Politecnico di Torino

Laurea Magistrale in Ingegneria Matematica

Analisi Comparativa e Strategie Future per lo Studio Epigenetico del Carcinoma Mammario

Revisione dei Lavori di Tesi
Data Pre-Processing Pipeline in GSE69914

Eleonora Guarnaschelli
Elisabetta Roviera

Contents

1 Correzione del bias tra sonde Infinium I e II **1**

2 Filtraggio tecnico **2**

3 Filtraggio delle CpG invarianti **3**

Abstract

MANCA L'ABSTRACT, LO SCRIVO DOPO.

Costruzione del dataset. Abbiamo acquisito il dataset `GSE69914` [1] in formato `.txt` contenente i valori di metilazione (β-values) e ottenuto le etichette di classe direttamente da `GEOparse` per garantirne la massima correttezza. La matrice è stata successivamente convertita in un file compresso `.parquet` con righe corrispondenti ai campioni e colonne alle sonde CpG. Infine, abbiamo eseguito i controlli di qualità per verificare l'assenza di valori mancanti e la correttezza del range dei β-values.

1 Correzione del bias tra sonde Infinium I e II

I dati grezzi di intensità di metilazione del dataset sono stati pre-processati dagli autori originali utilizzando il pacchetto `minfi` (v1.8.9) e la normalizzazione `BMIQ` (v1.4), come riportato in [1]. Questa pipeline include la **correzione del bias tra le sonde Infinium di tipo I e tipo II**, assicurando che le distribuzioni dei valori di metilazione (β-values) dei due design risultino comparabili prima delle analisi successive.

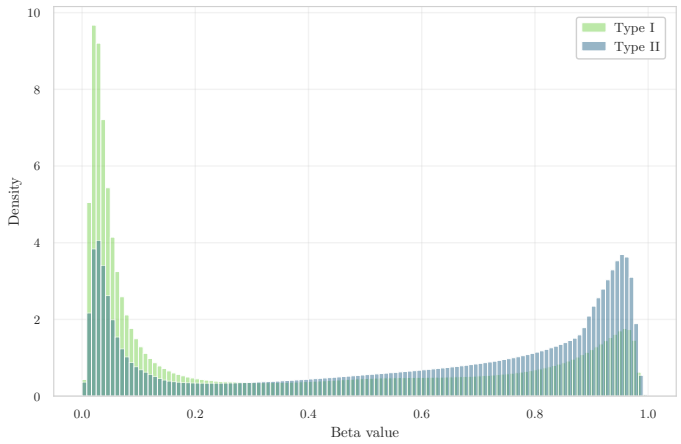
Abbiamo verificato in modo indipendente che tale correzione fosse effettivamente applicata. A tal fine, le distribuzioni dei β-values sono state stratificate per tipo di sonda (Type I vs. Type II) utilizzando il manifest ufficiale di Illumina 450K ottenuto dal pacchetto Bioconductor [2]. Seguendo l'approccio diagnostico proposto da Teschendorff et al. [3], abbiamo generato due grafici di controllo.

1. Le distribuzioni di densità dei β-values per le sonde Type I e Type II (Figure 1a) mostrano un'elevata sovrapposizione, senza evidenza di bias dovuto al design;

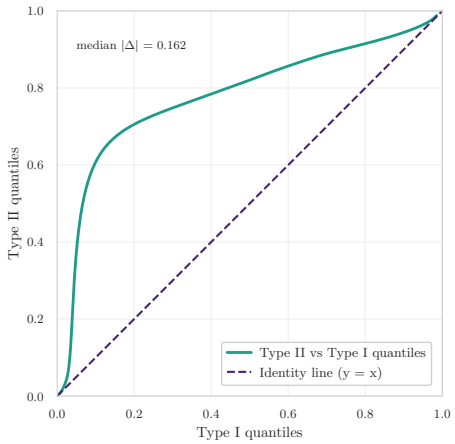
2. Il Q–Q plot che confronta i quantili delle sonde Type II rispetto alle Type I (Figure 1b) presenta un allineamento quasi diagonale, confermando l'efficacia della correzione.

Il pattern osservato dopo la normalizzazione riproduce fedelmente i risultati descritti da Teschendorff et al. [3] e da Wang et al. [4], che hanno dimostrato come la procedura BMIQ riduca sostanzialmente il bias tra sonde Infinium di tipo I e II, producendo distribuzioni sovrapposte e relazioni quantiliche bilanciate. ✓ Pertanto, la nostra verifica conferma che il

dataset **GSE69914** ha subito una corretta normalizzazione BMIQ e non presenta bias residui tra le sonde Infinium I/II prima dell'analisi.



