



---

## Contents

<b>1</b>	<b>Dataset construction and storage</b>	<b>1</b>
<b>2</b>	<b>Import and Data Structure</b>	<b>2</b>
<b>3</b>	<b>Data Validation and Integrity Check</b>	<b>2</b>
<b>4</b>	<b>Technical Filtering</b>	<b>3</b>
<b>5</b>	<b>Filtering of Invariant CpGs</b>	<b>3</b>
<b>6</b>	<b>Correction of Infinium I/II Probe Bias</b>	<b>3</b>
<b>7</b>	<b>Transformation to M-values</b>	<b>4</b>
<b>8</b>	<b>Batch-Effect Correction</b>	<b>4</b>
<b>9</b>	<b>Preliminary Statistical Filters</b>	<b>4</b>
<b>10</b>	<b>Correlation Pruning</b>	<b>4</b>
<b>11</b>	<b>Feature Standardization for Machine Learning</b>	<b>4</b>

---

### 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.

- 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  $\beta$ -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**  $\times$  **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  $\beta$ -values. ✓ verified: `String`, `Int8`, `Float32` schema detected.
- **Value range:** all  $\beta$ -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.** Next, I removed probes listed in curated exclusion sets that are known to produce biased or ambiguous signals:

- **SNP-affected probes:** excluded to avoid spurious methylation differences caused by underlying genetic polymorphisms.
- **Cross-reactive probes:** removed according to validated lists from Naeem et al. (2014) [4] and Pidsley et al. (2016) [5], which identify probes that hybridize to multiple genomic loci.
- **Sex chromosome probes:** optionally filtered if downstream analyses focus exclusively on autosomal loci.

[**Results pending:** number of probes removed by each list, remaining CpG count, updated dataset size.]

**Detection *p*-value filtering.** For each probe, the detection *p*-value measures the probability that its fluorescence intensity is indistinguishable from the background. Probes with detection *p*-value  $> 0.01\text{--}0.05$  in more than 1–5% of samples were excluded, as they likely represent unreliable hybridization or poor signal quality [6]. This filtering step follows Illumina’s quality control recommendations and standard practices for methylation array preprocessing.

[**Results pending:** summary of removed probes, percentage excluded, post-filter matrix dimensions.]

**Annotation-based filtering.** Finally, I performed a cross-check with official Illumina annotation files and updated curated probe databases [3] to ensure that only valid, well-characterized loci were retained. This step allows harmonization between older 450K annotations and updated genome builds (e.g., hg19  $\rightarrow$  hg38) and ensures consistent CpG naming across datasets.

[**Results pending:** number of probes retained after annotation matching, summary table or plot of final coverage.]

**Outcome.** After technical filtering, the dataset is expected to retain only high-confidence probes suitable for reliable normalization and subsequent biological interpretation.

[**Results pending:** summary of total retained CpGs, fraction of genome covered, histogram or density plot of detection *p*-values.]

## 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

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  $\beta$ -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).

## 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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1093/nar/gkw967) [Online]. Available: <https://academic.oup.com/nar/article/45/4/e22/2290937>
- [4] 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. 1, p. 51, 2014. DOI: [10.1186/1471-2164-15-51](https://doi.org/10.1186/1471-2164-15-51) [Online]. Available: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2164-15-51>

- [5] R. Pidsley, J. Y Wong, M. Volta, K. Lunnon, J. Mill, and L. C. Schalkwyk, “A data-driven approach to preprocessing illumina 450k methylation array data,” *Genome Biology*, vol. 17, no. 1, p. 84, 2016. DOI: [10.1186/s13059-016-1066-1](https://doi.org/10.1186/s13059-016-1066-1) [Online]. Available: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1066-1>
- [6] C. S. Wilhelm-Benartzi, D. C. Koestler, M. R. Karagas, J. M. Flanagan, B. C. Christensen, J. K. Wiencke, K. T. Kelsey, C. J. Marsit, and A. E. Houseman, “Review of processing and analysis methods for dna methylation array data,” *British Journal of Cancer*, vol. 109, no. 6, pp. 1394–1402, 2013. DOI: [10.1038/bjc.2013.496](https://doi.org/10.1038/bjc.2013.496) [Online]. Available: <https://www.nature.com/articles/bjc2013496>