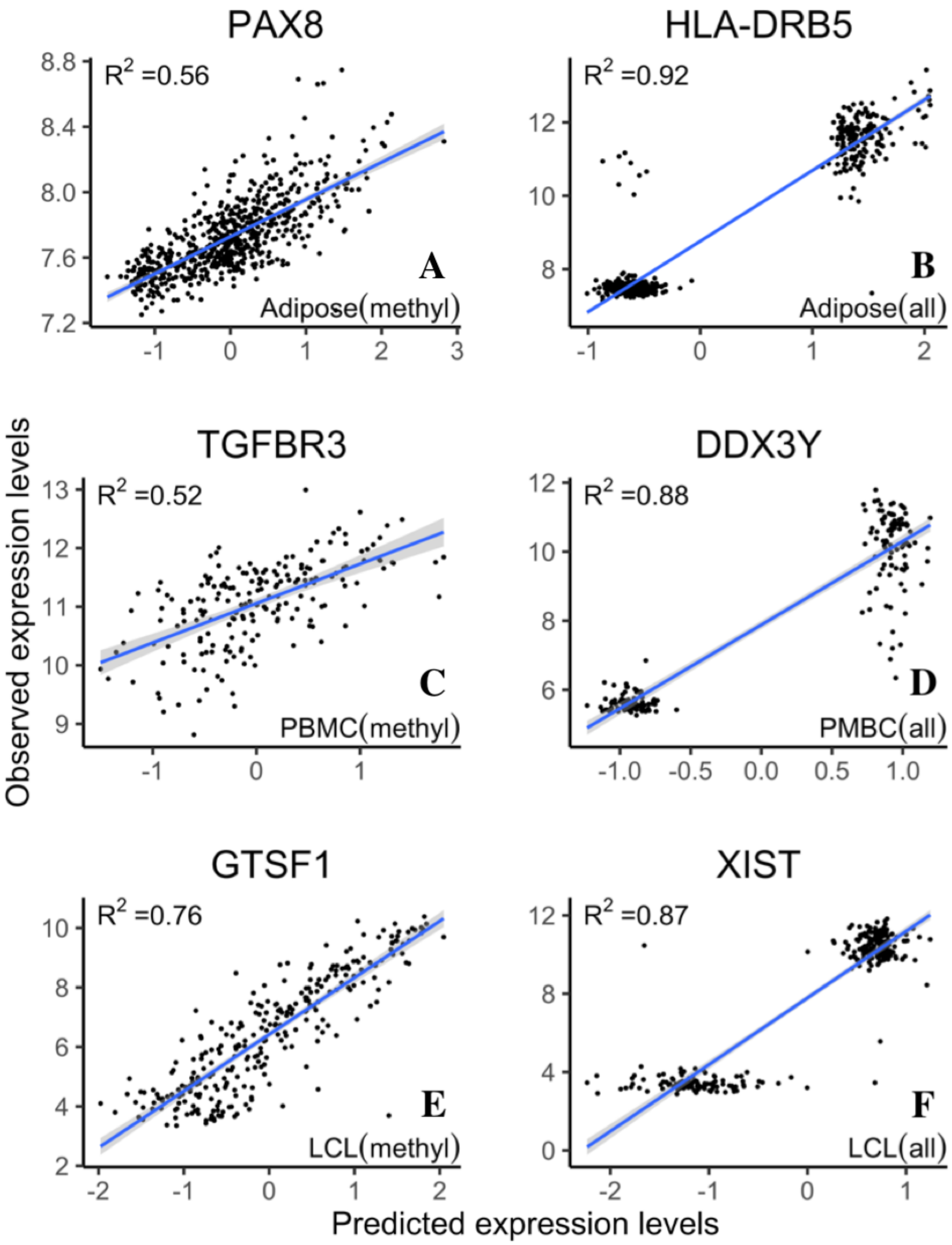

Figure 5 Example genes with high prediction power. R^2 is from LASSO regression models. Adipose, PBMC and LCL are the three datasets. *X*-axis indicates the predicted expression levels from LASSO regression models; *y*-axis indicates observed expression levels for each dataset (methyl, methylation probes; all, all probes).

