**Visualizing Soccer Game Dynamics with Social Network Graphs**

Missouri University of Science and Technology, Department of Mathematics and Statistics

Yik Fai Siu

Dr. Gayla Olbricht

April 3, 2017

*Soccer is one of the most popular team sports globally. An effective game strategy is key to success. However, analyzing the effectiveness of a strategy in a game is difficult due to the complexity of player interactions and ball flow. Having a good visualization of important aspects of the game dynamics could aid in interpreting player and team performance. In this work, we utilize social network graphs to visualize soccer game dynamics. These graphs incorporate individual player performance metrics as well as passing trends between players and can be compared between games to better understand differences in performance.*

**Introduction**

The human genome is the complete nucleic acid sequence which is encoded as DNA. DNA, deoxyribose nucleic acid, is made up of two dipolymer strands which coil around each other to form the well-known double helix shape. Four nucleotides make up the main structure of DNA: adenine, thymine, guanine and cytosine. Within the biopolymer strands, adenine pairs with thymine and guanine pairs with cytosine.

However, the code upon which human life is based is not so simplistic. Within the DNA strands are units of hereditary information, called genes. Genes are sequences of hundreds to millions of base pairs which control expression of characteristics. Genes may or may not be expressed. It has been shown that some gene expression can be controlled by the attachment of a methyl (CH3) group to the fifth carbon atom of a cytosine nucleotide. [1] This control mechanism is referred to as DNA methylation. This methylation occurs most often when a cytosine is immediately followed by a guanine, referred to as a CpG site [2].

In cancer biology research, is has been shown that atypical gains and losses of DNA methylation markers most commonly accompany the initiation and progression of cancer [2]. Thus, studying methylation patterns is not merely interesting. Rather, it offers the possibility of aiding in the development of therapies which reverse disease-associated epigenetic alternations.

Though the methylation patterns among one group of individuals may differ from another group of individuals (healthy vs. cancer groups, for instance, is a common group to group comparison), each individual also has a specific methylation pattern which varies idiosyncratically from the other individuals in their group. These individual level differences create statistical challenges when attempting to quantify differences between groups.

Much headway has been made in developing statistical techniques to combat these challenges. However, when presented with virtually countless ways with which to analyze a particular dataset, it is not often clear which statistical method is best.

In this study, the goals are two fold:

1. To employ statistical methods to identify locations with significant DNA methylation differences between patients with various stages of pre-cancerous or early cancerous cervical lesions in a group of HIV positive women.

2. To compare statistical methods for analyzing DNA methylation data in this dataset.

**Data Schema**

1. **Nature of Data & Source**

The data used in this study was obtained through researchers at the Fred Hutchinson Cancer Research Center in Seattle, Washington.

The data was collected through self-performed swabs from Kenyan women and tissue samples from Senegalese women. Each individual has an HIV positive status and some stage of pre or early cancerous cervical lesions.

The progression of the cervical lesions in each woman was assigned a rating based on their severity. The rating system is detailed below:

* **Low Grade:** 9 samples of low grade type were used in this study. Low grade is characterized as CIN1. CIN1 means abnormal cells can be found on the surface of the cervix. CIN1 is not cancer, but can become cancer and spread to nearby tissue. [3]
* **High Grade:** 11 samples of high grade type were employed in this study. The samples characterized as high-grade were either of type CIN2/3 or CIS. CIN2/3is characterized by the presence of moderately abnormal cells on the surface of the cervix which may become cancer if left untreated. [3] CIS cells are considered stage 0 cancer. That is, they are located only on the top layer of the cervix. [3]

1. **Illumina Infinium HumanMethylation450 BeadChip (450k array)**

The 450K array investigates a total number of 485,577 sites on the genome [4]. Sites investigated by the array include (but are not limited to)[[1]](#footnote-1) [5]:

* RefSeq genes
* CpG islands
* CpG shores
* CpG island shelves
* Non-CpG loci

The process of gathering methylation data using the 450K array goes as follows:

