Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Qingyu Wang Date

Genome-wide DNA methylation profile change in cancer cell lines under stresses

By

Qingyu Wang
Degree to be awarded: Master of Science in Public Health

Department of Biostatistics and Bioinformatics

Zhaohui "Steve" Qin, Ph.D.
Thesis Advisor

Xiangqin Cui, Ph.D.
Reader

Genome-wide DNA methylation profile change in cancer cell lines under stresses

By

Qingyu Wang

Bachelor of Science
Fudan University
2018

Thesis Advisor: Zhaohui "Steve" Qin, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2020

Abstract

Genome-wide DNA methylation profile change

in cancer cell lines under stresses

By Qingyu Wang

As one of the most studied epigenetic mechanisms, DNA methylation profile provides additional information beyond DNA sequence. Differentially methylated regions have been found to be associated with various diseases. DNA methylation has also been shown to be linked to tumorigenesis. In this study, we take HeLa cell line and D54 cell line as examples to analyze the genome-wide DNA methylation profile change in the cells when they go through different stress conditions including cold shock (TEMP), nutrition depletion (FBS, DPR), chemotherapy agent (MeP) and low glucose (GLUC). We conducted the experiment in two batches and collected the methylation status data (β values) from several replicated samples of each condition (control and stresses). After applying quality control and normalization to the data, we identified the differentially methylated positions (DMPs) in each pairwise comparison group (each stress vs. control) using the combinations of functions in the "minfi" package and a non-parametric test with balanced permutation, and found their nearby regions (genes). Then we tried to find the overlapped DMPs and overlapped differentially methylated (DM) genes among all these comparisons.

Hundreds of DMPs were identified for each pairwise comparison. Since the methylation changes in CTRL vs. TEMP and CTRL vs. DPR group are not very consistent with other data, we only focused on the overlaps between HeLa MeP, HeLa FBS, HeLa Glucose and D54 MeP groups. Only one overlapped DMP was found between HeLa MeP and FBS groups, while there was none overlapped DMP was found between D54 MeP and other HeLa stresses. There are 72, 21 and 12 overlapped DM genes between HeLa MeP and FBS, HeLa Glucose and MeP and FBS, and D54 MeP and other HeLa, respectively. Thus, in our study, we conclude that HeLa and D54 cell lines share DMRs under different stress conditions. Also we found that the 12 overlapped DMRs is not related to the distribution of mutations in cancer cells. For further studies, we plan to find biological properties they share (such as pathways) and explore more about the functions and roles the overlapped genes have.

Genome-wide DNA methylation profile change in cancer cell lines under stresses

By

Qingyu Wang

Bachelor of Science
Fudan University
2018

Thesis Advisor: Zhaohui "Steve" Qin, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2020

Acknowledgment

This thesis would be impossible without the instruction from Dr. Zhaohui (Steve) Qin and Dr. Xiangqin Cui. I give my deepest gratitude to them for their guidance and patience. I also thank Dr. Eric J. Sorscher and Disha Joshi for offering this extraordinary idea and opportunity to do this project, and thank Xiangning Xue for her excellent work and support.

I also thank my family and friends for their support and companionship, and thank faculties and staffs from our BIOS department. Your encouragement and help mean so much to me.

Contents

1. Introduction

With the rapid development of high throughput biotechnologies over the past decades, people have paid more attention on Epigenetics to get more information that DNA sequence cannot provide. Among the epigenetic mechanisms, DNA methylation is the most studied one [1]. In mammals including human beings, DNA methylation is nearly only found in CpG dinucleotides sites, with methyl groups on the cytosine at position C5 usually [2, 3]. The differentially methylated regions (DMRs), which are the genomic regions with different DNA methylation status across samples, are observed to be related with various diseases [4]. Through the gain or loss of DNA methylation, imprinting disorders can happen [5], which are associated with many diseases, including Beckwith–Wiedemann syndrome, Prader–Willi syndrome, Angelman syndrome and transient neonatal diabetes mellitus [2]. Increased DNA methylation is considered to have a role in diseases associated with repeat instability such as Fragile X syndrome [2]. Also, researchers found three genomic regions with significant DNA methylation changes in postmortem brain tissues from patients with autism spectrum disorder [6]. The differential methylation in two clusters potentially increase the risk for rheumatoid arthritis [7].

