Title

Computational method development and analysis for DNA methylome studies

Author

Guo, Wenbin

Publication Date

2024

UNIVERSITY OF CALIFORNIA

Los Angeles

Computational method development and analysis for DNA methylome studies

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioinformatics

by

Wenbin Guo

2024

ABSTRACT OF THE DISSERTATION

Computational method development and analysis for DNA methylome studies

by

Wenbin Guo

Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2024
Professor Matteo Pellegrini, Chair

DNA methylation underpins a wide range of biological processes and disease states, yet significant challenges persist in its computational analysis and practical applications. This dissertation presents advancements in three areas of DNA methylation research, including simulation method development and analysis in disease and aging. These advancements contribute to improved methodologies and a deeper understanding of the DNA methylome in health and disease, paving the way for fundamental research and clinical translation.

In the first part, we introduce BSReadSim, a novel bisulfite sequencing simulator that overcomes key limitations of existing tools, which often fail to capture the complexity of real-world data. By accurately integrating genetic variants, methylation profiles, and technical artifacts, BSReadSim produces synthetic datasets that closely resemble empirical observations. This versatile resource provides a robust framework for designing experiments, developing computational methods, and benchmarking analytical pipelines in DNA methylation research, ultimately enhancing the rigor and reliability of epigenetic studies.

The second part of this dissertation pioneers the use of saliva DNA methylomes for type 2 diabetes (T2D) biomarker discovery and risk assessment. By integrating comprehensive Whole Genome Bisulfite Sequencing (WGBS) and high-depth Targeted Bisulfite Sequencing (TBS), we developed a cost-effective, two-step research strategy for DNA methylation studies and identified T2D-associated methylation biomarkers in saliva. Importantly, we demonstrated that these epigenetic signatures are primarily intrinsic rather than driven by cell composition shifts, establishing saliva DNA methylome as a compelling non-invasive medium for T2D biomarker exploration. This approach holds substantial potential for both fundamental research and clinical applications, ultimately informing improved disease detection, monitoring, and personalized treatment strategies.

The third part of this work examines the interplay between epigenetic aging, cell composition, and breast cancer risk in normal breast tissue. By analyzing 181 normal breast samples, we revealed systematic biases in existing epigenetic clocks, highlighting the need for tissue-specific models to achieve accurate age predictions. Our findings established a clear link between epigenetic age acceleration and shifts in cell composition, particularly those associated with elevated breast cancer risk. Notably, we provided plausible molecular evidence connecting estrogen exposure to accelerated epigenetic aging and increased cancer susceptibility. These insights highlight the potential of epigenetic clocks as powerful tools for cancer risk assessment and stratification.

Together, these studies advance the field of DNA methylation by expanding our capacity to understand, interpret, and harness the wealth of information encoded within the DNA methylome. Through developing a realistic synthetic data simulator, exploring non-invasive avenues for disease biomarker discovery, and shedding light on the molecular underpinnings of epigenetic aging, this dissertation establishes a strong foundation for more rigorous, scalable, and impactful DNA methylation research. These advancements deepen our understanding of epigenetic regulation in health and disease, ultimately paving the way for transformative applications in diagnostics, risk assessment, and future epigenetic studies.

The dissertation of Wenbin Guo is approved.

Xia Yang

Xinshu Xiao

Jingyi Li

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2024

To my family

for their unconditional support and everlasting love

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the guidance, support, and encouragement of many incredible individuals.

First and foremost, I would like to express my deepest gratitude to my advisor, Matteo Pellegrini, for his unwavering support, boundless patience, and invaluable guidance throughout my PhD journey. His brilliance as a scientist, kindness as a mentor, and calm, solution-driven personality have been a constant source of inspiration, profoundly shaping both my academic and personal growth.

I am deeply grateful to Dr. Jingyi Jessica Li for her teaching and mentorship, for opening my eyes to the world of statistics, and for serving as a role model of academic excellence. I also extend my heartfelt thanks to my committee members: Dr. Xinshu Grace Xiao for her kind, encouraging words and constructive feedback, Dr. Xia Yang for her thought-provoking questions and inspiring scientific insights. To my undergraduate mentor Dr. Tao Wang, for inspiring my journey into bioinformatics and encouraging my professional growth. Their collective wisdom and support have been a tremendous privilege and blessing to me.

I want to thank my collaborators, including Dr. Lili Yang, Dr. Yanruide Li, Dr. Derek Lee, Dr. Mary Sehl, and Pranav Kannan, whose expertise and dedication have greatly enriched my research. To my current and former lab members—Dr. Colin Farrell, Dr. Mike Thompson, Dr. Marco Morselli, etc.—and my colleagues from JSB and so on, thank you for making this journey both rewarding and enjoyable.

To my cohort—Yi Ding, Cyrillus Tan, Nick Bayley, and Russell Littman—and the UCLA Bioinformatics Program, including Gene Gray and Eloy Lopez, thank you for fostering an inclusive and supportive community. I am also deeply grateful to the QCBio Collaboratory community for cultivating a collaborative environment and providing invaluable opportunities.

To my family—my mom, dad, sister, and extended family—your unconditional love and support have been my foundation through every challenge and triumph. To my friends from Southern California, San Francisco, Texas, and beyond, your presence has made my life brighter and more meaningful.

Lastly, as I reflect on this journey, I am reminded of a special moment from UCLA's centennial celebration:

> “ All of it is possible; generations have proven it, and now passing the torch to you, how will you light the way? ”

To all who have lighted my way, thank you.

VITA

2014-2018	B.S. in *Biological Science*, Wuhan University.
2016-2018	B.E. in *Computer Science*, Huazhong University of Science and Technology.
2018-2024	Graduate student researcher in *Bioinformatics*, University of California, Los Angeles.
2023-2024	Articulated M.S. student in *Statistics*, University of California, Los Angeles.

PUBLICATIONS

(* indicates equal contributions)

Guo, **W.** and Pellegrini, M. “BSReadSim: a versatile and efficient simulator to generate realistic bisulfite sequencing reads”. *bioRxiv* (2024)

Guo, **W.**, Morselli, M., Paul, K. C., Thompson, M., Ritz, B., and Pellegrini, M. “Type-2 diabetes biomarker discovery and risk assessment through saliva DNA methylome”. *medRxiv* (2024)

Sehl, M. E.*, **Guo**, **W.***, Farrell, C., Marino, N., Henry, J. E., Storniolo, A. M., Papp, J., Li, J. J., Horvath, S., Pellegrini, M., et al. “Systematic dissection of epigenetic age acceleration in normal breast tissue reveals its link to estrogen signaling and cancer risk”. *bioRxiv* (2024)

Li, Y.-R.*, Zhou, Y.*, Yu, J.*, Kim, Y. J., Li, M., Lee, D., Zhou, K., Chen, Y., Zhu, Y., Wang, Y.-C., et al. “Generation of allogeneic CAR-NKT cells from hematopoietic stem and progenitor cells using a clinically guided culture method”. *Nature Biotechnology* (2024), pp. 1–16

Lee, D.*, Dunn, Z. S.*, **Guo**, **W.***, Rosenthal, C. J., Penn, N. E., Yu, Y., Zhou, K., Li, Z., Ma, F., Li, M., et al. “Unlocking the potential of allogeneic Vδ2 T cells for ovarian cancer therapy through CD16 biomarker selection and CAR/IL-15 engineering”. *Nature Communications* 14.1 (2023), p. 6942

CHAPTER 1

Introduction

1.1 Overview of DNA methylation

1.1.1 DNA methylation: a key epigenetic regulator

DNA methylation, a key epigenetic modification discovered in the 20th century, is essential for regulating gene expression beyond the genetic code [1]. This biochemical process entails the addition of a methyl group to the 5th carbon of cytosine's pyrimidine ring, forming 5-methylcytosine (5mC) (Figure 1.1). In mammals, DNA methylation predominantly occurs at CpG dinucleotides [2], where a cytosine nucleotide is immediately followed by a guanine nucleotide in the 5' to 3' direction. Regions enriched with CpG sites, known as CpG islands, are often found in gene promoters. Methylating these regions generally represses gene expression [3], providing a crucial mechanism for regulating gene activity across tissues and developmental stages.

Beyond transcriptional regulation, DNA methylation is integral to numerous fundamental biological processes, such as genomic imprinting [4], X-chromosome inactivation [5], transposable element suppression [6], and embryogenesis [7]. Moreover, it orchestrates cellular differentiation and mediates the maintenance of cell identity [8], exemplifying Waddington's epigenetic landscape where cells with identical genetic material can follow distinct developmental trajectories to establish specialized functions [9]. The evolutionary conservation of DNA methylation across species underscores its fundamental importance in these processes [10], highlighting its indispensable role in biological systems.

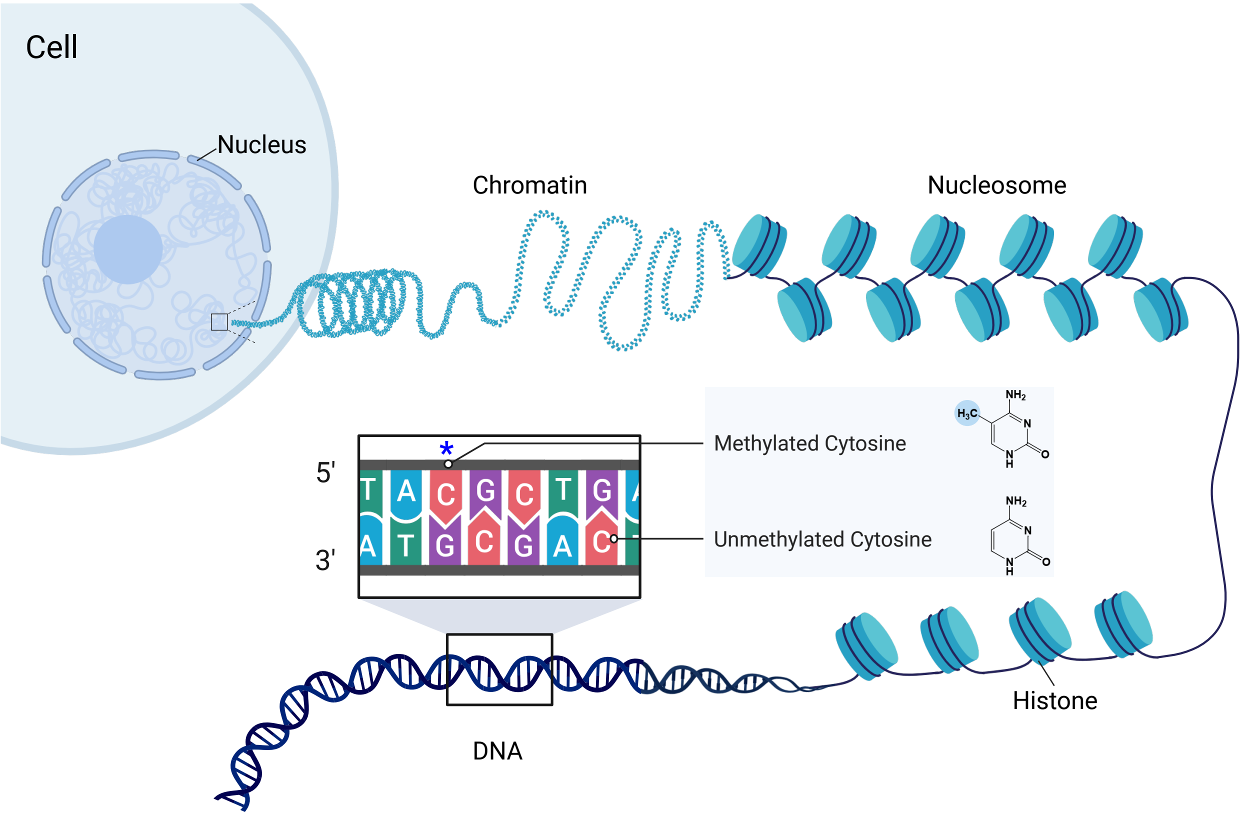


Figure 1.1: Overview of DNA methylation.

1.1.2 Establishment and maintenance of DNA methylation

The establishment and maintenance of DNA methylation is governed by DNA methyltransferases (DNMTs). DNMT1 preserves existing methylation patterns during DNA replication [11], ensuring the faithful inheritance of epigenetic marks during cell division. DNMT3A and DNMT3B establish de novo methylation patterns during development and in response to environmental stimuli [12]. Demethylation can occur passively or actively, with active removal mediated by ten-eleven translocation (TET) enzymes [13, 14]. This dynamic interplay between methylation and demethylation is essential for regulating gene expression, directing cell differentiation, and maintaining cellular identity. Dysregulation of these processes is implicated in pathological conditions, such as cancer, where global hypomethylation and site-specific hypermethylation of tumor suppressor genes are hallmark features [15, 16].

