

Data Analysis Report - DNA Methylation data analysis from Illumina HumanMethylation EPIC arrays

Service Request number - CFFMHS-CS-2685

Jyotirmoy Das*
email: jyotirmoy.das@liu.se

29 maj, 2023

Contact person: Jyotirmoy Das

Email: *jyotirmoy.das@liu.se*

Visiting Address:

Bioinformatics Unit,

Core Facility (KEF),

Floor 10,

Division of Cell Biology (CELLB)

SE-581 85 Linköping

Phone: 013- 28 4006

Mobile: +46 (0)79-065 53 87

Office address:

Room # 462.10.403, Entrance 71, US campus

*Bioinformatics Unit, Core Facility, Faculty of Medical and Health Sciences, Linköping University, Sweden

Contents

1	Service Request from iLab	3
1.1	General Information: Title and Short Description	3
1.2	Service includes	3
1.3	Data upload link	3
2	Data analysis	4
2.1	Data analysis logs	4
3	Data analysis method	5
	References	5
4	Data analysis result	6
4.1	List of result files	6
4.2	Tables	6
5	Session Information	7

CORE FACILITY

1 Service Request from iLab¹

1.1 General Information: Title and Short Description

In the Linköping Inflammatory response to Physical exertion (LIP) study, a standardized bicycle ergometer test have been used as a model for stress-induced inflammation in patients with a recent myocardial infarction (MI). Blood samples have been collected before and after the physical activity. Many different parameters have been measured, for example cortisol, ACTH, IL6 and IL18 in plasma as well as gene expression analysis in PBMC. Now we want to study if the DNA methylation levels, both globally and more specific for the glucocorticoid receptor genes, differ between cortisol responders (n=6) and cortisol non-responders (n=10) in this cohort. The age range of the 16 patients is 49-74 years.

Comment on the service request: In the “Sample sheet_LIP”, which is found in “link to the data”, a total of 8 confounding factors are included (where age and gender are two of them). Please, use that sample sheet instead of the one inserted above.

Explanation for the column smoking is: 0 = never smoked, 1 = previous smoker, 2 = smoker. gender: 1 = male, 2 = female

1.2 Service includes

- ☒ Quality control
- ☒ Normalization
- ☒ Variability analysis
- ☒ unsupervised clustering
- ☒ Dimension reduction
- ☒ Identification of differentially methylated positions (For DNA Methylation)

1.3 Data upload link

[dataLink](#)

¹Please read carefully the description before submit the service request.

2 Data analysis

2.1 Data analysis logs

- ☒ 23-05-26: Responded to Camilla's email query
- ☒ 23-05-26: Agreed on the service request on iLab. *Researcher needs financial approval. **Approved***
- ☒ 23-05-26: Data downloaded
- ☒ 23-05-26: Data unzipped
- ☒ 23-05-26: Sample_sheet prepared
- ☒ 23-05-26: Start running pipeline. ****ERROR: idat files missing! Emailed Camilla, waiting answer****
****Updated missing files****
- ☒ 23-05-26: Start running pipeline - Filtration.
 - Filtering probes with a detection p-value above 0.01. *****Removing 4232 probes.*****
 - Filtering probes with a beadcount <3 in at least 5% of samples. *****Removing 9513 probes*****
 - Filtering NoCG Start. Only Keep CpGs, *****removing 2959 probes from the analysis.*****
 - Filtering SNPs Start. Using general EPIC SNP list for filtering. Filtering probes with SNPs as identified in Zhou's Nucleic Acids Research Paper 2016. *****Removing 96245 probes from the analysis.*****
 - Filtering MultiHit Start. Filtering probes that align to multiple locations as identified in Nordlund et al *****Removing 11 probes from the analysis.*****
 - Filtering XY Start. Filtering probes located on X,Y chromosome, *****removing 16573 probes from the analysis.*****

