Supplemental Information for *Human methylome variation across Infinium 450K data on the Gene Expression Omnibus*

Sean K. Maden Reid F. Thompson Kasper D. Hansen Abhinav Nellore

Abstract

This document contains supplemental information for the manuscript *Human methylome variation across Infinium 450K data on the Gene Expression Omnibus*, including details about data access, aggregation, and analyses. Large data files are located at the recount.bio website. The `recountmethylation` R/Bioconductor companion package provides access to databases of compiled DNA methylation array data, including a user's guide and a data analyses vignette that reproduces several analyses from the manuscript.

Contents

Gene Expression Omnibus (GEO) data **2**
- Sample identification . . . 2
- Data acquisition . . . 2
- DNAm assay data . . . 2

Sample metadata **2**
- Preprocessing – GSE-wise annotations . . . 2
- Postprocessing – Learning formatted labels with regular expressions . . . 2
- Sample type predictions . . . 3
- DNAm model-based predictions . . . 3
- Metadata column descriptions . . . 3

Quality and summary metrics **4**
- BeadArray metrics . . . 4
- Detection p-values . . . 4
- Methylated and unmethylated log2 median signals . . . 4
- Genetic relatedness . . . 5

Statistical analyses **5**
- Summary statistics and statistical tests . . . 5
- Principal component analyses . . . 5
- GEO year-wise sample and study availability . . . 5
- Metadata concordance analyses and plot . . . 5

Works cited **6**

Gene Expression Omnibus (GEO) data

Sample identification

DNA methylation (DNAm) array samples were identified as published to the Gene Expression Omnibus (GEO) and available in the GEO Data Sets database as of March 31, 2019. The Entrez Utilities software (v.10.9) was used to quantify DNAm array sample availability by year and platform for 3 major Illumina BeadArray platforms (HM450K, HM27K, and EPIC/HM850K, Figures 1 and S1, see script `eqplot.py`).

Data acquisition

We used the Python programming language to develop a download management system to handle and version file downloads from GEO. We used this job management system to obtain GSM IDATs, GSE SOFT files, and other data using batch queries to the GEO Data Sets repository.

DNAm assay data

DNAm assay data were read from IDATs. Signals were read from sample IDATs into an R session using the `minfi` package (v.1.29.3, Aryee et al. (2014)). Descriptions of the principal data types are as follows:

- Red signal: Red color channel signal(s) for beads, read from sample IDATs.
- Green signal: Green color channel signal(s) for beads, read from sample IDATs.
- Methylated signal: Methylated signal amount calculated from red and green signals.
- Unmethylated signal: Unmethylated signal amount, calculated from red and green signals.
- Beta-value: Fraction of DNAm (M/[U + M + e]) calculated from red and green signals.

Raw/unnormalized signals were used for quality metric calculations, with red and green signals for BeadArray metrics and the log2 array medians of methylated and unmethylated signals. DNAm model-based predictions used the noob-normalized Beta-values (see below). These DNAm array data were stored as databases and are accessible using the companion `recountmethylation` R/Bioc package.

Sample metadata

Metadata preprocessing and postprocessing was performed for downloaded GSE SOFT files, which resulted in the metadata table used for manuscript analyses (Table S1, see `metadata` directory and `md_postprocessing.R` script in the repo).

Preprocessing – GSE-wise annotations

Preprocessing was performed on free text tags and labels extracted from sample characteristics columns in the GSE SOFT files. This was preformed on a GSE-wise basis, where the most frequent and informative labels were retained. Care was taken to retain only sample-specific information (see file `md-preprocess.rda`).

Postprocessing – Learning formatted labels with regular expressions

Preprocessed metadata were postprocessed by string pattern matching with regular expressions (Table S1, see file `md-postprocess.rda` and script `md_postprocessing.R`). Labels were applied hierarchically, wherein specific labels have general labels applied automatically and separated by semicolons. Resulting labels are all lowercase, and underscores replace spaces for applicable labels. For example, a label of "Peripheral blood" is converted into "peripheral_blood;blood".