Full-size DOI: 10.7717/peerj.6757/fig-5

multiple regression has increased power when all CpG sites in each gene are included as predictors; however, it has substantial over-fitting problem for genes with large number of CpGs. The shrinkage based LASSO regression 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 avoiding model overfit as well as variable selection and model interpretability. In this study, not all expressed genes have LASSO models because LASSO fails to select informative predictors in some genes even with minimum penalization, which indicates that no predictive information exists in the DNA methylation data at these genes. LASSO is not the only shrinkage-based regression method. There are other penalty regression models, such as the Ridge (*Hoerl & Kennard, 1970*), elastic net (*Zou & Hastie, 2005*), elastic net with rescaled-coefficients and grouped lasso (*Yuan & Lin, 2006*; *Meier, Van De Geer & Bühlmann, 2008*). Further evaluation is needed for their merits in improving prediction in this setting.

The prediction power from DNA methylation in our analysis seems to be much lower than that from DNA sequence variants evaluated in different human tissues (*Gamazon et al., 2015*). One potential reason for relative low prediction power we observed from DNA methylation 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*). Some more comprehensive tools, such as FEM (*Jiao, Widschwendter & Teschendorff, 2014*) and ROADMAP (*Kundaje et al., 2015*), may help integrate the influences of the other factors on gene expression. 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 variabilities; therefore, it is not surprising to see lower prediction power in the PBMC and adipose datasets compared to the LCL dataset. The potential of DNA methylation alone as surrogate for gene expression is likely to be limited in general, especially in the tissues with mixed cell types, such as PBMC, which is used widely in human population studies. The combination of DNA methylation and genotype, should be more powerful for this purpose, as indicated by the increased prediction power when SNP-containing probes were included in the prediction models (Figs. 3 and 4). This can be a promising future direction.

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. All three datasets we analysed show relatively low prediction power. The better predictive genes are dataset specific and their function varies in different tissues or cell types. 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
LCL	lymphoblastoid cell
MDS	myelodysplastic syndrome
GO	gene ontology

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Degui Zhi was partially supported by NIH Grant R01 HG008115; Xiangqin Cui was partially supported by NIH 2P60AR048095. Huan Zhong was supported by Hong Kong Baptist University's strategic development fund SDF15-1012-P04 to Yiji Xia. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
NIH: R01 HG008115, 2P60AR048095.
Hong Kong Baptist University's strategic development: SDF15-1012-P04.

Competing Interests

Degui Zhi and Xiangqin Cui are Academic Editors for PeerJ. The authors declare there are no competing interests.

Author Contributions

- Huan Zhong analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Soyeon Kim analyzed the data, authored or reviewed drafts of the paper, approved the final draft.
- Degui Zhi and Xiangqin Cui conceived and designed the experiments, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

For the Adipose Dataset, the gene expression data is available at E-TABM-1140; and the DNA methylation data can be found at E-MTAB-1866. For the PBMC Dataset, gene expression data is available at GSE40732; and the DNA methylation data is available at GSE40576. For the LCL Dataset, gene expression data is available at GSE23120; and the DNA methylation data is available at GSE36369. The software package 'MethylXcan' is available at https://github.com/dorothyzh/MethylXcan.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.6757#supplemental-information.

REFERENCES

Ball MP, Li JB, Gao Y, Lee J-H, LeProust EM, Park I-H, Xie B, Daley GQ, Church GM. 2009. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnology* **27**:361–368 DOI 10.1038/nbt.1533.

Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology* **12**:R10 DOI 10.1186/gb-2011-12-1-r10.

Berg AH, Scherer PE. 2005. Adipose tissue, inflammation, and cardiovascular disease. *Circulation Research* **96**:939–949.

Deelen P, Zhernakova DV, De Haan M, Van der Sijde M, Bonder MJ, Karjalainen J, Van der Velde KJ, Abbott KM, Fu J, Wijmenga C, Sinke RJ, Swertz MA, Franke L. 2015. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Medicine* **7**:30 DOI 10.1186/s13073-015-0152-4.

Del Rey M, O'Hagan K, Dellett M, Aibar S, Colyer HAA, Alonso ME, Diez-Campelo M, Armstrong RN, Sharpe DJ, Gutierrez NC. 2013. Genome-wide profiling of methylation identifies novel targets with aberrant hypermethylation and reduced expression in low-risk myelodysplastic syndromes. *Leukemia* **27**:610–618 DOI 10.1038/leu.2012.253.

Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL, Cox NJ. 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**:1091–1098 DOI 10.1038/ng.3367.

