



**Politecnico  
di Torino**

# 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

1	Dataset construction and storage	2
2	Import and Data Structure	2
3	Data Validation and Integrity Check	2
4	Correction of Infinium I/II Probe Bias	3
5	Technical Filtering	4
6	Filtering of Invariant CpGs	4
7	Comparison of Beta-value and M-value Quantifications	5
8	Batch-Effect Correction	6
9	Preliminary Statistical Filters	6
10	Correlation Pruning	6
11	Feature Standardization for Machine Learning	6
A	Filtering lists	7
A.1	Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the humanmethylation450 array . . . . .	7
A.2	Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray . . . . .	7
A.3	Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling . . . . .	7
A.4	Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes	8
A.5	Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip . . . . .	8

---

## 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  $CpG \times sample$  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  $sample \times CpG$ . 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 ( $\sim 485k$  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 Correction of Infinium I/II Probe Bias

Raw methylation intensity data (*IDAT* files) from the **GSE69914** dataset were processed using the **minfi** package (v1.8.9) and **BMIQ** normalization (v1.4), as reported in [4]. This preprocessing pipeline includes the **Bias Correction between Infinium Type I and Type II probes**, ensuring that  $\beta$ -value distributions from both probe designs are comparable prior to downstream analysis.

To verify that the bias correction was successfully applied, the  $\beta$ -value distributions were stratified by probe design (Type I vs. Type II) using the official Illumina 450K manifest obtained from the **IlluminaHumanMethylation450kanno.ilmn12.hg19** Bioconductor package [5]. Following the framework proposed by Teschendorff et al. [6], two diagnostic plots were generated.

1. **Left panel:** the  $\beta$ -value density distributions of Type I and Type II probes are shown to overlap substantially, indicating no evident design-driven bias in the normalized data;
2. **Right panel:** a Q-Q plot of Type II versus Type I quantiles demonstrates a median absolute deviation from the identity line of  $|\Delta| = 0.162$ , confirming good alignment between probe types after BMIQ normalization.

The observed deviation value ( $|\Delta| \approx 0.16$ ) is consistent with well-corrected datasets, as discussed by Teschendorff et al. [6] and by Yousefi et al. [7], who report that post-BMIQ alignment typically reduces the median deviation below  $\sim 0.20$ , while uncorrected data often exceed 0.25–0.30. Hence, the present dataset shows **no residual Infinium I/II bias**, and no further probe-type correction was required. **MANCA L'INTERVALLO GIUSTO (SOTTO I 0.18)**

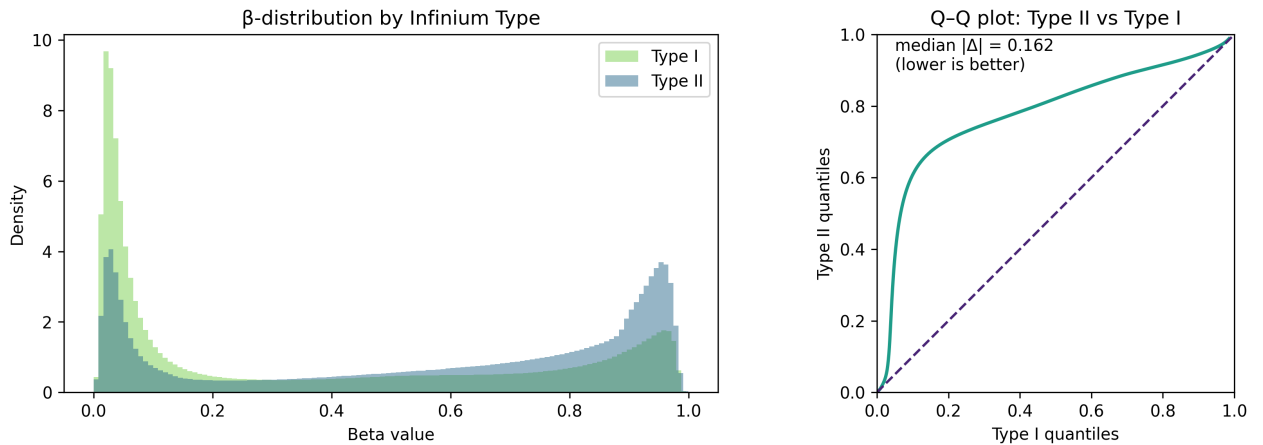


Figure 1: Diagnostic assessment of Infinium I/II probe-type bias. **(Left)** Density of  $\beta$ -values by probe design shows well-overlapping distributions. **(Right)** Q-Q plot (Type II vs. Type I) with median deviation  $|\Delta| = 0.162$ , confirming effective BMIQ normalization.

