Bioinformatics
Supervisor: Dr. Andrew Roth
December 12, 2020
Directed Studies Report (CPSC 448)

Non-parametric Bayesian Model

Multi-modal Pan-Cancer Stratification Using DNA Methylation Data

[1]Kevin Yang

[1]Computer Science and Statistics, University of British Columbia, Vancouver, V6T 1Z4, Vancouver

Abstract

Motivation: DNA methylation is an epigenetic mark that is suspected to have regulatory roles in a broad range of biological processes and diseases. There is evidence that DNA methylation levels are associated with transcriptional repression which may be closely related to cancer. Technology is now available to reliably measure DNA methylation in large samples and genome-wide, allowing for large scale analysis of methylation data. A common objective is to identify a latent structure shared across cancers from different tissue types reflecting commonly altered gene pathways. These latent structures stratify cancer patients into functionally similar groups. This process of stratification is commonly done using gene expression data; now that DNA methylation data is available in large quantities, this study investigates using both data types in stratification. We first perform important preprocessing steps on raw methylation data; specifically, normal approximations are used to transform methylation values and feature selection methods are applied to extract the most informative genes. Then we perform cluster analysis using a Dirichlet Process Mixture Model.

Results: Our study suggests that DNA methylation provides key information for stratification. Cluster analysis on DNA methylation data independently shows latent structures among cancer patients of differing types. In particular, we discover a shared latent structure among ovarian, colorectal and breast cancer patients. For joint analysis, two types of RNA expression data are used: RNA-seq and RNA-array. We combine both types with DNA methylation and perform a cluster analysis. When DNA methylation and RNA-seq data are combined, lung and brain cancer patients cluster together; whereas, when DNA methylation and RNA-array data are combined, brain and blood cancer patients cluster together. These results motivate further study on using DNA methylation as an additional feature to RNA expression.

1 Introduction

Cancers develop via the acquisition of genomic changes. These changes in turn alter the cells harbouring them, leading to changes in cell state (phenotype). One common effect of these mutations is to induce a series of modifications to the genome, which do not alter the encoded DNA but rather the ability of the DNA to be read and processed into protein. These heritable non-genetic changes are broadly referred to as epigenetic changes. In recent years, research has shown that epigenetic changes may play a central role in cancer causation, progression and treatment. There are significant connections between epigenetic alterations and disease; according to Baylin and Jones (2011), "Epigenetic alterations are leading candidates for the development of specific markers for cancer detection, diagnosis and prognosis". One such epigenetic change is DNA methylation which is unambiguously linked with transcriptional repression. When present in promoter regions, DNA methylation correlates negatively with gene expression; furthermore, characteristic changes in DNA methylation have been reported for cancer. Research indicates that gene promoter CpG islands acquire abnormal hypermethylation resulting in transcriptional silencing in cancer (Bock, 2012). It is important to note that 70-80% of CpGs in the human genome are affected from DNA methylation; therefore, if the effects of DNA methylation in cancer are better understood, considerable progress will be made for finding an effective remedy for cancer in humans. The potential relationship between DNA methylation and cancer opens up a new avenue of exploration. We are interested if viewing cancer as a disease of both epigenetic and genetic abnormalities can reveal insights into the development of cancer. Advances in next-generation sequencing and microarray technology allow for analysis on DNA methylation in large samples and genome-wide; currently, DNA methylation is the only epigenetic mark that can be measured reliably. As a result, DNA methylation can be profiled accurately using high throughput sequencing and is a potential feature for categorizing cancer

patients into functionally similar groups. This process of grouping patients, or stratification, is most commonly done using gene expression data; with DNA methylation data now available, we intend to use both gene expression and DNA methylation data for stratification. Our hypothesis is that using DNA methylation as an additional feature for stratification will improve power to detect biologically meaningful patient groups.

Section 1.1 contains a brief overview of previous work done on Dirichlet Processes. A brief description of the model can be found in Section 2. Methods that were used during the analysis are detailed in Section 3. Important results are presented in Section 4. Finally, discussion and suggestions for future research are given in Section 6.

1.1 Previous Work

Unsupervised learning aims to discover the latent structure in a dataset. According to Griffith and Ghahramani (2011), "One of the key problems faced by unsupervised learning algorithms is thus determining the amount of latent structure - the number of clusters, dimensions, or variables - needed to account for the regularities expressed in the data." Griffith and Ghahramani indicate that this is often addressed as a model selection problem, where the model with the dimensionality that results in the best performance is chosen. Model selection is very difficult even with procedures such as cross validation; the Bayesian nonparametric approach is an effective alternative to model selection. By using a model with an unbounded complexity, underfitting is mitigated, while the Bayesian approach of computing or approximating the full posterior over parameters mitigates overfitting (Teh, 2010). The Dirichlet Process is a widely used Bayesian nonprametric model that allows for an unbounded number of dimensions which side steps the problem of determining the optimal number of features. The Dirichlet Process sets a Dirichlet prior which has a wide support, effectively broadening the scope and type of inferences that can be made. Teh notes a limitation to the model; namely, that draws from the Dirichlet process are limited to discrete distributions. The development of Markov Chain Monte Carlo techniques has now made it tractable to use more general priors. Radford Neal proposes an MCMC technique to sample from the posterior distribution using Gibbs sampling when conjugate priors are used (Neal, 2000). Dahl notes that although Gibbs samplers effectively update model parameters, they can have difficulty in updating the clustering of the data (Dahl, 2005); he introduces a merge-split sampler to addresses this issue.

