# 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 $\beta$-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** × **CpG loci**. ✓ confirmed as expected: 407 samples and 485,512 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.

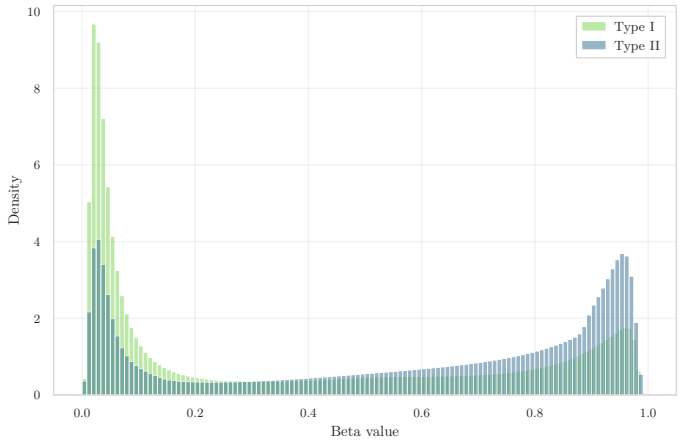
# 4 Correction of Infinium I/II Probe Bias

Raw methylation intensity data (*IDAT* files) from the **GSE69914** dataset were processed by the original authors 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.

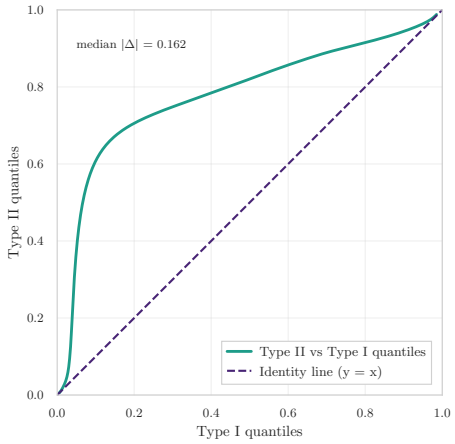
I independently verified that this probe-type bias correction had been successfully applied. To this aim, the $\beta$-value distributions were stratified by probe design (Type I vs. Type II) using the official Illumina 450K manifest obtained from the Bioconductor package [5]. Following the diagnostic framework proposed by Teschendorff et al. [6], two diagnostic plots were generated:

1. The $\beta$-value density distributions of Type I and Type II probes (Figure 1a) show extensive overlap, with no evident design-driven bias in the normalized data;
2. The Q–Q plot comparing Type II versus Type I quantiles (Figure 1b) exhibits a near-diagonal alignment, confirming effective bias correction.

The observed post-normalization pattern closely reproduces the characteristic outcome described by Teschendorff et al. [6] and Wang et al. [7], who demonstrated that BMIQ substantially reduces Infinium Type II probe bias, yielding overlapping $\beta$-value distributions and near-diagonal Q–Q relationships. ✓ Hence, this verification confirms that the GSE69914 dataset underwent proper BMIQ normalization, and no residual Infinium I/II bias was detected prior to analysis.




(a) $\beta$-value distribution by Infinium probe design (Type I vs. Type II). The near-overlapping shapes indicate effective correction of probe-type bias.




(b) Q–Q plot comparing quantiles of Type II vs. Type I probes. The alignment to the diagonal ($y = x$) confirms balanced signal distributions after normalization.

Figure 1: Diagnostic evaluation of Infinium Type I/II probe bias correction in the GSE69914 dataset. The consistent $\beta$-value and quantile distributions demonstrate the effectiveness of the BMIQ normalization previously applied.

# 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 (`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, following the recommendations of Zhou et al. [3] and Pidsley et al. [8], who emphasize the importance of excluding non-CpG or improperly mapped loci using the official Illumina manifest.