1. **Bisulfite Treatment/Conversion:** The DNA sample is treated with bisulfite. In the presence of bisulfite, unmethylated cytosines are deaminated to uracil while methylated cytosines are not affected by the presence of bisulfite and remain cytosines. This step takes approximately 16 hours and involves a thermocycling process [6].
2. **Whole Genome Amplification (WGA):** Following bisulfite conversion, the genome is amplified. That is, copies of the genome are made [6].
3. **Enzymatic Fragmentation:** The copies of the genome are fragmented into predetermined lengths and sections [6].
4. **Hybridization:** During hybridization, the bisulfite converted sections of DNA anneal to locus specific DNA oligomers linked to bead types (methylated or unmethylated) [6]. The hybridization directly relates to the assay scheme, which is detailed further below as its design directly affects the statistical procedures used to analyze the data.
5. **Fluorescent Staining:** Following the hybridization step, the array is fluorescently stained and scanned. The stain gives off either a red or green color based on the state of methylation at a specific locus. The intensities read during scanning are converted to beta values, which are quantitative readings utilized in further analysis [6].

As mentioned above, the 450K assay scheme directly affects the statistical procedures used during analysis. The Infinium Methylation 450K Assay scheme involves the inclusion of two different assays. Termed Infinium I and II, the two assays have significantly different designs. Each assay outputs a color intensity after staining (either red or green) which reflects the methylation status of the interrogated site.

The following details the design of the Infinium I assay:

The Infinium I assay interrogates 135,501 cytosine sites [4].Two bead types correspond to each CpG locus. One bead type corresponds to the methylated state of the CpG site and one bead type corresponds to unmethylated state of the CpG site. The design of this assay assumes that the methylation status of adjacent CpG sites is correlated [7]. The design of the assay is detailed in more depth in Figure 1.

The following details the design of the Infinium II assay:

The Infinium II assay interrogates 350,076 cytosine sites [4]. One bead type corresponds to each CpG locus. Methylation status is detected by single-base extension. Each locus will be detected in two colors. The design of the assay does not assume that adjacent CpG sites are correlated [7]. The assay design is detailed in greater depth in Figure 1.

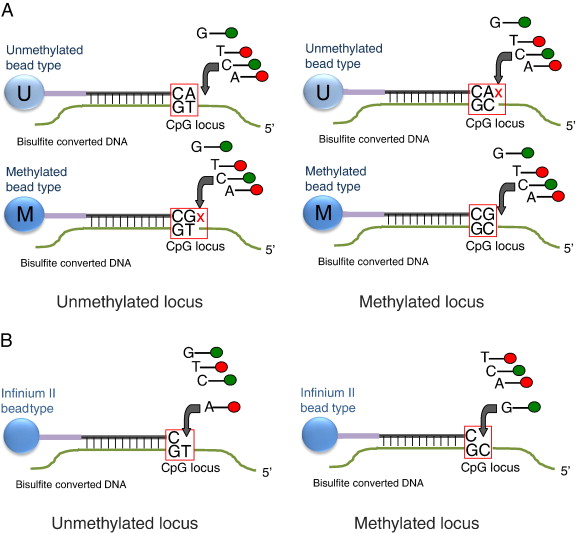


Figure 1- The figure above details the probe designs associated with the Infinium I and II assays. ‘A’ details the Infinium I assay design and ‘B’ details the Infinium II design [7]

Because of the difference in probe designs, the methylation measurements taken by the Infinium I and II probes are not identical [8]. Thus, these differences must be rectified during data processing.

1. **Standard Data**

The results of the implementation of the combined assay are converted from intensity measurements to a standard output [6]. The standard output is a β value which is calculated from the intensities of the methylated and unmethylated allele. The β can be represented mathematically by:

where Max(M,0) represents the maximum number of methylated sites, Max(U,0) represents the maximum of unmethylated sites and the addition of 100 stabilizes the estimate [1].

Conceptually, the β value represents the methylation level (percentage) for a specific location. Thus, β values vary from 0 to 1. A value of 0 indicates that no copy of the CpG site is methylated. Meanwhile, a β value of 1 indicates that every copy of the CpG site is methylated [1].

An alternative measurement of the methylation level of a site does exist. It is referred to as the M value. M values can vary from negative infinity to positive infinity. An M value approaching 0 suggests that the methylation rate is 50% (that is, 50% of the copies are methylated). Thus, positive M values suggest a methylation rate greater than 50% and a negative M value suggests a methylation rate less than 50% [1]. Mathematically, the M value is given to be [1]:

Β values can be converted to M values via the following relationship: [1]

The decision to use M or β values lies with the researcher as a matter of preference.