Grundberg E, Meduri E, Sandling JK, Hedman ÅK, Keildson S, Buil A, Busche S, Yuan W, Nisbet J, Sekowska M. 2013. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *The American Journal of Human Genetics* **93**:876–890 DOI 10.1016/j.ajhg.2013.10.004.

Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, Bell JT, Yang T-P, Meduri E, Barrett A. 2012. Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature Genetics* **44**:1084–1089 DOI 10.1038/ng.2394.

Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, Falconnet E, Bielser D, Gagnebin M, Padioleau I, Borel C, Letourneau A, Makrythanasis P, Guipponi M, Gehrig C, Antonarakis SE, Dermitzakis ET. 2013. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2013**:1–18 DOI 10.7554/eLife.00523.

Hoerl AE, Kennard RW. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**:55–67 DOI 10.1080/00401706.1970.10488634.

Huang DW, Sherman BT, Lempicki RA. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**:44–57.

Jabbari K, Bernardi G. 2004. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* **333**:143–149 DOI 10.1016/j.gene.2004.02.043.

Janowski BA, Huffman KE, Schwartz JC, Ram R, Hardy D, Shames DS, Minna JD, Corey DR. 2005. Inhibiting gene expression at transcription start sites in chromosomal DNA with antigene RNAs. *Nature Chemical Biology* **1**:216–222 DOI 10.1038/nchembio725.

Jeziorska DM, Murray RJS, De Gobbi M, Gaentzsch R, Garrick D, Ayyub H, Chen T, Li E, Telenius J, Lynch M, Graham B, Smith AJH, Lund JN, Hughes JR, Higgs DR, Tufarelli C. 2017. DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proceedings of the National Academy of Sciences of the United States of America* **114**:E7526–E7535 DOI 10.1073/pnas.1703087114.

Jiao Y, Widschwendter M, Teschendorff AE. 2014. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics* **30**:2360–2366 DOI 10.1093/bioinformatics/btu316.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics* **13**:484–492 DOI 10.1038/nrg3230.

Jones B. 2015. Gene expression: layers of gene regulation. *Nature Reviews Genetics* **16**:128–129 DOI 10.1038/nrg3918.

Kaikkonen MU, Lam MTY, Glass CK. 2011. Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovascular Research* **90**:430–440 DOI 10.1093/cvr/cvr097.

Krueger F, Kreck B, Franke A, Andrews SR. 2012. DNA methylome analysis using short bisulfite sequencing data. *Nature Methods* **9**:145–151 DOI 10.1038/nmeth.1828.

**Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning A, Wang X, ClaussnitzerYaping Liu M, Coarfa C, Alan Harris R, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, David Hawkins R, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Scott Hansen R, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Abdennur N, Adli M, Akerman M, Barrera L, Antosiewicz-Bourget J, Ballinger T, Barnes MJ, Bates D, Bell RJA, Bennett DA, Bianco K, Bock C, Boyle P, Brinchmann J, Caballero-Campo P, Camahort R, Carrasco-Alfonso MJ, Charnecki T, Chen H, Chen Z, Cheng JB, Cho S, Chu A, Chung W-Y, Cowan C, Athena Deng Q, Deshpande V, Diegel M,

Ding B, Durham T, Echipare L, Edsall L, Flowers D, Genbacev-Krtolica O, Gifford C, Gillespie S, Giste E, Glass IA, Gnirke A, Gormley M, Gu H, Gu J, Hafler DA, Hangauer MJ, Hariharan M, Hatan M, Haugen E, He Y, Heimfeld S, Herlofsen S, Hou Z, Humbert R, Issner R, Jackson AR, Jia H, Jiang P, Johnson AK, Kadlecek T, Kamoh B, Kapidzic M, Kent J, Kim A, Kleinewietfeld M, Klugman S, Krishnan J, Kuan S, Kutyavin T, Lee A-Y, Lee K, Li J, Li N, Li Y, Ligon KL, Lin S, Lin Y, Liu J, Liu Y, Luckey CJ, Ma YP, Maire C, Marson A, Mattick JS, Mayo M, McMaster M, Metsky H, Mikkelsen T, Miller D, Miri M, Mukame E, Nagarajan RP, Neri F, Nery J, Nguyen T, O'Geen H, Paithankar S, Papayannopoulou T, Pelizzola M, Plettner P, Propson NE, Raghuraman S, Raney BJ, Raubitschek A, Reynolds AP, Richards H, Riehle K, Rinaudo P, Robinson JF, Rockweiler NB, Rosen E, Rynes E, Schein J, Sears R, Sejnowski T, Shafer A, Shen L, Shoemaker R, Sigaroudinia M, Slukvin I, Stehling-Sun S, Stewart R, Subramanian SL, Suknuntha K, Swanson S, Tian S, Tilden H, Tsai L, Urich M, Vaughn I, Vierstra J, Vong S, Wagner U, Wang H, Wang T, Wang Y, Weiss A, Whitton H, Wildberg A, Witt H, Won K-J, Xie M, Xing X, Xu I, Xuan Z, Ye Z, Yen C, Yu P, Zhang X, Zhang X, Zhao J, Zhou Y, Zhu J, Zhu Y, Ziegler S, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L-H, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L-H, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**:317–330 DOI 10.1038/nature14248.