Cancer is a global human health problem, and tied to stability of genes that control normal cell functions. DNA methylation is also a contributing factor to cancer with a specific

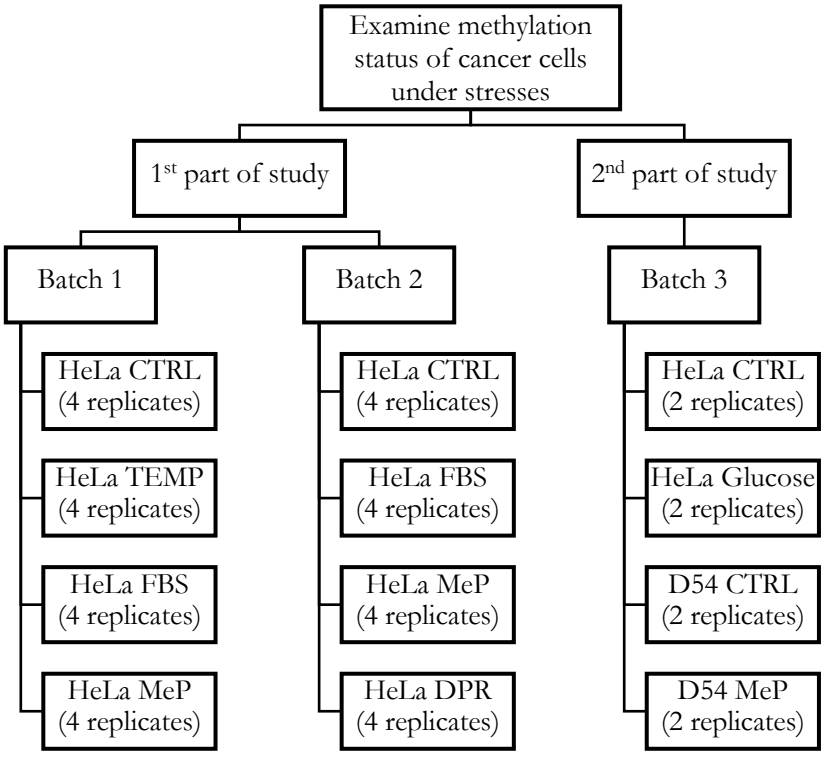
pattern: genome-wide hypomethylation and gene-specific hypermethylation happening at the same time in the same cell [2]. For example, differential methylation occurring in CpGs in gene DNMT3B contributes to the rising risk of cancers including breast and lung adenocarcinoma [8, 9]. Hypermethylation of promoter regions of several candidate genes has been suggested to be related to colon cancer carcinogenesis [10].

Although extensive research has been conducted on the association between differential methylation and the risk of cancer, few studies are on the differential methylation within cancer cells under external stress conditions. Due to the importance of DMRs in diseases, many studies have focused on developing statistical methods to identify them [11]. In this study, we take the HeLa cell line as an example to explore the differential methylations using statistical analysis methods. We will analyze the genome-wide DNA methylation profile change in the HeLa cell line when the cells go through different stress conditions including cold shock, nutrition depletion, chemotherapy agent and low glucose. We want to explore where differentially methylated regions are under these conditions and whether there is overlap among these regions.

2. Methods

2.1. Methylation Data

We did two parts of experiments to collect the methylation status data (β values) from replicated samples of each condition (control and stresses).



In the first part of this study, the first two batches of HeLa cells were divided to 4 groups respectively: for the first batch, there were one control group (CTRL) and three treatment groups under three different stress conditions (cold shock (TEMP), nutrition depletion (FBS), and chemotherapy agent (MeP)); for the second batch, there were one control group (CTRL) and another three treatment groups under three different stresses (nutrition depletion (FBS), chemotherapy agent (MeP), and nutrition restorage after depletion (DPR)).

In the control groups, HeLa (ATCC® CCL-2™) cervical carcinoma cells were grown in DMEM supplemented media with 10% fetal bovine serum, at 37°C with 5% CO_2, which is the standard HeLa cell culture condition. Each group in the first part was replicated for four times. Then the third batch of experiment was conducted in the second part of this study. In this part, there were 2 groups of HeLa cell line and 2 groups of D54 cell line. To be specific, there were one control group of HeLa cells, one low Glucose group (0.78% Glucose) of HeLa cells, one control group of D54 cells and one group with chemotherapy agent (MeP) of D54 cells, with two replicated samples of each. Genomic DNA extraction was conducted for each sample and Infinium EPIC 850K methylation array (Illumina Inc.) was used to measure the methylation status of 865,859 CpG sites, with input requirements of 500ng high molecular weight gDNA per sample. Then the Bioconductor minfi package [14] was used to process the raw data (IDAT files), as well as to do quality control, normalization and analyses.

