Politecnico di Torino

Master's Degree in Mathematical Engineering

Material for Thesis

2– About Data Visualization and Exploration

Elisabetta Roviera s328422

Contents

Note The papers summarized in this report currently represent the main references consulted for the study of data visualization and DNA methylation analysis. Additional papers may be reviewed in the future if further methodological or biological details become necessary. For comprehensive information and methodological specifics, refer to the original publications, all of which are cited in the bibliography of this document.

1 SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips

Keywords CpG, CpG island, Illumina HumanMethylation450 (450k), Infinium I/II, SWAN, subset-quantile within-array normalization, β-values, M-values, technical variation, ROC, minfi [1]

CpG vs CpG island A **CpG** is a cytosine followed by guanine (5'–C–phosphate–G–3') whose cytosine can be methylated; CpG methylation modulates gene regulation. **CpG islands** (CGIs) are genomic regions with high CpG density, often near promoters, exhibiting distinct methylation patterns from non-island regions. On the 450k array, probe design enriches Infinium I probes within CpG-dense regions (islands) more than Infinium II, contributing to different intensity and β-value distributions across probe types.

Why normalization is needed (Infinium I vs II) The 450k platform mixes two chemistries: *Infinium I* (two probes, single color) and *Infinium II* (one probe, two colors). Empirically, Infinium II shows compressed β ranges versus Infinium I, yielding different distributions driven by both technical effects (detection scheme) and biology (distinct genomic targeting). Naïve between-type quantile normalization is inappropriate because the probe types interrogate different genomic subsets; a method must equalize technical differences without erasing true biology.

SWAN: definition, rationale, and outcome **SWAN** (Subset-quantile Within-Array Normalization) is a *within-array* normalization that uses CpG-content–matched subsets of Infinium I/II probes to estimate a common quantile target per channel (M/U) and then *interpolates* the remaining probes to that target. Assumption: probes with the same number of CpGs in the 50 bp body have similar intensity distributions, so matching on CpG count preserves biological differences while removing type-specific technical bias. Result: β-value distributions from Infinium I and II become more consistent; replicate concordance increases; and power to detect differential methylation improves (better ROC and more true positives against RRBS ground truth).

Algorithmic sketch (per array, per channel) (1) Partition probes by type (I/II) and by probe-body CpG count (1,2,3). (2) Randomly select equal-sized subsets across these strata (size N is the minimum stratum count; ∼11,303 if no filtering). (3) Quantile-normalize the paired I/II subset by averaging quantiles. (4) Linearly interpolate all remaining probes to the subset's reference distribution (capped/extrapolated at extremes as specified). Compute $\beta = M/(M + U)$ or M-values downstream.

Motivation & validation SWAN targets the probe-type artifact (distributional shifts and peak misalignment) while maintaining biological structure (e.g., CGI vs non-CGI). It reduces median differences between I/II β distributions across fully methylated, hemi-methylated, and unmethylated standards; improves KS concordance and correlations in technical replicates; aligns 450k distributions closer to 27k (Infinium I-only) for shared CpGs; and increases detection of RRBS-supported DMPs (higher TPR across q-value thresholds; better ROC).

Plot types in the paper (and how to replicate on your data)

- *Density of $\log_2$-intensities by channel and probe type.* Kernel densities of M and U channels, split by Infinium I vs II, highlight type-wise distributional differences pre-normalization; after SWAN, intensities become more comparable within CpG-count strata. *Replication:* compute $\log_2(M)$ and $\log_2(U)$ per probe, stratify by type, and overlay densities; optionally facet by CpG-count (1/2/3) to visually support SWAN's assumption.

- *β-value density overlays by probe type and standard.* Overlaid β densities (Infinium I solid, II dashed) for unmethylated, hemi-, and fully methylated controls show compression in Infinium II that SWAN mitigates; report median/IQR alignment and peak shifts (ΔP_U, ΔP_M). *Replication:* compute β per probe, group by type and by biological group (e.g., Normal/Adjacent/Tumor), plot densities, and quantify peak positions and median gaps before/after SWAN.