**Methodology**

1. **Representation of Data in R**

The methylation data is first stored in IDAT files. Each sample (person) has two IDAT files which correspond to their methylation status. One file is a ‘green’ file and one is a ‘red’ file. The ‘green’ file conveys the methylation status of each investigated site in terms of the green intensities which were output. The ‘red’ file does the same but in relation to the red signals. The files for methylation levels are binary [9]. That is, they consist of strings of 0’s and 1’s.

It is necessary to convert the IDAT files into usable methylation measurements (β or M values). This is done by converting the files first into a different data class. The data is first transferred from IDAT to files to an RGChannelSet class. The class represents the unprocessed data. The class allows the user to access the data of each color channel in matrix format [10]. Though using the data in such a format does not allow direct access of usable methylation measurements, it does allow the user to feed the data through quality control and processing commands which could not be accomplished with 2 individual files for each sample.

In later processing steps, it is of great importance that the user is able to assess the methylation status of each loci by converting the data to usable methylation measurements. This is accomplished by converting the data from the RGChannelSet class to the MethylSet-Class [10]. The β and M values at each loci may be accessed from this class of data.

For the purpose of data analysis and processing steps, the Bioconductor and minfi extension packages of R were employed.

1. **Quality Control**

The importance of quality control cannot be overstated. It allows the researcher to:

* Detect and remove low quality samples
* Perform corrections to the data which normalization procedures will not catch
* Improve downstream results

In the minfi package, quality control is performed through the output of a quality control report [8]. Contained in the report are important plots such as the beta density plot, bean plots, and control probe plots.

Beta density plots show the density of the methylation beta values for all samples. The plot is generally colored by sample group. They are useful for identifying deviant samples. However, it may be difficult to pinpoint the deviant sample [8]. A follow-up analysis of other quality control plots is necessary. An example of a beta density plot is given below in Figure 2.



Figure 2- Shown above is an example of a standard beta density plot. It can be seen that the plot possesses two modes—one near 0% methylation and one near 100% methylation, which is expected. [8]

Bean plots allow the researcher to identify specific ‘problem’ samples [8]. An example bean plot is given below in Figure 3.



Figure 3-Shown above is a standard bean plot. Bean plots allow researchers to pinpoint deviant samples. The samples are colored by group [8].

It should be noted that a normal sample is characterized by 2 modes—one close to 0% methylation and a second close to 100% methylation [8]. Thus, a deviant sample is defined to be one lacking in shape (likely lacking 2 modes) similar to the other samples.

Control probe plots vary in type and significance. However, in practice, the most important are the bisulfite conversion plots [11]. The plots detail the success (or failure) of the control probes to be converted correctly during the bisulfite conversion step detailed above. [12] That is, on the BeadChip array, there are quality control probes whereby the site interrogated is known to be methylated (and should thus remain a cytosine as per the bisulfite conversion step) [12]. The quality of the sample can be assessed by measuring outcome from the quality control probes. If the intensity level at a site known to be methylated does not reflect the outcome predicted, the quality of the sample may be called into question. A problem may have occurred during bisulfite conversion. In this way, this gives researchers another method with which to detect deviant samples. A standard bisulfite conversion plot is depicted in Figure 4 below.



Figure 4-Standard bisulfite conversion plot [8]

1. **Preprocessing/Normalization**

The preprocessing and normalization process is paramount to correct for differences which may propagate incorrect identification of differentially methylated regions and/or sites during downstream analysis. Normalization involves the removal of random noise and technical and systematic variation caused by microarray technology [13].

Within the minfi package, there are three preprocessing/normalization options which were focused on in this study:

* **Raw preprocessing method:** performs no correction to the data. (This is not normally recommended in any setting) [8].
* **preprocessIllumina method**: performs control term normalization and background correction. The following are brief descriptions of these procedures [8].

**-Control Term Normalization:** the probe intensity is multiplied by a constant normalization factor (NF). The program requires the use of a reference array to calculate the normalization factor [13, 14]. The user can select the reference array using the minfi package [10].

