Orthogonalization: A Computational Approach to Determining Methylation and Cell Composition Effects

Kevin Luk

Background Information Many diseases are a consequence of alterations in gene expression patterns. They are often caused by DNA sequence mutations that retard cell function. However, it has recently become apparent that in addition to DNA sequence mutations, epigenetic mutations can an also lead to changes in gene expression. Consequently, the mechanisms and effects of these epigenetic mutations have become of interest. DNA methylation is an example of one of these epigenetic mutations. The frequency and locations of DNA methylation varies between individuals. While it is believed that genetic inheritance plays a role in DNA methylation, much of the variation between individuals is still unattributed.

Objective Using dynamic programing, I sought to (1) derive the patient's cell composition profile from individual patient methylation data, (2) optimize this approach, and (3) evaluate its effectiveness.

Design, Setting, and Participants Using the Illumina 450k assay, CpG methylation levels were measured across 36690 sites in 6 male patient's (age 38 ± 13.6 years) blood cells. From this data we simulated 100 patients' blood methylation data. The novel cell composition approach, Deconvolution, was used to attain the methylation levels of a given pure cell type and a patient's cell composition profile. I developed a program that would optimize Deconvolution's solution called Orthogonalization. Since the Illumina 450k measures methylation most frequently at two distinct levels, I evaluated the quality of Deconvolution's solutions by calculating the sum of the minimum difference between the Deconvolution's methylation levels and the Illumina 450k's most frequently read methylation levels. This was used as a quantitative metric to score the quality of the results. My dynamic algorithm minimized this score by repeatedly taking the Deconvolution's solution and randomly perturbing its values until a smaller sum was achieved. After, the effectiveness of my program was evaluated using the Wilcoxon signed-rank test and linear regressions against the simulated data. The optimization of Orthogonalization was conducted empirically by running Orthogonalization with varying base parameters.

Results The coefficient and weight scores decreased on average after running the Orthogonalization algorithm on Deconvolution solutions (p=3.91 x 10^{-3} and 1.95 x 10^{-3}, respectively). The predicted cell proportions fit closely against the simulated cell proportions (r^2 = 0.74, 0.85, 0.81, for CD4, CD8, CD14, respectively). Orthogonalization gives relatively consistent results with a standard deviation of 2.2% and 1.5% for coefficient and weight scores, respectively. The optimum values for the perturbation factor and illegal weight and coefficient percentage thresholds were 10%, 10%, and 1, respectively.

Discussion Orthogonalization was able to reliably derive a patient's cell composition profile from simulated individual patient methylation data. Further research is needed to support the methylation composition model. More testing is needed on non-simulated data on other cell types to determine if the approach still remains effective. The Orthogonalization tool may prove evaluable in helping explain individual DNA methylation variation across populations and ultimately further our understanding of the epigenetic aetiology of diseases.

Background Information

Many diseases are a consequence of alterations in gene expression patterns[1]. With the completion of the Human Genome Project, we have a nearly complete list of human genes. However, simply having a catalogue of genes does not inform us on how mutations will alter gene expression. The situation is far more complex. There is a system that cells use to determine when and where a particular gene will be expressed during development. This system is overlaid on DNA in the form of epigenetic marks that are heritable during cell division but do not alter the DNA sequence. Consequently, these epigenetic alterations can also cause changes to gene expression leading to disease.

The only currently known epigenetic modification of DNA in mammals is methylation of cytosine at position C5 in CpG dinucleotides[2]. The frequency and locations of DNA methylation varies between individuals. Research into patterns of methylation variation across the genome is still in its infancy[2]. This research is of particular interest as DNA methylation is starting to be used as a quantitative biological marker to complex diseases. For example cancer cells can be identified by their hypomethylated genome relative to their normal counterparts[3]. As such being able to explain the variation between individuals will help generate hypothesises regarding the mechanisms and effects of DNA methylation in diseases and phenotypes of interest. Viewing a patient's methylation levels as a product of their cell composition profile and the methylation level of a given pure cell type, I sought to computationally derive these components using dynamic programming. I also sought to optimize and evaluate the effectiveness of this approach.