## 5 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 [3], [8].
- **Cross-reactive probes:** probes with off-target/multi-mapping hybridisation [9], [10], [3].
- **Design-/platform-specific masks:** consolidated MASK\_\* flags for mapping, SNP windows, non-CpG probes, and optional sex-chromosome probes [3].
- **Naeem hierarchical QC (450K):** discard logic for multi-mapping, repeats, INDEL, and disruptive SNPs [11].

**List provenance used here.** Naeem et al. 2014 [11]; Chen et al. 2013 [9]; Pidsley et al. 2016 [8]; Zhou et al. 2016 [3]; McCartney et al. 2016 [10]. ✓ **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 (IllumID);
- 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.** DA AGGIUNGERE LA REFERENCE

## 6 Filtering of Invariant CpGs

CpG sites exhibiting minimal methylation variability across samples provide no discriminatory information and may inflate the multiple-testing burden. Following the empirically driven approach of Edgar *et al.* [12], probes with low dispersion were identified using the inter-decile **beta-range**:  $r_\beta = P90(\beta) - P10(\beta)$ , which quantifies methylation variability in a robust and outlier-resistant manner.

**Whole-cohort variability analysis.** Across all tissue types, the  $r_\beta$  distribution was highly right-skewed, with most loci showing limited variability (Figure 2, left). Three candidate thresholds ( $r_\beta < 0.01, 0.02, 0.05$ ) were evaluated, corresponding to: ✓  $r_\beta < 0.01$ : 1,782 CpGs removed, ✓  $r_\beta < 0.02$ : 28,569, ✓  $r_\beta < 0.05$ : 79,433.

The conservative cutoff  $r_\beta \geq 0.01$  was selected, removing only invariant probes while retaining loci with meaningful biological variation. – MANCA DA DECIDERE

**Normal and adjacent-tissue subset.** To investigate whether tumor-related heterogeneity influenced the variability profile, the same analysis was repeated using only normal and normal-adjacent tissues (labels 0, 1, 3). As expected, the overall dispersion decreased (Figure 2, right), yielding: ✓  $r_\beta < 0.01$ : 3,972 CpGs removed, ✓  $r_\beta < 0.02$ : 40,342, ✓  $r_\beta < 0.05$ : 102,473.

This confirms that tumor samples contribute most of the global variability, while non-neoplastic tissues exhibit tighter methylation clustering.

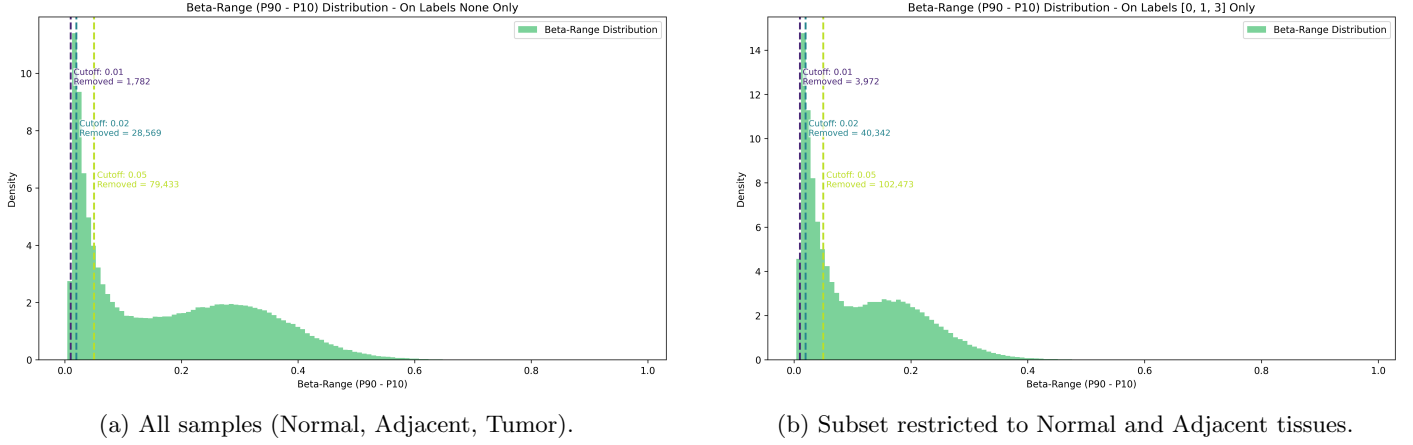


Figure 2: Distribution of beta-range (P90 – P10) across CpGs with candidate thresholds (0.01, 0.02, 0.05). The right panel shows reduced variability when excluding tumor samples, consistent with tissue-specific stability reported by Edgar *et al.* [12].