2.2. Quality Control

After reading in the raw data and loading the 850K Annotation, the control probes in the Annotation were removed and several matrices were calculated. Then six steps of data filtering process were conducted. The first step was based on detection p-values. A probe was considered as failed if its p-value is above 0.01. Second, probes with < 3 beads in at least 5% of samples per probe were filtered out. Third, non-CpG probes contained in the dataset were filtered out. In the fourth and fifth steps, all SNP-related [12] probes and all multi-hit

probes [13] were filtered out. At last, all probes located in chromosome X and Y were filtered out because the difference between chromosome X and Y may make the differential CpG sites on them inaccurate. After filtering, 717,934 CpG sites remained for analyses. The beta values (β) of the CpGs denoted the estimate of their methylation level.

2.3. Quantile Normalization

Generally, probes can generate some biases **(Figure 1)**. In this case, normalization is paramount to adjust the bias. Quantile normalization method was used in this study. First, the ranks of the beta-values were determined among each group. Second, the means of beta-values from different groups for each rank were calculated. Third, the means were filled in to substitute the original values.

2.4. Methylation Change Analyses

After quantile normalization, two distinct methods for monitoring DMPs in pairwise comparisons were applied genome wide. The first method was to use the "dmpFinder" function in a minfi package to detect DMPs by testing each genomic position for association between methylation and a phenotype (control or stresses) with linear regression and the second method was to use a non-parametric test with balanced permutation, which resamples the data when doing tests.

For the first part of the study, both methods were applied. Balanced permutation method was used for resampling. For example, for the low FBS stress, two data frames, data 1 and data 2, were set up for the 4 control groups and 4 FBS groups in each batch. Firstly, permutation was done on data 1 and 70 combinations were generated. For each combination of data 1, secondly, all 70 combinations in data 2 were generated. Taken together, there were 4900 permutation combinations in total. Then t statistics were calculated for each combination and p-values were calculated for the true group labels. The thresholds for the minfi method was false discovery rate q-value thresholds while a nominal permutation p-value was used as the threshold for the non-parametric test.

For the second part of the study, since there were only one batch with two samples in each group, there are too few combinations for the balanced permutation to generate satisfying results. Thus, only the "dmpFinder" function in the minfi package was used to detect DMPs with the threshold of p-value < 0.01.

Since we wanted to know whether different stresses affect same DMPs and genes, we examined the overlaps of results from different pairwise comparison groups. Once we had identified the DMPs, we annotated them with their nearest genes. Then we explored the DMPs and differentially methylated genes shared among stresses.

3. Results

In the first part of the study, after using minfi package and the balanced permutation method, according to the threshold that q-value (calculated by test from "minfi") < 0.05 and p-value (calculated by balanced permutation test) < 0.001, there are 348 DMPs for CTRL vs. FBS comparison, and 475 DMPs for CTRL vs. MeP. Most of DMPs in the comparisons of HeLa FBS and HeLa MeP are hypomethylation. According to the threshold that q-value < 0.05 and p-value < 0.05/number of observations, there are 6009 DMPs for CTRL vs. temperature comparison, and 57078 DMPs for CTRL vs. DPR. Since the methylation data of low temperature have much more risks to be biased, and the data of DPR might not be consistent with other stresses, we focus on the overlaps between CTRL vs. MeP group and CTRL vs. FBS group. The numbers of genes found to be differentially methylated are 579 for CTRL vs. FBS, 772 for CTRL vs. MeP and 72 genes are overlapped between the two comparisons. The p-value distribution plots and QQ plots are shown for CTRL vs. FBS **(Figure 2(a))**, CTRL vs. MeP **(Figure 2(b))** and CTRL vs. DPR **(Figure 2(c))** comparisons. There is only one CpG site, named "cg24166386" in Chromosome 13, found as DMP in both CTRL vs. MeP and CTRL vs. FBS.