1.1.3 Clinical and therapeutic relevance

DNA methylation has gained substantial attention in clinical research for its potential to reflect disease status and elucidate disease mechanisms. Environmental factors such as diet and lifestyle can induce lasting epigenetic modifications, linking DNA methylation to various health outcomes and diseases [17, 18]. Aberrant methylation patterns have been linked to numerous diseases, serving as both drivers and consequences of pathogenesis [19, 20]. These discoveries have fueled the development of DNA methylation-based biomarkers, which were utilized for disease detection, and monitoring [21, 22].

On the other hand, epigenetic therapies targeting DNA methylation, such as DNMT inhibitors azacitidine and decitabine, are already in clinical use for certain hematological cancers [23–25]. These therapies aim to reverse aberrant methylation patterns by epigenetic reprogramming, offering promise for disease treatment. Besides cancer, DNA methylation-based therapies are being explored in neurodegenerative diseases [26], autoimmune disorders [27], and age-related conditions [28, 29], highlighting their potential for a broad spectrum of epigenetically driven diseases. As our knowledge of DNA methylation expands, it is increasingly evident that studying this epigenetic mechanism is crucial for advancing biological discovery and developing innovative diagnostic and therapeutic strategies.

1.2 DNA methylation profiling

DNA methylation profiling encompasses a variety of techniques to measure methylation status across the genome [30]. Among these, bisulfite sequencing [31] has emerged as an essential tool, providing unparalleled precision to differentiate methylated and unmethylated cytosines at single-nucleotide resolution. This approach involves treating genomic DNA with sodium bisulfite, where unmethylated cytosines are converted into uracil and subsequently read as thymine during sequencing, while methylated cytosines are left intact. By comparing the treated DNA sequence to a reference sequence, researchers can reconstruct the original methylation patterns on the DNA sequences. In practice, bisulfite sequencing can be performed in different forms according to the research objective:

- **Whole Genome Bisulfite Sequencing (WGBS)**: This approach profiles the entire genome, providing a comprehensive view of DNA methylation. While WGBS delivers an unbiased and complete methylation landscape, it generates a large volume of data, incurring higher costs and demanding substantial computational resources for processing and analysis.
- **Reduced Representation Bisulfite Sequencing (RRBS)**: To optimize costs and reduce data complexity, RRBS focuses on CpG-rich regions of the genome, such as CpG islands. By using enzymatic digestion and size selection to enrich these regions, RRBS offers a more targeted analysis while capturing key methylation patterns.
- **Targeted Bisulfite Sequencing (TBS)**: TBS employs probes or primers to capture and sequence specific regions of interest. This method achieves deep coverage of selected loci, making it particularly suited for validating results from broader studies or investigating candidate regions linked to disease.

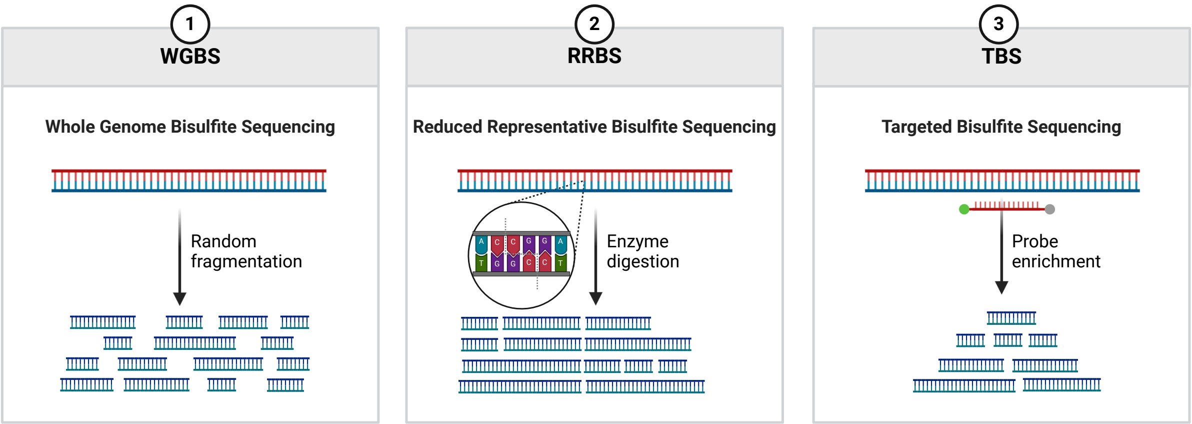

1
WGBS
Whole Genome Bisulfite Sequencing
Random fragmentation
2
RRBS
Reduced Representative Bisulfite Sequencing
Enzyme digestion
A C C G G A
T G G C C T
3
TBS
Targeted Bisulfite Sequencing
Probe enrichment

Figure 1.2: Three types of bisulfite sequencing technology.

Over the past two decades, bisulfite sequencing has seen remarkable advancements, significantly improving its accuracy, efficiency, and research applicability. High-throughput platforms like Illumina and PacBio now enable large-scale, cost-effective methylation studies with high resolution. Enhanced library preparation methods [32, 33], including strategies to mitigate bisulfite-induced DNA degradation, have further improved data quality and reliability, solidifying bisulfite sequencing as the gold standard for DNA methylome profiling. Beyond bisulfite sequencing, other innovative approaches such as EPIC microarray [34], nanopore sequencing [35], and TAPS [36] have been developed, and the advent of single-cell methylation profiling techniques [37–39] has further propelled the field. Collectively, these technological advancements enable more precise DNA methylation profiling, advance the study of epigenetic regulation, and broaden the applications of DNA methylation research in both research and clinical applications.

1.3 Downstream analysis of DNA methylation

The DNA methylation profiling techniques provide a quantitative framework for assessing methylation status at specific genomic loci, enabling a detailed exploration of the epigenetic landscape. By comparing the methylation-supporting measure (e.g., signal intensity for microarray data or methylated counts for sequencing data) to the total measure (e.g., combined signal intensity or total read counts covering a site), a ratio or fraction can be calculated, representing the average methylation level at each site across the cell(s) in the sample. The collection of DNA methylation levels (DNA methylome) enables a wide range of downstream analyses and allows researchers to explore the alternation and regulation in detail.

1.3.1 Differential methylation analysis

Differential methylation analysis aims to identify methylation changes associated with variables of interest. This analysis typically involves testing methylation level differences across conditions (e.g., disease versus non-disease), either at individual cytosines or within specific genomic regions. Depending on the data characteristics, parametric tests like the Student's t-test or nonparametric tests like the Mann-Whitney U test or the Kolmogorov-Smirnov (KS) test can be applied [40]. Advanced methods such as DSS [41] and MACAU [42] used hierarchical models with beta-binomial distributions to model the methylated and unmethylated read counts. These models account for sampling variations and provide robust estimates of differentially methylated sites or regions, achieving higher statistical power to detect small effect sizes.

1.3.2 Epigenome-wide association studies

Epigenome-wide association studies (EWAS) aim to identify associations between DNA methylation changes and specific traits or diseases across large populations. The analysis examines the entire DNA methylome to uncover methylation changes correlating with complex phenotypes, such as disease risk, environmental exposures, etc. Regression frameworks are typically used in statistical tests, e.g. linear mixture models [43], to manage the vast number of sites and adjust for potential confounders like population stratification and cell-type heterogeneity. Additionally, regression frameworks utilizing beta-binomial framework [42] are also used to account for technical and biological variability observed in bisulfite sequencing data. By incorporating covariates and controlling for potential confounders, these methods ensure precise identification of methylation changes associated with complex traits or diseases, thereby providing insights into the underlying epigenetic mechanisms and allowing for the discovery of potential diagnostic or prognostic biomarkers.

1.3.3 Cell type abundance deconvolution

Current DNA methylation profiling at the single-cell level still faces challenges such as high costs and low sequencing coverage. Consequently, most DNA methylation profiling is performed on bulk samples, which are composed of a mixture of cell types. Accurately determining the proportion of each cell type within these bulk samples is essential for interpreting epigenetic data and understanding the underlying biology of complex tissues.

The deconvolution process typically involves using reference profiles of known cell types, where cell-type specific methylation signatures can be identified. By applying algorithms such as Non-negative Least Squares [44], the proportions of each cell type in the bulk sample can be estimated by minimizing the construction loss for the observed bulk methylation data. Advanced models, such as CIBERSORT [45] and EpiDISH [46], further enhance accuracy through machine learning and Bayesian approaches, collectively enabling precise profiling of cell proportional dynamics in complex tissues.

1.3.4 Biomarker discovery and disease states prediction

DNA methylation's stable nature and critical role in reflecting biological and disease states make it a compelling target for biomarker study in aging and a wide range of diseases. With base-resolution profiling techniques, current research focuses on uncovering disease-specific methylation patterns and leveraging them for early diagnosis, disease monitoring, and personalized medicine. For example, hypermethylation of tumor suppressor genes is employed for early cancer detection [47], while methylation changes in circulating free DNA (cfDNA) are tracked through liquid biopsies to monitor disease progression [48]. Furthermore, machine learning approaches are increasingly utilized to develop methylation-based risk scores for complex diseases [49]. These advancements pave the way for personalized medicine, enabling prevention and treatment strategies tailored to an individual's epigenetic profile.

1.4 Structure of the dissertation

A comprehensive understanding of DNA methylation is pivotal for advancing epigenetic research and enhancing clinical applications. This dissertation contributes to this field by addressing several computational and analytical challenges in DNA methylome studies. It focuses on developing a novel bisulfite sequencing simulator, identifying noninvasive biomarkers, and understanding the role of DNA methylation changes in aging and disease.

chapter 1 introduces the foundational concepts of DNA methylation, covering its biological functions, establishment and maintenance, and relevance in both basic research and clinical applications. It offers an overview of methylation profiling techniques, with a particular focus on bisulfite sequencing, and discusses the methodologies used to analyze methylation data. By establishing the necessary knowledge background, it sets the stage for the more specialized discussions in the following chapters.

chapter 2 introduces BSReadSim, a novel generative framework for simulating realistic bisulfite sequencing reads. Unlike existing tools, BSReadSim can integrate genetic variants and methylation profiles, enabling profile-based simulations while accounting for technical variabilities. By addressing limitations in current methodologies, it facilitates the experiment design for DNA methylome studies, development and benchmark of computational tools. By improving the realism and flexibility of bisulfite sequencing simulations, BSReadSim can enhance the reliability and rigor of method development for DNA methylation analysis.

chapter 3 investigates the potential of using saliva DNA methylome for type 2 diabetes (T2D) biomarker discovery and risk assessment. By combining comprehensive WGBS with high-depth TBS, this study identifies and profiles diabetes-specific epigenetic signals in saliva, while demonstrating a practical and cost-effective research scheme for epigenetic biomarker discovery that achieves both broad coverage and targeted precision in DNA methylation profiling. It validates, for the first time, saliva DNA methylation as a reliable, non-invasive biomarker for T2D, offering a promising alternative for future research and clinical diagnostics.

chapter 4 examines the molecular and cellular changes in aging breast tissue and their connection to cancer risk. By analyzing DNA methylation and gene expression from 181 normal breast samples, the study evaluates eight epigenetic clocks, highlighting their inherent biases in age estimation and the need for refined definitions of age acceleration. It reveals how age-related shifts in cell composition and CpG site methylation, particularly those enriched for estrogen receptor binding, link accelerated aging to cancer risk. These findings underscore the importance of addressing model biases and cellular heterogeneity when interpreting epigenetic age estimates and highlight the potential of age acceleration metrics for cancer risk stratification and prevention.

chapter 5 summarizes the main findings of this dissertation, focusing on contributions to developing computational tools, validating non-invasive biomarkers, and investigating epigenetic changes in T2D disease and aging. It also outlines potential future research directions to advance computational methods and analytical approaches in DNA methylome studies.

1.5 References

[1] Mattei, A. L., Bailly, N., and Meissner, A. "DNA methylation: a historical perspective". *Trends in Genetics* 38.7 (2022), pp. 676–707.

[2] Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., Hetzel, J., Jain, J., Strauss, S. H., Halpern, M. E., et al. "Conservation and divergence of methylation patterning in plants and animals". *Proceedings of the National Academy of Sciences* 107.19 (2010), pp. 8689–8694.

[3] Moore, L. D., Le, T., and Fan, G. "DNA methylation and its basic function". *Neuropsychopharmacology* 38.1 (2013), pp. 23–38.

[4] Ferguson-Smith, A. C. "Genomic imprinting: the emergence of an epigenetic paradigm". *Nature Reviews Genetics* 12.8 (2011), pp. 565–575.

[5] Mohandas, T., Sparkes, R., and Shapiro, L. "Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation". *Science* 211.4480 (1981), pp. 393–396.