*****All filterings are Done, now you have 736385 probes and 16 samples.*****

- ☒ 23-05-26: Start running pipeline -QC.
 - Plots generated - Hierarchical clusters, MDS for top 1000 probes and beta value density distribution.
- ☒ 23-05-26: Start running pipeline - Normalization -BMIQ.
 - normalized table saved.
- ☒ 23-05-26: Cell type heterogeneity - PBMC.
- ☒ 23-05-26: Start running pipeline - SVD.
 - batch effects found from Slides (Sentrix_ID), BMI, Gender, Hypertension_treatment.
- ☒ 23-05-26: Start running pipeline - Correcting batch effects. ****BMI is a confounding covariate with Sample Group** Emailed Camilla - waiting for answer. **meeting on Monday****
- ☒ 23-05-26: Start running pipeline - finding DMCs.
 - *****You have found 459 significant MVPs with a BH adjusted P-value below 0.05.***** Table generated.
- ☒ 23-05-26: Dimensional analysis - PCA.
 - before and after batch correction, figure generated PDF and PNG format.
- ☒ 23-05-29: Meeting with Camilla - set the BMI to 3 quadrant, Q1 (<25%), Q2 (>=25% to <75%) and Q3 (>=75%) and rerun the analysis of DMC. **0% 25% 50% 75% 100% 17.060 23.215 25.855 26.945 30.670**
- ☒ 23-05-29: the sample is collected from Whole Blood. The cell deconvolution is performed on the whole blood cells. **Cell type deconvolution run on 7 different blood cell types, B, NK, CD4T, CD8T, Monocytes, Neutrophils, Eusinophils.**
- ☒ 23-05-29: tables should be sent on Excel format.
- ☒ 23-05-29: After cell type deconvolution, SVD run found 3 confounding factors, gender, batches, and Hypertension_treatment (not BMI any more)
- ☒ 23-05-29: After batch effect correction with above mentioned factors, DMC analysis run with BH-corrected p-value < 0.05, yielded 0 significant DMCs. Further $p - value_{BH}$ set to 0.2 and yielded a total of ****You have found 24888 significant MVPs with a BH adjusted P-value below 0.2.****

3 Data analysis method

The IDAT files from Illumina® HumanMethylation EPIC arrays were analyzed using R (*v4.2.1*) (R Core Team, 2019) and bioconductor packages (*v3.16*), Chip Analysis Methylation Pipeline (ChAMP) analysis package (*v2.28.0*) (Tian *et al.*, 2017).

The files were pre-processed to filter out CpGs with detection p-value > 0.01, as well as SNP CpGs, unbound and multi-hit CpGs and CpGs from XY chromosome. A quality assessment on the filtered data was performed and using beta-mixture quantile normalization (BMIQ) function, normalized within dataset was calculated. The β - and M-values for each CpG per sample was estimated (**Figure- distribution plot; Table - QuantileNormalizedBetaValueFile850K**).

Since the samples were collected from Whole Blood cells, a cell type deconvolution was performed using the EpiDISH package (*v2.16*) (Zheng *et al.*, 2018) (**figure - CellTypeFractionation; Table - 1) CellTypeCorrectedBetaFile, 2) CellTypeFractionationDataFile**)

To reduce the batch effect in relation to biological variation on the data matrix, deconvolution (singular value decomposition, SVD) was performed on the normalized data using runCombat function (**Figure - SVD; Table - NormalizedFile_CellTypeCorrected_BatchCorrected**) and corrected against the confounding factors (e.g., Gender, Slides, Hypertension_treatment).

The differential methylation analysis was done on the corrected data with the linear modeling (lmFit) and eBayes algorithm between two sample groups (**Cortisol_responder_Vs_Cortisol_nonresponder**). The differentially methylated CpGs (DMCs) were considered significant at the Benjamini-Hochberg (BH)-corrected p-value ($p - value_{BH} < 0.2$). The resulted DMCs were annotated using AnnotationDbi package (*v1.60.2*) (Pages *et al.*, 2017) (Human Genome version 38) using the in-house script. (**Table- DMCresult_CellTypeCorrected_BatchCorrected_BHcorrected02**)

The hierarchical cluster analysis (**Figure hclust**) was performed using the Euclidean distance calculation within the ape package (*v5.7*) (Paradis *et al.*, 2004).

The principal component analysis (**Figure PCA**) was performed using FactoMineR (*v2.8*) and factoExtra (*v1.0.7*) packages with in-house R script.

All differences with a $p - value_{BH} < 0.05$ were considered significant if not otherwise stated. We calculated family-wise error rate (FWER) using the Benjamini-Hochberg (BH)-correction method. All analyses were performed in R (*v4.2.1*) with the mentioned packages.

References

- Pages, H. *et al.* (2017) Package ‘AnnotationDbi’. *Bioconductor Packag. Maint.*
- Paradis, E. *et al.* (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.
- R Core Team (2019) R: A language and environment for statistical computing.
- Tian, Y. *et al.* (2017) ChAMP: Updated methylation analysis pipeline for illumina BeadChips. *Bioinformatics*, **33**, 3982–3984.
- Zheng, S.C. *et al.* (2018) Identification of differentially methylated cell types in epigenome-wide association studies. *Nature methods*, **15**, 1059–1066.