**-Background Correction:** the mechanics of the technique vary from one software package to another. Regardless of mechanics, the correction helps to remove nonspecific signals from the total signal and corrects for between-array artefacts [13].

* **SWAN method:** SWAN stands for subset-quantile within array normalization. Technical differences between Infinium I and II have been discussed above. This technical variability can be substantially reduced by employing SWAN. The method takes into account biologic likelihood of multiple CpGs within 50 base pairs of one another. The method then uses a subset quantile normalization approach to adjust the intensity readings of each array [8].

1. **Downstream Analysis**

There exist what seems to be countless statistical methods to employ when analyzing DNA methylation data. For this study, the focus has been placed on two methods which are detailed below.

* + 1. **F-test**

Consider an F-test in a one-way analysis of variance model. This test is used to determine whether the expected values of a quantitative variable differ between groups [15].

Within the context of this project, a one-way analysis of variance (ANOVA) model can be used to determine whether the expected values of methylation level differ between high and low grade samples at a cytosine site level.

Therefore, let the null hypothesis be: , and let the alternative hypothesis be: where represent the average methylation level across Group A and Group B at an arbitrary cytosine site, respectively. The null hypothesis is thus that the average methylation levels across the groups at one cytosine site are not statistically different. The alternative hypothesis is thus that the methylation levels at one cytosine site are significantly different from one another. This test would be recursively performed across all cytosine sites.

The ANOVA table in the case of high and low grade HIV positive samples can be represented as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum of Squares | Degrees of Freedom | MS | F-Statistic |
| Grade (High or Low) |  | 1 |  |  |
| Error |  | 18 |  |
| Total |  | 19 |  |

Table 1-One way ANOVA table

Under the null hypothesis, the F-statistic follows Snedecor’s F distribution with 1 degree of freedom in the numerator and 18 degrees of freedom in the denominator. This is equivalent to a squared t-distribution with 18 degrees of freedom.

In the context of minfi, this F-test can be run recursively over the entire experiment by employing the dmpFinder function (with the variance shrinkage off) which detects differentially methylated positions [10].

* + 1. **Moderated F-test (empirical Bayes)**

The moderated F-test uses an empirical Bayes method to shrink the variance of the data [16]. In minfi, using the dmpFinder function with variance shrinkage set to true, empirical Bayes posterior means are calculated to the shrink the variance [10].

The premise of the empirical Bayes method is that better information about a parameter can be obtained by pooling information across an experiment. Thus, in the context of our experiment, the test statistic will utilize data from all cytosine sites to estimate the variance.

First, we assume some prior distribution on . In this case, we use a scaled distribution .

Assume: where represents the degrees of freedom and represents the estimate of the variance and k represents the kth cytosine site investigated.

In the empirical Bayes method, we estimate from the data using method of moments estimation from all cytosines.

A posterior mean is calculated based on our assumptions above giving:

In this way, we are able to account for global ( and local ( ) variances. As gets larger, the global variance estimate provides more influence.

Thus, this where the term “variance shrinkage” comes from in the minfi documentation [8].

Assuming independence, for samples of size 6 and 12, the power of the empirical Bayes method is higher than that of the F-test across a range of truly methylated cytosines [1].

Assuming independence across a range of truly methylated regions, the stability of the empirical Bayes method outranks the F-test for samples of size 6 and 12 [1].

For our purposes, these results can be extrapolated to samples of size 9 and 11.

* + 1. **Multiple testing correction**

Upon identification of differentially methylated regions through the aforementioned tests, it is important to employ a multiple testing correction procedure. This reduces the likelihood of identifying false-positive loci. The procedure adjusts confidence measures to account for the number of tests performed [1].

|  |  |  |  |
| --- | --- | --- | --- |
|  | Number not rejected | Number rejected |  |
| True null hypothesis | T | U | m0 |
| Non-true null hypothesis | V | W | m-m0 |
| Total | m-R | R | m |

Table 2-Possible outcomes from m hypothesis tests [1]

Considering Table 2 above, we can define the false discovery rate to be:

[1]

A second form of the false discovery rate is referred to as the proportion of false discoveries. It can be defined as:

[1]

Two different procedures can be used to control either the FDR or the pFDR. The first is the Benjamini-Hochberg procedure. The Benjamini-Hochberg procedure gives strong control of the false discovery rate at the significance level, α. Similarly, the Storey’s q-value procedure uses point estimators of the total number of true null hypotheses, m0. The Storey’s q-value procedure has higher power than the Benjamini-Hochberg procedure and controls pFDR at significance level α when the test statistics are independent or weakly correlated [1].