[6] Deniz, Ö., Frost, J. M., and Branco, M. R. "Regulation of transposable elements by DNA modifications". *Nature Reviews Genetics* 20.7 (2019), pp. 417–431.

[7] Fouse, S. D., Shen, Y., Pellegrini, M., Cole, S., Meissner, A., Van Neste, L., Jaenisch, R., and Fan, G. "Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation". *Cell stem cell* 2.2 (2008), pp. 160–169.

[8] Moris, N., Pina, C., and Arias, A. M. "Transition states and cell fate decisions in epigenetic landscapes". *Nature Reviews Genetics* 17.11 (2016), pp. 693–703.

[9] Parry, A., Rulands, S., and Reik, W. "Active turnover of DNA methylation during cell fate decisions". *Nature Reviews Genetics* 22.1 (2021), pp. 59–66.

[10] Catania, S., Dumesic, P. A., Pimentel, H., Nasif, A., Stoddard, C. I., Burke, J. E., Diedrich, J. K., Cooke, S., Shea, T., Gienger, E., et al. "Evolutionary persistence of DNA methylation for millions of years after ancient loss of a de novo methyltransferase". *Cell* 180.2 (2020), pp. 263–277.

[11] Goyal, R., Reinhardt, R., and Jeltsch, A. "Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase". *Nucleic acids research* 34.4 (2006), pp. 1182–1188.

[12] Yagi, M., Kabata, M., Tanaka, A., Ukai, T., Ohta, S., Nakabayashi, K., Shimizu, M., Hata, K., Meissner, A., Yamamoto, T., et al. "Identification of distinct loci for de novo DNA methylation by DNMT3A and DNMT3B during mammalian development". *Nature communications* 11.1 (2020), p. 3199.

[13] Kohli, R. M. and Zhang, Y. "TET enzymes, TDG and the dynamics of DNA demethylation". *Nature* 502.7472 (2013), pp. 472–479.

[14] Vincent, J. J., Huang, Y., Chen, P.-Y., Feng, S., Calvopiña, J. H., Nee, K., Lee, S. A., Le, T., Yoon, A. J., Faull, K., et al. "Stage-specific roles for tet1 and tet2 in DNA demethylation in primordial germ cells". *Cell stem cell* 12.4 (2013), pp. 470–478.

[15] Feinberg, A. P., Ohlsson, R., and Henikoff, S. "The epigenetic progenitor origin of human cancer". *Nature reviews genetics* 7.1 (2006), pp. 21–33.

[16] Yang, J., Xu, J., Wang, W., Zhang, B., Yu, X., and Shi, S. "Epigenetic regulation in the tumor microenvironment: molecular mechanisms and therapeutic targets". *Signal transduction and targeted therapy* 8.1 (2023), p. 210.

[17] Martin, E. M. and Fry, R. C. "Environmental influences on the epigenome: exposure-associated DNA methylation in human populations". *Annual review of public health* 39.1 (2018), pp. 309–333.

[18] Cavalli, G. and Heard, E. "Advances in epigenetics link genetics to the environment and disease". *Nature* 571.7766 (2019), pp. 489–499.

[19] Jin, Z. and Liu, Y. "DNA methylation in human diseases". *Genes & diseases* 5.1 (2018), pp. 1–8.

[20] Greenberg, M. V. and Bourc'his, D. "The diverse roles of DNA methylation in mammalian development and disease". *Nature reviews Molecular cell biology* 20.10 (2019), pp. 590–607.

[21] Li, W. and Zhou, X. J. "Methylation extends the reach of liquid biopsy in cancer detection". *Nature Reviews Clinical Oncology* 17.11 (2020), pp. 655–656.

[22] Yousefi, P. D., Suderman, M., Langdon, R., Whitehurst, O., Davey Smith, G., and Relton, C. L. "DNA methylation-based predictors of health: applications and statistical considerations". *Nature Reviews Genetics* 23.6 (2022), pp. 369–383.

[23] Flotho, C., Claus, R., Batz, C., Schneider, M., Sandrock, I., Ihde, S., Plass, C., Niemeyer, C., and Lübbert, M. "The DNA methyltransferase inhibitors azacitidine, decitabine and zebularine exert differential effects on cancer gene expression in acute myeloid leukemia cells". *Leukemia* 23.6 (2009), pp. 1019–1028.

[24] Wang, Y., Tong, C., Dai, H., Wu, Z., Han, X., Guo, Y., Chen, D., Wei, J., Ti, D., Liu, Z., et al. "Low-dose decitabine priming endows CAR T cells with enhanced and persistent antitumour potential via epigenetic reprogramming". *Nature communications* 12.1 (2021), p. 409.

[25] Kelly, T. K., De Carvalho, D. D., and Jones, P. A. "Epigenetic modifications as therapeutic targets". *Nature biotechnology* 28.10 (2010), pp. 1069–1078.

[26] Zhang, L., Liu, Y., Lu, Y., and Wang, G. "Targeting epigenetics as a promising therapeutic strategy for treatment of neurodegenerative diseases". *Biochemical pharmacology* 206 (2022), p. 115295.

[27] Mekinian, A., Zhao, L. P., Chevret, S., Desseaux, K., Pascal, L., Comont, T., Maria, A., Peterlin, P., Terriou, L., D'Aveni Piney, M., et al. "A Phase II prospective trial of azacitidine in steroid-dependent or refractory systemic autoimmune/inflammatory

disorders and VEXAS syndrome associated with MDS and CMML". *Leukemia* 36.11 (2022), pp. 2739–2742.

[28] Selvarani, R., Mohammed, S., and Richardson, A. "Effect of rapamycin on aging and age-related diseases—past and future". *Geroscience* 43 (2021), pp. 1135–1158.

[29] Reale, A., Tagliatesta, S., Zardo, G., and Zampieri, M. "Counteracting aged DNA methylation states to combat ageing and age-related diseases". *Mechanisms of Ageing and Development* 206 (2022), p. 111695.

[30] Yong, W.-S., Hsu, F.-M., and Chen, P.-Y. "Profiling genome-wide DNA methylation". *Epigenetics & chromatin* 9 (2016), pp. 1–16.

[31] Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning". *Nature* 452.7184 (2008), pp. 215–219.

[32] Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R. V., Branco, M. R., and Reik, W. "Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data". *Genome biology* 19 (2018), pp. 1–19.

[33] Miura, F., Shibata, Y., Miura, M., Sangatsuda, Y., Hisano, O., Araki, H., and Ito, T. "Highly efficient single-stranded DNA ligation technique improves low-input whole-genome bisulfite sequencing by post-bisulfite adaptor tagging". *Nucleic acids research* 47.15 (2019), e85–e85.

[34] Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., and Clark, S. J. "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling". *Genome biology* 17 (2016), pp. 1–17.

[35] Liu, Y., Rosikiewicz, W., Pan, Z., Jillette, N., Wang, P., Taghbalout, A., Foox, J., Mason, C., Carroll, M., Cheng, A., et al. “DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation”. *Genome biology* 22 (2021), pp. 1–33.

[36] Liu, Y., Siejka-Zielińska, P., Velikova, G., Bi, Y., Yuan, F., Tomkova, M., Bai, C., Chen, L., Schuster-Böckler, B., and Song, C.-X. “Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution”. *Nature biotechnology* 37.4 (2019), pp. 424–429.

[37] Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. “Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity”. *Nature methods* 11.8 (2014), pp. 817–820.

[38] Clark, S. J., Smallwood, S. A., Lee, H. J., Krueger, F., Reik, W., and Kelsey, G. “Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq)”. *Nature protocols* 12.3 (2017), pp. 534–547.

[39] Cao, Y., Bai, Y., Yuan, T., Song, L., Fan, Y., Ren, L., Song, W., Peng, J., An, R., Gu, Q., et al. “Single-cell bisulfite-free 5mC and 5hmC sequencing with high sensitivity and scalability”. *Proceedings of the National Academy of Sciences* 120.49 (2023), e2310367120.

[40] Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., and Hoffmann, S. “Metilene: Fast and Sensitive Calling of Differentially Methylated Regions from Bisulfite Sequencing Data”. *Genome Research* 26.2 (2016), pp. 256–262.

[41] Park, Y. and Wu, H. “Differential Methylation Analysis for BS-seq Data under General Experimental Design”. *Bioinformatics* 32.10 (2016), pp. 1446–1453.

[42] Lea, A. J., Tung, J., and Zhou, X. "A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data". *PLoS genetics* 11.11 (2015), e1005650.

[43] Rahmani, E., Yedidim, R., Shenhav, L., Schweiger, R., Weissbrod, O., Zaitlen, N., and Halperin, E. "GLINT: a user-friendly toolset for the analysis of high-throughput DNA-methylation array data". *Bioinformatics* 33.12 (2017), pp. 1870–1872.

[44] Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. "DNA Methylation Arrays as Surrogate Measures of Cell Mixture Distribution". *BMC Bioinformatics* 13.1 (2012), p. 86.

[45] Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. "Robust enumeration of cell subsets from tissue expression profiles". *Nature methods* 12.5 (2015), pp. 453–457.

[46] Teschendorff, A. E., Breeze, C. E., Zheng, S. C., and Beck, S. "A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies". *BMC bioinformatics* 18 (2017), pp. 1–14.

[47] Topaloglu, O., Hoque, M. O., Tokumaru, Y., Lee, J., Ratovitski, E., Sidransky, D., and Moon, C.-s. "Detection of promoter hypermethylation of multiple genes in the tumor and bronchoalveolar lavage of patients with lung cancer". *Clinical cancer research* 10.7 (2004), pp. 2284–2288.

[48] Li, S., Zeng, W., Ni, X., Liu, Q., Li, W., Stackpole, M. L., Zhou, Y., Gower, A., Krysan, K., Ahuja, P., et al. "Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring". *Proceedings of the National Academy of Sciences* 120.28 (2023), e2305236120.

[49] Nabais, M. F., Gadd, D. A., Hannon, E., Mill, J., McRae, A. F., and Wray, N. R. "An overview of DNA methylation-derived trait score methods and applications". *Genome Biology* 24.1 (2023), p. 28.

CHAPTER 2

BSReadSim: a versatile and efficient simulator to generate realistic bisulfite sequencing reads

Abstract

Realistic bisulfite sequencing simulators are crucial for advancing method development in computational epigenetics. However, existing tools often fall short due to oversimplified generative models that fail to capture the complexity of real data. We present BSReadSim, an efficient and versatile simulator that generatesrealistic bisulfite sequencing reads. BSReadSim excels in integrating reference genetic variants and methylation profiles, offering unmatched versatility across multiple sequencing technologies, including WGBS, RRBS, and TBS. By accurately modeling methylation patterns, sampling biases, sequencing errors, and leveraging optimized implementation, BSReadSim efficiently generates realistic synthetic datasets tailored to specific experimental needs while maintaining computational feasibility. By enhancing the realism and flexibility of bisulfite sequencing simulations, BSReadSim supports improved experiment design, method development, and benchmarking of computational tools, ultimately advancing the reliability and rigor of DNA methylation analysis tools.

Key words: DNA methylation; Bisulfite sequencing; Simulation; Synthetic data; Computational tools;

2.1 Introduction

DNA methylation is a crucial epigenetic modification involving the addition of a methyl group to the fifth carbon of cytosine's pyrimidine ring. In mammals, this modification predominantly occurs at CpG dinucleotides and plays a critical role in regulating gene expression [1], maintaining genomic stability [2], and underpinning fundamental biological processes such as cellular differentiation [3], development [4], and responses to environmental stimuli [5]. Aberrant DNA methylation has been linked to various diseases [6], establishing it as a central focus in biomedical research for uncovering disease mechanisms and advancing innovative diagnostic and therapeutic strategies.

Bisulfite sequencing (BS-seq) is widely recognized as the gold standard for profiling DNA methylation at single-base resolution. In this approach, genomic DNA is treated with sodium bisulfite, where unmethylated cytosines (C) are converted to (U) and subsequently read as thymines (T) during sequencing. In contrast, methylated cytosines (mC) remain unaltered, preserving their sequence identity (C) [7]. This chemical distinction allows for precise differentiation between methylated and unmethylated cytosines, enabling accurate quantification of methylation levels at individual cytosine sites. Despite its unparalleled resolution and accuracy, BS-seq data present significant analytical challenges due to the inherent complexity of DNA methylation dynamics, bisulfite-induced base changes, and technical variability in the sequencing process.

Since the development of bisulfite sequencing, various computational tools have been developed to analyze the bisulfite sequencing data, including specialized read aligners [8–18], SNP-callers [14, 17–23], as well as tools for identifying allelic-specific methylation [22, 24, 25], with nearly every method claiming to achieve the best performance. The pressing need for rigorously benchmarking these tools calls for a reliable simulator to generate realistic bisulfite sequencing data with ground truth. Additionally, a versatile simulator would be invaluable in experimental design. By simulating various scenarios, researchers can determine the best sequencing strategy and optimal sequencing depth needed to obtain accurate methylation measurements, ensuring that experiments are both cost-effective and well-powered. The dual utilities highlight the profound importance of developing a comprehensive bisulfite sequencing simulator to advance both computational tool development and epigenetic research.