In the second part, at the beginning, correlation plots are generated to show the consistency of the experiment with the first part **(Figure 3)**. Then according to the threshold that p-value < 0.01, 5748 DMPs for HeLa CTRL vs. Glucose and 7131 DMPs for D54 CTRL vs.

MeP were found. The numbers of nearby genes are 3717 genes for HeLa and 4445 genes for D54. As for the overlaps between the second part of the study and the first part, we switched to identifying overlapping DMRs (the nearest genes of the DMPs) instead of DMPs since few overlapped DMPs exist. For instance, there is no overlapped CpG site found as DMP in all HeLa stresses (Glucose, FBS and MeP). Thus, the differentially methylated genes from HeLa Glucose pairwise comparison group vs. HeLa FBS, HeLa Glucose vs. HeLa MeP, and HeLa Glucose vs. Other HeLa stresses (72 overlapped differentially methylated genes between HeLa FBS group and HeLa MeP group, from previous results) are compared. Respectively, 154, 190 and 21 differentially methylated genes are found to be overlapped. There are 207 overlapped differentially methylated genes between D54 MeP vs. D54 CTRL comparison group from the second part of the study and the HeLa MeP vs. HeLa CTRL group from the first part. There are 12 overlapped differentially methylated genes **(Table 1)** in total for all stresses in HeLa (Glucose, FBS, MeP) and D54 (MeP). In order to see whether the number of the global overlaps (12 genes) is higher than the expected number of overlaps by chance, we applied a hypergeometric test and obtained a p-value of 0.0018. Since the p-value is smaller than 0.05 (significant level), we could say that the number of overlaps is significantly higher than expected. DMPs from D54 MeP vs. HeLa MeP in the first part (475 DMPs, from previous results) are compared and 6 overlaps have been found.

4. Discussion

4.1. Interpretation of Results

In this study, we aim to identify the differentially methylated CpG sites and regions when the cancer cells (HeLa cell line and D54 cell line) went through several stress conditions (TEMP, FBS, MeP, DPR and 0.78% Glucose). In addition, we would like to find the overlapped DMPs or DMRs between different stress conditions. We wonder if there are shared DMPs or DMRs under each stress and within both HeLa and D54 cell lines. As a result, we found a large number of DMPs and differential methylated genes (treated as DMRs here) for each stress vs. CTRL. However, there is no DMP that is common for all comparison groups. As for differential methylated genes, we found 12 overlaps between every comparison groups in both HeLa and D54 cell lines.

4.2. Further Studies

Some previous studies were focused on how certain differential methylated genes affect the risk of a healthy individual getting a cancer [8]. However, few study was about what genome-wide DNA methylation profile could be changed by outside stresses in a cancer cell. Our study addressed more about this problem, which could be meaningful for further treatment researches of human cancer.

For further studies, we should explore whether the differentially methylated genes share a common pathway. If so, it will get us a better knowledge about cancer and human genes. In addition, to further investigate the related properties and meanings of the overlapped DMPs and DMRs, we can also examine the counts of SNPs nearby the overlapped CpG sites. We can divide the DNA into distinct "bins" of a 5 kb length each and use the data of SNP counts from 1000 Genomes Project to obtain the number of SNPs in the bins that contained the overlapped DMPs. We will get more information about the overlaps by comparing the SNP counts in each "DMP bin" with the average SNP counts. We can try other cancer cell lines to compare with the current ones to verify the connection between DNA methylation and cancer as well.

4.3. Limitations

One limitation for the current study is that the results for now were not sufficient for us to conclude that the 12 overlapped genes play an important role in human cancer. We did not see whether the DMPs are "stable" in the standard condition or they would change their methylation status from time to time. Another limitation is that the amount of replicated samples of each group was not enough. Based on only two replicates for each group in the second part of the study, it was hard to calculate the convincing q-values (showing the false discovery rate). That might cause some biases when we were looking for DMPs.