(a) Distribuzione dei β-values per tipo di sonda (Type I vs. Type II). Le curve quasi sovrapposte indicano una correzione efficace del bias.



(b) Q–Q plot tra i quantili delle sonde Type II e Type I. L'allineamento alla diagonale ($y = x$) conferma la correzione del bias di design.

Figure 1: Valutazione diagnostica della correzione del bias tra sonde Infinium Type I e II nel dataset **GSE69914**. Le distribuzioni coerenti dei β-values e dei quantili dimostrano l'efficacia della normalizzazione BMIQ applicata.

2 Filtraggio tecnico

Il filtraggio tecnico ha lo scopo di rimuovere le sonde inaffidabili o affette da artefatti biologici o di piattaforma prima delle fasi di normalizzazione e modellazione statistica. Questo passaggio consente di ridurre il rumore, migliorare la riproducibilità delle analisi e mantenere esclusivamente i siti CpG ad alta affidabilità.

Rimozione delle sonde affette da artefatti tecnici. Abbiamo applicato diversi insiemi di filtri consolidati presenti in letteratura per eliminare sonde problematiche o non univocamente mappate. In particolare, sono stati rimossi:

- **SNP-affected probes:** sonde contenenti varianti comuni nel sito CpG, nella base di estensione o all'interno della sequenza della sonda [5], [6];
- **Sonde cross-reattive:** sonde con ibridazione multipla o mappatura non specifica [5], [7], [8];
- **Maschere di design o piattaforma:** flag `MASK_*` relativi a errori di mappatura, SNP adiacenti, sonde non-CpG e sonde su cromosomi sessuali opzionali [5];
- **Controlli gerarchici Naeem (450K):** esclusione di sonde multi-mappate, ripetute o affette da varianti strutturali (INDEL) e SNP interferenti [9].

La Table 1 sintetizza le principali categorie di artefatti tecnici considerate e le corrispondenti fonti bibliografiche utilizzate per la loro rimozione: Naeem et al. 2014 [9], Chen et al. 2013 [7], Pidsley et al. 2016 [6], Zhou et al. 2016 [5] e McCartney et al. 2016 [8]. ✓ Complessivamente, sono state rimosse **225.426 sonde CpG**.

Table 1: Categorie tecniche coperte da ciascuna risorsa di filtraggio.

Categoria	**Naeem**	**Chen**	**Pidsley**	**Zhou**	**McCartney**
Ibridazione multipla / multi-mapping	Sì	Sì	Sì	`MASK_mapping`	Liste 2–3
SNP al sito CpG / base di estensione	Sì	–	Sì	`MASK_snp5, MASK_extBase`	–
SNP adiacenti tollerati	Sì	–	–	–	–
INDEL / varianti strutturali	Sì	–	–	–	–
Sonde non-CpG	–	–	Sì	`MASK_nonCG`	Lista 3

Filtraggio basato su annotazione. Abbiamo successivamente verificato la coerenza del dataset con il file di manifest ufficiale di Illumina (*HumanMethylation450 v1.2 Manifest File*), assicurando il mantenimento esclusivo delle sonde valide e ben caratterizzate. Sono state eliminate sonde non-CpG (identificativi con prefisso "ch").

Questo passaggio garantisce la consistenza tra i dati sperimentali e l'annotazione ufficiale Illumina, prevenendo disallineamenti genomici nelle analisi successive. Seguendo le raccomandazioni di Zhou et al. [5] e Pidsley et al. [6]. ✓ Sono state rimosse ulteriori **875 sonde non-CpG**.