Several bisulfite sequencing read simulators have been developed over the years, each tailored to specific technologies with distinct limitations. Among these, Sherman [26], BSBolt [27], and BSSim [28] focus on Whole Genome Bisulfite Sequencing (WGBS). Sherman provides basic simulation functionality but lacks support for genetic variant input and complex methylation profiles. BSBolt allows input of methylation profiles but fails to preserve site-specific methylation levels during simulation. BSSim can support limited genetic variant input through SNP frequency tables, yet it does not incorporate individual genotype data or handle indels. For Reduced Representation Bisulfite Sequencing (RRBS), RRBSsim [29] offers limited support for genetic variants but cannot integrate methylation profiles. Lastly, MethylFASTQ [30] supports both WGBS and Targeted Bisulfite Sequencing (TBS). However, its inability to differentiate between Watson and Crick strands in TBS compromises its utility for targeted sequencing applications. Additionally, like other tools, it does not support genetic variant input or methylation profile simulations. A comprehensive summary of the capabilities and limitations of these simulators is provided in Table 2.1.

Despite their varying focuses, existing bisulfite sequencing simulators share several critical limitations. First, they fail to fully integrate genetic and epigenetic profiles into the simulated bisulfite sequencing data, constraining their utility in computational tool development, benchmarking, and experimental design. Additionally, these simulators rely on oversimplified generative models, where DNA fragments, base quality scores, and sequencing errors are sampled using uniform probability models. This approach neglects the inherent complexity of bisulfite sequencing data, resulting in less realistic simulations. Furthermore, many of these tools struggle with computational efficiency, and none can simulate data across all three technologies (WGBS, RRBS, and TBS), further limiting their practicality for large-scale applications.

To address these limitations, we propose a novel bisulfite sequencing simulator, BSReadSim, incorporating advanced features such as detailed genetic variant and methylation profile inputs, allele-specific methylation, non-uniform coverage sampling, and quality score and sequencing error modeling. Designed with high-efficiency implementation, the simulator generates realistic bisulfite sequencing data within a practical timeframe. By supporting multiple bisulfite sequencing technologies (WGBS, RRBS, and TBS), BSReadSim provides a robust platform for benchmarking and validating bioinformatics tools under realistic conditions, facilitating experimental design and serving as a valuable resource for computational epigenomic research.

2.2 Methods

2.2.1 BSReadSim overview

BSReadSim is designed to generate realistic reads that mimic biological and technical variations observed in real bisulfite sequencing data. The simulator's workflow, depicted in Figure 2.1, illustrates the integration of genetic variants, methylation profile, and technical artifacts to produce high-fidelity simulated data. The simulation framework works as follows:

Haplotypes generation: The simulator begins with a reference genome sequence from a FASTA file, duplicating each chromosome to represent a diploid organism. Genetic variants, including single nucleotide polymorphisms (SNPs) and short insertions or deletions (indels), can be introduced either by specifying a mutation rate for random mutations or using a predefined VCF file. When phased genetic variants are provided in the VCF file, the haplotypes are constructed to accurately reflect the phasing, preserving the true genetic and allelic structure. These haplotypes form the foundation for subsequent simulation steps, including the simulation of allele-specific methylation (ASM).

Methylation database construction: Following haplotype generation, a methylation database is created by scanning the methylable bases along the haplotypes and recording their positions and sequence contexts (e.g., CG, CHG, CHH). Methylation levels are assigned to these positions if provided in a CGmap or ASM file. For positions lacking predefined methylation data, context-specific beta distributions, estimated from real methylation profiles, are used to simulate methylation levels, ensuring biologically realistic representation.

Fragmentation: The simulator supports three bisulfite sequencing technologies—WGBS, RRBS, and TBS—each employing a tailored fragmentation strategy. DNA fragments are sampled from both haplotypes within a predefined length range (default: 100 to 1000 base pairs). In WGBS, fragmentation sites are randomly distributed across the genome, simulating an unbiased approach. RRBS uses enzyme digestion (e.g., MspI, which recognizes CCGG sites) to concentrate on CpG-rich regions. Thus, DNA fragments are generated between the restriction enzyme sites. For TBS, fragments are centered around provided probe locations, with the exact positions simulated using a Gaussian distribution to account for variability.

Methylation pattern assignment: After generating DNA fragments, the simulator assigns methylation patterns to each fragment using the previously constructed methylation database. For each cytosine in a CpG context within the fragment, the methylation level is retrieved from the database, and the methylation state (methylated or unmethylated) is determined using one of two models: an independent Bernoulli model or a bidirectional Long Short-Term Memory (LSTM) network. The Bernoulli model independently assigns each cytosine's methylation state based on the retrieved methylation level, simulating random methylation patterns. In contrast, the LSTM model incorporates the sequence context and surrounding bases to predict methylation states on the read, leveraging patterns learned from real biological data to simulate biologically realistic methylation states.

Bisulfite conversion: After assigning methylation patterns to the DNA fragments, the simulator performs *in silico* bisulfite conversion. In this process, unmethylated cytosines are converted to thymines, while methylated cytosines remain unchanged. To reflect the imperfect nature of bisulfite treatment in real experiments, the simulator applies a fixed conversion success rate, mimicking incomplete conversion. This accounts for scenarios where unmethylated cytosines fail to convert and remain unchanged, ensuring that the final simulated sequences accurately represent both the methylation status and the stochastic nature of bisulfite conversion.

Sequencing quality/error assignment: Following bisulfite conversion, the simulator generates a pair of sequencing reads from both ends of the DNA fragments based on the specified read length. Base quality scores are assigned to each base on the reads, either set uniformly across the read or simulated by a Markovian chain using the quality state transition matrix. Sequencing errors are then introduced, either uniformly at a specified error rate or using a quality-specific confusion matrix, ensuring realistic error patterns in the simulated reads.

Reads Output and Processing: After sequencing error assignment, the simulator compiles the sequencing reads into standard FASTQ files, including both the nucleotide sequences and their corresponding quality scores. The read name encodes the origin of each read, specifying the chromosome, start, and end positions. Additionally, annotations in the comment line of each read record base changes, such as genetic variants, incomplete bisulfite conversions, or sequencing errors. These detailed annotations provide ground truth, ensuring traceability for each base observation and facilitating downstream benchmarking and analysis.

The generated synthetic reads can be processed and analyzed using the same procedure as the real bisulfite sequencing reads. This may include alignment to a reference genome, methylation and snp calling, or other analyses pertinent to bisulfite sequencing studies. By accurately mimicking the characteristics of real sequencing data, these outputs provide a robust foundation for testing and validating bioinformatics tools and pipelines under controlled conditions. The following sections will detail the modeling component of the simulator to provide a thorough understanding of the simulator's capabilities and utility, including data sources, modeling details and parameters, as well as implementation specifics.

2.2.2 DNA fragment sampling model

To accurately replicate the distinct characteristics of different bisulfite sequencing technologies (such as WGBS, RRBS, and TBS), BSReadSim employs DNA fragment generation and sampling processes tailored for each technology. While it also supports the basic uniform sampling approach, as other simulators did, BSReadSim can also offer a profile-based sampling model, where the probability of sampling each DNA fragment is determined by specific fragment features. This allows for more nuanced and realistic data generation.

1. Whole Genome Bisulfite Sequencing (WGBS):

In WGBS, the GC content of a DNA fragment—measured as the proportion of G or C bases—can directly impact its over- or under-representation of fragments in the sequencing output, a phenomenon known as GC bias [31]. The simulator leverages this relationship by using the GC ratio as a primary predictor of sampling probability, expressed as

$$p_i = f_1(GC_i) \tag{2.1}$$

where GC_i represents the GC ratio of the fragment i, with sampling probability p_i. We utilized a previously published WGBS dataset from PGP-UK [32] (Sample Accession ID: ERR2359938) to estimate the empirical function f_1. Specifically, the WGBS reads were processed and aligned to the reference genome using BSBolt, which was then divided into 100-base pair windows. For each window, the GC ratio and sequencing depth were calculated. To address variability in sequencing depth, the GC ratio spectrum was divided into 100 bins. Within each bin, the interquartile range (IQR) method was applied to identify and remove regions with extreme sequencing depths, which likely represent alignment artifacts or repetitive elements. The remaining depth values were normalized to scale between 0 and 1 (relative depth), serving as the sampling probability for each region (Figure 2.6). During simulation, the GC ratio for each fragment is calculated, and rejection sampling is applied based on the corresponding sampling probability. This approach mimics the coverage biases observed in real WGBS data.

2. Reduced Representation Bisulfite Sequencing (RRBS):

RRBS targets CpG-rich regions on the genome by utilizing restriction enzymes such as MspI, which cut at specific recognition sites (e.g., CCGG). The simulator replicates this process by first identifying the restriction sites on the haplotypes. It then generates all possible DNA fragments within the predefined fragment length range as candidates based on these restriction sites. The sampling probability for each fragment i is modeled as a function of its GC ratio (GC_i), fragment length (L_i), and the number of restriction enzyme sites contained within the fragment ($Count_i$).

$$p_i = f_2(GC_i, L_i, Count_i) \tag{2.2}$$

To learn the function f_2, we utilized a previously published RRBS dataset [29]. After processing and aligning the reads to the reference genome, DNA fragments were identified using the read pairs from the RRBS data. For each fragment, sequencing depth, GC ratio, fragment length, and the number of restriction sites were counted. Outliers were removed, and relative depths were calculated using an approach similar to WGBS data. To model the relationship between these features and the observed relative depths, a multivariate spline was fitted (Figure 2.7), allowing the simulator to estimate the sampling probability for each potential fragment candidate. During the simulation, each fragment was assigned a sampling probability predicted by the model and was subsequently sampled with these probabilities.

3. Targeted Bisulfite Sequencing (TBS):

TBS uses probes to enrich specific genomic regions of interest, with varying capture efficiencies that influence the enrichment of target regions (Figure 2.8). In BSReadSim, this variability is incorporated by assigning different sampling probabilities to the targeted regions. These probabilities can either be directly provided or empirically estimated from real TBS data to reflect probes' efficiency or target regions' accessibility. For empirical estimation, the depth of each probe region is calculated and normalized to generate a relative depth value. During simulation, DNA fragments are sampled from the targeted regions according to their assigned sampling probabilities, ensuring that the simulated data realistically reflects the enrichment and depth variations of real TBS experiments.

2.2.3 Methylation pattern model

sampled DNA fragments can utilize one of two models to assign methylation states to methylable sites, reflecting varying levels of complexity and realism in simulating methylation patterns.

1. Independent Bernoulli model

The Independent Bernoulli Model is a straightforward approach in which each methylable cytosine site within a DNA fragment is independently assigned a methylation state based on a Bernoulli distribution. The methylation level for site j, denoted as m_j, determines the probability of being methylated. The methylation state for site j on read i, denoted as y_{ij}, is then determined by:

$$y_{ij} \sim \text{Bernoulli}(m_j) \tag{2.3}$$

This model is computationally efficient and well-suited for generating baseline methylation patterns. However, it does not consider dependencies between neighboring sites' methylation states or the influence of genomic context, resulting in less realistic simulated methylation patterns.

2. Bidirectional Long Short-Term Memory (LSTM) model

To account for site-site dependency, we model Y_i, the methylation states of all sites on a read i, as being simultaneously sampled from an unknown distribution g. This distribution is determined by relevant features, including the methylation levels of sites on the read (M_i), inter-site distances (D_i), and genomic context (C_i), represented as the one-hot embedding of the surrounding sequences. Formally, this can be expressed as:

$$Y_i \sim g(M_i, D_i, C_i) \tag{2.4}$$

In this work, we utilize a bidirectional LSTM (BiLSTM) model to implicitly learn this function, capturing the intricate relationships among these features. By leveraging its bidirectional architecture, the BiLSTM integrates upstream and downstream sequence and methylation context, allowing it to account for both local and long-range dependencies. This enables the BiLSTM to accurately simulate methylation patterns that reflect the dependencies and variability observed in real biological systems.

To ensure the output methylation states maintain a marginal probability aligned with the predefined methylation levels, we designed a composite loss function combining Binary Cross-Entropy (BCE) loss and Mean Squared Error (MSE) loss. The BCE loss evaluates the accuracy of predicted binary methylation states for each site, while the MSE loss ensures that the averaged prediction state of a read matches the input methylation level. Together, these loss functions guide the BiLSTM in capturing dependencies between adjacent sites while maintaining input methylation levels. This ability to simulate realistic methylation patterns on a read while maintaining methylation level fidelity at individual sites sets our simulator apart from others that lack this capability.