References

1. Portela A, Esteller M. Epigenetic modifications and human disease. Nat. Biotechnol. 2010;28:1057–68.
2. Robertson, K. D. DNA methylation and human disease. Nature Reviews Genetics 2005; 6(8): 597-610.
3. Bird, A. DNA methylation patterns and epigenetic memory. Genes. Dev. 2002;16:6–21.
4. Mallik, S., Odom, G. J., Gao, Z., Gomez, L., Chen, X., & Wang, L. An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays. Briefings in Bioinformatics 2018;20(6):2224–2235.
5. Feinberg, A. P., Cui, H. & Ohlsson, R. DNA methylation and genomic imprinting: insights from cancer into epigenetic mechanisms. Sem. Canc. Biol. 2002;12:389–398.
6. Ladd-Acosta C, Hansen KD, Briem E, et al. Common DNA methylation alterations in multiple brain regions in autism. Mol. Psychiatry 2014;19:862-71.
7. Liu Y, Aryee MJ, Padyukov L, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat. Biotechnol. 2013;31:142-7.
8. Shen, H., Wang, L., Spitz, M.R., Hong, W.K., Mao, L., and Wei, Q. A novel polymorphism in human cytosine DNA-methyltransferase-3B promoter is associated with an increased risk of lung cancer. Cancer Res. 2002;62:4992–4995.

9. You, J. S. and P. A. Jones. Cancer genetics and epigenetics: two sides of the same coin?. Cancer cell 2012;22(1):9-20.

10. Lao V.V., Grady W.M.. Epigenetics and colorectal cancer. Nat. Rev. Gastroenterol. Hepatol. 2011;8:686-700.

11. Mallik, S., Odom, G. J., Gao, Z., Gomez, L., Chen, X., & Wang, L.. An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays. Briefings in Bioinformatics 2018;20(6):2224–2235.

12. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes. Nucleic Acids Research 2017;45(4):e22.

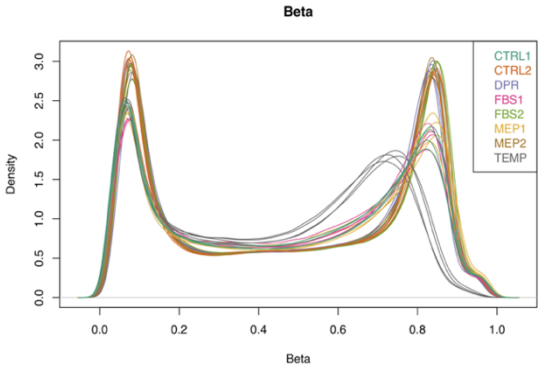
13. Nordlund J, Bäcklin CL, Wahlberg P, et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. Genome Biology 2013;14(9):r105.

14. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. Bioinformatics 2014;30(10):1363–1369.

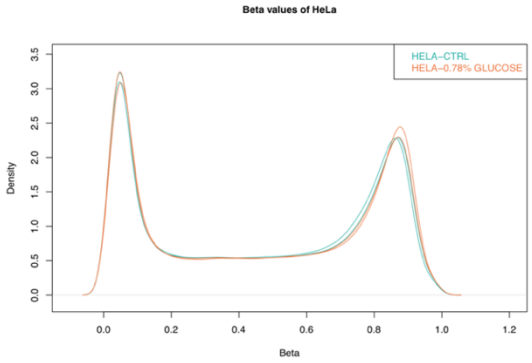
15. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Andrew F, Teschendorff AE. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics 2017;33(24):3982–3984.

