Politecnico di Torino

Master's Degree in Mathematical Engineering

Material for Thesis

b– Data Pre-processing Pipeline in GSE69914 — Illumina 450K

Elisabetta Roviera s328422

Contents

Abstract

MANCA L'ABSTRACT, LO SCRIVO DOPO.
AGGIUNGI I LINK AI NOTEBOOK - GITHUB SIA PER DATASET CONSTRUCTION AND STORAGE CHE PER IL DATASET PRE PROCESSING.

1 Dataset construction and storage

I constructed the working methylation matrix in three stages, prioritizing speed, low memory usage, and reproducible I/O.

1. **Ingestion and transposition.** I parsed the GEO Series Matrix for `GSE69914`, skipping the 73-line metadata header. The source table is organized as *CpG* × *sample* with `ID_REF` as CpG identifiers; I coerced all sample columns to numeric (invalid entries set to `NaN`) and then transposed the matrix to the analysis layout *sample* × *CpG*. After transposition, I promoted the original column names (GSM accessions / basenames) to a dedicated identifier column named `id_tissue` and kept CpG probe columns only (prefix `cg` or `ch`).
2. **Label derivation and append-only write.** I derived the class label directly from GSM metadata by parsing the field *status(0=normal, 1=normal-adjacent, 2=breast cancer, 3=normal-BRCA1, 4=cancer-BRCA1)*, producing a numeric `label` in $\{0, 1, 2, 3, 4\}$. To avoid an in-memory join on a very wide table (∼485k CpGs), I streamed through the transposed file once and appended `label` row-wise, preserving row order and ensuring constant memory usage.
3. **Columnar storage and typed schema.** For long-term access, I wrote the labeled table to columnar `Parquet` with a fixed schema: `label` as `Int8` and probe intensities as `Float32`. I applied a lazy, regex-based column projection (`cg|ch`) to cast all probe columns in one pass and compressed the file with lossless LZ4. This yields fast full-table reads and efficient column projection (both in Polars and pandas) without repeatedly parsing large CSV text.

This procedure produces a compact, typed matrix that enables rapid downstream preprocessing (technical filtering, normalization) and modeling without incurring large RAM overhead or costly re-ingestion steps.

2 Import and Data Structure

The processed dataset is imported from the LZ4-compressed `.parquet` file generated in the previous step. The structure is already optimized for analysis.

- **File format:** columnar `Parquet` (LZ4 compression) for fast I/O.
- **Rows:** samples (one per tissue).
- **Columns:**
  - `id_tissue`: unique sample identifier.
  - `label`: numeric class code (`Int8`).
  - `cg`, `ch`: methylation probes (`Float32`).
- **Import method:** read via Polars (or pandas) with column projection for efficient partial loading.

Precision and data representation. Because β-values are strictly bounded within $[0, 1]$ [1], and methylation differences of biological interest typically occur at magnitudes between 10^{-2} and 10^{-3}, single-precision floating point (`float32`, machine $\epsilon \approx 10^{-7}$) provides more than adequate numerical accuracy while significantly reducing memory usage and I/O time. This representation is further supported by recent large-scale genomics frameworks that process molecular features, including DNA methylation data, entirely in `float32` precision [2].

Moreover, this format ensures minimal memory usage and extremely fast access for all downstream preprocessing and analysis tasks.

3 Data Validation and Integrity Check

Data Validation. I validated the structural integrity of the processed dataset to ensure that its layout, types, and values were correctly preserved after conversion and compression.

- **Dimensions:** the dataset contains $(407, 485.514)$ entries, corresponding to **samples** × **CpG loci**. ✓ confirmed as expected: 407 samples and 485.514 probes.
- **Data types:** `id_tissue` is stored as `String`, `label` as `Int8`, and probe intensities as `Float32`, ensuring **compact representation** and sufficient precision for β-values. ✓ verified: `String, Int8, Float32` schema detected.
- **Value range:** all β-values fall within the valid range $0 \leq \beta \leq 1$, confirming their correct interpretation as methylation proportions. ✓ The observed range was [0.000000, 0.997110].

Missing Value Analysis. Next, I performed a comprehensive check for missing values (NaN), as these can severely impact model performance and must be addressed before training.