Methodology

Understanding Methylation Levels as a Composition

Each patient's methylation level at a given probe site was viewed as a product of the proportion of a given cell type and the methylation level at that given site of a cell that is composed only that cell type. This is defined below:

$$\beta_{ij} = \sum_{c=1}^{k} X_{ic}\, w_{jc}$$

$$X_{ic} \sim N(\mu_{ic\,,}\ \sigma_{ic}^{2})$$

Where:
β_{ij} = Measured methylation at site *i* in individual *j*
X_{ic} = Measured methylation for a cell of 100% of type *c*
w_{jc} = Proportion of cell type *c* in individual *j*
$\mu_{i,c}$ and $\sigma^{2}_{i,c}$ = Mean and variance of methylation at site *i* in cell type *c*

A novel cell composition approach called Deconvolution was used to attain the methylation levels of a pure cell type (X_{ic}) and a patient's cell composition (w_{jc}). Since each β_{ij} is a composition of X_{ic} and w_{jc}, this data can be represented as matrices at the population level as shown below:

$$P_{n\times k} = W_{n\times s} \cdot C_{s\times k}$$

Where:
W = weight matrix
C = coefficient matrix
P = product matrix
n = number of patients
s = number of sites where methylation is measured
k = number of cell types

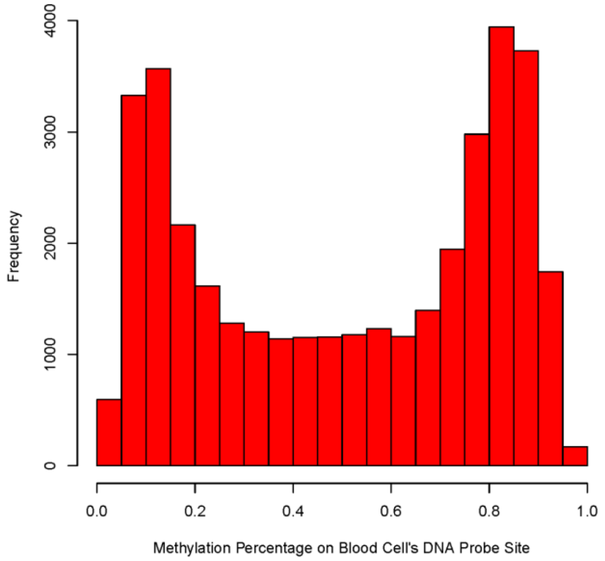
Developing Orthogonalization

Since the Illumina 450k measures methylation most frequently at two distinct levels, I quantitatively scored the quality of the Deconvolution's solutions by calculating the minimum difference between the Deconvolution's methylation levels and the most frequently read methylation levels. In this case the most commonly read levels are 0.1 and 0.9 (see figure 1). This is mathematically defined as:

$$\sum_{i=1}^{s}\sum_{l=1}^{k}\min(\left|C_{i,l}-0.9\right|,\left|C_{i,l}-0.1\right|)$$

To improve the Deconvolution's solutions I developed a dynamic algorithm, Orthogonalization, that minimized this difference by repeatedly taking the Deconvolution solutions and randomly perturbing its values until a smaller sum was achieved. The perturbed matrices were created by randomly generating transformation matrices, T_{kxk}, and multiplying them by *C*. The new *W* matrix was generated via the dot product of the *P* matrix and the inverse perturbed *C* matrix. Both the changed *W* and *C* matrices had all their values roughly restricted to 0 and 1. This process was repeated multiple times until the score was minimized. Orthogonalization's pseudocode is outlined in the provided in the supplementary information.

Figure 1: Histogram of measured Blood Methylation Levels on Illumina 450k



Creating Simulated Data for Testing

Blood methylation data was obtained from Reinus et al study. Using the Illumina 450k methylation bead assay, six healthy male blood donors' (age 38 ± 13.6 years) methylation levels were measured across 36690 sites in their blood cell's genome. I simulated 100 patients' blood methylation data by taking the average methylation value for each cell type across these 6 replicates and then randomly generating 100 patients' cell composition profiles. Each individual's methylation values at a given site was calculated as a weighted product of their cell composition profile and the average methylation value for that given cell type at that DNA probe site.