2.2.4 Sequencing quality and error model

Depending on the user's needs, the sequencing quality and error on a read can be generated using two approaches: a uniform model and an advanced state transition model. The uniform model assigns a consistent quality score and introduces errors at a constant rate, offering simplicity and computational efficiency. For users requiring greater realism, the advanced model contains the following two parts and captures dependencies across sequential base-calling cycles, providing more realistic simulations.

1. Quality transition matrix:

We adopt the same strategy as pIRS [33] and use a quality transition matrix to model quality scores across sequencing cycles. Each element in the matrix represents the probability of transitioning from a quality score in one sequencing cycle to a specific quality score in the subsequent cycle. This approach accounts for the observation that the quality of base calls often depends on the quality of preceding calls, particularly under conditions where the sequencing quality deteriorates along the read. In our simulator, we constructed the quality score transition matrix for read1 and read2, respectively, from the WGBS data, effectively representing the potential difference for the read pairs. (Figure 2.9)

During the simulation, the quality score for the initial five bases is randomly drawn from the empirical discrete distribution constructed from the real data. For subsequent bases, the simulator uses the quality-transition matrix to determine the following quality score based on the score of the preceding base. This method effectively captures the progressive nature of quality deterioration characteristic of many sequencing platforms, particularly for longer reads.

2. Sequencing error generation:

Each quality score has a specific sequencing error profile, with lower quality scores generally indicating higher probabilities of errors and different base errors having different error rates. In our simulator, we empirically derived the base transition matrices for each quality score from real sequencing data. Unlike whole genome sequencing, where error rates can be more directly estimated by comparing the aligned reads to the reference genome, bisulfite sequencing poses additional challenges due to bisulfite conversion, where the observed difference between reads and reference genome can be attributed to either sequencing error or bisulfite conversion. To address this issue, we focus on the overlapped bases of read1 and read2 in paired-end reads. Given the paired reads from the same DNA fragments, any observed discrepancies in the overlapped region must be due to sequencing errors, thus providing a reliable means of estimating errors and minimizing the confounding effects of bisulfite conversion. The estimated error rate profile for each sequencing quality score is presented in Figure 2.10, effectively capturing the relationship between quality scores and their corresponding error rates.

During simulation, once a quality score is determined for each base, the corresponding base transition matrices are applied to introduce potential sequencing errors using a discrete distribution. This method ensures that the simulated reads realistically reflect the error characteristics observed in actual sequencing experiments.

2.2.5 Computational optimization strategies

One major bottleneck of bisulfite sequencing read simulation lies in the computational speed. To mitigate this issue, we implemented several computational optimization strategies to ensure efficient processing, enabling the simulation of large datasets within a reasonable timeframe. These strategies are particularly crucial for making the tool accessible and practical for researchers working with high-throughput sequencing data. The following points summarize the key techniques employed:

High-efficiency implementation: One key optimization was implementing computational and memory-intensive components of the simulator in C/C++, such as haplotype generation, methylation database construction, and fragment sampling. This lower-level, high-performance language offers better control over memory management and enables efficient data structures, significantly improving computational efficiency. For instance, haplotype construction and the parsing of genetic variants and methylation profiles were implemented using HTSLIB [34], a highly optimized C++ library specifically designed for handling next-generation sequencing data. Additionally, the methylation database was constructed using a customized data structure that utilizes pointers and vectors, ensuring efficient storage and rapid access to site-specific methylation data.

Bit encoding and operations: To fully leverage the advantage of C++ and optimize performance, we adopted the bit encoding and operation framework from WGSIM [35], a tool to efficiently simulate whole-genome sequencing reads. By representing each nucleotide (A, C, G, T) as a 2-bit binary value, this encoding reduces the memory footprint by a factor of four compared to traditional byte-based representations. Additionally, bitwise operations—such as AND, OR, XOR, and shifts—are used to perform computations directly on these binary-encoded sequences. These operations are inherently faster than equivalent arithmetic operations by directly manipulating the bits at the hardware level, reducing the time required for tasks such as mutation introduction, fragment generation, and fragment feature extraction. With the compact encoding and the use of fast bitwise operations, BSReadSim not only reduces memory usage but also accelerates computational processes. This dual benefit is particularly important when simulating large genomic datasets with limited computing resources, where both memory efficiency and processing speed are critical.

Memory and data access optimization: Efficient memory management was a key focus in our simulator's design to handle large-scale simulations efficiently and effectively. One of the critical optimizations involved sorting DNA fragments before retrieving their corresponding methylation levels from the methylation database. By sorting fragments, we increased data locality, meaning that related data is accessed sequentially, which significantly reduces cache misses and thus improves processing speed. This approach also exemplifies the trade-off of space for time, as the temporary storage required for sorting is outweighed by the performance gains achieved during data retrieval and processing. On the other hand, the simulator processes genome fragments one chromosome at a time and generates sequencing reads chunk by chunk on the fly, rather than holding all chromosomes and read data in memory simultaneously, reducing the memory footprint and enabling the efficient processing.

Algorithmic function optimizations: We identified several frequently used functions in BSReadSim and optimized them for improved performance, such as simulating vectors from Bernoulli and discrete distributions. The Bernoulli distribution is heavily utilized in methylation state assignment and bisulfite conversion; we re-implemented the function by comparing a vector of random numbers to the target probabilities and directly mapping the True/False results to 0/1 states. This optimization achieved a 30-fold speed increase compared to the standard `bernoulli.rvs()` method for vectors of length 150. The discrete distributions are frequently used for simulating sequencing quality and errors, we optimized the sampling process by generating a uniform random number between 0 and 1 and comparing it against the precomputed cumulative distribution function (CDF) derived from the discrete probabilities. The monotonic nature of the CDF enables efficient identification of the corresponding discrete class using binary search. This optimization reduces computational overhead, significantly accelerating the sampling process, achieving a speed-up of approximately 13.5 times faster than the standard `np.random.choice()` method.

Other optimization endeavors: Beyond the strategies outlined above, we also implemented several additional techniques. The simulator's scalable and modular design allows efficient handling of datasets ranging from small targeted sequencing to large-scale whole-genome studies, with individual components optimized as needed. Data compression and I/O optimization, such as gzip compression, reduce storage demands and improve data access by processing compressed data directly in memory. Parallel processing further accelerates performance by distributing computational tasks across multiple CPU cores, significantly reducing runtime. These strategies ensure the simulator is both efficient and robust, addressing the demands of high-throughput sequencing simulations.

Customizable trade-offs: Recognizing the diverse needs of researchers, we offer users the flexibility to balance computational trade-offs. Users can choose between high-fidelity simulations that prioritize accuracy at the cost of higher resource demands or lower-fidelity simulations optimized for speed. This adaptability allows researchers to tailor simulations to their specific goals and resource availability.

These optimizations significantly enhance the simulator's performance, making it a valuable tool for the DNA methylome community. By effectively balancing speed and resource efficiency, our simulator provides a robust platform for simulating bisulfite sequencing reads across multiple technologies, experiment design, and the development and testing of bioinformatics methods with high fidelity and realism.

2.2.6 Code availability

The code is freely available at `https://github.com/wbvguo/BSReadSim.git`

2.3 Results

2.3.1 Faithful incorporation of reference genetic variants

To evaluate the effectiveness of our bisulfite sequencing reads simulator in integrating genetic variants, we conducted a profile-based simulation using BSReadSim with a customized VCF file aiming at a sequencing depth of 20. After simulation, the synthetic reads were aligned to the reference genome and inspected using the Integrative Genomics Viewer (IGV) [36]. Figure 2.2 demonstrates the simulator's ability to faithfully incorporate specified genetic variants into simulated bisulfite sequencing reads.

Specifically, the VCF file contains a homozygous SNP on chromosome 10 at position 90,937, with the reference allele A and the alternate allele T, as shown in the top panel of Figure 2.2. In the BAM alignment track, all bases at the SNP locus are T, confirming the accurate incorporation of the genetic variant as defined in the VCF file. These results establish our simulator as the first tool to seamlessly integrate predefined genetic variants into bisulfite sequencing reads, offering a reliable platform for advanced epigenomic research. This capability is particularly advantageous for studies that require integrating genetic variants into bisulfite sequencing data, such as developing and benchmarking tools for bisulfite SNP calling, allele-specific methylation, and methylation QTL simulation.

2.3.2 Accurate preservation of reference methylation profiles

To further evaluate our simulator, we assessed its ability to accurately preserve the reference methylation profiles—a critical feature for generating synthetic data that closely mimics real data and facilitating experimental design. The simulation used a reference genome and a prespecified methylation profile from a CGmap file, targeting at sequencing depth of 20. The generated reads were then aligned to the reference genome, and methylation levels were quantified. Fidelity was evaluated by comparing the designed methylation levels with the estimated levels derived from the simulated data.

We repeat the same simulation procedure for both BSBolt and BSReadSim. The results revealed that BSBolt failed to preserve the reference methylation profile accurately. As shown in Figure 2.3, the widespread distribution of dots indicates significant discrepancies between the designed and simulated methylation levels. This limitation arises because BSBolt randomly assigns methylation values to CG sites from the reference profile, disregarding their specific genomic location information. In contrast, BSReadSim exhibited much higher fidelity in replicating the reference methylation profile. While minor deviations were observed, primarily due to stochastic variations in sequencing depth, most simulated methylation levels closely aligned with the reference profile, with data points clustering near the diagonal. These results validate BSReadSim as a reliable tool for simulating bisulfite sequencing data, particularly for applications requiring accurate preservation of input methylation profiles.

2.3.3 Effective capture of site-site dependency

To evaluate the ability of our BiLSTM-based model to capture site-site dependency, we compared the entropy-distance relationship observed in real data, BiLSTM-simulated data, and Bernoulli-simulated data (Figure 2.4). Entropy was calculated based on the joint state probabilities of adjacent sites (00, 01, 10, 11), which provides a measure of the methylation concordance in adjacent sites and reflects the site-site dependency. Higher entropy indicates lower concordance between adjacent sites, reflecting weaker site-site dependency. To ensure robust measurement, only site pairs with read counts exceeding 20 were included for analysis.

In the real data, a weak but significant positive correlation between entropy and distance was observed ($R^2 = 0.11$), reflecting the gradual weakening of site-site dependencies with increasing distance, consistent with previous finding [37]. The BiLSTM-simulated data closely replicated this trend, showing a comparable positive correlation ($R^2 = 0.09$), demonstrating the model's ability to capture realistic dependency structures. However, the entropy in the BiLSTM-simulated data is consistently lower than in the real data, indicating that the real data exhibits greater stochasticity than the model assumes [38]. Future work can consider refining the model to capture the additional variability in the real data not fully captured by the BiLSTM model.

On the other hand, the Bernoulli model assumes that sites are independent, leading to no concordance between adjacent sites. As expected, this results in consistently high entropy that does not vary with distance ($R^2 = 0.01$). This behavior highlights the limitation of the Bernoulli model in representing the spatial dependencies inherent in real methylation data.

These results emphasize the BiLSTM model's capability to effectively preserve site-site dependency, making it a valuable tool for generating methylation patterns that reflect biological systems. By capturing both local and long-range dependencies, the BiLSTM-based simulator offers significant advantages over the simpler independent Bernoulli model.

2.4 Discussion

The field of bisulfite sequencing simulation has witnessed the development of several simulators designed to generate synthetic data for various applications. However, existing tools such as Sherman, BSBolt, and MethylFASTQ are limited in their ability to fully integrate genetic and methylation profiles, often producing synthetic data that lacks the complexity of real bisulfite sequencing. These tools also suffer from computational inefficiencies, making them impractical for large-scale studies. To address these limitations, BSReadSim was developed with advanced features, including detailed genetic variant input, allele-specific methylation, and context-aware sequencing error modeling. Results show that BSReadSim can faithfully incorporate reference genetic and methylation profiles while effectively preserve the site-site dependency as observed in real data, providing a robust platform for generating realistic bisulfite sequencing data.

Building on these strengths, BSReadSim can enhance the fidelity of simulations and ensure synthetic data closely mirrors real-world sequencing outputs, making it particularly valuable for benchmarking bioinformatics tools (Figure 2.5) and designing experiments.

1. **Benchmarking Bisulfite Sequencing Aligners**: Accurate alignment is crucial for downstream tasks; however, methylation and bisulfite conversion introduce an additional layer of complexity, complicating the alignment process and requiring specialized handling. In the past, a number of aligners have been developed to tackle this challenge. BSReadSim's ability to generate realistic reads with known fragment origins can provide a rigorous benchmarking framework for bisulfite sequencing aligners, helping identify and refine the most effective alignment tools.