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

# 6 Filtering of Invariant CpGs

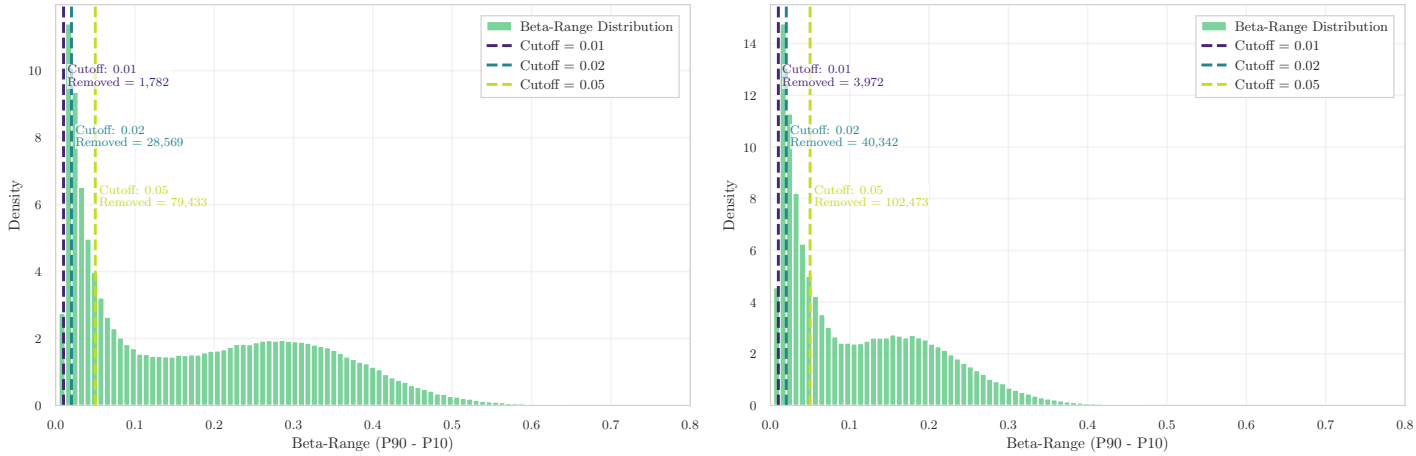
CpG sites exhibiting minimal methylation variability across samples provide no discriminatory information and may inflate the multiple-testing burden. Following the empirically driven approach of Edgar *et al.* [12], probes with low dispersion were identified using the inter-decile **beta-range**: $r_\beta = \mathrm{P90}(\beta) - \mathrm{P10}(\beta)$, 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 2a). 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.

**Normal and adjacent-tissue subset.** Given the final objective of identifying CpGs informative for distinguishing normal from adjacent tissues, this filtering step was applied specifically to the subset including **normal** (label 0), **adjacent** (label 1), and **BRCA1-modified normal** (label 3) samples. As expected, restricting to these tissues decreases overall dispersion (Figure 2b), 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 methylation variability, consistent with the increased stochastic variation reported in cancer tissues compared to normal [13]. ✓ Accordingly, the reference threshold proposed by Edgar *et al.* [12], $r_\beta < 0.05$, was adopted as a pragmatic cutoff and applied to the Normal–Adjacent subset, leaving a final matrix of **99 samples** × **156,738 CpGs**. CpGs below this threshold in the restricted cohort were removed from the

full dataset. If subsequent analyses reveal that this choice is overly permissive—resulting in the exclusion of an excessive number of informative loci—the criterion will be re-evaluated and adjusted toward a more conservative threshold (e.g., 0.02 or 0.01). This approach ensures that only loci exhibiting stable methylation within non-tumor tissues are discarded, preserving CpGs with potential biological relevance for the Normal–Adjacent classification task.




(a) Whole cohort including Normal, Adjacent, and Tumor samples. The distribution is right-skewed, with most CpGs showing limited variability.

(b) Subset restricted to Normal, Adjacent, and BRCA1-modified Normal tissues. Excluding tumor samples reduces overall dispersion across CpGs.

Figure 2: Distribution of inter-decile beta-range ($r_\beta$ = P90 − P10) across CpG sites. The analysis highlights reduced variability in non-tumor tissues, supporting the use of $r_\beta < 0.05$ as a filtering criterion in the Normal–Adjacent subset.

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

Illumina Infinium reports two background-corrected signals per CpG, methylated $y^{(M)}$ and unmethylated $y^{(U)}$. The **Beta-value** is the proportion of methylated intensity,

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

while the **M-value** is the log-ratio,

$$M_i = \log_2\left(\frac{\max(y_i^{(M)}, 0) + \alpha}{\max(y_i^{(U)}, 0) + \alpha}\right) = \log_2\left(\frac{\beta_i}{1 - \beta_i}\right).$$

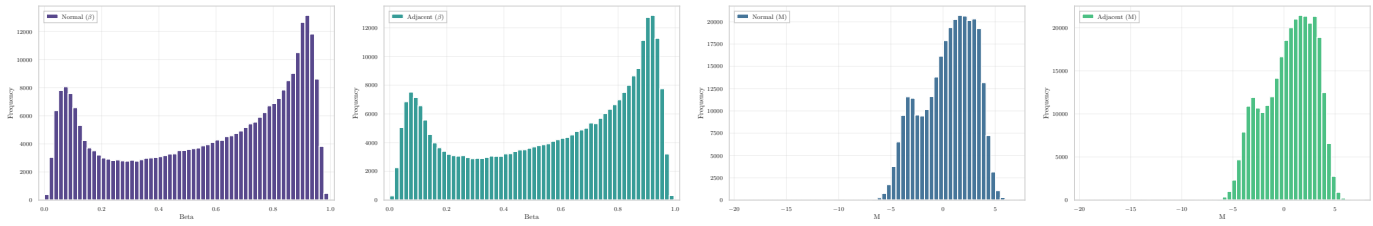
**Rationale.** As shown by Du *et al.* [14], $\beta$ is *bounded* and *heteroscedastic* (its variance depends on the mean, especially near 0 and 1), whereas $M$ is approximately *homoscedastic* and therefore more suitable for linear modeling and $t$-tests. However, $\beta$ retains higher interpretability since it directly reflects the methylation fraction. Consequently, $M$-values are adopted for inferential analyses, while $\beta$-values are preferred for reporting and visualization.