Clustering methods have seen great success in single tissue analysis. Makretsov et al. (2004) showed that using hierarchical clustering analysis of multiple immunomarkers (protein expression profiles) improved prognostication in patients with invasive breast cancer. They also found that hierarchical clustering grouped breast cancers into classes with clinical relevance and was superior to individual prognostic markers. Crook et al. (2018) applied a new clustering method called SUGSVarSel on an expression dataset for breast cancer tumour data from The Cancer Genome Atlas (TCGA) and obtained 5 clusters which corresponded to known breast cancer subgroups. These types of results motivate our study on whether clustering methods are effective for multiple tissue analysis.

2 Model

We use Neal's MCMC technique and Dahl's merge-split sampler in our Dirichlet Process Mixture Model. This model aims to identify a latent structure reflecting commonly altered gene pathways among cancers. However, clustering cancer patients based on this latent structure is difficult because the tissue of origin's gene expression effect often dominates the latent structure effect. Therefore, we account for the tissue effect and cluster based on the latent structure using a Dirichlet Process prior. The model learns the tissue parameters in a supervised learning setting, while simultaneously learning the latent structure, based on the resulting residuals, in an unsupervised setting. Figure 1 shows the Bayesian network diagram for the model.

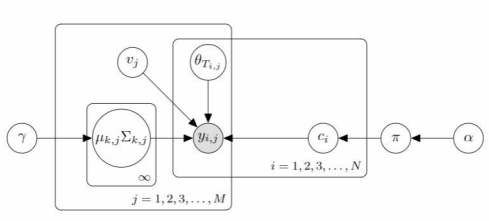

Fig. 1. Bayesian Network for DPGMM

$$
\begin{aligned}
v_j &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\
\theta_{T_{i,j}} &\sim \mathcal{N}(\mu_1, \sigma_1^2) \\
\alpha &\sim Gamma(\alpha_1, \beta_1) \\
\{\mu_{k,j}, \Sigma_{k,j}\} &\sim NG(\{\mu, \lambda, \alpha_0, \beta_0\}) \\
\pi|\alpha, \{\mu_{k,j}, \Sigma_{k,j}\} &\sim DP(\alpha, \{\mu_{k,j}, \Sigma_{k,j}\}) \\
c_i|\pi &\sim Discrete(\pi_1, \ldots, \pi_k) \\
y_{i,j}|\{\mu_{c_i,j}, \Sigma_{c_i,j}\}, \theta_{T_{i,j}}, v_j &\sim \mathcal{N}(\mu_{c_i,j} + \theta_{T_{i,j}} + v_j, \Sigma_{c_i,j})
\end{aligned}
$$

- $y_{i,j}$ log normalized gene expression for patient i gene j
- θ_{T_i} tissue parameters (supervised)
- $\{\mu_{c_i}, \Sigma c_i\}$ mean and precision parameters for cluster c_i
- ν_j baseline gene parameter
- c_i cluster assignments

Remark: We still have conjugacy between the base distribution and likelihood. In a nutshell,

$$e^{(y_{i,j} - \theta_{T_i,j} - \nu_j - \mu_{c_i,j})^2} = e^{(B_{i,j} - \mu_{c_i,j})^2}$$

will still have the form of a normal distribution where $B_{i,j} = y_{i,j} - \theta_{T_i,j} - \nu_j$ is called the shifted data.

3 Methods

3.1 Pseudo-Code

The inference procedure is described in Algorithm 1. 500 burnin iterations and 10000 inference iterations were used for each run. All Equations are defined in Section 8.3. Code can be found at https://github.com/keviny2/MethylationDPGMM.