**Results**

1. **Quality Control**

When working with the high grade and low grade HIV positive samples, it was necessary to perform the aforementioned quality control procedures. The first quality control results suggested that there existed two deviant samples. This result was confirmed by each of the three standard plots. The beta plot, bean plot and bisulfite conversion plots are shown below in figures 5-7.

In all subsequent plots, low-grade samples are denoted as Group A (or simply ‘A’) and high-grade samples are denoted as Group B (or simply ‘B’).



Figure 5-Beta plot of low and high-grade HIV+ samples

Above, in the beta plot (Figure 5), it is easy to see one sample from Group A which seems to be clearly deviant, lacking the expected bi-modal shape. Less obvious, however, is the existence of a second deviant sample from Group B. The existence of a second deviant sample can be made by looking at the corresponding bean plot.



Figure 6- Bean plot of high and low-grad HIV+ samples

It is demonstrated in the bean plot above (Figure 6) that samples P2 and P8 deviate significantly from the expected bimodal shape. The deviation of these samples is confirmed by the quality control probe bisulfite conversion plot.



Figure 7-Bisulfite conversion control probe plot

Shown above is the bisulfite conversion control probe plot (Figure 7) which corresponds to the data. It can bee seen that P2 and P8, the suspected deviant samples, are deviant with respect to the control probes. This indicates that perhaps there was a problem during the bisulfite conversion step of DNA processing.

Based on the results detailed above, the samples P2 and P8 were removed from the dataset and thus excluded from the analysis.

Quality control reports generated after removing P2 and P8 showed no further deviant samples.

1. **Preprocessing/Normalization**

In order to evaluate which preprocessing method may be best the dataset, each procedure was run and the success was evaluated via plots. The results are detailed below in figures 8-10.

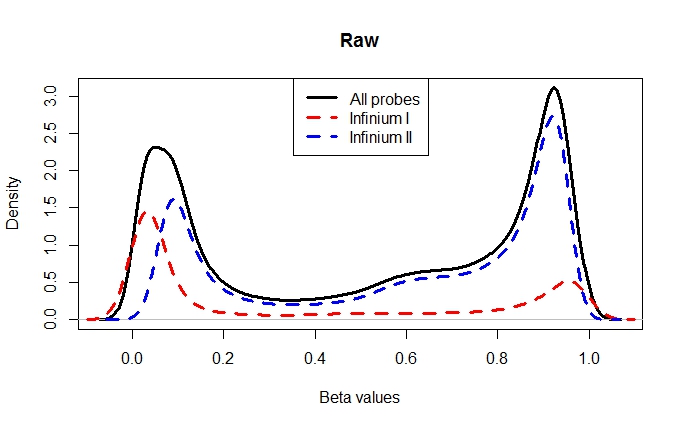


Figure 8-Plot of raw preprocessed beta values



Figure 9-Plot of illuminaPreProcess beta values

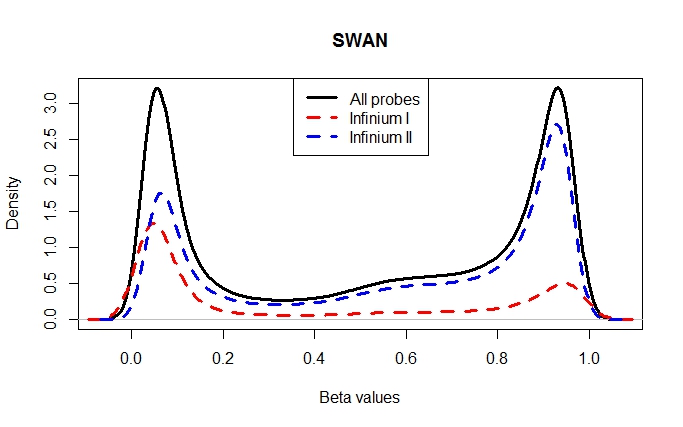


Figure 10--SWAN normalized beta values

By parsing through the graphical results above, the best preprocessing method for the data was chosen.