3 Filtraggio delle CpG invarianti

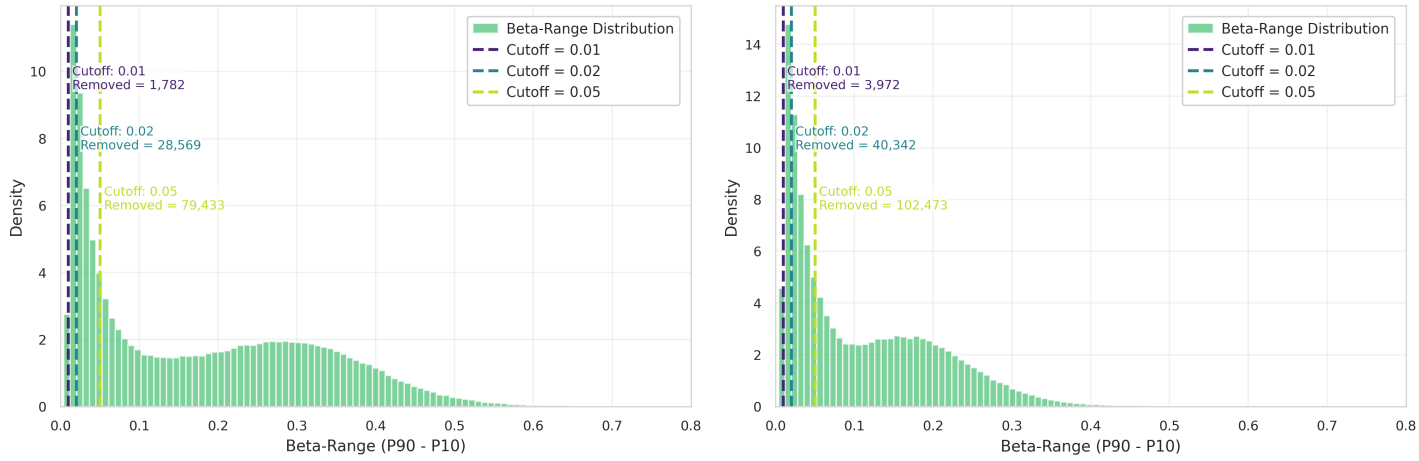
I siti CpG che mostrano una variabilità di metilazione minima tra i campioni non forniscono informazioni discriminanti e possono aumentare inutilmente il carico di test multipli. Seguendo l'approccio empirico proposto da Edgar *et al.* [10], abbiamo identificato le sonde a bassa dispersione utilizzando l'intervallo inter-decile dei valori di metilazione, definito come: $r_\beta = \mathrm{P90}(\beta) - \mathrm{P10}(\beta)$, una misura robusta e resistente agli outlier della variabilità di metilazione.

Analisi di variabilità sull'intero dataset. Considerando tutti i tipi di tessuto, la distribuzione di r_β risulta fortemente asimmetrica a destra, con la maggior parte dei loci che mostrano una variabilità limitata (Figure 2a). Abbiamo valutato tre soglie candidate ($r_\beta < 0.01$, 0.02, 0.05), ottenendo: ✓ $r_\beta < 0.01$: 1.782 CpG rimosse, ✓ $r_\beta < 0.02$: 28.569, ✓ $r_\beta < 0.05$: 79.433.

Sottoinsieme Normal e Adjacent. Poiché l'obiettivo finale è identificare le CpG informative per distinguere i tessuti **normali** (label 0) da quelli **adiacenti** (label 1), abbiamo applicato il filtraggio in modo specifico a questo sottoinsieme. Come atteso, limitando l'analisi a tessuti non tumorali la dispersione complessiva diminuisce (Figure 2b), con i seguenti risultati: ✓ $r_\beta < 0.01$: 3.972 CpG rimosse, ✓ $r_\beta < 0.02$: 40.342, ✓ $r_\beta < 0.05$: 102.473.

Questo conferma che i campioni tumorali contribuiscono in modo predominante alla variabilità globale della metilazione, in accordo con quanto riportato da Hansen et al. [11], che descrivono un'aumentata variabilità stocastica nei tessuti neoplastici rispetto a quelli normali. ✓ Abbiamo pertanto adottato come soglia di riferimento il valore proposto da Edgar *et al.* [10], $r_\beta < 0.05$, applicandolo al sottoinsieme Normal–Adjacent, ottenendo una matrice finale di **99 campioni × 156.740 CpG**. Le CpG al di sotto di tale soglia nel sottoinsieme ristretto sono state rimosse anche dal dataset completo.

Nel caso in cui analisi successive evidenzino che questa scelta risulti eccessivamente permissiva — portando all'esclusione di un numero rilevante di loci potenzialmente informativi — il criterio verrà rivalutato in senso più conservativo (ad esempio soglie di 0.02 o 0.01). Questo approccio garantisce che vengano scartati solo i siti con metilazione stabile nei tessuti non tumorali, preservando invece le CpG con possibile rilevanza biologica per la distinzione tra tessuti Normal e Adjacent.



(a) Distribuzione dell'intervallo inter-decile r_β considerando l'intera coorte (Normal, Adjacent e Tumor). La distribuzione è asimmetrica, con la maggior parte delle CpG a bassa variabilità.

(b) Distribuzione dell'intervallo inter-decile r_β limitata ai tessuti Normal e Adjacent. L'esclusione dei campioni tumorali riduce la dispersione complessiva.

Figure 2: Distribuzione dell'intervallo inter-decile dei valori di metilazione ($r_\beta = \mathrm{P90} - \mathrm{P10}$) nei siti CpG. L'analisi evidenzia una minore variabilità nei tessuti non tumorali, supportando l'uso di $r_\beta < 0.05$ come criterio di filtraggio per il sottoinsieme Normal–Adjacent.

References

[1] 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.

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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/