Final dataset retained CpGs with  $r_\beta \geq 0.01$ , following the tissue-specific non-variability filtering approach of Edgar *et al.*, ensuring the removal of non-informative loci while preserving biologically relevant variation. – MANCA DA DECIDERE

## 7 Comparison of Beta-value and M-value Quantifications

Illumina Infinium arrays measure two fluorescence intensities per CpG: methylated ( $y^{(M)}$ ) and unmethylated ( $y^{(U)}$ ) signals, both background-corrected and normalized. From these signals, two equivalent representations of methylation level can be derived [13].

**Definitions.** The **Beta-value** ( $\beta$ ) represents the proportion of methylated signal:

$$\beta_i = \frac{\max(y_i^{(M)}, 0)}{\max(y_i^{(M)}, 0) + \max(y_i^{(U)}, 0) + \alpha}, \quad 0 \leq \beta_i \leq 1,$$

while the **M-value** ( $M$ ) is defined as the  $\log_2$ -ratio of methylated to unmethylated intensities:

$$M_i = \log_2 \left( \frac{\max(y_i^{(M)}, 0) + \alpha}{\max(y_i^{(U)}, 0) + \alpha} \right), \quad M_i \in (-\infty, \infty).$$

The two measures are related by the logit transform:

$$M_i = \log_2 \left( \frac{\beta_i}{1 - \beta_i} \right) \iff \beta_i = \frac{2^{M_i}}{1 + 2^{M_i}}.$$

**Rationale.** As established by Du *et al.* [13],  $\beta$ -values are bounded and heteroscedastic—variance depends on the mean, especially near 0 or 1—whereas  $M$ -values are approximately homoscedastic and symmetrically distributed, making them statistically more suitable for linear-model and  $t$ -test analyses. For interpretability,  $\beta$  remains preferred for reporting because it directly corresponds to a percentage of methylation.

**Implementation in this dataset.** The **GSE69914** matrix provides  $\beta$ -values only; raw intensities ( $y^{(M)}, y^{(U)}$ ) are unavailable. Hence, I derived  $M$  using the standard logit transform:

$$M = \log_2 \left( \frac{\beta + \varepsilon}{1 - \beta + \varepsilon} \right),$$

with a small  $\varepsilon$  for numerical stability. This reproduces the expected scale and statistical properties of  $M$ , enabling robust modeling even without raw intensity data.

✓ All subsequent statistical analyses are performed in  $M$ -space (for inference and testing), and results are converted back to  $\beta$  for visualization and biological interpretation.

## 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.

Table 1: Technical categories covered by each filtering resource.

Category	Naeem A.1	Chen A.2	Pidsley A.3	Zhou A.4	McCartney A.5
Cross-hybridisation / multi-mapping	Yes (Steps 1–2)	Yes	Yes	<b>MASK_mapping</b>	Table 2+3
SNP at CpG / extension base	Yes (Step 5)	–	Yes	<b>MASK_snp5</b> , <b>MASK_extBase</b>	–
Nearby SNP tolerated	Conditional (Step 6)	–	–	–	–
INDEL / structural variant	Yes (Step 3)	–	–	–	–
Non-CpG probes	–	–	Yes	<b>MASK_nonCG</b>	Table 3

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

Naeem *et al.* [11] proposed a structured filtering framework to reduce false discoveries by sequentially discarding probes based on hybridisation specificity and genomic integrity (Figure 3). The main exclusion steps are:

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

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.* [9] 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  $\sim 29$ k 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.* [8] 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  $\beta$  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;

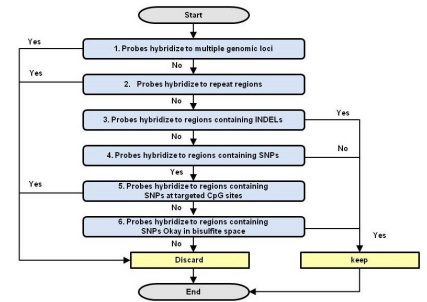


Figure 3: Workflow for determining affected Probes.