4 Data analysis result

4.1 List of result files

1. QC - MDS plot, beta value density plot, hierarchical cluster as PDF format.
2. Normalized data- Normalized beta value table as TEXT format. Normalized distribution plots as PDF.
3. Variability analysis -
 - batchEffectCorrection and
 - Cell type heterogeneity.
4. Unsupervised clustering - before and after batch correction.
5. Dimension reduction - Principal Component Analysis, before and after batch correction.
6. DMPs - Table with **24888 BH-corrected significant CpGs**.
 - **Annotation version:** HG38.
 - **Significant level:** 0.2
 - **p-value correction method:** Benjamini-Hochberg (BH)

4.2 Tables

1. Normalized File - *QuantileNormalizedBetaValueFile850K*
2. Cell type correction - 1) *CellTypeCorrectedBetaFile*, \ 2) *CellTypeFractionationDataFile*
3. batch Effect Corrected - *NormalizedFile_CellTypeCorrected_BatchCorrected*
4. DMC - *DMCresult_CellTypeCorrected_BatchCorrected_BHcorrected01*

5 Session Information

R version 4.2.1 (2022-06-23)

Platform: x86_64-pc-linux-gnu (64-bit)

Running under: Ubuntu 20.04.6 LTS

Matrix products: default

BLAS: /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3

LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3

locale:

```
[1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C              LC_TIME=sv_SE.UTF-8      LC_COLLATE=en_US.UTF-8
[6] LC_MESSAGES=en_US.UTF-8  LC_PAPER=sv_SE.UTF-8     LC_NAME=C                 LC_ADDRESS=C
[11] LC_MEASUREMENT=sv_SE.UTF-8 LC_IDENTIFICATION=C
```

attached base packages:

```
[1] parallel stats4 stats graphics grDevices utils datasets methods base
```

other attached packages:

```
[1] viridis_0.6.2                viridisLite_0.4.1        dplyr_1.1.1
[4] magrittr_2.0.3              factoextra_1.0.7         ggplot2_3.4.1
[7] FactoMineR_2.6              dendextend_1.16.0       RColorBrewer_1.1-3
[10] ape_5.7-1                   stringr_1.5.0            ChAMP_2.26.0
[13] RPM_1.25                    cluster_2.1.4            DT_0.27
[16] IlluminaHumanMethylationEPICmanifest_1.0.0 Illumina450ProbeVariants.db_1.32.0 DMRcate_2.10.0
[19] ChAMPdata_2.28.0            minfi_1.42.0             bumhunter_1.10.0
[22] locfit_1.5-9.7              iterators_1.0.14         foreach_1.5.1
[25] Biostrings_2.66.0           XVector_0.38.0           SummarizedExperiment_1.20.0
[28] Biobase_2.58.0              MatrixGenerics_1.10.0    matrixStats_0.11.0
[31] GenomicRanges_1.50.2        GenomeInfoDb_1.34.9      IRanges_2.32.0
[34] S4Vectors_0.36.2           BiocGenerics_0.44.0
```

loaded via a namespace (and not attached):

```
[1] Hmisc_4.7-1                  svglite_2.1.0
[3] Rsamtools_2.12.0            crayon_1.5.2
[5] MASS_7.3-58.1               rhdf5filters_1.8.0
[7] nlme_3.1-159                backports_1.4.1
[9] sva_3.44.0                  impute_1.70.0
[11] rlang_1.1.0                 limma_3.54.2
[13] DSS_2.44.0                  filelock_1.0.2
[15] BiocParallel_1.30.4         rjson_0.2.21
[17] globaltest_5.50.0          bit64_4.0.5
[19] glue_1.6.2                  isva_1.9
[21] rngtools_1.5.2              methylumi_2.42.0
[23] AnnotationDbi_1.60.0       tidyselect_1.2.0
[25] XML_3.99-0.14              nleqslv_3.3.3
[27] tidyr_1.3.0                 zoo_1.8-11
[29] ggpubr_0.4.0                GenomicAlignments_1.32.1
[31] xtable_1.8-4                evaluate_0.20
[33] cli_3.6.0                   zlibbioc_1.44.0
[35] rstudioapi_0.14            doRNG_1.8.6
[37] rpart_4.1.16                ensembledb_2.20.2
```