Lokk K, Modhukur V, Rajashekar B, Märtens K, Mägi R, Kolde R, Koltšina M, Nilsson TK, Vilo J, Salumets A. 2014. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biology* **15**:3248 DOI 10.1186/gb-2014-15-4-r54.

Meier L, Van De Geer S, Bühlmann P. 2008. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**:53–71 DOI 10.1111/j.1467-9868.2007.00627.x.

Razin A, Riggs AD. 1980. DNA methylation and gene function. *Science* **210(4470)**:604–610 DOI 10.1126/science.6254144.

Song M-A, Brasky TM, Marian C, Weng D, Taslim C, Dumitrescu RG, Llanos AA, Freudenheim JL, Shields PG. 2015. Racial differences in genome-wide methylation profiling and gene expression in breast tissues from healthy women. *Epigenetics* **10(12)**:1177–1187 DOI 10.1080/15592294.2015.1121362.

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 Research* **20**:440–446 DOI 10.1101/gr.103606.109.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**:267–288.

Ting AH, McGarvey KM, Baylin SB. 2006. The cancer epigenome—components and functional correlates. *Genes and Development* **20**:3215–3231 DOI 10.1101/gad.1464906.

Ting AH, Schuebel KE, Herman JG, Baylin SB. 2005. Short double-stranded RNA induces transcriptional gene silencing in human cancer cells in the absence of DNA methylation. *Nature Genetics* **37**:906–910 DOI 10.1038/ng1611.

Verdin E, Ott M. 2015. 50 years of protein acetylation: from gene regulation to epigenetics, metabolism and beyond. *Nature Reviews Molecular Cell Biology* **16**:258–264 DOI 10.1038/nrm3931.

Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. 2014. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology* **15**:R37 DOI 10.1186/gb-2014-15-2-r37.

Wu H, Coskun V, Tao J, Xie W, Ge W, Yoshikawa K, Li E, Zhang Y, Sun YE. 2010. Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science* **329(5990)**:444–447 DOI 10.1126/science.1190485.

Xie R, Wen J, Quitadamo A, Cheng J, Shi X. 2017. A deep auto-encoder model for gene expression prediction. *BMC Genomics* **18(Suppl 9)**:845 DOI 10.1186/s12864-017-4226-0.

Yang IV, Pedersen BS, Liu A, O'Connor GT, Teach SJ, Kattan M, Misiak RT, Gruchalla R, Steinbach SF, Szefler SJ. 2015. DNA methylation and childhood asthma in the inner city. *Journal of Allergy and Clinical Immunology* **136**:69–80 DOI 10.1016/j.jaci.2015.01.025.

Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. 2014. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**:577–590 DOI 10.1016/j.ccr.2014.07.028.

You JS, Jones PA. 2012. Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell* **22**:9–20 DOI 10.1016/j.ccr.2012.06.008.

Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**:49–67 DOI 10.1111/j.1467-9868.2005.00532.x.

Zeng P, Zhou X, Huang S. 2017. Prediction of gene expression with cis-SNPs using mixed models and regularization methods. *BMC Genomics* **18**:368 DOI 10.1186/s12864-017-3759-6.

Zheng Y, Joyce BT, Liu L, Zhang Z, Kibbe WA, Zhang W, Hou L. 2017. Prediction of genome-wide DNA methylation in repetitive elements. *Nucleic Acids Research* **45**:8697–8711 DOI 10.1093/nar/gkx587.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**:301–320 DOI 10.1111/j.1467-9868.2005.00503.x.