- No missing entries (NaN) were detected across any CpG probe. ✓ Total NaN count: 0 — Overall missing rate: 0%.
- The methylation matrix is therefore **complete**, requiring **no filtering or imputation** procedures. ✓ Dataset confirmed fully complete.
- For future datasets:
  - If the overall missing rate is $< 1\%$, imputation may be considered as an optional step.
  - If probe missingness exceeds 5% or sample missingness exceeds 10%, the affected entities should be discarded, following standard preprocessing practices [3].

This validation confirms the dataset is structurally sound, numerically consistent, and complete, enabling unbiased downstream variance modeling, differential methylation testing, and batch correction without any further cleaning

4 Technical Filtering

Technical filtering aims to remove unreliable or biologically confounded probes before normalization and statistical modeling. This step reduces noise, improves downstream reproducibility, and ensures that only high-confidence CpG loci are retained for analysis.

Exclusion of technical probe sets. I excluded probes using curated resources that operationalise known technical artefacts (see Appendix A for details).

- **SNP-affected probes:** probes with common variation at the interrogated CpG, at the single-base extension site, or within the probe body [4], [5].
- **Cross-reactive probes:** probes with off-target/multi-mapping hybridisation [6], [7], [4].
- **Design-/platform-specific masks:** consolidated `MASK_*` flags for mapping, SNP windows, non-CpG probes, and optional sex-chromosome probes [4].
- **Naeem hierarchical QC (450K):** discard logic for multi-mapping, repeats, INDEL, and disruptive SNPs [8].

List provenance used here. Naeem et al. 2014 [8]; Chen et al. 2013 [6]; Pidsley et al. 2016 [5]; Zhou et al. 2016 [4]; McCartney et al. 2016 [7]. ✓ Total CpGs removed: 225,426.

Annotation-based filtering. I performed a cross-check with the official Illumina manifest file (*HumanMethylation450 v1.2 Manifest File*) to ensure that only valid and well-characterized loci were retained. This annotation-based filtering step validates probe integrity using the manufacturer's reference genome mapping (hg19) and removes:

- probes with invalid or missing chromosome information (`CHR`);
- probes with undefined or non-positive genomic positions (`MAPINFO`);
- duplicated probe identifiers (`IlmnID`);
- non-CpG-targeting probes (i.e., IDs beginning with "ch").

This step guarantees consistency between the experimental dataset and the official Illumina annotation, harmonizing CpG identifiers across datasets and preventing misaligned genomic coordinates in downstream analyses. The remaining probes thus represent a validated subset of the HumanMethylation450 array.

✓ After manifest-based validation, **875 non-CpG probes** were removed, leaving a final matrix of **407 samples** × **259,213 CpGs**.

5 Filtering of Invariant CpGs

Remove CpGs with very low variance (e.g., variance $< 1 \times 10^{-4}$), as they carry no discriminative information (Naeem et al., 2014).

6 Correction of Infinium I/II Probe Bias

QUA VORREI FAR VEDERE CHE QUESTA COSA è STATA FATTA -¿ MI BASTA QUINDI UN PLOT PER DIMOSTRARE CHE NEI RAW DATA è STATA FATTA Integrate probe design information (Type I / Type II) from Illumina annotation files and inspect density distributions.
Apply **Peak-Based Correction (PBC)** when clear bimodal peaks (near 0 and 1) are visible (Teschendorff et al., 2013). In parallel, apply **BMIQ normalization** as a robust alternative and compare results across normalization strategies.

7 Transformation to M-values

Convert β-values to M-values using $M = \log_2\left(\frac{\beta}{1-\beta}\right)$ to stabilize variance and improve suitability for linear modeling (Du et al., 2010).

CpG Variability Diagnosis (post M-value):
Compute variance or interquartile range (IQR) across samples for each CpG to obtain a diagnostic ranking of variable loci. Highly variable CpGs are typically more informative for distinguishing normal and normal-adjacent tissues. (Naeem et al., 2014; Phipson et al., 2014). Optionally, visualize the variance distribution or perform PCA on top-variable CpGs to assess early group separation.

8 Batch-Effect Correction

Remove inter-array technical variation using **ComBat** (Johnson et al., 2007) or its methylation-specific extension **ComBat-met** (Wang et al., 2025).
If batch effects are confounded with biological groups, include the batch variable as a covariate in linear modeling (e.g., *limma*).