Algorithm 1

function INFERENCE($y = (y_1, \ldots, y_n), iter = l, burnin = b$)
  Sample $\phi = (\phi_c : c \in \{c_1, \ldots, c_n\})$ from $\mathcal{N}(0, 1)$
  Sample α, ν from $\mathcal{N}(0, 1)$
  Sample α from $Gamma(0.1, 100)$
  Sample G_0 from $NG(0, 1, 1, 1)$
  Sample cluster assignments $c = (c_1, \ldots, c_n)$ from
  Let $\pi = \{S_1, \ldots, S_q\}$ be the set of all components
  for $j = 1, \ldots, b$ **do**
    Perform burnin
  end for
  for $iter = 1, \ldots, l$ **do**
    for $i = 1, \ldots, n$ **do**
      Draw a new value from $c_i | c_{-i}, y_i$ as defined by (i) and (ii)
    end for
    Choose two random indices i, j corresponding to two data points
    if i, j are from the same component **then**
      Let S be the component containing i and j
      Let $S^i = \{i\}, S^j = \{j\}$
      for $k \in S$ **do**
        Add k to S^i with probability defined by (iii)
        Otherwise, add k to S^j
      end for
      Let $\pi^* = \pi \cup \{S^i, S^j\} \backslash \{S\}$
    else
      Let S^i and S^j be components such that $i \in S^i$ and $j \in S^j$
      Let $S = S^i \cup S^j$
      Let $\pi^* = \pi \cup \{S\} \backslash \{S^i, S^j\}$
    end if
    Compute Metropolis-Hastings ratio and accept π^* with probability given by (iv)
    Update ν
    Update ϕ
    Update α
    Adapt ν, ϕ, α proposals
  end for
end function

3.2 Normal Approximations

Methylation values are beta values (Equation i in Section 8.1) which we assumed to follow a beta distribution. The Dirichlet Process Mixture Model requires normality; therefore, normal approximations to the beta distribution were explored. We applied each normal approximation on 200000 methylation values from ovarian, breast and colorectal cancer patient data obtained through ICGC. Each normal approximation produced a new dataset with transformed methylation values.

3.2.1 Wise's Approximation

Wise suggested an extremely simple normal approximation to the beta distribution (Wise, 1960). See Equation (1).

Let $X \sim Beta(\alpha, \beta)$ and $\alpha \geq \beta$, then y is more nearly normal than X if

$$y = -\log X^{\frac{1}{3}} \tag{1}$$

3.2.2 Peizer and Pratt's Approximation

Peizer and Pratt proposed a normal approximation that works for beta, binomial, negative binomial, F, t, Poisson, gamma, chi square and Pascal distributions (Peizer and Pratt, 1968). The approximation does not include any complex arithmetic and can be performed on a simple pocket calculator. Moreover, when compared to other normal approximations, it appeared to be far more accurate. Unfortunately, it did not perform well for beta distributions that are J or U shaped. See Equation (2).

Let $Y \sim Beta(\alpha, \beta)$, then z is normally distributed if

$$z = d_2 \frac{1 + qg(\frac{S}{np}) + pg(\frac{T}{nq})^{\frac{1}{2}}}{(n + \frac{1}{6})pq} \tag{2}$$

$S = \beta - \frac{1}{2}, \quad T = \alpha - \frac{1}{2}, \quad n = \alpha + \beta - 1$
$p = 1 - y, \quad q = y$
$d_1 = S + \frac{1}{6} - (n + \frac{1}{3})p$
$d_2 = d_1 + 0.02(\frac{q}{S+.5} - \frac{p}{T+.5} + \frac{q-.5}{n+1})$
$g(x) = (1 - x)^{-2}(1 - x^2 + 2x \ln(x))$

3.2.3 Alfers and Dinge's Approximation

Alfers and Dinges built off the work done by Peizer and Pratt; they presented a normal approximation specifically for beta and gamma distributions (Alfers and Dinges, 1984). The approximation compared very well with existing approximation methods. Moreover, theoretical bounds for the error showed that the precision was especially great in the extreme tails. Similar to the approximation proposed by Peizer and Pratt, the approximation method only used elementary functions that can be found on a pocket calculator. See Equation (3).

Let $Y \sim Beta(\alpha m, (1 - \alpha)m)$, then

$$\mathcal{L}(A(\alpha, Y)) \sim \mathcal{N}(0, \frac{1}{m}) \tag{3}$$

$A(\alpha, p) = \frac{\alpha - p}{\sqrt{pq}} \sqrt{1\pi(\alpha, p)}$
$\pi(\alpha, p) = 2ph(1 - \frac{\beta}{q}) + 2qh(1 - \frac{\alpha}{p})$
$h(x) = \frac{1}{x} + \frac{(1-x)\ln(1-x)}{x^2} - \frac{1}{2}$
$q = 1 - p, \quad \beta = 1 - \alpha$

3.3 Normality Tests

Normality tests were used to empirically assess the effectiveness of each normal approximation. Each test produced a test statistic and p-value which we used to check for normality. These normality tests were conducted on transformed methylation values. The functions shapiro.test() from the nortest package and ad.test() from the stats package in R were applied to conduct the Shapiro-Wilk and Anderson-Darling tests respectively. The shapiro.test() function only accepted vectors with length between 3 and 5000; therefore, values were obtained by averaging across 500 random samples of 5000 methylation values each. There were no restrictions regarding vector length for ad.test().