2. **Benchmarking Bisulfite Sequencing SNP Callers**: The identification of SNPs in bisulfite sequencing data is complicated by sequencing errors and bisulfite-induced changes. BSReadSim enables detailed benchmarking of SNP callers by providing synthetic reads that faithfully incorporate genetic variants and provide traceable changes,

thereby offering the ground truth necessary for reliable benchmarking of these tools.

3. **Probe Design for Targeted Bisulfite Sequencing and Methylation Arrays**: Designing probes for TBS and methylation arrays requires careful consideration of repetitive elements and potential off-target effects. BSReadSim enables researchers to optimize probe design by simulating TBS data and aligning it back to the reference genome, identifying potential off-target or multi-mapped probes. By providing the feedback loop, BSReadSim can serve to refine the probe sets, thereby improving the reliability and effectiveness of targeted sequencing studies.

Despite its advancements, several areas remain for improvement and further exploration. BSReadSim currently preserves both genetic and methylation profiles, offering valuable realism for simulating bisulfite sequencing data. However, further testing is needed to assess its unique capability in simulating allele-specific methylation (ASM) [39], which is critical for developing and benchmarking ASM detection tools. Additionally, comprehensive testing and comparisons with existing simulators are necessary to fully evaluate BSReadSim's computational efficiency and advantages. For site-site dependency modeling, exploring advanced techniques such as Gaussian processes [40] could be further explored to enhance the prediction accuracy. Finally, leveraging BSReadSim to benchmark other bisulfite sequencing tools will be an important step in demonstrating its utility across diverse applications and advancing computational epigenetics research.

In summary, BSReadSim fills a critical gap in bisulfite sequencing by offering a versatile and high-fidelity simulator capable of generating realistic bisulfite sequencing data efficiently. With its unique advantages, BSReadSim supports various applications, including benchmarking alignment and SNP calling tools and optimizing probe design for targeted sequencing experiments. These features highlight its value to the epigenomics research community. Further refinements, including evaluating allele-specific methylation, extensive testing against other simulators, and its use in benchmarking bisulfite sequencing tools, will enhance its impact on computational genetics and epigenomic research.

2.5 Acknowledgements

I thank the public data sources that made this work possible. I want to thank Harray Zhang, Hongxiang Fu, and Junxi Feng for their valuable input and discussion and the Hoffman2 at UCLA for providing essential computational support. Additionally, I acknowledge ChatGPT for its assistance with debugging and writing. Figure 2.1 was created using BioRender, thus acknowledge it by courtesy.

2.6 Tables and figures

Table 2.1: Summary of existing bisulfite sequencing read simulators

Features	**Sherman**	**BSBolt**	**BSSim**	**MethylFASTQ**	**RRBSsim**
Sequencing technology	WGBS	WGBS	WGBS	WGBS/TBS*	RRBS
Genetic variant input	No	No	Yes*	No	Yes*
Haplotype-aware	No	No	No	No	No
Methylation profile input	No	Yes*	No	No	No
Site-site dependency	No	No	No	No	No
Allelic-specific methylation	No	No	No	No	No
Adjustable bisulfite conversion rate	Yes	No	Yes	No	Yes
GC-bias/non-uniform coverage	No	No	No	No	No
Multi-thread support	No	No	Yes	Yes	No

(*) denotes limited support:

- MethylFASTQ cannot distinguish between the Watson and Crick strands for Targeted Bisulfite Sequencing (TBS).
- BSSim and RRBSsim only accept SNP input using a frequency table. They cannot faithfully utilize given genotypes, preserve haplotype information, or handle indel variants.
- BSBolt randomly picks a value from the methylation reference input for simulation. As a result, for a particular CG site, the simulated data and the reference profile will likely have different methylation levels.

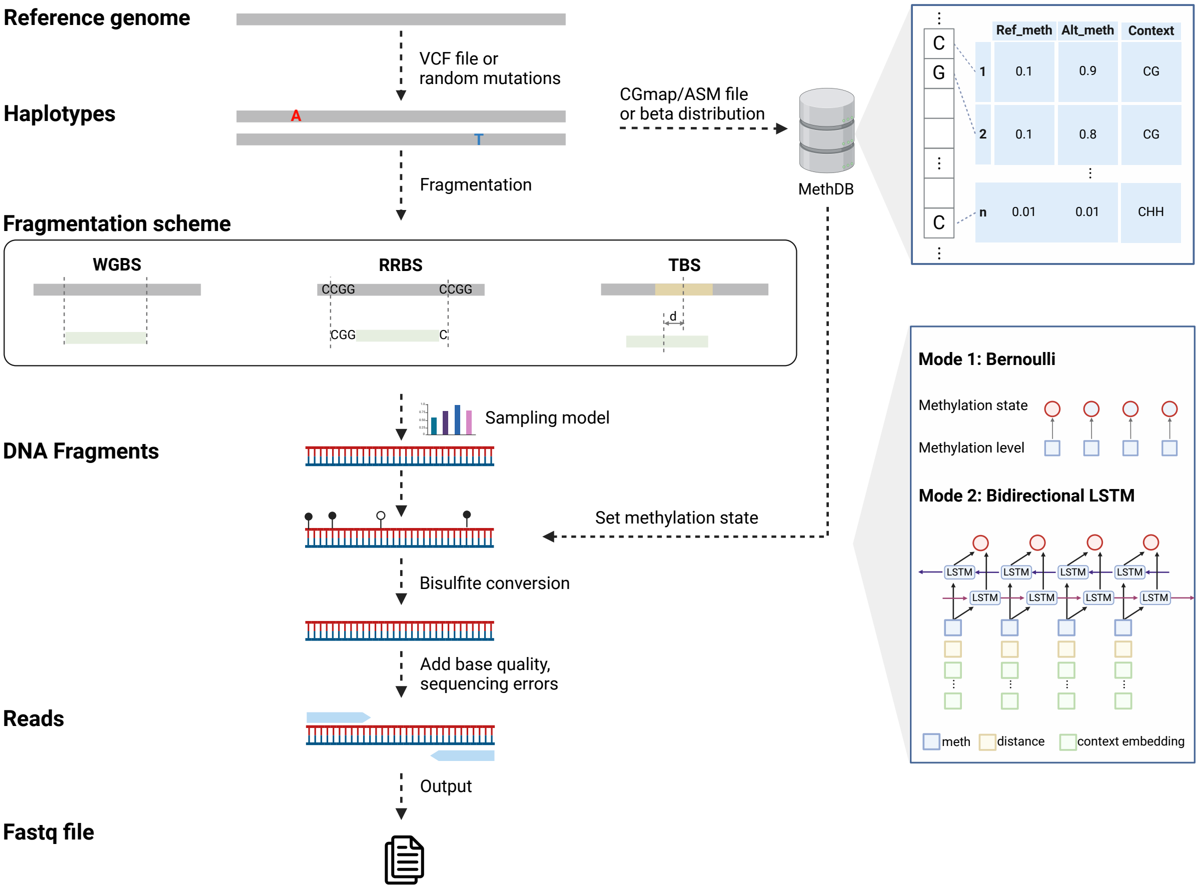
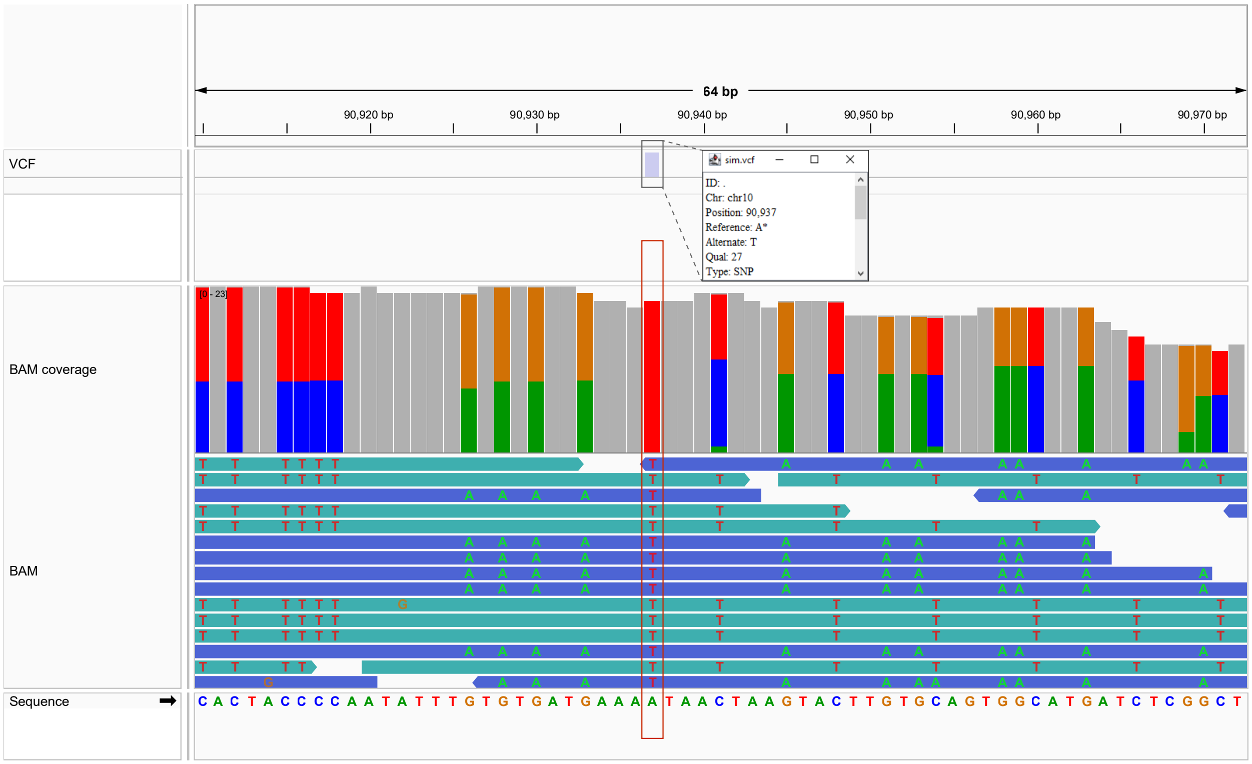


Figure 2.1: Overview of the Bisulfite Sequencing read Simulation (BSReadSim) Framework. Workflow illustrating the simulation process for bisulfite sequencing data. The process begins with the reference genome, from which haplotypes are generated either through a provided VCF file or by randomly introducing mutations. A methylation database (MethDB) is then constructed, leveraging the methylable bases of the haplotypes and a specified methylation profile (sourced from a CGmap/ASM file or context-specific beta distributions). Subsequently, the haplotypes undergo fragmentation and sampling according to the selected sequencing strategy—WGBS, RRBS, or TBS—to generate DNA fragments. The methylation state of each cytosine within these fragments is determined using a Bernoulli or bidirectional LSTM model. Following the assignment of methylation states, DNA fragments undergo *in silico* bisulfite conversion, read generation, and the addition of base quality scores and sequencing errors to produce realistic bisulfite sequencing reads. Finally, the read data are output in the standardized Fastq file format and ready for downstream analysis.


64 bp
90,920 bp
90,930 bp
90,940 bp
90,950 bp
90,960 bp
90,970 bp
VCF
sim.vcf
BAM coverage
[0 - 23]
BAM
Sequence
C A C T A C C C C A A T A T T T G T G T G A T G A A A A T A A C T A A G T A C T T G T G C A G T G G C A T G A T C T C G G C T

ID: .
Chr: chr10
Position: 90,937
Reference: A*
Alternate: T
Qual: 27
Type: SNP

Figure 2.2: BSReadSim incorporates genetic variants to simulated read data. IGV visualization of read alignment for simulated read data, highlighting the faithful incorporation of predefined genetic variants. The top panel (VCF) displays the predefined VCF file, indicating a homozygous SNP on chromosome 10 at position 90,937, with the reference allele A and the alternate allele T. The middle panel (BAM coverage) illustrates the read coverage at this region, with color-coded bars representing the proportion of reads supporting each base (A: green; C: blue; G: orange; T: red) alongside the sequencing depth distribution. The bottom panel (BAM) shows individual read alignments, where the presence of T alleles at the SNP site is clearly visible, reflecting the homozygous SNP introduced in the simulation and consistent with the input VCF file. The sequence at the bottom of the figure provides the reference sequence context around the SNP. This figure demonstrates an example of the simulator's capability to faithfully incorporate predefined genetic variants into simulated bisulfite sequencing data.