- *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.* [3] 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: probes with low or inconsistent mapping quality (non-unique alignment or presence of INDELs);
- MASK.typeINextBaseSwitch: Type-I probes carrying a SNP in the extension base that causes a color-channel switch (CCS probes);
- MASK.extBase: probes whose extension base is inconsistent with the expected color channel or CpG context;
- MASK.sub30.copy: probes with non-unique 30-bp 3' subsequences (potential cross-hybridisation);
- MASK.snp5.common: probes overlapping a SNP within  $\pm 5$  bp of the interrogated CpG (even with global MAF  $\geq 1\%$ );
- MASK.snp5.GMAF1p: probes overlapping SNPs with global MAF  $\geq 1\%$ ;
- MASK.general: recommended composite mask integrating mapping, SNP, and cross-reactivity filters for general use;
- MASK.rmsk15: probes overlapping RepeatMasker regions (not recommended for exclusion in standard workflows).

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

McCartney *et al.* [10] assessed probe design artifacts and published supplementary tables that include:

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

## 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," *Epigenetics & Chromatin*, vol. 9, no. 37, 2016. DOI: [10.1186/s13072-016-0084-1](https://doi.org/10.1186/s13072-016-0084-1) [Online]. Available: <https://academic.oup.com/nar/article/45/4/e22/2290930>
- [4] National Center for Biotechnology Information (NCBI), *Gse69914 - dna methylation profiles in breast tissue samples*, Processed using minfi v1.8.9 and BMIQ v1.4 as reported in the GEO metadata, Gene Expression Omnibus (GEO), 2015.
- [5] M. D. Robinson, G. K. Smyth, K. D. Hansen, and et al., *IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for illumina's 450k methylation arrays*, R package version 0.6.0, Bioconductor, 2015. [Online]. Available: <https://bioconductor.org/packages/IlluminaHumanMethylation450kanno.ilmn12.hg19>
- [6] A. E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, D. Tegner, J. Gomez-Cabrero, and S. Beck, "A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450k dna methylation data," *Bioinformatics*, vol. 29, no. 2, pp. 189–196, 2013. DOI: [10.1093/bioinformatics/bts680](https://doi.org/10.1093/bioinformatics/bts680) [Online]. Available: <https://academic.oup.com/bioinformatics/article/29/2/189/199637>



- [7] P. Yousefi, I. Huen, C. Schmid-Siegert, H. B. Staehli, and et al., “Normalization of 450k methylation array data improves replication in large epidemiological studies,” *Epigenetics*, vol. 8, no. 1, pp. 79–87, 2013. DOI: [10.4161/epi.23018](https://doi.org/10.4161/epi.23018) [Online]. Available: <https://www.tandfonline.com/doi/full/10.4161/epi.23018>
- [8] R. Pidsley, E. Zotenko, T. J. Peters, M. G. Lawrence, G. P. Risbridger, P. Molloy, S. Van Dijk, 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](https://doi.org/10.1186/s13059-016-1066-1) [Online]. Available: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1066-1>
- [9] 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](https://doi.org/10.4161/epi.23470) [Online]. Available: <https://www.tandfonline.com/doi/full/10.4161/epi.23470>
- [10] 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](https://doi.org/10.1080/15592294.2016.1146858) [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S221359601630071X>
- [11] 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](https://doi.org/10.1186/1471-2164-15-514) [Online]. Available: <https://bmcbgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-514>
- [12] R. D. Edgar, M. J. Jones, M. J. Meaney, G. Turecki, and M. S. Kobor, “An empirically driven data reduction method on the human 450k methylation array to remove tissue-specific non-variable cpGs,” *BMC Genomics*, vol. 18, no. 1, p. 690, 2017. DOI: [10.1186/s12864-017-4129-5](https://doi.org/10.1186/s12864-017-4129-5) [Online]. Available: <https://bmcbgenomics.biomedcentral.com/articles/10.1186/s12864-017-4129-5>
- [13] P. Du, X. Zhang, C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin, “Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis,” *BMC Bioinformatics*, vol. 11, no. 1, p. 587, 2010. DOI: [10.1186/1471-2105-11-587](https://doi.org/10.1186/1471-2105-11-587) [Online]. Available: <https://doi.org/10.1186/1471-2105-11-587>