3.3.1 Shapiro-Wilk Test

Shapiro and Wilk created a statistical procedure that tested a complete sample for normality (Shapiro and Wilk, 1965). The procedure produced a test statistic W (Equation 4), where smaller values of W indicated non-normality. W was computed as follows:

- Let $\{x_1, x_2, \ldots, x_n\}$ be our sample
- Order observations to obtain an ordered sample $y_1 \leq y_2 \leq \ldots \leq y_n$
- Compute

$$S^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

- If n is even, $n = 2k$, compute

$$b = \sum_{i=1}^{k} a_{n-i+1}(y_{n-i+1} - y_i)$$

If n is odd, $n = 2k + 1$ and

$$b = a_n(y_n - y_1) + \ldots + a_{k+2}(y_{k+2} - y_k)$$

- Compute

$$W = \frac{b^2}{S^2} \tag{4}$$

Refer to Table 5 and Table 6 in *An analysis of variance test for normality* (Shapiro and Wilk, 1965) for values of a and points of the distribution of W respectively.

3.3.2 Anderson-Darling Test

Anderson and Darling designed a distribution-free test which examined the likelihood a sample was drawn from a particular distribution (Anderson and Darling, 1954). The procedure produced a test statistic W_{n^2} (Equation 5), where larger values of W_{n^2} indicated non-normality. W_{n^2} was computed as follows:

- Let $F(x)$ be the cumulative distribution function for the distribution of interest
- Order the sample such that $x_1 \leq x_2 \leq \ldots \leq x_n$
- Compute

$$W_{n^2} = -n - \frac{1}{n}\sum_{j=1}^{n}(2j-1)[\ln(u_j) + \ln(1 - u_{n-j+1})] \tag{5}$$

3.4 Clustering

3.4.1 Data

We first performed cluster analysis on DNA methylation and RNA-seq data individually with 50 ovarian, 50 breast and 50 colorectal cancer patient samples. We also performed two joint cluster analyses: one with DNA methylation and RNA-seq data, another with DNA methylation and RNA-array data. Paired DNA methylation and RNA-seq data contained 28 blood, 29 brain and 42 lung cancer patient samples; paired DNA methylation and RNA-array data contained 30 blood, 36 brain and 5 lung cancer patient samples.

3.4.2 Clustermaps

Clustermaps were used to visualize results; we interpret the clustermaps as follows: color bars indicate the patient's cancer type and squares represent which patients were clustered together. The more patients in the same cluster, the larger the squares. It is obvious that each patient appears in the same cluster as itself, so the clustermap will always have squares on the diagonal. The intensity of a square describes the proportion of iterations the patients of interest appeared in the same cluster. Higher intensities indicate that a group of patients appeared in the same cluster more often and lower intensities indicate that a group of patients appeared in the same cluster less often. Refer to Figure 2 for an example clustermap.

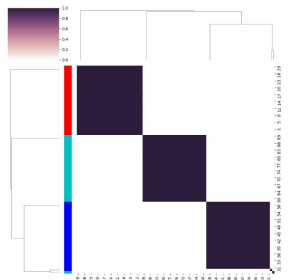

Fig. 2. Example clustermap. Color labels indicate cancer types: Type I, Type II, Type III

3.4.3 Concatenation

To concatenate DNA methylation and RNA-seq or RNA-array data, we performed the following steps:

- Assume we have an $n \times m_1$ DNA methylation matrix and an $n \times m_2$ RNA-seq or RNA-array matrix
- Perform PCA on the DNA methylation matrix to produce an $n \times p$ matrix, where p represents the number of principal components
- Perform MAD on the RNA-seq or RNA-array matrix to produce an $n \times d$ matrix
- Concatenate the two matrices to produce a new matrix with dimension $n \times p + d$

3.5 Feature Selection Methods

Various feature selection methods were used to extract the most informative genes while minimizing information loss.

3.5.1 Mean Absolute Deviation (MAD)

The mean absolute deviation was computed in Python. We subsetted on the top 100 genes with the largest mean absolute deviation. Refer to Equation (6) for the definition.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \tag{6}$$

3.5.2 Principal Component Analysis (PCA)

The well known feature selection method called Principal Component Analysis was used to extract a subset of genes. We used the sklearn library for Python to perform PCA; the top 100 components were kept for the analysis. Refer to *Principal Component Analysis: a review and recent developments* (Jollife and Cadima, 2016) for more information on PCA.

4 Results

4.1 Normal Approximations

Peizer and Pratt's approximation was unstable for methylation data. Estimates for the beta distribution parameters α and β revealed that $\alpha + \beta < 1$. This resulted in taking the natural logarithm of a negative number (because $n = \alpha + \beta - 1 < 0$) which is undefined; therefore, no results were obtained using this approximation method. Histograms (Figure 3) were constructed for the other two approximations using ggplot2 in R.