Testing the Effectiveness of Orthogonalization and Quality Control

In each performance and quality control test the sum of squares of the differences between the simulated data and the solutions were calculated. This score was used to quantitatively evaluate the quality of the results. For example for a given solution to C, the coefficient matrix, the score is calculated by

$$\sum_{i=1}^{s}\sum_{l=1}^{k}\left(C_{i,l}-C'_{i,l}\right)^{2}$$

Orthogonalization uses three parameters in its computations: a perturbation factor used in generating the random *T*, and the % of illegal (less than 0 or greater than 1) weights and coefficients allowed. All tests used a baseline of perturbation factor of 1 and 10% for illegal weights and coefficients allowed.

The effectiveness of Orthogonalization was evaluated by performing a Wilcoxon signed-rank test on 10 different trials of Deconvolution to determine if there was a statistically significant decrease in score after running Orthogonalization. A linear regression between the simulated data and the Orthogonalization's predicted data was also performed as a secondary way to determine its effectiveness. To test Orthogonalization's reliability, the program was run 10 times on the same set of data using different random seeds and the scores were compared. The optimum percentage of illegal $C_{i,l}$ and $W_{j,i}$ was identified empirically by running and graphing the scores. The perturbation factor used in generating the random T matrices was also identified in a similar fashion.

Results

	T+	P-Value
Coefficient Scores	54	3.91×10^{-3}
Weight Scores	55	1.95×10^{-3}

Figure 2: Wilcox Signed Rank Test to Evaluate Effectiveness of Orthogonalization after Deconvolution

	Average	Std. Deviation	% Std Dev
Weight Score	0.402	0.074	1.5
Coefficient Score	125.8	19.6	2.2

Figure 3: Variability due to Randomness Between 10 Identical Trials of Orthogonalization on Same Set of Data