Appendix. Figures and Tables

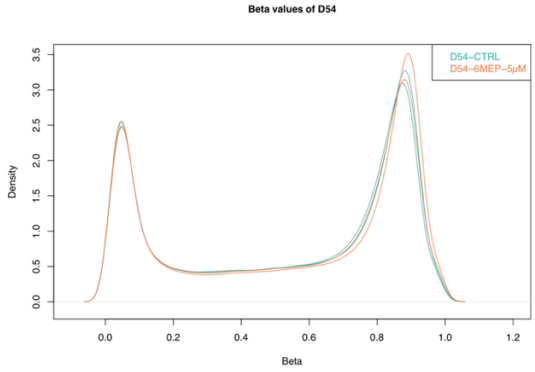
Figure 1. Beta-value distribution Plot



(a). for batch 1 & 2

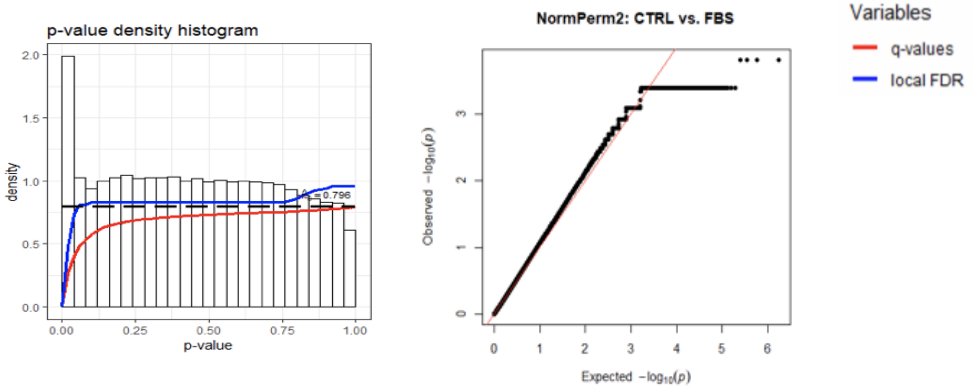


(b). for batch 3 – HeLa cell line

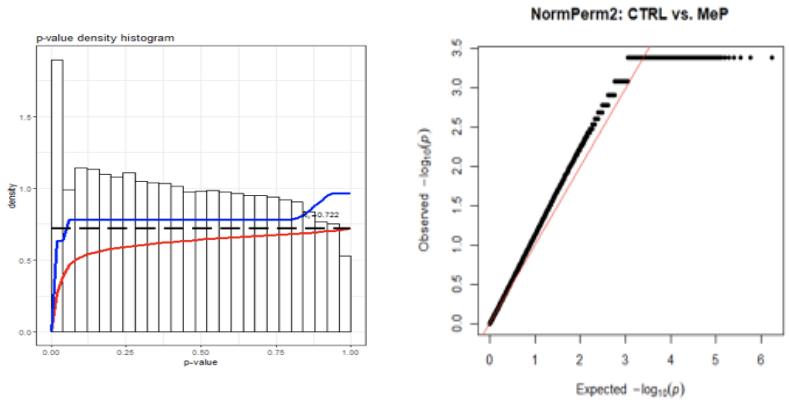


(c). for batch 3 – D54 cell line

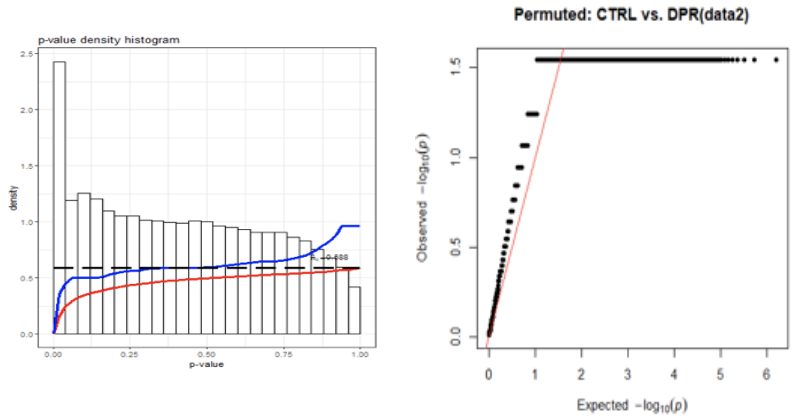
Figure 2. P-value distribution Plot and QQ plot