**Implementation in this dataset.** The GSE69914 matrix provides $\beta$ only; therefore, $M$ was derived as

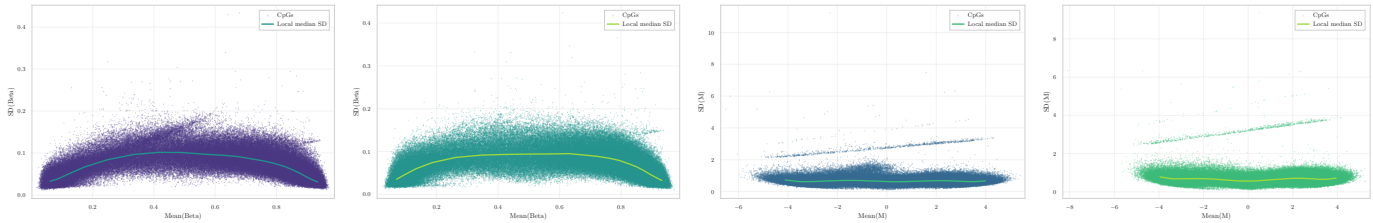
$$M = \log_2\left(\frac{\beta + \varepsilon}{1 - \beta + \varepsilon}\right),$$

using a small $\varepsilon = 10^{-6}$ to ensure numerical stability for extreme $\beta$ values. ✓ The working matrix was converted to $M$-values for all subsequent statistical analyses.

**Distributional comparison.** Figure 3 illustrates the distributional differences between $\beta$ and $M$ quantifications across Normal and Normal-adjacent tissues. As evident in the upper panels, $\beta$ exhibits a bounded, U-shaped distribution, while $M$ yields approximately symmetric profiles. The lower panels confirm the pronounced mean-dependent variance of $\beta$ and the near-homoscedastic behavior of $M$, ✓ in agreement with Du *et al.* [14].



(a) Histogram of $\beta$ values in Normal samples (b) Histogram of $\beta$ values in Normal-adjacent samples (c) Histogram of $M$ values in Normal samples (d) Histogram of $M$ values in Normal-adjacent samples



(e) Mean–SD trend for $\beta$ in Normal samples (f) Mean–SD trend for $\beta$ in Normal-adjacent samples (g) Mean–SD trend for $M$ in Normal samples (h) Mean–SD trend for $M$ in Normal-adjacent samples

Figure 3: **Distributional properties of Normal vs Normal-adjacent tissue methylation.** Top row: $\beta$-value (bounded fractional) and $M$-value (log-ratio) histograms illustrate the characteristic U-shaped $\beta$ distributions and the more symmetric $M$ profiles. Bottom row: Mean–SD relationships across CpGs reveal pronounced mean-dependent heteroscedasticity in $\beta$, which is substantially reduced after logit transformation to $M$, confirming the homoscedasticity advantage described by Du *et al.* [14]. All plots were computed on the current working set of 156,738 CpGs remaining after the filtering steps applied so far.

## 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

TODOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO

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

TODOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO

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 | `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 |

## 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 4). 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*.

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

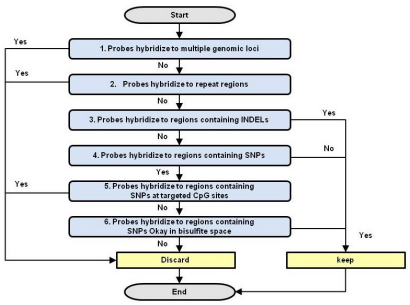



Figure 4: Workflow for determining affected Probes.

## 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 ∼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.* [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;

– *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 ±5 bp of the interrogated CpG (even with global MAF ¡ 1%);
- `MASK.snp5.GMAF1p`: probes overlapping SNPs with global MAF ¿ 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 [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, Feb. 2016, Published online 24 Oct 2016. DOI: 10.1093/nar/gkw967 [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. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69914

[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 [Online]. Available: https://academic.oup.com/bioinformatics/article/29/2/189/199637

[7] T. Wang, W. Guan, J. Lin, W. Chen, X. Zhu, X. Zhang, S. Haider, and ..., "A systematic study of normalization methods for infinium 450k methylation data using whole-genome bisulfite sequencing as the gold standard," *Epigenetics*, vol. 10, no. 6, pp. 536–545, 2015. DOI: 10.1080/15592294.2015.1057384 [Online]. Available: https://doi.org/10.1080/15592294.2015.1057384

[8] 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

[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 [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 [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 [Online]. Available: https://bmcgenomics.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 [Online]. Available: https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-017-0320-z

[13] K. D. Hansen, W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry, and A. P. Feinberg, "Increased methylation variation in epigenetic domains distinguishes tumour from normal tissue," *Nature Genetics*, vol. 43, no. 8, pp. 768–775, 2011. DOI: 10.1038/ng.865 [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3145050/

[14] 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 [Online]. Available: https://doi.org/10.1186/1471-2105-11-587