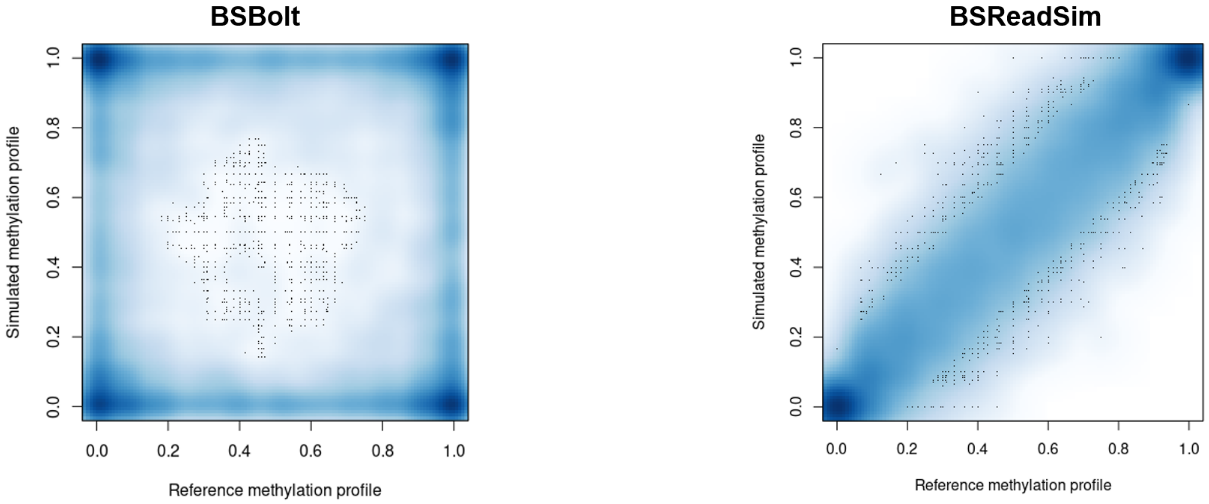


Figure 2.3: BSReadSim preserves methylation profile in simulated read data. Figure comparing the fidelity of methylation profile preservation between two simulators, BSBolt (left) and BSReadSim (right). Each dot represents a methylable base in the genome, with the x-axis depicting the reference methylation profile and the y-axis showing the methylation profile calculated from the simulated bisulfite sequencing reads. In the BSBolt panel, the widespread distribution of dots indicates a significant loss of location information. Conversely, the BSReadSim panel shows dots closely aligned along the diagonal, demonstrating BSReadSim's ability to accurately replicate the reference methylation profile and maintain site-specific methylation patterns across the genome. It's important to note that randomness in sequencing depth can introduce slight variability in the estimated methylation levels, rendering the dots do not perfectly align on the diagonal in the right panel.

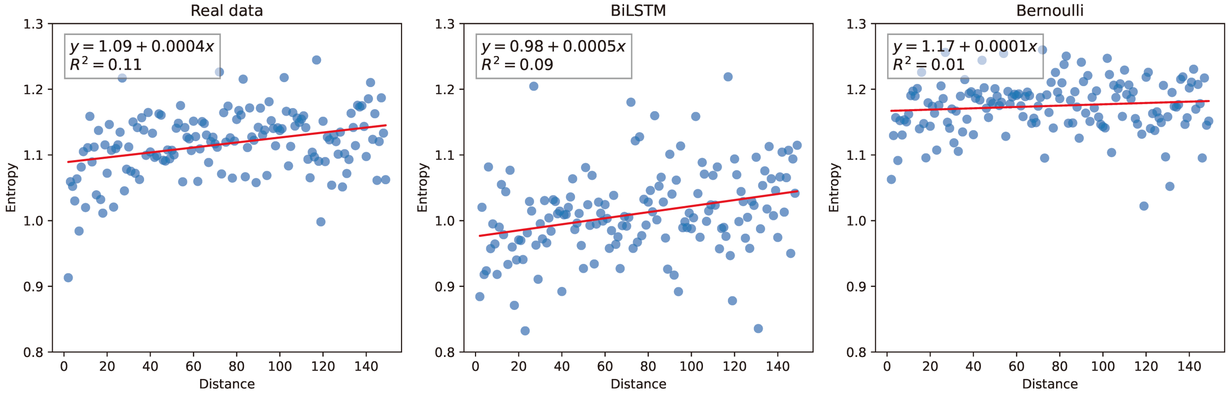


Figure 2.4: BSReadSim captures of site-site dependency in real data. Comparison of entropy-distance relationships in real data, BiLSTM-simulated data, and Bernoulli-simulated data. The scatterplots depict the entropy of methylation states as a function of the genomic distance between adjacent sites. The red line represents a linear regression fit to the data, with the equation and R^2 value shown in each panel. The left panel shows real data, where a weak but consistent positive correlation is observed ($R^2 = 0.11$). The middle panel represents data simulated using the BiLSTM model, which closely approximates the real data pattern ($R^2 = 0.09$), demonstrating its ability to capture site-site dependencies. The right panel shows data generated by the Bernoulli model, which lacks dependency between adjacent sites and exhibits minimal correlation ($R^2 = 0.01$), highlighting its limitation in reflecting realistic methylation patterns.

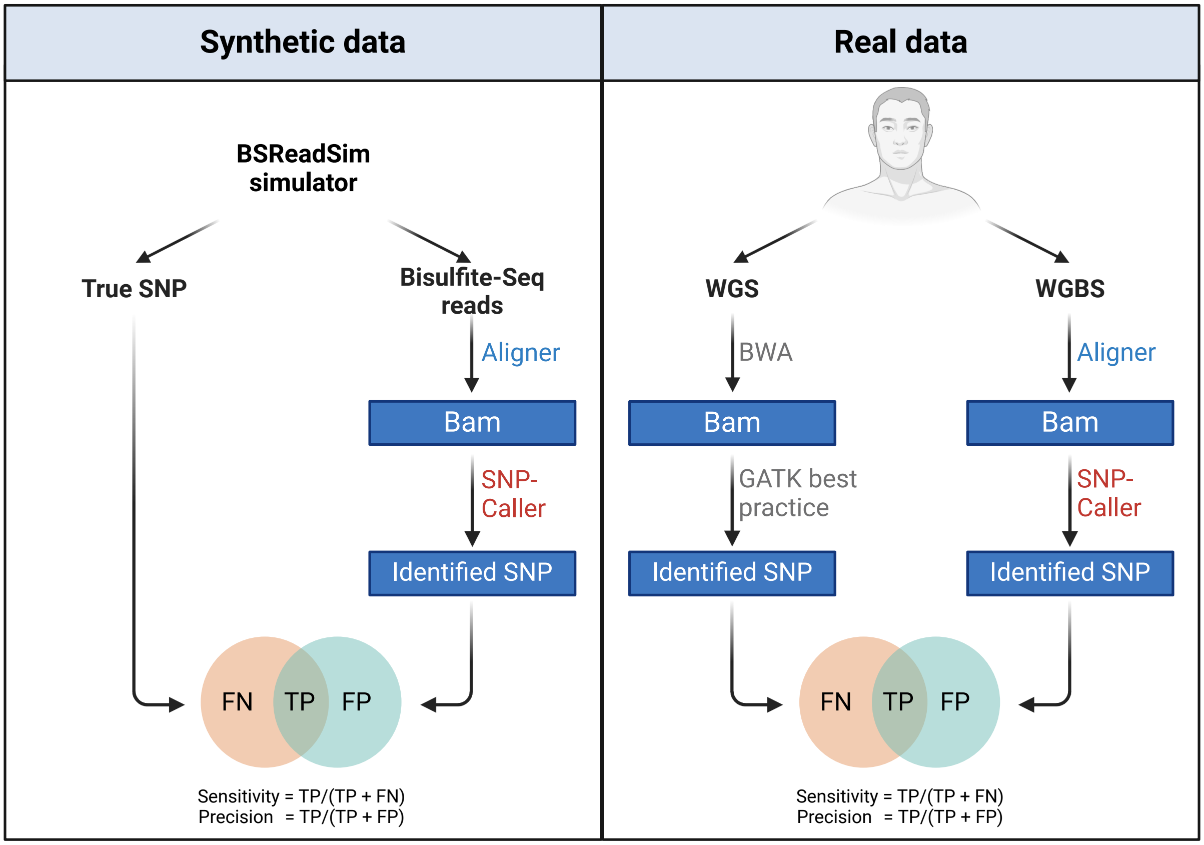


Figure 2.5: Potential applications of BSReadSim. The left panel illustrates synthetic data generation by BSReadSim with known ground truth, including the true SNPs and the true origin of reads. The synthetic data follows standard bisulfite sequencing read processing steps, including alignment and SNP identification. Each aligner and SNP-caller's performance can be assessed by comparing the analyzed results to the designed ground truth (benchmarking aligners by comparing the aligned locations to their true origins and benchmarking SNP-callers by comparing identified SNPs to the true SNPs.). The right panel complements this by applying a similar framework to real data from two sequencing modalities (WGS and WGBS), allowing the evaluation of aligner and SNP-calling tools in real-world scenarios. Together, these approaches provide a comprehensive benchmarking framework that integrates both synthetic and real data.

2.7 Supplementary materials

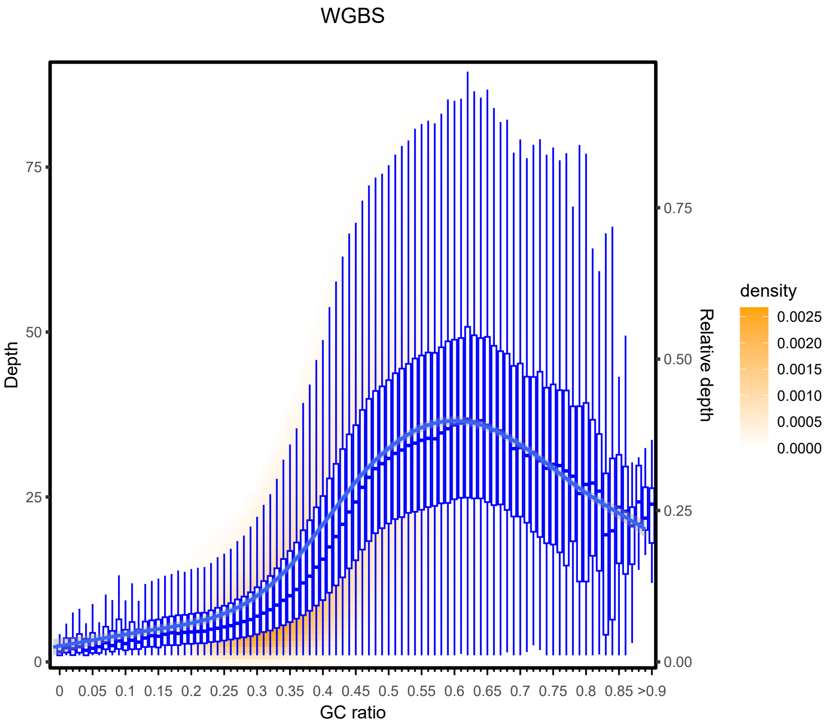


Figure 2.6: Fragment sampling model for WGBS data. Figure showing the relationship between GC content (x-axis) and relative sequencing depth (right y-axis) in WGBS data. The genome is segmented into 100 bp bins, and the GC ratio and depth are calculated for each bin. Blue boxplots represent the distribution of sequencing depth across different GC ratios, with the central line indicating the median depth, the blue curve representing the local trend, the box representing the interquartile range (IQR), and the whiskers extending to 1.5 times the IQR. The relative depth (right y-axis) is normalized to range between 0 and 1, and the color gradient indicates fragment density (orange for higher density). The figure highlights the GC bias in WGBS, where fragments with intermediate GC content have higher sequencing depth than those with very low or high GC content.