(a). for HeLa CTRL vs. FBS

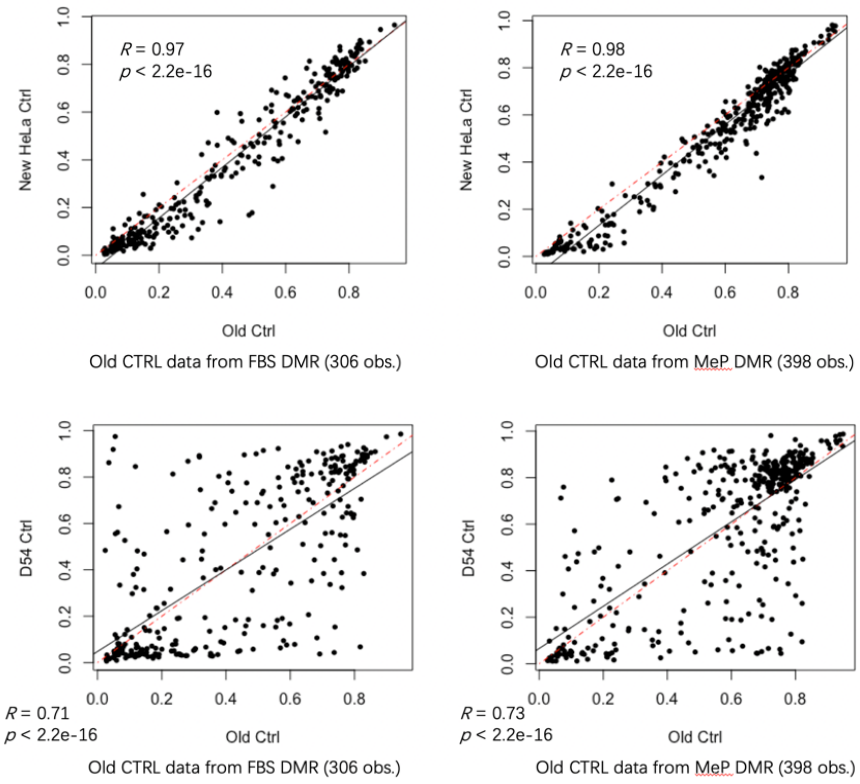


(b). for HeLa CTRL vs. MeP

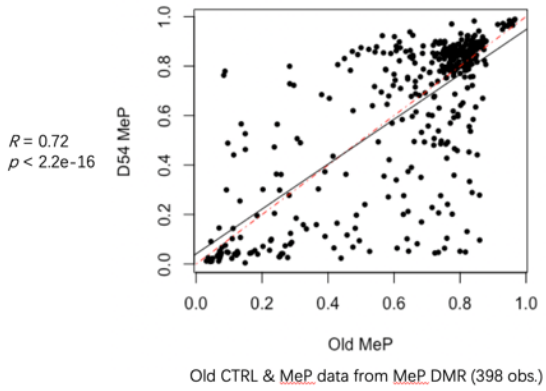


(c). HeLa CTRL vs. DPR

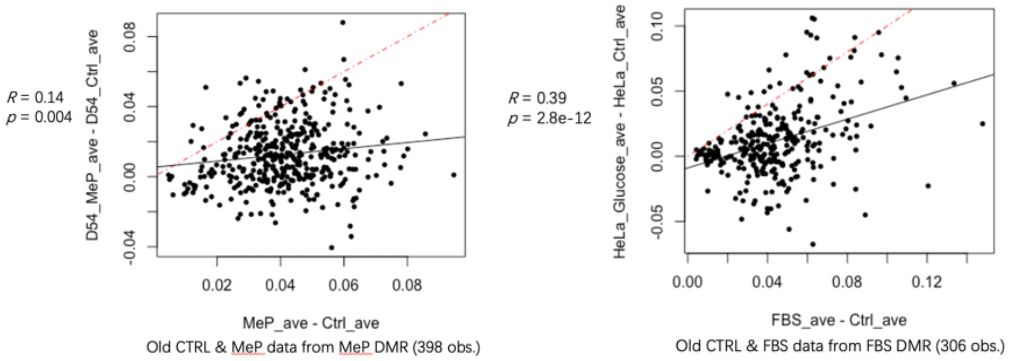
Figure 3. Correlation Plots between the 1st part experiments and the 2nd part experiments



(a). correlations between CTRL groups in the 1st part of study and the 2nd part of study



(b). correlations between HeLa MeP group and D54 MeP group



(c). correlations between methylation changes (beta difference compare to CTRL) of different stresses; the left one is the correlation plot of beta differences of D54 MeP vs. beta differences of HeLa MeP group; the right one is the correlation plot of beta differences of HeLa Glucose vs. beta differences of HeLa FBS group