9 Preliminary Statistical Filters

Levene / Brown–Forsythe test: assesses homogeneity of variances across groups.
DiffVar: empirical Bayes model for differential variance detection (Phipson et al., 2014).
limma: moderated linear model for differential methylation, suitable for M-values and inclusion of covariates (Ritchie et al., 2015).
iEVORA: extension of EVORA for identifying CpGs with increased epigenetic instability in pre-neoplastic or field-defect tissues (Teschendorff et al., 2012).

10 Correlation Pruning

Remove redundant CpGs with high inter-correlation (e.g., Pearson $|r| > 0.9$) within local genomic regions to reduce collinearity (Gatev et al., 2020; Bommert et al., 2022).

11 Feature Standardization for Machine Learning

Standardize features (e.g., z-score transformation or StandardScaler) on M-values to ensure comparable scales across CpGs. Fit the scaler on the training fold and apply it to the test fold to prevent data leakage. (Friedman et al., 2010; Aref-Eshghi et al., 2025).

A Filtering lists

This appendix details the major technical filtering lists and annotations applied in Illumina 450K and EPIC methylation arrays to remove unreliable or ambiguous CpG probes.

A.1 Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the humanmethylation450 array

Naeem *et al.* [8] proposed a structured filtering framework to reduce false discoveries by sequentially discarding probes based on hybridisation specificity and genomic integrity (Figure 1). The main exclusion steps are:

- Multi-mapping or cross-hybridising probes ⇒ *discard*.
- Probes overlapping repetitive elements (LINE/SINE/ALU) ⇒ *discard*.
- Probes targeting regions with INDELs ⇒ *discard*.
- Probes overlapping SNPs:
  - SNP at CpG or extension site ⇒ *discard*.
  - SNP near target but not interfering in bisulfite space ⇒ *keep*.

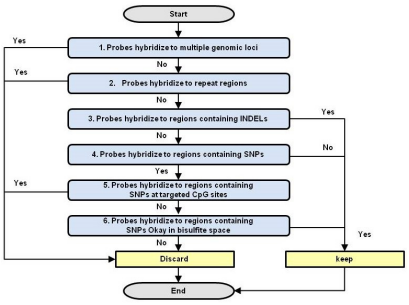

Figure 1: Workflow for determining affected Probes.

The resulting "discard" set therefore removes probes affected by cross-reactivity, structural polymorphisms (INDELs), and SNPs that alter CpG interrogation.

A.2 Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray

Chen *et al.* [6] empirically identified probes in the 450K array that exhibit off-target hybridisation or overlap with common SNPs. For this project, only the **cross-reactive list** is available and used. This list enumerates ∼29k multi-mapping probes whose methylation signal cannot be uniquely attributed to one genomic locus.

A.3 Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling

Pidsley *et al.* [5] provide a platform-level assessment of the EPIC array, with explicit *annotated probe lists* that flag probes whose methylation signal may be confounded by design- or genome-related artefacts. The focus is on probe categories and positions where technical bias arises, rather than on a universal, prescriptive blacklist. In particular, they catalogue:

- **Cross-hybridising CpG-targeting probes:** CpG probes showing sequence homology (off-target matches) to additional genomic loci, yielding non-unique hybridisation and potentially inflated or ambiguous β signals.
- **Cross-hybridising non-CpG-targeting probes:** off-target issues among CNG/non-CpG probes; these are typically excluded in CpG-centric analyses due to limited interpretability and higher risk of artefacts.
- **Probes overlapping common genetic variation:** annotation of variants from population data at three critical positions:
  - *At the interrogated CpG* (polymorphic CpG) — directly affects the presence of the CpG dinucleotide and the measured methylation state;
  - *At the single-base extension (SBE) site* (Type I) — perturbs extension and dye chemistry, biasing intensity ratios;
  - *Within the probe body* — can reduce binding affinity or alter hybridisation kinetics, especially for common variants.