Figure 4: Weight and Coefficient Scores at Various Perturbation Factor Levels

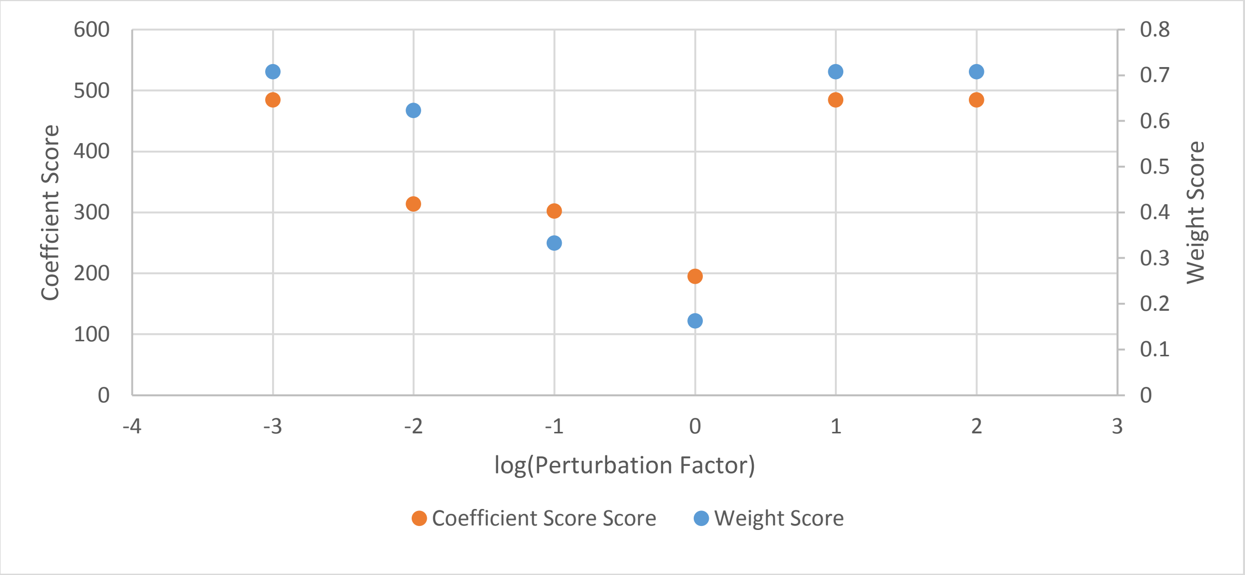


Figure 5: Weight Scores at Various Illegal Weight Thresholds