Table 1. The 12 overlapped DM genes and their nearby DMPs

Gene Name	CpG Name	Methylation Change (Beta difference compared to CTRL)			
		HeLa MeP	HeLa FBS	HeLa GLUC	D54 MeP
CDH11	cg10853728	0.0091	0.0249	-0.0139	0.0580
	cg11271299	0.0256	0.0546	0.0929	0.0379
	cg18265326	0.0137	0.0175	-0.0145	0.0002
	cg18637626	0.0138	0.0130	-0.0521	0.0102
CDKAL1	cg06851325	-0.0224	-0.0118	-0.0852	-0.0050
	cg07196044	-0.0125	0.0072	0.0632	-0.0214
	cg08559711	-0.0287	0.0031	0.0322	-0.0326
	cg10474435	0.0341	0.0211	0.0158	0.0589
	cg14546778	0.0146	0.0244	-0.0048	-0.0027
	cg16076758	-0.0019	0.0134	0.0509	-0.0607
	cg17154807	0.0124	0.0369	-0.0008	-0.0765
	cg20790163	-0.0035	0.0724	-0.0252	-0.0122
	cg21909643	0.0487	0.0515	0.0481	0.0783
DLGAP2	cg00837987	0.0198	0.0032	-0.0003	0.0364
	cg02709139	0.0010	0.0052	0.0148	-0.0303
	cg11734845	0.0092	0.0009	-0.0667	0.0200
	cg11966432	0.0154	0.0298	0.0478	-0.0017
	cg20791311	0.0207	0.0215	0.0312	0.0353
	cg23042540	-0.0184	0.0017	-0.0539	0.0071
GALNT9	cg00012529	0.0426	0.0348	-0.0266	-0.0104
	cg01421019	0.0161	0.0129	-0.0115	0.0152
	cg07782603	0.0086	-0.0016	-0.0282	-0.0283
	cg09085411	-0.0471	-0.0300	0.1333	-0.0765
	cg10047181	0.0445	0.0278	0.0214	0.0372
	cg16767842	-0.0285	-0.0069	0.0273	-0.0253
	cg19325331	0.0362	0.0219	-0.0229	0.0600
	cg19748840	0.0090	-0.0203	-0.0722	0.0397
	cg23777173	-0.0092	0.0083	0.0080	0.0799
	cg23914694	0.0420	-0.0097	-0.1030	0.0200
MCF2L	cg07310677	0.0112	0.0059	-0.0057	0.0486
	cg12829183	0.0160	0.0245	0.0137	0.0326
	cg16959787	-0.0012	-0.0067	-0.0280	0.0641
	cg19641924	0.0010	-0.0052	-0.0156	-0.0612
	cg20487860	0.0022	-0.0026	-0.0595	0.0977
	cg20908740	0.0343	0.0075	-0.0329	0.0052
	cg26269162	0.0047	0.0089	-0.0096	0.0660

NFIA	cg07310946	-0.0375	-0.0374	-0.0767	0.0326
	cg12115385	0.0391	0.0670	0.0025	-0.0142
	cg14093103	-0.0038	-0.0091	-0.0459	0.0055
	cg24403479	0.0404	0.0994	0.0240	0.0372
	cg25369553	0.0088	0.0005	-0.0233	0.0805
SEMA3A	cg01571974	-0.0247	-0.0137	-0.0358	0.0602
	cg09943050	0.0496	0.0186	0.0091	-0.0077
	cg12862820	0.0092	0.0870	0.0262	-0.0034
	cg16346212	0.0119	0.0096	-0.0784	0.0184
	cg27307829	-0.0355	0.1231	0.0779	0.0180
TENM2	cg06659019	-0.0011	0.0187	0.0423	-0.0423
	cg17525385	0.0287	0.0011	-0.0349	0.0094
	cg19450817	-0.0099	0.0017	0.0772	0.0129
	cg23998048	0.0148	0.0050	-0.0342	0.0085
TNS1	cg01089914	0.0002	-0.0052	-0.0644	-0.0163
	cg05434674	0.0227	0.0238	-0.0425	0.0527
	cg07072618	0.0402	0.0154	-0.0268	0.0118
	cg20217899	0.0308	0.0320	-0.0380	-0.0061
	cg25928629	0.0374	0.0101	-0.0295	0.0010
TRIM2	cg08444115	0.0636	0.0316	-0.0309	-0.0122
	cg08651153	-0.0156	-0.0234	-0.0388	-0.0639
	cg16910246	-0.0307	-0.0249	0.0882	-0.0340
	cg20347377	0.0033	0.0128	-0.0009	-0.0439
UBASH3B	cg12883617	0.0305	-0.0048	0.0557	0.0531
	cg15826891	-0.0271	0.0701	0.1052	0.0081
	cg19535509	0.0597	0.0098	-0.0330	0.0303
	cg24387544	0.0129	0.0038	0.0866	0.0255
UNC5C	cg12832751	0.0025	0.0077	0.0104	0.0565
	cg13177026	0.0297	0.0421	0.0468	0.0058
	cg20947434	0.0269	0.0097	0.0271	-0.0151