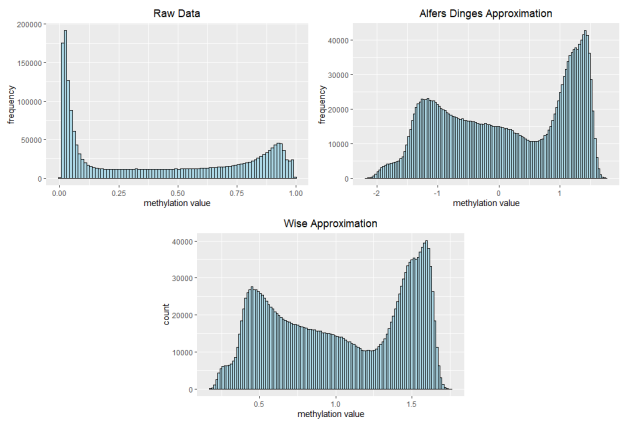

Fig. 3. Histograms for normal approximations to methylation values

4.2 Normality Tests

We first constructed Q-Q plots (Figure 4) for a simple visual to examine the performance of each approximation; Peizer and Pratt's approximation was left out for the reasons already stated. The points formed a curve resembling a logit function for both Q-Q plots, which implied the resulting distribution was non-normal. The histograms already revealed that the distribution was bimodal, so these results were not surprising.

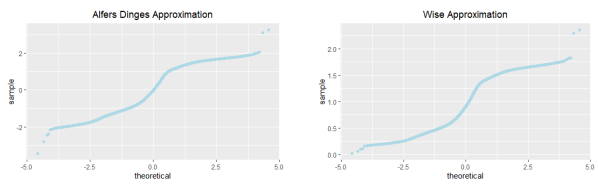

Fig. 4. Q-Q plots for normal approximations to methylation values

Both normality tests indicated that Alfers and Dinge's approximation performed better than Wise's approximation. The Shapiro-Wilks test yielded $W = 0.933$ for Alfers and Dinge's approximation and $W = 0.926$ for Wise's approximation; lower values of W for the Shapiro-Wilks test indicate non-normality. The Anderson-Darling test yielded $W_{n^2} = 119.075$ for Alfers and Dinge's approximation and $W_{n^2} = 131.7$ for Wise's approximation; higher values of W_{n^2} indicate non-normality (See Table 1). However, p-values for both approximation methods were extremely small, which suggested that both approximations produced non-normal distributions.

Table 1. Results from the Shapiro-Wilks (S-W) and Anderson-Darling (A-D) tests

	Test Statistic	p-value
Alfers-Dinges	S-W: 0.933 A-D: 119.075	S-W: 2.06e-42 A-D: 3.7e-24
Wise	S-W: 0.926 A-D: 131.7	S-W: 4.39e-44 A-D: 3.7e-24

4.3 Clustering

Various configurations of feature selection methods and tissue effect formats (on or off) were experimented; DNA methylation and RNA-seq data were analyzed separately. We tried the following combinations: MAD with tissue effect off, MAD with tissue effect on, and PCA with tissue effect on.

4.3.1 Mean Absolute Deviation (MAD)

MAD with tissue effect off: Both DNA methylation and RNA-seq clustered cancer patients by tissue (Figure 5). Some colorectal patients formed a cluster separate from the majority of colorectal patients.

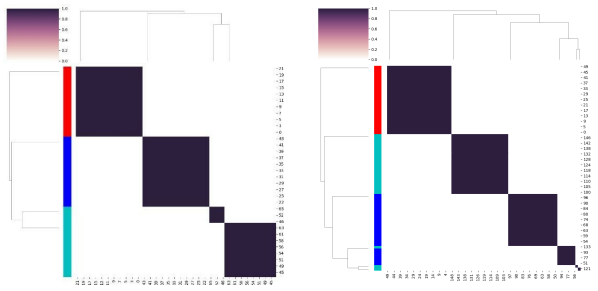

Fig. 5. Clustermap for methylation (left) and RNA-seq (right), tissue effect off using MAD. Color labels indicate cancer types: ovarian, breast, colorectal

MAD with tissue effect on: For DNA methylation, breast and ovarian cancer patients formed one cluster, while breast and colorectal patients formed another cluster. For RNA-seq, we observed clustering between patients from all three cancer types. Some breast cancer patients also formed their own cluster. Refer to Figure 6.

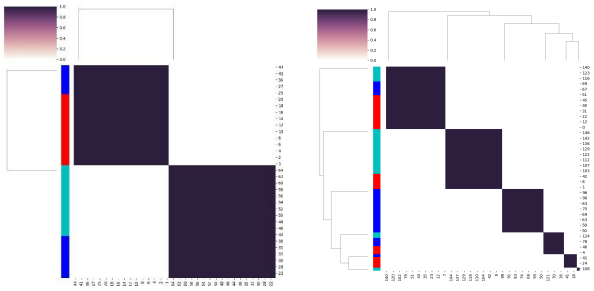

Fig. 6. Clustermap for methylation (left) and RNA-seq (right), tissue effect on using MAD. Color labels indicate cancer types: ovarian, breast, colorectal

4.3.2 Principal Component Analysis (PCA)

PCA with tissue effect on: For both data types, we observed clustering between patients from all three cancer types (Figure 7). DNA methylation produced four smaller clusters while RNA-seq produced one large and one small cluster.