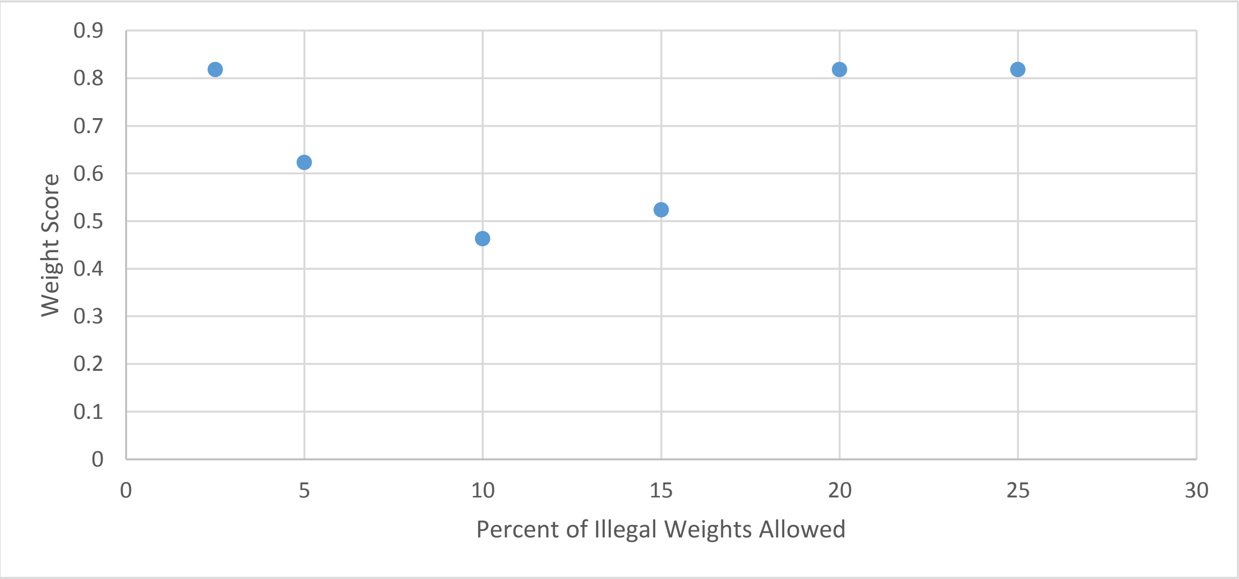


Figure 6: Coefficient Score at Various Illegal Coefficient Thresholds

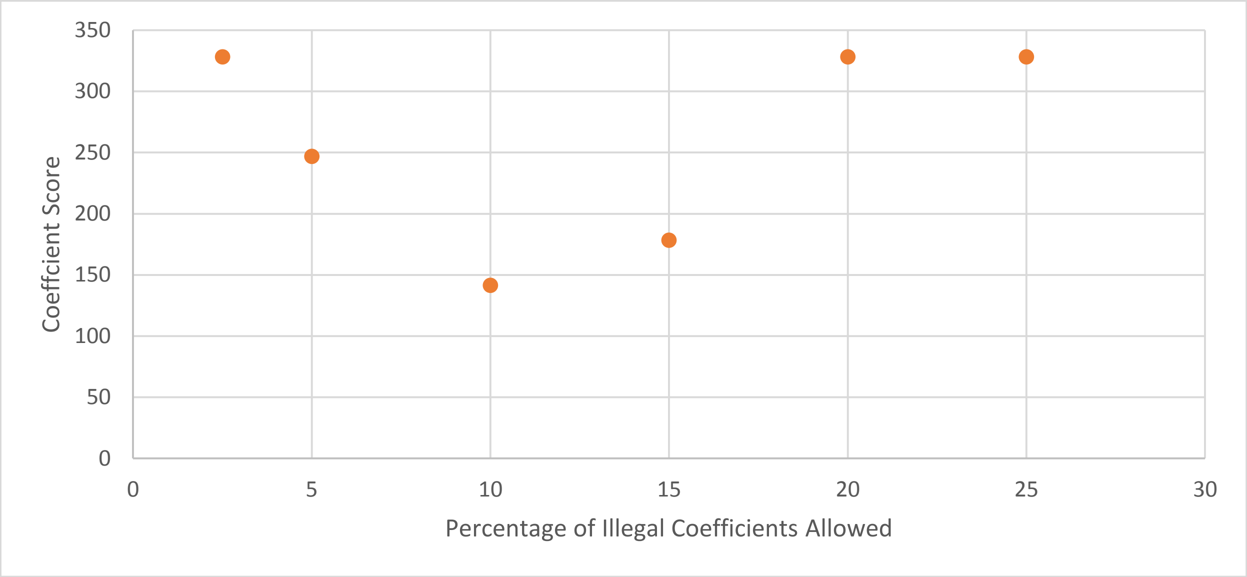
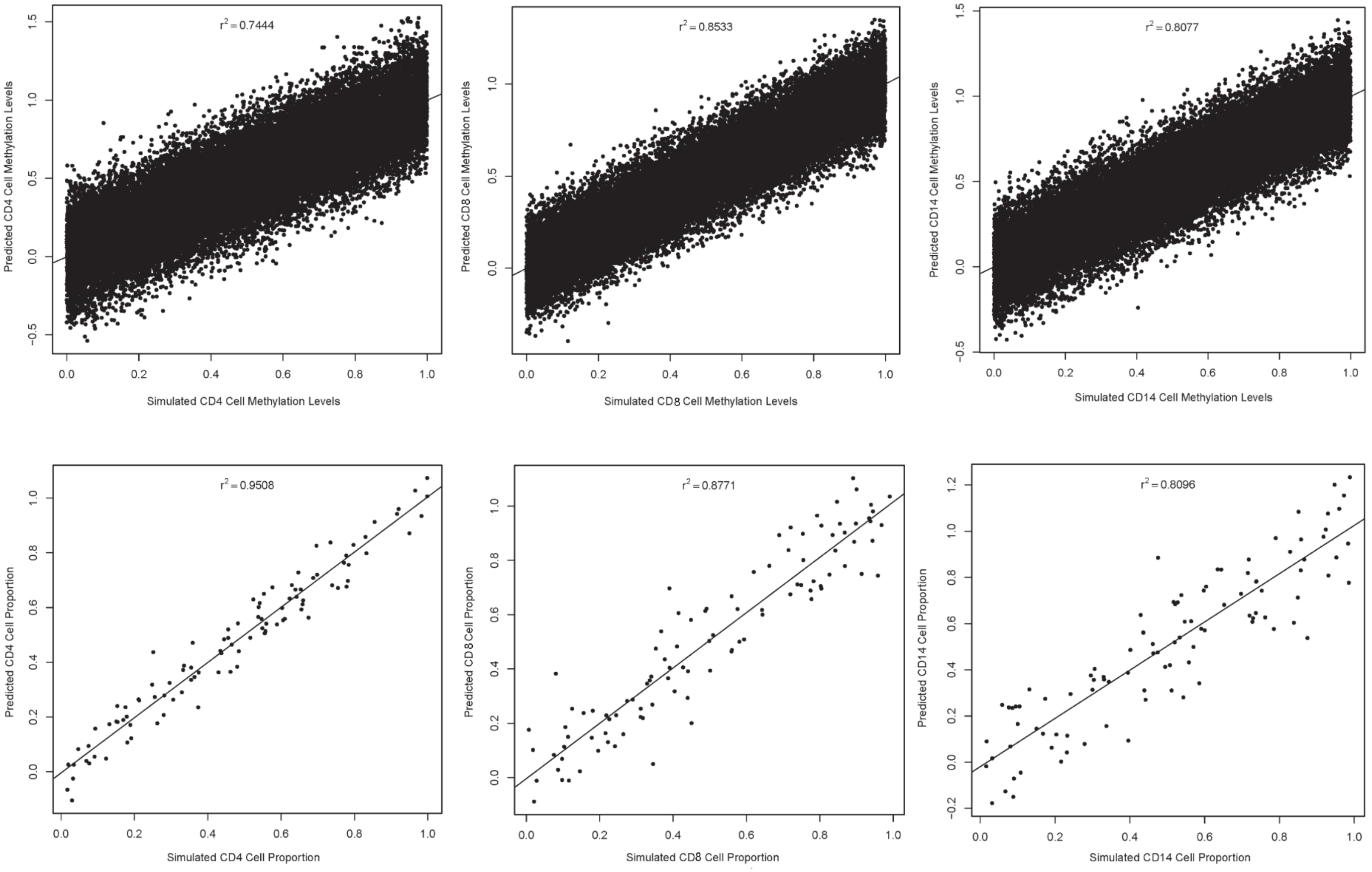