Raw preprocessing (Figure 8) is rarely used and not suggested based on standard research practice. This is because taking the data without any correction does not account for differences in array design, etc. Thus, the data received from raw preprocessing cannot be accurately analyzed group to group because corrections have not been applied. Therefore, the raw preprocessing method was not carried through to the analysis step.

The preProcessIllumina method (Figure 9) seemed to overcorrect the data and place inordinate weight upon the mode near 0% methylation. To avoid an overcorrection, the preProcessIllumina method was not carried through to the analysis step.

It was of paramount importance that the data be corrected to account for differences between the Infinium I and II probes, thus making the data comparable. Because of this need and because the SWAN method did not make an apparent overcorrection (Figure 10), the SWAN method processed beta values were chosen to be carried through the data analysis portion.

1. **Downstream Analysis**

**Results of Standard F-test:**

Because two samples were removed during quality control, the one-way ANOVA table for any cytosine level test can be modeled as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum of Squares | Degrees of Freedom | MS | F-Statistic |
| Grade (High or Low) |  | 1 |  |  |
| Error |  | 16 |  |
| Total |  | 17 |  |

Table 3-Adjusted One way ANOVA table

The results of the application of this test to the data returned no differentially methylated cytosine between Group A and B among 485,512 positions tested. Utilizing the Benjamini-Hochberg multiple testing correction procedure yielded a lowest p-value of .6882. Using the Storey’s q-value procedure yielded a lowest q-value of .4874.

**Results of Moderated F-test:**

Despite the additional power of this test over the standard F-test, no differentially methylated cytosine between high and low grade samples were detected among the 485,512 positions tested. Utilizing the Benjamini-Hochberg procedure gave a lowest p-value of .6653. Using the Storey’s q-value procedure yielded a lowest q-value .47951.

**Discussion**

Researchers at the Fred Hutchinson Cancer Research Center completed an independent analysis of this data set. The findings of the research and analysis presented here were confirmed during their analysis. Thus, the results above have been confirmed independently.

The results suggest that the methylation profiles of HIV positive individuals with different stages of pre-cancerous lesions (or early cancerous in the case of CIS samples) are not statistically different. This is an interesting result as methylation upon certain positions have been shown to be key in the progression of various cancers. Despite differences in power, neither statistical test employed detected any differentially methylated positions.

By analyzing carefully the 450K array setup, the researcher is able to recognize the importance of corrective measures for differences in probe types.

Similarly, though there are hundreds of different valid methods of statistical analysis, there is not a ‘one size fits all approach’ when dealing with DNA methylation data. One single approach does not correct all the statistical problems which one may encounter when dealing with DNA methylation data. That is, the researcher must make sacrifices in some areas for gains in power or stability in another method. Ultimately, many decisions must be made by the researcher about which statistical method is best for their dataset.

While great strides have been made in the creation of methods to properly analyze DNA methylation in a way which honors the underlying statistical theories and notions, there is still much work to be done.

**Future Work**

Work completed in the future pertaining to this project may include the use of a bumphunting method which detects differentially methylated regions rather than positions. The bumphunting method accounts for correlation among nearby methylated sites [10], which has been shown to be a reasonable assumption in DNA methylation data. The method has higher power to detect differentially methylated regions.

**Nomenclature**

Given below are definitions of terms referenced throughout the paper which may be in need of clarification:

* **Allele:** an allele is an alternative form of a gene at a given locus on the chromosome [17].
* **CpG island:** areas of a high density of cytosines and the corresponding base pair, guanine. There are roughly 45,000 CpG islands in humans and are usually found near the 5’ end of genes. The islands can be several hundred to several thousand base pairs long [17].
* **CpG island shelves:** the region less than 2,000 base pairs away from a CpG shore in either direction on the DNA strand. [5]
* **CpG site:** a cytosine followed immediately by a guanine in the 5’-3’ direction on the DNA strand [2].
* **CpG shore:** the region less than 2,000 base pairs away from a CpG island in either direction on the DNA strand [5].
* **RefSeq genes:** genes whose cytosines are interrogated based on a collection of previously sequenced reference genes. The RefSeqGene database defines genomic sequences that can be used as reference standards for well-studied genes [18].
* **Stability:** refers to the standard deviation of total discoveries during differential methylation analysis [1]

**Appendix**

Included in this appendix is a copy of the annotated R code used to process the data.