[39]	IlluminaHumanMethylationEPICanno.ilm10b4.hg19_0.6.0	shiny_1.7.4
[41]	xfun_0.37	askpass_1.1
[43]	clue_0.3-61	multtest_2.52.0
[45]	KEGGREST_1.38.0	tibble_3.2.1
[47]	interactiveDisplayBase_1.34.0	ggrepel_0.9.3
[49]	base64_2.0.1	biovizBase_1.44.0
[51]	scrime_1.3.5	png_0.1-8
[53]	permute_0.9-7	reshape_0.8.9
[55]	withr_2.5.0	lumi_2.48.0
[57]	bitops_1.0-7	plyr_1.8.8
[59]	AnnotationFilter_1.20.0	JADE_2.0-3
[61]	coda_0.19-4	pillar_1.9.0
[63]	GlobalOptions_0.1.2	cachem_1.0.7
[65]	GenomicFeatures_1.48.4	multcomp_1.4-20
[67]	scatterplot3d_0.3-42	DelayedMatrixStats_1.18.0
[69]	vctr_0.6.1	ellipsis_0.3.2
[71]	generics_0.1.3	tools_4.2.1
[73]	foreign_0.8-82	munsell_0.5.0
[75]	emmeans_1.8.1-1	DelayedArray_0.24.0
[77]	fastmap_1.1.1	compiler_4.2.1
[79]	abind_1.4-5	httpuv_1.6.9
[81]	rtracklayer_1.56.1	geneLenDataBase_1.32.0
[83]	ExperimentHub_2.4.0	beanplot_1.3.1
[85]	Gviz_1.40.1	plotly_4.10.1
[87]	GenomeInfoDbData_1.2.9	gridExtra_2.3
[89]	DNAcopy_1.70.0	edgeR_3.40.2
[91]	lattice_0.20-45	deldir_1.0-6
[93]	utf8_1.2.3	later_1.3.0
[95]	BiocFileCache_2.4.0	jsonlite_1.8.4
[97]	affy_1.74.0	scales_1.2.1
[99]	carData_3.0-5	sparseMatrixStats_1.8.0
[101]	estimability_1.4.1	genefilter_1.78.0
[103]	lazyeval_0.2.2	promises_1.2.0.1
[105]	car_3.1-0	doParallel_1.0.17
[107]	latticeExtra_0.6-30	R.utils_2.12.0
[109]	goseq_1.48.0	checkmate_2.1.0
[111]	sandwich_3.0-2	rmarkdown_2.20
[113]	nor1mix_1.3-0	statmod_1.4.37
[115]	webshot_0.5.4	siggenes_1.70.0
[117]	dichromat_2.0-0.1	BSgenome_1.64.0
[119]	HDF5Array_1.24.2	bsseq_1.32.0
[121]	survival_3.4-0	yaml_2.3.7
[123]	systemfonts_1.0.4	htmltools_0.5.4
[125]	memoise_2.0.1	VariantAnnotation_1.42.1
[127]	BiocIO_1.6.0	quadprog_1.5-8
[129]	digest_0.6.31	mime_0.12
[131]	leaps_3.1	rappdirs_0.3.3
[133]	BiasedUrn_1.07	RSQLite_2.3.0
[135]	data.table_1.14.8	blob_1.2.4
[137]	R.oo_1.25.0	preprocessCore_1.58.0
[139]	fastICA_1.2-3	shinythemes_1.2.0
[141]	splines_4.2.1	Formula_1.2-4

[143] labeling_0.4.2	Rhdf5lib_1.18.2
[145] illuminaio_0.38.0	AnnotationHub_3.4.0
[147] ProtGenerics_1.28.0	RCurl_1.98-1.10
[149] broom_1.0.1	hms_1.1.3
[151] rhdf5_2.40.0	colorspace_2.1-0
[153] base64enc_0.1-3	BiocManager_1.30.20
[155] shape_1.4.6	nnet_7.3-17
[157] GEOquery_2.64.2	Rcpp_1.0.10
[159] mclust_6.0.0	mvtnorm_1.1-3
[161] circlize_0.4.15	multcompView_0.1-8
[163] fansi_1.0.4	tzdb_0.3.0
[165] R6_2.5.1	grid_4.2.1
[167] lifecycle_1.0.3	curl_5.0.0
[169] kpmt_0.1.0	ggsignif_0.6.3
[171] affyio_1.66.0	Matrix_1.5-1
[173] qvalue_2.30.0	TH.data_1.1-1
[175] ROC_1.72.0	org.Hs.eg.db_3.15.0
[177] IlluminaHumanMethylation450kmanifest_0.4.0	htmlwidgets_1.6.2
[179] biomaRt_2.52.0	purrr_1.0.1
[181] missMethyl_1.30.0	marray_1.74.0
[183] rvest_1.0.3	mgcv_1.8-40
[185] flashClust_1.01-2	openssl_2.0.6
[187] htmlTable_2.4.1	codetools_0.2-18
[189] IlluminaHumanMethylation450kanno.ilmn12.hg19_0.6.1	G0.db_3.16.0
[191] gtools_3.9.3	prettyunits_1.1.1
[193] dbplyr_2.3.2	R.methodsS3_1.8.2
[195] gtable_0.3.3	DBI_1.1.3
[197] watermelon_2.2.0	httr_1.4.5
[199] KernSmooth_2.23-20	stringi_1.7.12
[201] progress_1.2.2	reshape2_1.4.4
[203] farver_2.1.1	annotate_1.74.0
[205] xml2_1.3.3	combinat_0.0-8
[207] kableExtra_1.3.4	restfulr_0.0.15
[209] interp_1.1-3	readr_2.1.2
[211] BiocVersion_3.15.2	bit_4.0.5
[213] jpeg_0.1-9	pkgconfig_2.0.3
[215] rstatix_0.7.0	knitr_1.42