- *Boxplots/point-ranges of medians and IQRs.* Summaries of β medians and IQRs by probe type pre/post SWAN visualize improved alignment. *Replication:* per sample, compute median and IQR of β for type I vs II and plot paired segments to show reduction after SWAN.

- *QQ plots by CpG-count strata.* QQ plots show greater linearity when probes are grouped by equal CpG counts, supporting SWAN's matching strategy. *Replication:* for each CpG-count stratum, produce type-vs-type QQ plots of intensities, pre/post SWAN.

- *27k vs 450k β densities.* For shared CpGs, SWAN improves peak alignment of 450k (I/II) to 27k. *Replication:* if you have legacy 27k or external truth, overlay densities for intersecting sites; otherwise simulate a "gold" subset using high-confidence loci.

- *ROC curve and TPR vs q-value curves.* Using RRBS to define true positives/negatives, SWAN yields higher AUC and higher %TP across thresholds. *Replication:* if you lack RRBS, approximate with spike-in controls or cross-tissue contrasts with orthogonal evidence; otherwise show internal consistency (replicate concordance).

Take-home for your pipeline Use SWAN early (within-array), then proceed with between-array normalization if needed; compute both β and M (analysis often on M, visualization often on β). Always stratify QC plots by probe type and CpG-count to verify SWAN's assumptions on your dataset. Expect smaller type-driven artifacts, tighter replicates, and improved DMP discovery without erasing CGI vs non-CGI biology.

2 Charting Differentially Methylated Regions in Cancer with Rockermeth

Keywords DMR, AUC, ROC, heterogeneous HMM, segmentation, CpG, CpG island, WGBS, 450k, TCGA, PMD, hypo-blocks, PSFSE, single-cell methylation [2].

What is Rocker-meth (idea) Rocker-meth is a platform-agnostic method to call *differentially methylated regions* (DMRs) by converting per-site tumor–normal methylation contrasts into ROC-based AUC scores and then *segmenting* the genome with a heterogeneous HMM. Unlike per-site tests, Rocker-meth emphasizes regional signals (from hundreds of bp to Mbp), capturing focal hypermethylation and large hypomethylated blocks ("hypo-blocks") that align with partially methylated domains (PMDs). It yields both cohort-level DMR catalogs and sample-level scores (burden/PSFSE) useful for subtype discovery and single-sample assessment.

Why it is useful (purpose) In cancer, biologically meaningful methylation alterations often span regions (promoters, enhancers, PMDs), not just single CpGs. Rocker-meth increases sensitivity to such patterns across arrays (HM450) and sequencing (WGBS, scMeth), improves functional interpretability (e.g., links to under/over-expression by genic context), and provides harmonized DMR catalogs across tumor types for downstream integration.

How it works (algorithm)

1. **Site-wise AUC:** For each CpG, compute ROC AUC contrasting tumor vs. normal beta values. AUC ≈ 1 indicates hypermethylation in tumor; AUC ≈ 0 indicates hypomethylation.
2. **Segmentation via heterogeneous HMM:** Model the AUC track with a 3-state HMM (hypo / neutral / hyper). Emissions are truncated Gaussians on [0,1]; transition probabilities incorporate genomic distance to favor longer, coherent segments.
3. **Intra-segment test:** Assess homogeneity/differential signal with a Wilcoxon–Mann–Whitney test on beta values within segments; adjust by FDR to retain robust DMRs.
4. **Sample-wise calling (optional):** Derive per-sample Z-scores over the cohort DMR set to quantify how strongly an individual supports each DMR (PSFSE/"burden"); enables single-sample interpretation and heterogeneity analyses.