##Connect with bioconductor and install minfi extension and manifest files source(https://bioconductor.org/biocLite.R) biocLite("minfi") require(minfi) biocLite("IlluminaHumanMethylation450kmanifest")

##set working directory to file where the data is stored setwd("S:/Fred Hutch Data")

##view the files to verify proper data in file list.files()

##Convert the sample sheet with group data and basenames to a data frame df<- read.csv("InfiniumHDMeth\_Sample\_Sheet\_24Samples-1\_VP\_12-13-12\_FINAL.csv",header=TRUE)

##Display the data frame to verify proper reading df

##Read all Red/Green IDAT files into an Reset class ##targets is basename column in dataframe RGset<-read.450k.exp(targets=df)

##display details on RGset. RGset pd<-pData(RGset) pd[,1:4]

##generate quality control report pdf qcReport(RGset, sampNames=pd$Sample\_Name,sampGroups=pd$Sample\_Group, pdf="qcReport.pdf")

##run raw preprocessing procedure MSet.raw<-preprocessRaw(RGset)

##assign raw values to new variable MsetEx<-MSet.raw

##assign copy of RGset to new variable RGsetEX<-RGset

##preprocess data using preprocessIllumina method MsetEx<-preprocessIllumina(RGset)

##set seed for SWAN procedure set.seed

##preprocess data using SWAN procedure Mset.swan<-preprocessSWAN(RGsetEX, MsetEx)

##output beta values associated with SWAN procedure getBeta(Mset.swan)

##Generate Raw plot plotBetasByType(Mset.raw[,1], main=“Raw”)

##Generate preprocessIllumina plot plotBetaByType(MSetEx[,1], main=“PreprocessIllumina”)

##Generate SWAN plot plotBetasByType(Mset.swan[,1],main="SWAN")

##View details of SWAN corrected data Mset.swan

#calculate detection p-values pVals=detectionP(RGset)

##filter out probes that have failed in one or more samples based on detection p-value MSetSw=Mset.swan[rowSums(pVals<= 0.01)==ncol(pVals),]

#find differentially methylated positions using EB procedure dmp=dmpFinder(MSetSw, pheno=as.vector(RGsetEX$Sample\_Group), type="categorical", shrinkVar=TRUE)

#find differentially methylated positions using standard F test dmp1=dmpFinder(MSetSw, pheno=as.vector(RGsetEX$Sample\_Group), type="categorical", shrinkVar=FALSE)

#save output to text files write.table(dmp, "S:/DMP.txt") write.table(dmp1,"S:/DMP1.txt")

#calculate Benjamini-Hochberg adjusted p-values d<-dmp[,”pval”] pvalue<-p.adjust(d, “BH”) d1<-dmp1[,”pval”] pvalue1<-p.adjust(d1,”BH”)

**Acknowledgements**

Many thanks must be given to the Fred Hutchinson Cancer Research Institute for their willingness to share their data.

This project would not have been possible without the guidance and knowledge of Dr. Gayla Olbricht.