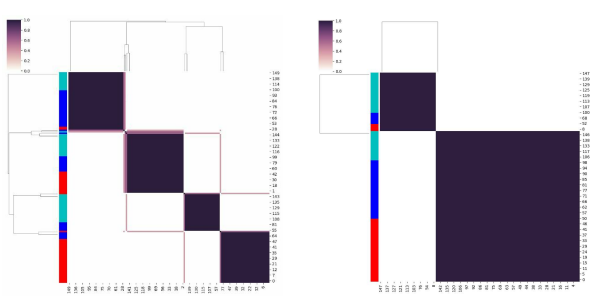

Fig. 7. Clustermap for methylation (left) and rna (right), tissue effect on using PCA. Color labels indicate cancer types: ovarian, breast, colorectal

4.3.3 Joint Analysis

We used scatterplots to visualize the concatenated data. Each (x, y) pair on the scatterplot corresponds to the methylation and RNA-seq or RNA-array value for a single patient and gene. The plots showed that there was weak correlation between methylation and RNA expression values (Figure 8). The correlation coefficient was approximately 0.13.

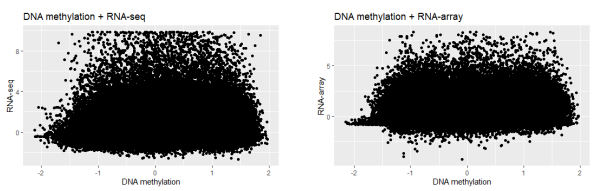

Fig. 8. Scatterplot of DNA methylation + RNA-seq (left) and DNA methylation + RNA-array (right)

We also experimented with configurations for concatenated data.

DNA methylation + RNA-seq data with tissue effect off: Interestingly, all cancer patients formed one large cluster.

DNA methylation + RNA-array data with tissue effect off: Cancer patients cluster by tissue (See Figure 9 for clustermaps)

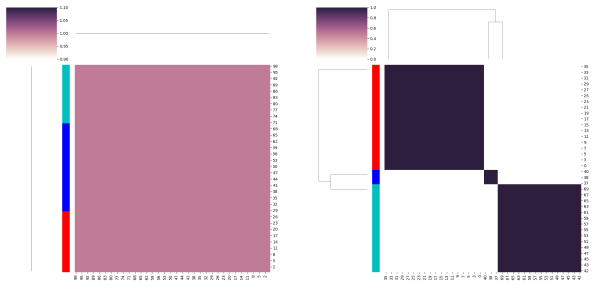

Fig. 9. Clustermap for methylation + RNA-seq (left) and methylation + RNA-array (right), tissue effect off, color labels indicate cancer types: brain, lung, blood

DNA methylation + RNA-seq with tissue effect on: We observed clustering between brain and lung cancer patients; blood cancer patients formed their own cluster.

DNA methylation + RNA-array with tissue effect on: We observed clustering between brain and blood cancer patients: lung cancer patients formed their own cluster. (See Figure 10 for clustermaps)

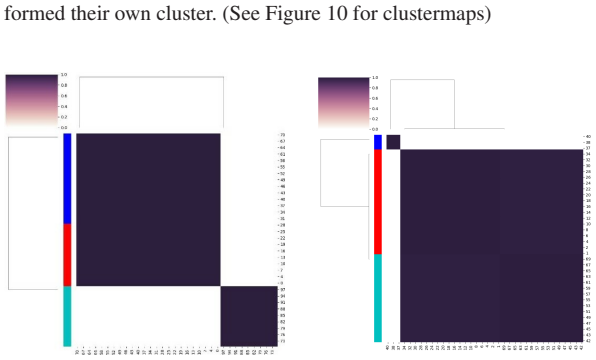

Fig. 10. Clustermap for methylation + RNA-seq (left) and methylation + RNA-array (right): Tissue Effect On - Color labels indicate cancer types: brain, lung, blood