A.4 Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes

Zhou *et al.* [4] released a unified probe annotation for both 450K and EPIC arrays with multiple binary mask columns (`MASK_*`). Each mask flags probes to be removed due to specific technical artifacts:

- `MASK_mapping`: cross-hybridisation or ambiguous alignment;
- `MASK_snp5`: SNP within ±5 bp of the CpG;
- `MASK_extBase`: SNP at the single-base extension;
- `MASK_commonSNPs`: overlap with common polymorphisms;
- `MASK_nonCG`: non-CpG-targeting probes;
- `MASK_chrXY`: probes located on X or Y chromosomes (optional).

This resource provides a programmatic and reproducible way to perform fine-grained probe masking.

A.5 Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip

For the MethylationEPIC array, McCartney *et al.* [7] assessed probe design artifacts and published supplementary tables that include:

- cross-hybridising CpG-targeting probes;
- cross-hybridising non-CpG-targeting probes.

A.6 Comparative Overview

Table 1: Technical categories covered by each filtering resource.

Category	Naeem14	Chen13	Pidsley16	Zhou16	McCartney16
Cross-hybridisation / multi-mapping	Yes (Steps 1–2)	Yes	Yes	`MASK_mapping`	Table 2+3
SNP at CpG / extension base	Yes (Step 5)	–	Yes	`MASK_snp5, MASK_extBase`	–
Nearby SNP tolerated	Conditional (Step 6)	–	–	–	–
INDEL / structural variant	Yes (Step 3)	–	–	–	–
Non-CpG probes	–	–	Yes	`MASK_nonCG`	Table 3
Sex-chromosome probes (X/Y)	Yes	Yes	Yes	`MASK_chrXY`	–

References

[1] L. Weinhold, S. Wahl, S. Pechlivanis, P. Hoffmann, and M. Schmid, "A statistical model for the analysis of beta values in dna methylation studies," *BMC Bioinformatics*, vol. 17, no. 1, p. 480, 2016. DOI: 10.1186/s12859-016-1347-4 [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1347-4

[2] B. P. de Almeida, G. Richard, H. Dalla-Torre, C. Blum, L. Hexemer, P. Pandey, S. Laurent, M. Lopez, A. Laterre, M. Lang, and et al., "A multimodal conversational agent for dna, rna and protein tasks," *Nature Machine Intelligence*, 2025. DOI: 10.1038/s42256-025-01047-1 [Online]. Available: https://www.nature.com/articles/s42256-025-01047-1

[3] W. Zhou, P. W. Laird, and H. Shen, "Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes," *Nucleic Acids Research*, vol. 45, no. 4, e22, 2016. DOI: 10.1093/nar/gkw967 [Online]. Available: https://academic.oup.com/nar/article/45/4/e22/2290937

[4] W. Zhou, P. W. Laird, and H. Shen, "Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes," *Epigenetics & Chromatin*, vol. 9, no. 37, 2016. DOI: 10.1186/s13072-016-0084-1 [Online]. Available: https://academic.oup.com/nar/article/45/4/e22/2290930

[5] R. Pidsley, E. Zotenko, T. J. Peters, M. G. Lawrence, G. P. Risbridger, P. Molloy, S. Van Djik, B. Muhlhausler, C. Stirzaker, and S. J. Clark, "Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling," *Genome Biology*, vol. 17, no. 1, p. 208, 2016. DOI: 10.1186/s13059-016-1066-1 [Online]. Available: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1066-1

[6] Y. A. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, and R. Weksberg, "Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray," *Epigenetics*, vol. 8, no. 2, pp. 203–209, 2013. DOI: 10.4161/epi.23470 [Online]. Available: https://www.tandfonline.com/doi/full/10.4161/epi.23470

[7] D. L. McCartney, R. M. Walker, S. W. Morris, A. M. McIntosh, D. J. Porteous, and K. L. Evans, "Identification of polymorphic and off-target probe binding sites on the illumina infinium methylationepic beadchip," *Epigenetics*, vol. 11, no. 2, pp. 118–128, 2016. DOI: 10.1080/15592294.2016.1146858 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221359601630071X

[8] H. Naeem, N. C. Wong, Z. Chatterton, M. K. Hong, J. S. Pedersen, N. M. Corcoran, C. M. Hovens, and G. Macintyre, "Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the humanmethylation450 array," *BMC Genomics*, vol. 15, no. 514, 2014. DOI: 10.1186/1471-2164-15-514 [Online]. Available: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-514