Figure 7: Linear Regression of Simulated vs Predicted Values



Discussion

Orthogonalization was able to successful derive a patient's cell composition profile from simulated patient methylation data. As shown in Figure 2, the coefficient and weight scores decrease on average after running the Orthogonalization algorithm on Deconvolution solutions ($p=3.91 \times 10^{-3}$ and 1.95×10^{-3}, respectively), implying Orthogonalization is an effective technique to optimize Deconvolution. The solution not only had lower weight and coefficient scores, but the Orthogonalization's solution matrices closely fitted the simulated data. The coefficient of determination for the simulated cell proportion data versus Orthogonalization's predictions was 0.95, 0.88, 0.81, for cell types, CD4, CD8, CD14, respectively, implying Orthogonalization was able to accurately predict the cell proportion of each patient. We see similar results for the predicted pure cell type methylation levels with r^2 of 0.74, 0.85, 0.81, for CD4, CD8, CD14, respectively.

Performance-wise, Orthogonalization gives relatively consistent results despite its random and non-deterministic nature. As seen in Figure 3, between 10 identical Orthogonalization trials on same set of data, the coefficient and weight scores differed minimally from each other with a standard deviation of 2.2% and 1.5% respectively.

After setting a parameter baseline and repeatedly running Orthogonalization it was discovered that the optimum values for the perturbation factor and illegal weight and coefficient percentage thresholds were 10%, 10%, and 1, respectively. Figures 4, 5, and 6, each show a roughly parabolic relationship between the score and parameter. Thus, by roughly identifying the peaks, the lowest scores, I was able to approximately determine these optimum values.

While successful at optimizing Deconvolution's solution, Orthogonalization is technically limited. Firstly, Orthogonalization allows some illegal value weight and coefficient values, rendering some patient data useless. Secondly, Orthogonalization was only tested using three cell types at once. It is predicted that adding more cell types may make the computations more difficult and less reliable. Thirdly, while the Orthogonalization data closely predicted the simulated data's components, it does not assure that this will work for real methylation. Orthogonalization is based on a model where individual methylation values are a product of various factors. Since I tested Orthogonalization on simulated patient methylation data that was intentionally created as a product of known factors, I can only say that my algorithm is successful in deriving two components from a product. Orthogonalization is still relies on a relatively novel recent composition methylation model[4].

Further testing is still needed on Orthogonalization. The optimum parameters were found from setting a baseline and altering one single parameter. It does not account for the other two parameters changing as well. Consequently more Orthogonalization runs that change all three parameters in combination are needed to establish the relationship between these parameters and determine the optimum value combination. It is also essential to test Orthogonalization on non-simulated data of other cell types where the components are known. In doing so one can compare Orthogonalization's predicted data to the known components and determine if this approach still remain effective.

Intensive research efforts are being devoted to studying DNA methylation systems involved in epigenetic gene regulation. Defects in methylation machinery have been shown to cause autoimmune disease such as systemic lupus erythematosus and immunodeficiency syndrome[2]. Thus understanding the variation in patterns of methylation between individuals may further our understanding of epigenetic aetiology of diseases. The Orthogonalization tool may prove evaluable in the future in helping derive and explain variation in DNA methylation between individuals.

References

1. Conerly M, Grady WM. Insights into the role of DNA methylation in disease through the use of mouse models. *Disease Models & Mechanisms*. 2010;3(5-6):290-297. doi:10.1242/dmm.004812.
2. Robertson K. DNA Methylation and human disease. *Nature Reviews Genetics*. 2005:6:597-610. doi:10.1038/nrg1655
3. Baylin S. DNA Methylation and gene silencing in cancer. *Nature Clinical Practice Oncology*. 2005:2:S4-S11. doi:10.1038/ncponc0354
4. Houseman A, Kelsey K, Wiencke J, Marsit C. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. *BMC Bioinformatics*. 2015:16:95. doi:10.1186/s12859-015-0527-y