5 Discussion

The results suggest that using DNA methylation as an additional feature for stratification is likely to improve power to detect biologically meaningful patient groups. The cluster analysis results show that independently, DNA methylation reveals a latent structure among cancer patients of differing types; specifically, when the tissue effect is accounted for, DNA methylation reveals a shared latent structure among ovarian, breast and colorectal cancer patients. The clustermaps show, for the most part, that DNA methylation and RNA-seq data cluster patients in a similar fashion. However, the number and size of these clusters differ between the two data types, which suggest that slightly different latent structures are being detected. When the two data types are jointly considered, latent structures also emerge. Our results demonstrate how DNA methylation, if analyzed alongside RNA expression data, can reveal insights into the development of cancer. We now discuss some limitations of our study. Some batch effects might have been overlooked. The only batch affect we addressed was using different BeadChips to measure DNA methylation values; therefore, investigating other potential batch effects will lead to more accurate results. Moreover, the beta distribution does not model methylation beta values effectively. The histogram of the raw data shows that methylation beta values come from a mixture. The distribution is skewed heavily to the right, but also contains considerable values close to 1. This explains the bimodal distribution produced by the normal approximation methods. Nonell and Gonzalez propose a simplex distribution instead to model methylation beta values (Nonell and Gonzalez, 2019). Another concern relates to the scatterplot, which shows very weak correlation. The correlation coefficient is less than 0.15, implying that DNA methylation does not have an obvious relationship with either RNA-seq or RNA-array. Perhaps the sample was too small and increasing sample size would reveal a correlation. An exciting area for future research pertains to an emerging direction in machine learning known as multi-view learning. Multi-view learning, also known as data integration from multiple feature sets, uses data represented by multiple distinct feature sets for machine learning. (Sun, 2013). The aim of multi-view learning is to improve generalization performance by considering multiple distinct feature sets, called views (Zhao et al., 2017). It has seen great success in recent years and is suitable for our application. In this study, we performed a naive concatenation of DNA methylation and RNA expression data for stratification. We expect that using a sophisticated framework such as multi-view learning instead of a naive concatenation will improve performance.

6 Acknowledgements

We thank Andrew Roth for providing guidance throughout this project and carefully reading this report. We are also grateful for past work done by Matteo Lepur on the Dirichlet Process Mixture Model and for his help along the way.

7 References

Bock, C. (2012). Analysing and interpreting DNA methylation data. Nature Reviews, Genetics, 13.

Teschendrff, A.E. and Reton, C.L. (2018). Statistical and integrative system-level analysis of DNA methylation data.*Nature Reviews, Genetics, 19.*

Bock, C. et al.(2010). Quantitatve comparison of genome-wide DNA methylation mapping technologies. *Nature Biotechnology.*

Sun, S. (2013). A survey of multi-view machine learning. *Nerual Comput and Applic.*

Korthauer, K. (2020). Lecture 17 - DNA Methylation. STAT/BIOF/GSAT 540.

Peizer, D.B. and Pratt, J.W. (1968). A Normal Approximation for Binomial, F, Beta, and Other Common, Related Tail Probabilities, I. *Journal of the American Statistical Association.*, **63**, 324, pp. 1416-1456

Alfers, D. and Dinges, H. (1984). A Normal Approximation for Beta and Gamma Tail Probabilities.

Baylin, S.B. And Jones, P.A. (2011). A decade of exploring the cancer epigenome - biol ogical and translational implications. *Nature Reviews - Cancer*.

Jolliffe, I.T. and Cadima, J. (2016). Principal Component Analysis: a review and recent developments. Phil. Trans. R. Soc. A 374:20150202.

Dahl, D.B. (2005). Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models.

Neal, R.M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, **9**, 2, pp. 249-265.

Teh, Y.W. Dirichlet Process.

Griffiths, T.L. and Ghahramani, Z. (2011). The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research*, **12**, pp. 1185-1224.

Escobar, M.D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, **90**, 430, pp. 577-588.

Nonell, L. and Gonzalez, J.R. (2019). Are methylation beta-values simplex distributed?

Wise, M.E. (1960). On normalizing the incomplete beta-function for fitting to dose-response curves. *(Biometrika)*, **47**, pp. 173-175.

Makretsov, N.A. et al. (2004). Hierarchical Clustering Analysis of Tissue Microarray Immunostaining Data Identifies Prognostically Significant Groups of Breast Carcinoma. *(Clinical Cancer Research)*, **10**, pp.6143-6151.

Crook, O.M. et al. (2018). Fast approximate inference for variable selection in Dirichlet process mixtures, with an application to pan-cancer

proteomics. *(Statistical Applications in Genetics and Molecular Biology)*

Anderson, T.W. and Darling, D.A. (1965). A Test of Goodness of Fit. *Journal of the American Statistical Association*, **49**, pp.765-769.

Shapiro, S.S. and Wilk, M.B.. An Analysis of Variance Test for Normality. *(Biometrika)*, **52**, pp.591-611.

8 Appendix

8.1 Data

The raw methylation data obtained from the ICGC portal contains 20 columns and many million rows. The clinical data obtained from the ICGC portal contains 21 columns and each row corresponds to the information for a single donor. For the raw methylation data, only 4 columns are relevant:

- icgc_donor_id - string, unique identifier for each donor
- probe_id - string, unique identifier for each methylation probe
- methylation_value - number, the measured beta value
- array_platform - string, type of assay used

For clinical data, only 2 columns are relevant:

- icgc_donor_id - string, unique identifier for each donor
- project_code - string, project the data is from

Methylation values are measured using beta values (See Equation i). By definition, beta values are between 0 and 1. Data is shifted prior to analysis (See Equation ii).

$$\beta = \frac{\textit{methylated intensity}}{\textit{methylation intensity + unmethylated}} \tag{i}$$

$$y_{shifted} = y - \nu - T^{\intercal} \cdot \phi \tag{ii}$$

Figure 11, Figure 12 and Figure 13 show heatmaps of the data.