# References

|  |  |
| --- | --- |
| [1] | D. Li, Z. Xie, M. Le Pape and T. Dye, "An evaluation of statistical methods for DNA methylation microarray data analysis," *BMC Bioinformatics,* vol. 16, no. 217, 2015. |
| [2] | M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen and R. A. Irizarry, "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays," *Bioinformatics,* vol. 30, no. 10, pp. 1363-1369, 2014. |
| [3] | N. I. o. Health, "NCI Dictionary of Cancer Terms," National Institute of Health, [Online]. Available: www.cancer.gov/publications/dictionaries/cancer-terms. [Accessed March 2016]. |
| [4] | R. Pidsley, C. C. Y. Wong, M. Volta, K. Lunnon, J. Mill and L. C. Schalkwyk, "A data-driven approach to preprocessing Illumina 450K methylation array data," *BMC Genomics,* vol. 14, no. 293, 2013. |
| [5] | M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, J.-B. Fan and R. Shen, "High density DNA methylation array with single CpG site resolution," *Genomics,* vol. 98, pp. 288-295, 2011. |
| [6] | D. J. Weisenberger, D. Van Den Berg, F. Pan, B. P. Berman and P. W. Laird, "Comprehensive DNA Methylation Analysis on the Illumina Infinium Assay Platform," illumina, 2008. |
| [7] | illumina, "HumanMethylation450 BeadChip Achieves Breadth of Coverage Using Two Infinium Chemistries," 2012. [Online]. Available: http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote\_hm450\_data\_analysis\_optimization.pdf. [Accessed September 2015]. |
| [8] | K. D. Hansen and M. J. Aryee, "The minfi User's Guide Analyzing Illumina 450k Methylation Arrays," 9 October 2011. [Online]. Available: http://www.bioconductor.org/packages/2.11/bioc/vignettes/minfi/inst/doc/minfi.pdf. [Accessed 2015]. |
| [9] | M. L. Smith, K. A. Baggerly, H. Bengtsson, M. E. Ritchie and K. D. Hansen, "illuminaio: An open source IDAT parsing tool for Illumina microarrays," 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3968891/. |
| [10] | K. D. Hansen, M. Aryee, R. A. Irizarry, A. E. Jaffe, J. Maksimovic, E. A. Houseman, J.-P. Fortin, T. Triche and S. V. Andrews, "Analyze Illumina's methylation arrays," 2016. [Online]. Available: https://bioconductor.org/packages/3.3/bioc/manuals/minfi/man/minfi.pdf. [Accessed 2016]. |
| [11] | J.-P. Fortin and K. D. Hansen, "Minfi tutorial BioC2014," July 2014. [Online]. Available: https://www.bioconductor.org/help/course-materials/2014/BioC2014/minfi\_BioC2014.pdf. [Accessed February 2016]. |
| [12] | CORDIS, "Community Research and Development Information Service," [Online]. Available: http://cordis.europa.eu/docs/results/279/279143/final1-final-report-combined-tables-and-figures-and-references.pdf. |
| [13] | C. S. Wilhelm-Benartzi, D. C. Koestler, M. R. Karagas, J. M. Flanagan, B. C. Christensen, K. T. Kelsey, C. J. Marsit, E. A. Houseman and R. Brown, "Review of processing and analysis methods for DNA methylation array data," *British Journal of Cancer,* pp. 1394-1402, 27 August 2013. |
| [14] | illumina, "GenomeStudio Methylation Module v1.8 User Guide," November 2010. [Online]. Available: http://support.illumina.com/content/dam/illumina-support/documents/myillumina/90666eaa-0c66-48b4-8199-3be99b2b3ef9/genomestudio\_methylation\_v1.8\_user\_guide\_11319130\_b.pdf. [Accessed March 2016]. |
| [15] | Laerd Statistics, "Laerd Statistics," [Online]. Available: https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php. |
| [16] | G. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology,* vol. 3, no. 1. |
| [17] | National Institute of Health, "Genetics Home Reference," 21 March 2016. [Online]. Available: https://ghr.nlm.nih.gov/glossary. [Accessed 23 March 2016]. |
| [18] | National Center for Biotechnology Information, "RefSeqGene," 1 October 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/refseq/rsg/about/. [Accessed March 2016]. |
| [19] | J. D. Storey, "False Discovery Rates," 2010. |
| [20] | M. D. Robinson, A. Kahraman, C. W. Law, H. Lindsay, M. Nowicka, L. M. Weber and X. Zhou, "Statistical methods for detecting differentially methylated loci and regions," *Frontiers in Genetics,* vol. 5, no. 324, September 2014. |
| [21] | C. Bock, "Analysing and interpreting DNA methylation data," *Genetics,* vol. 13, pp. 705-719, October 2012. |
| [22] | J. Cloud, "Why your DNA Isn't Your Destiny," 6 January 2010. [Online]. Available: http://content.time.com/time/magazine/article/0,9171,1952313,00.html. [Accessed May 2015]. |

1. Further definitions of the above mentioned investigated sites are given in the nomenclature section [↑](#footnote-ref-1)