Sample type predictions

GSM metadata were scraped from downloaded SOFT files into JSON-formatted files. GSE-specific data were removed, like experiment title and methods text. For example, the filtered JSON file for sample GSM937258 was:


```
[
{
  "!Sample_characteristics_ch1": "tissue site: Bone",
  "!Sample_characteristics_ch1.1": "tissue type: Tumor",
  "!Sample_characteristics_ch1.2": "sample id (short form for plotting): 24-6b",
  "!Sample_characteristics_ch1.3": "description: Xiphoid Met",
  "!Sample_source_name_ch1": "Prostate cancer metastasis",
  "!Sample_title": "Prostate cancer metastasis 16032"
}
]
```


JSON files were run using a fork of the MetaSRA-pipeline software repo (Bernstein et al. (2017)). This produced mapped ontology term labels and predictions for a series of sample types (Figure S2, see file `mdmap-gsm_35k.rda`). The most likely sample types were retained under the "sampletype" column in the postprocessed metadata. For applicable samples, cell line annotations were appended from the Cellosaurus resource (see file `ccformat.txt`).

DNAm model-based predictions

Model-based predictions of sex, age, and cell type fractions were derived using noob-normalized DNAm (Beta-values, see below) compiled from available GSM IDATs. Sex and cell fractions were predicted using the `getSex` and `estimateCellCounts` functions from the `minfi` package, and age predictions were performed using the `agep` function from the `wateRmelon` (v.1.28.0) package. The `getSex` and `estimateCellCounts` functions perform light preprocessing that may be influenced by input sample order. Repeated calculations with randomized sample ordering showed $> 95\%$ concordance for both predicted variables across repetitions.

Metadata column descriptions

Descriptions of the sample metadata columns (Table S1, file `md-postprocess.rda`) are as follows:

1. gsm : Sample record
2. gsm_title : Sample record title
3. gseid : Study Record ID
4. disease : Learned disease or study group term
5. tissue : Learned tissue type term
6. sampletype : Most likely sample type from MetaSRA-pipeline (`msraptype` is type label, `msrapconf` is prediction confidence, and `ccid`, `ccacc`, and `cccat` are cellosarus id, accession, and attribute info where applicable)
7. arrayid_full: Full array ID (format: `Sentrix ID_chip coordinate`)
8. basename : Basename for sample IDATs (e.g. common name for paired IDAT files)
9. age : Age from mined sample metadata (`valm` is age value, `unitm` is age units)
10. predage : Age (years) predicted from norm. Beta-values using wateRmelon::agep()
11. sex : Sex from mined sample metadata (M = male, F = female)

12. predsex : Sex predicted from unnorm. signal using minfi::getSex() (M = male, F = female)
13. predcell.CD8T : Predicted CD8T cell fraction from minfi::estimateCellCounts()
14. predcell.CD4T : Predicted CD4T cell fraction from minfi::estimateCellCounts()
15. predcell.NK : Predicted Natural Killer cell fraction from minfi::estimateCellCounts()
16. predcell.Bcell : Predicted Bcell fraction from minfi::estimateCellCounts()
17. predcell.Mono : Predicted Monocyte fraction from minfi::estimateCellCounts()
18. predcell.Gran : Predicted Granulocyte fraction from minfi::estimateCellCounts()
19. storage : Sample storage procedure, annotated from metadata (`FFPE` or `F`)

Quality and summary metrics

Several types of quality and summary metrics were generated from the DNAm assay data (Table S2). These were used to determine metric performances, performance difference across preparations, and study-wise sample sub-threshold frequencies (Figures 2, S3, and S4).

BeadArray metrics

Seventeen BeadArray metrics were calculated from red and green signals (Table S3, see script `beadarray_cgctrlmetrics.R`). Metric formulae and thresholds were determined by consulting the platform documentation ((“Illumina Genome Studio Methylation Module V1.8” 2010), (“BeadArray Controls Reporter Software Guide” 2015)) and related functions in the `ewastools` package (Heiss and Just (2018)), with the following notes on calculations:

- Use extension Grn A/T probes for system background.
- For metrics where denominators would be 0 for some samples, use a uniform denominator offset of 1.
- Use C and U 1-3 and 4-6 for Bisulfite Conversion I.
- Use probe address “34648333” and “43603326” (DNP, Biotin subtype) for Biotin Staining Background.
- For Specificity I, use PM, MM 1-3 for green signal, 4-6 for red signal.
- For Specificity II, use just probes S1-3, as probe S4 unavailable in control probe annotation.

Detection p-values

The quantities of probes below 3 detection p-value cutoffs (0.01, 0.05, and 0.1) were determined. Detection p-values were calculated using the `detectionP` function from the `minfi` package.

Methylated and unmethylated log2 median signals

Low log2 medians for M and U signal may indicate poor sample quality (Aryee et al. (2014)). Methylated (M) and unmethylated (U) signal log2 medians were calculated from the raw/unnormalized signals.