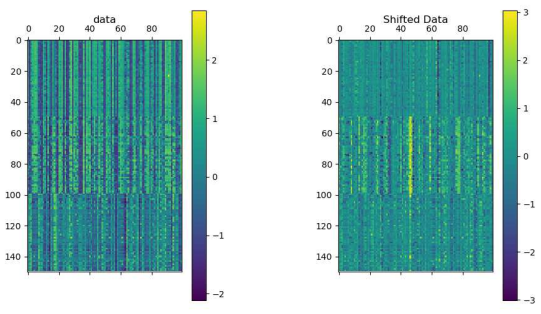

Fig. 11. Raw methylation data (left) and shifted methylation data (right)

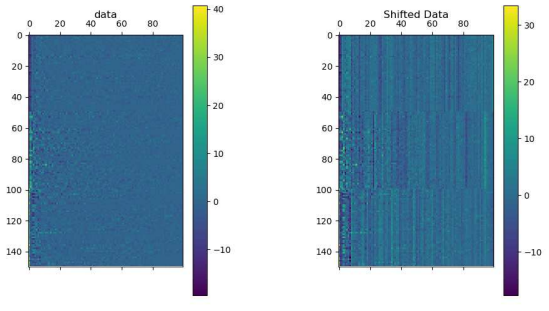

Fig. 12. Methylation data after PCA is performed prior shifting (left), post shifting (right)

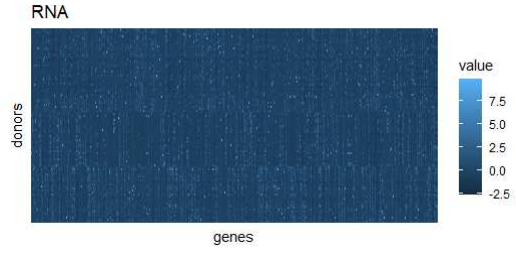

Fig. 13. Concatenated data without shifting

8.2 Probe to Gene Mapping

Each methylation value is associated with a probe ID. The gene symbol of the closest gene to each probe was determined using the R package FDb.InfiniumMethylation.hg19. The preprocessing pipeline can be found in the 'methylationfun' R package. Installation instructions can be found at: https://github.com/keviny2/methylationfun. The pipeline performs the following tasks:

- Load the data
- Filter out probes that are not from the specified array platform (default is to keep InfiniumMethylation450)
- Average methylation values for duplicate probe ids
- Filter out probe ids which are not shared across all donors
- Convert probe ids to gene closest to it
- Transform methylation values using a normal approximation to the beta distribution
- Average methylation values for duplicate genes
- Filter out genes which are not shared across all donors
- Returns a list of dataframes containing the methylation data, donor ids, gene symbols and tissue assignments

8.3 Equations in Psuedo-Code

$$P(c_i = c|c_{-i}, y_i) = b\frac{n_{-i,c}}{n-1+\alpha}\int F(y_i,\phi)\, dH_{-i,c}(\phi) \tag{i}$$

$$P(c_i = c|c_{-i}, y_i) = b\frac{\alpha}{n-1+\alpha}(y_i,\phi)\, dG_0(\phi) \tag{ii}$$

$$P(k \in S^i|S^i, S^j, \mathbf{y}) = \frac{|S^i|\int p(y_k|\phi)p(\phi|\mathbf{y}_{S^i})\, d\phi}{|S^i|\int p(y_k|\phi)p(\phi|\mathbf{y}_{S^i})\, d\phi + |S^j|\int p(y_k|\phi)p(\phi|\mathbf{y}_{Sj})\, d\phi} \tag{iii}$$

$$a(\pi^*|\pi) = min[1, \frac{p(\pi^*|y)q(\pi|\pi^*)}{p(\pi|y)q(\pi^*|\pi)}] \tag{iv}$$

- $n_{-i,c}$ is the number of elements in component c after removing i
- b is the normalization constant
- $F(y_i, \phi)$ is the likelihood
- $H_{-i,c}$ is the posterior distribution of ϕ based on the prior G_0 and all observations y_j for which $j \neq i$ and $c_j = c$
- $p(\phi|\mathbf{y}_S)$ is the posterior distribution of a component location ϕ based on the prior $F_0(\phi)$ and the data corresponding to the indices in S
- $p(\pi|y)$ is the density of the partition posterior distribution evaluated at π
- $q(\pi|\pi^*)$ is the probability of proposing π from the state π^*
- **Refer to *Markov Chain Sampling Methods for Dirichlet Process Mixture Models* (Neal, 2000) for full definitions of (i) and (ii)
- **Refer to *Sequatially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models* (Dahl, 2005) for full definitions of (iii) and (iv)