When to apply (use cases)

- Tumor vs. adjacent-normal comparisons on HM450/WGBS.
- Pan-cancer catalogs and subtype stratification (e.g., PRAD subtypes, PAM50 in BRCA).
- Detection of large-scale hypomethylation (hypo-blocks/PMDs) and focal promoter hypermethylation.
- Single-cell methylation: project sparse scMeth data onto cohort-level DMRs to recover disease status and clonal structure.

Main findings Rocker-meth demonstrates superior performance compared to several existing methods, achieving higher precision and recall even under low signal-to-noise conditions in simulated datasets, with a specificity exceeding 0.99. When applied to TCGA whole-genome bisulfite sequencing (WGBS) data, the method revealed extensive hypomethylation spanning large genomic regions, corresponding to partially methylated domains (PMDs), while maintaining the ability to detect focal hypermethylation events. Analyses conducted on Illumina 450k data confirmed that these large-scale trends are consistent across platforms, underscoring the method's robustness. Functionally, Rocker-meth showed that hypermethylated DMRs are frequently located near transcription start sites and 5' untranslated regions, where they are associated with gene under-expression, whereas hypomethylated regions tend to occur within intronic and intergenic domains. By integrating chromatin-state information, the study further revealed that distinct chromatin contexts contribute to different patterns of transcription factor deregulation. Finally, when extended to single-cell methylation data, Rocker-meth successfully separated tumor from normal cells, capturing patient-specific methylation signatures and revealing that hypo-blocks exhibit limited intra-tumoral heterogeneity.

How to replicate the paper's plots

- **Performance/benchmark plots:** Precision–recall/F1 vs. methods on synthetic datasets; reproduce by simulating contrasts or using held-out truth labels.
- **DMR burden (lollipop/bar/dot):** Count gain (hyper) vs. loss (hypo, hypo-block) DMRs per tumor type or cohort; visualize genome fraction covered.
- **Length distributions (box/violin):** Segment-length distributions by event class (hyper/hypo/hypo-block) and genic vs. intergenic annotation.

- **PMD overlap (bar/stacked):** Fractional overlap of hypo-blocks with PMD/HMD catalogs; useful to confirm large-scale demethylation.
- **Cross-platform concordance (scatter):** Average $\Delta\beta$ agreement for overlapping DMRs between WGBS and HM450; add Pearson R and p.
- **Sharing/recurrence (line/area):** Fraction of DMRs shared across tumor types vs. DMSs to highlight regional vs. site-wise prevalence.
- **UMAP/PCA on DMR space:** Dimensionality reduction using per-DMR $\Delta\beta$ to show tumor-type/subtype structure.
- **Genomic context (density around TSS):** Density of DMR midpoints within ± 10 kb of TSS for hyper/hypo/hypo-blocks.
- **Genic/chromatin annotation (box/dot):** Enrichment across promoter, 5'UTR, intron, intergenic; ChromHMM state fractions per DMR class.
- **Expression integration (odds-ratio dot plots):** Fisher OR for hyper⇒under-expressed and hypo⇒over-expressed by genic context.
- **TF-focused analyses (lollipop/bars):** Enrichment of TF genes among concordant DMR–DEGs; distribution by TF families; chromatin-state contrasts (TssA vs. ReprPC).
- **Single-cell projections (heatmap/UMAP):** Heatmap of per-cell beta over DMRs and UMAP colored by lineage; PCFSE distributions per patient.

3 DNA methylation outlier burden, health, and ageing in Generation Scotland and the Lothian Birth Cohorts of 1921 and 1936

Keywords DNA methylation, epigenetic drift, outlier burden, ageing, survival, cohort studies, Lothian Birth Cohorts, Generation Scotland, CpG, stochastic epigenetic mutations [3]