Genetic relatedness

Sample genetic identities were calculated for GSE records using the `call_genotypes` function from the `ewastools` package (Heiss and Just (2018)). This method uses a probabilistic model with data from high-frequency SNPs probed by the HM450K platform to determine whether samples share the same genetic identity. GSM IDs from the same GSE record that shared genetic identity were determined.

Statistical analyses

Summary statistics and statistical tests

Sample data were obtained, read, and analyzed programmatically using the R (v.4.0.0) and Python (v.3.7.0) languages in a CentOS 7 remote server environment. IDAT signals were read into `SummarizedExperiment` objects using the `minfi` package. Summary statistics were generated using base R functions. Statistical tests were performed using the `stats` (v.4.0.0) R package. Correlation tests used the Spearman method by setting `method = "spearman"` in the `cor.test` function. Analyses of variance (ANOVAs) were performed using the `anova` function. Label enrichments were tested using Binomial with the `binom.test` function, and T-tests used the `t.test` function. Principal component analyses (PCA) used the `prcomp` function. Unless noted otherwise, p-value adjustments used the Benjamini-Hotchberg method by setting `method = "BH"` in the `p.adjust` function. Plots used either base R, `ggplot2` (v.3.1.0), or `ComplexHeatmap` (v.1.99.5).

Principal component analyses

Approximate array-wide principal component analysis was performed using noob-normalized Beta-values from autosomal probes. Feature hashing, or the "hashing trick" (Weinberger et al. (2010)), was used as an intermediate dimensionality reduction step (see data files in [`pca_fh1k_all_gsm35k.zip` at recount.bio]https://recount.bio/data/recountmethylation_manuscript_supplement/data/. Whole-array Beta-value matrices (>480,000 columns) were collapsed into 1,000 hashed feature columns before cluster analysis. This was implemented with the following `feature_hash` Python function, where the `arr` argument is the sample data and `target_dim=1000` specifies the target hashed feature dimension:

```
def feature_hash(arr, target_dim=1000):
    low_d_rep = [0 for _ in range(target_dim)]
    for i, el in enumerate(arr):
        hashed = mmh3.hash(str(i))
        if hashed > 0:
            low_d_rep[hashed % target_dim] += arr[i]
        else:
            low_d_rep[hashed % target_dim] -= arr[i]
    return low_d_rep
```

GEO year-wise sample and study availability

Yearly GSM and GSE record quantities were obtained by platform using the Entrez Utilities software. Data were plotted using the `ggplot2` package.

Metadata concordance analyses and plot

Concordances for sex and age were calculated between samples with both the mined and predicted labels available (Table 1). For mined and predicted sex, the fraction concordances were calculated. For mined and

predicted ages, ANOVAs were used to calculate covariate variance percentages, and p-values, for multivariate models of predicted/epigenetic age consisting of chronological age, GSE ID, cancer status, and predicted sample type.

Predicted age and chronological age were regressed to calculate R-squared values and correlated using Spearman's test. These were plotted with the regression model using `ggplot2` (Figure 1b). See example 1 in the `data_analyses` vignette of the `recountmethylation` package for details.

Works cited

Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. 2014. "Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays." *Bioinformatics* 30 (10): 1363–9. https://doi.org/10.1093/bioinformatics/btu049.

"BeadArray Controls Reporter Software Guide." 2015, October. https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/infinium_assays/infinium_hd_methylation/beadarray-controls-reporter-user-guide-1000000004009-00.pdf.

Bernstein, Matthew N., AnHai Doan, Colin N. Dewey, and Jonathan Wren. 2017. "MetaSRA: Normalized Human Sample-Specific Metadata for the Sequence Read Archive." *Bioinformatics* 33 (18): 2914–23. https://doi.org/10.1093/bioinformatics/btx334.

Heiss, Jonathan A., and Allan C. Just. 2018. "Identifying Mislabeled and Contaminated DNA Methylation Microarray Data: An Extended Quality Control Toolset with Examples from GEO." *Clinical Epigenetics* 10 (June). https://doi.org/10.1186/s13148-018-0504-1.

"Illumina Genome Studio Methylation Module V1.8." 2010. Illumina. https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/genomestudio/genomestudio-2011-1/genomestudio-methylation-v1-8-user-guide-11319130-b.pdf.

Weinberger, Kilian, Anirban Dasgupta, Josh Attenberg, John Langford, and Alex Smola. 2010. "Feature Hashing for Large Scale Multitask Learning." *arXiv:0902.2206 [Cs]*, February. http://arxiv.org/abs/0902.2206.