Description of plots

- The study presents scatter and contour plots that visualize the relationship between age and $\log_{10}$-transformed DNA methylation outlier burden across both cross-sectional and longitudinal cohorts. Regression lines are superimposed to highlight the positive association between age and burden, before and after adjusting for covariates such as blood cell proportions and smoking status.
- Longitudinal line plots illustrate individual trajectories of methylation outlier burden across multiple waves of data collection. Each line corresponds to a participant, showing within-person variation and a general trend of increasing burden with advancing age.
- Density and histogram plots display the distribution of outlier counts per individual, emphasizing the strong right-skewness of raw data and the normalization achieved after $\log_{10}$ transformation.
- QQ plots are used to evaluate the distributional assumptions of the statistical models and to confirm the presence of heteroscedasticity across age groups.
- Kaplan–Meier survival curves compare individuals with high versus low outlier burden, revealing a modest reduction in survival probability among those with greater methylation irregularity, even after controlling for sex and age.
- Flow diagrams summarize the overall analytical workflow, from data preprocessing and outlier definition to regression modeling and survival analysis, providing a clear and reproducible structure for similar methylation studies.

Together, these visualizations provide a comprehensive depiction of how DNA methylation outlier burden evolves with age, demonstrating its stochastic accumulation and its potential use as an indicator of biological ageing.

4 A urine-based DNA methylation assay, ProCUrE, to identify clinically significant prostate cancer

Keywords Prostate cancer, DNA methylation, biomarkers, ProCUrE assay, HOXD3, GSTP1, ROC curve, LASSO model, PSA, urine test, risk stratification [4]

Description of plots

- Box and scatter plots visualize the methylation levels (percent methylated reference, PMR) of candidate genes across benign and cancer patients, showing a clear upward shift for tumor samples and identifying outliers through individual data points.
- Distribution plots display the comparative methylation frequencies of six biomarkers, highlighting the wide range of methylation levels among genes such as GSTP1 and HOXD3.
- Receiver Operating Characteristic (ROC) curves compare individual gene classifiers and the combined ProCUrE model, demonstrating improved diagnostic performance with higher area under the curve (AUC) values for the two-gene model.
- Bar plots illustrate the sensitivity and specificity of each biomarker and the ProCUrE composite score, clarifying its stronger predictive balance for aggressive prostate cancer.
- Comparative percentage plots show the proportions of true-positive and false-positive detections between ProCUrE and age-adjusted PSA, visually emphasizing the superior precision of the methylation-based assay.
- Risk-stratified bar charts display the percentage of patients testing positive across clinically insignificant versus significant disease categories based on multiple scoring systems (Gleason, CAPRA, D'Amico), offering a visual summary of diagnostic performance.
- Scatter and regression visualizations illustrate the additive diagnostic value of combining ProCUrE with PSA or PCPT scores, showing an upward shift in discriminative metrics (c-statistics) compared to PSA alone.

Together, these visualizations collectively demonstrate how the ProCUrE assay integrates DNA methylation biomarkers into a robust, urine-based framework capable of improving diagnostic accuracy and risk stratification for clinically significant prostate cancer.

References

[1] J. Maksimovic, L. Gordon, and A. Oshlack, "Swan: Subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips," *Genome Biology*, vol. 13, R44, 2012, Open Access. DOI: 10.1186/gb-2012-13-6-r44. [Online]. Available: http://genomebiology.com/2012/13/6/R44.

[2] M. Benelli et al., "Charting differentially methylated regions in cancer with rocker-meth," *Communications Biology*, vol. 4, no. 1, p. 1249, 2021. DOI: 10.1038/s42003-021-02761-3.

[3] A. Seeboth et al., "Dna methylation outlier burden, health, and ageing in generation scotland and the lothian birth cohorts of 1921 and 1936," *Clinical Epigenetics*, vol. 12, no. 49, 2020. DOI: 10.1186/s13148-020-00838-0. [Online]. Available: https://doi.org/10.1186/s13148-020-00838-0.

[4] F. Zhao et al., "A urine-based dna methylation assay, procure, to identify clinically significant prostate cancer," *Clinical Epigenetics*, vol. 10, no. 147, 2018. DOI: 10.1186/s13148-018-0575-z. [Online]. Available: https://doi.org/10.1186/s13148-018-0575-z.