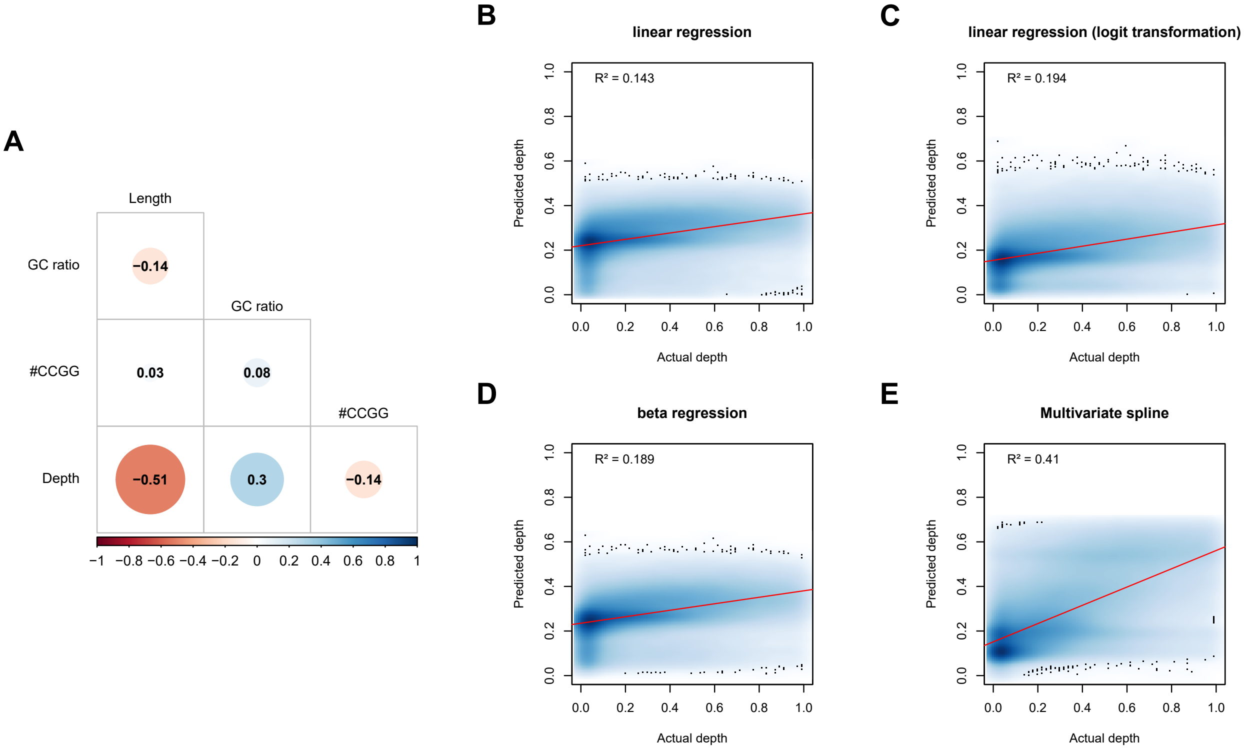


Figure 2.7: Fragment sampling model for RRBS data. Analysis of factors influencing sampling probability in RRBS data and comparison of predictive models. (A) Correlation matrix showing the relationships between sequencing depth and fragment features (GC ratio, fragment length, restriction site count) in RRBS data. The size and color intensity of the circles indicate the strength and direction of the correlations, with depth showing a negative correlation with fragment length and a moderate positive correlation with GC ratio and the number of restriction sites. (B-E) Comparison of different models for predicting relative depth from the fragment features using linear regression (B), linear regression with logit transformation (C), beta regression (D), and multivariate spline (E). The multivariate spline model (E) shows the best fit, with an R^2 of 0.41, indicating its superior performance in capturing the complex relationship between sequencing depth and fragment features in RRBS data.

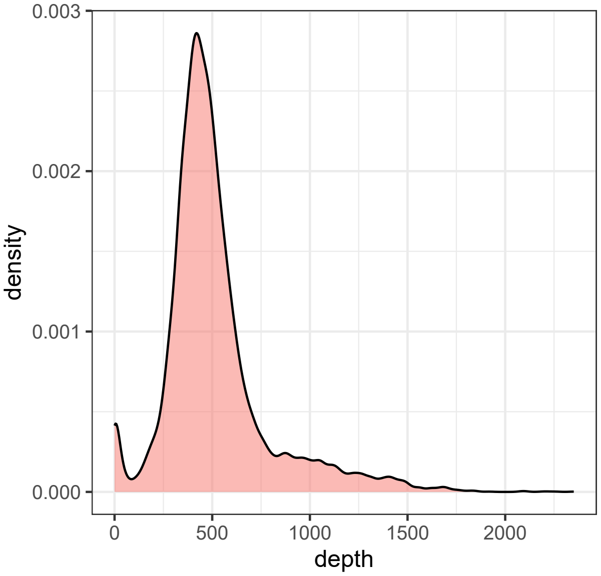


Figure 2.8: Fragment sampling model for TBS data. Density plot of sequencing depth for targeted regions in Targeted Bisulfite Sequencing (TBS). The plot shows the distribution of sequencing depth across probe-enriched regions, reflecting the variability in capture efficiency of different probes. In the simulation, this variability is modeled by assigning sampling probabilities to targeted regions based on either provided values or empirical estimates from real TBS data. This approach ensures that the simulated reads accurately represent the regional enrichment and depth variations observed in actual TBS data.

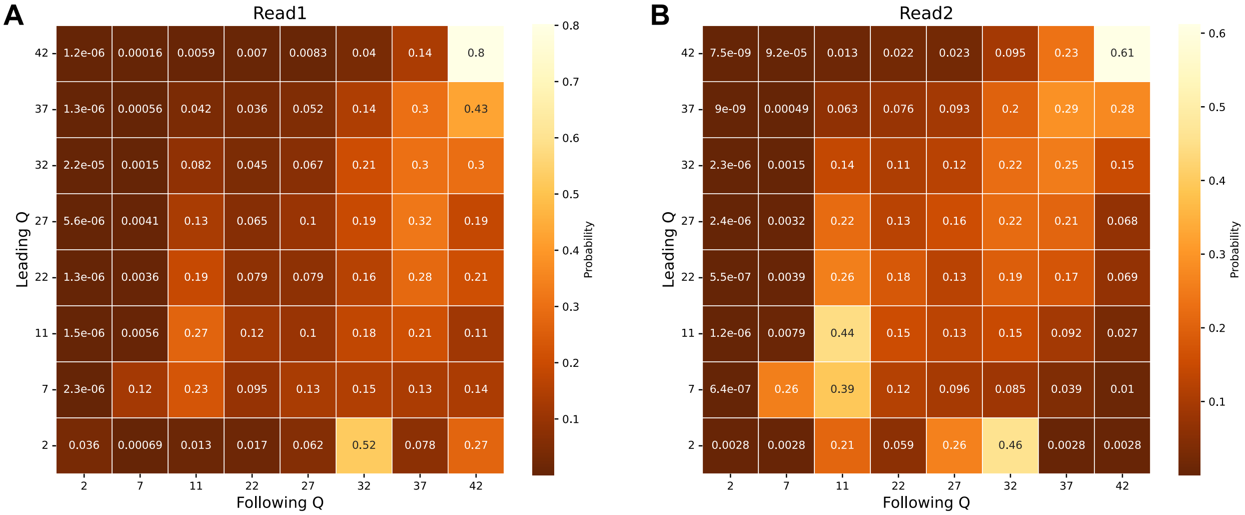


Figure 2.9: Heatmap of base quality transition probabilities. Quality-transition matrices for Read1 (A) and Read2 (B) in Whole Genome Bisulfite Sequencing (WGBS) data. These matrices represent the probability of transitioning from a preceding quality score (Leading Q) to a subsequent score (Following Q) across sequencing cycles on the read. Each element shows the probability of a quality score change between cycles, with color intensity indicating transition probability. The matrices capture the dependency of sequencing quality on preceding bases and account for quality degradation along read length.

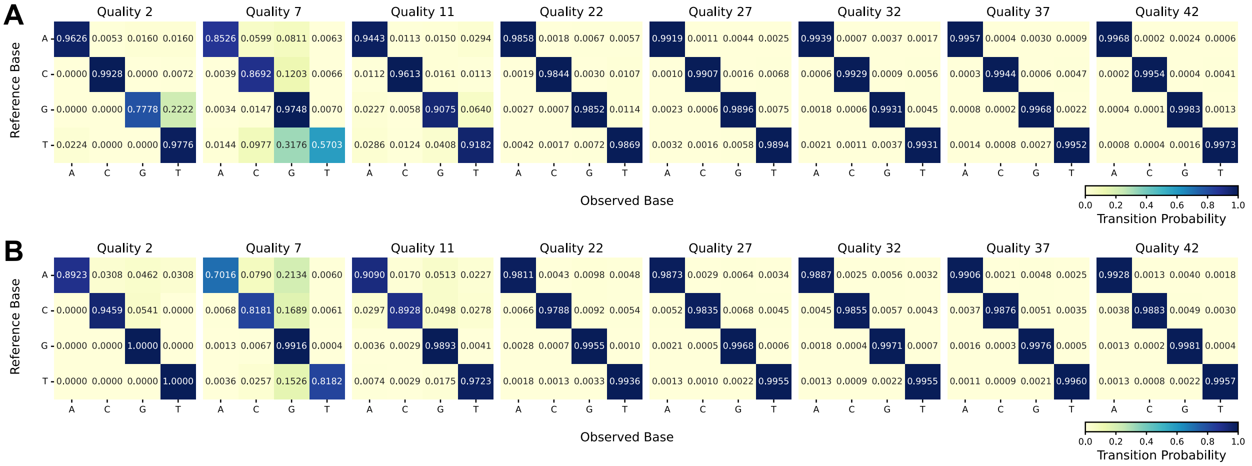


Figure 2.10: Sequencing error profiles across base quality. Base transition matrices for varying quality scores in bisulfite sequencing data. Panels A and B show the probability of observed bases (A, C, G, T) for given reference bases across different quality scores (2, 7, 11, 22, 27, 32, 37, 42) for Read1 and Read2, respectively. Lower quality scores correspond to higher error probabilities.

2.8 References

[1] Mattei, A. L., Bailly, N., and Meissner, A. "DNA methylation: a historical perspective". *Trends in Genetics* 38.7 (2022), pp. 676–707.

[2] Smith, Z. D. and Meissner, A. "DNA Methylation: Roles in Mammalian Development". *Nature Reviews Genetics* 14.3 (2013), pp. 204–220.

[3] Farlik, M., Halbritter, F., Müller, F., Choudry, F. A., Ebert, P., Klughammer, J., Farrow, S., Santoro, A., Ciaurro, V., Mathur, A., et al. "DNA methylation dynamics of human hematopoietic stem cell differentiation". *Cell stem cell* 19.6 (2016), pp. 808–822.

[4] Greenberg, M. V. and Bourc'his, D. "The diverse roles of DNA methylation in mammalian development and disease". *Nature reviews Molecular cell biology* 20.10 (2019), pp. 590–607.

[5] Moore, L. D., Le, T., and Fan, G. "DNA methylation and its basic function". *Neuropsychopharmacology* 38.1 (2013), pp. 23–38.

[6] Jin, Z. and Liu, Y. "DNA methylation in human diseases". *Genes & diseases* 5.1 (2018), pp. 1–8.

[7] Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning". *Nature* 452.7184 (2008), pp. 215–219.

[8] Xi, Y. and Li, W. "BSMAP: whole genome bisulfite sequence MAPping program". *BMC bioinformatics* 10 (2009), pp. 1–9.

[9] Krueger, F. and Andrews, S. R. "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". *bioinformatics* 27.11 (2011), pp. 1571–1572.

[10] Lim, J.-Q., Tennakoon, C., Li, G., Wong, E., Ruan, Y., Wei, C.-L., and Sung, W.-K. "BatMeth: improved mapper for bisulfite sequencing reads on DNA methylation". *Genome biology* 13 (2012), pp. 1–14.

[11] Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M. Q., Chen, P.-Y., and Pellegrini, M. "BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data". *BMC genomics* 14 (2013), pp. 1–8.

[12] Pedersen, B. S., Eyring, K., De, S., Yang, I. V., and Schwartz, D. A. "Fast and accurate alignment of long bisulfite-seq reads". *arXiv preprint arXiv:1401.1129* (2014).

[13] Harris, E. Y., Ounit, R., and Lonardi, S. "BRAT-nova: fast and accurate mapping of bisulfite-treated reads". *Bioinformatics* 32.17 (2016), pp. 2696–2698.

[14] Merkel, A., Fernández-Callejo, M., Casals, E., Marco-Sola, S., Schuyler, R., Gut, I. G., and Heath, S. C. "gemBS: high throughput processing for DNA methylation data from bisulfite sequencing". *Bioinformatics* 35.5 (2019), pp. 737–742.

[15] Zhang, Y., Park, C., Bennett, C., Thornton, M., and Kim, D. "Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N". *Genome Research* 31.7 (2021), pp. 1290–1295.

[16] Sena Brandine, G. de and Smith, A. D. "Fast and memory-efficient mapping of short bisulfite sequencing reads using a two-letter alphabet". *NAR Genomics and Bioinformatics* 3.4 (2021), lqab115.

[17] Farrell, C., Thompson, M., Tosevska, A., Oyetunde, A., and Pellegrini, M. "BiSulfite Bolt: A bisulfite sequencing analysis platform". *GigaScience* 10.5 (2021), giab033.

[18] Zhou, W., Johnson, B. K., Morrison, J., Beddows, I., Eapen, J., Katsman, E., Semwal, A., Habib, W. A., Heo, L., Laird, P. W., et al. "BISCUIT: an efficient, standards-compliant tool suite for simultaneous genetic and epigenetic inference in bulk and single-cell studies". *Nucleic Acids Research* 52.6 (2024), e32–e32.

[19] Liu, Y., Siegmund, K. D., Laird, P. W., and Berman, B. P. "Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data". *Genome biology* 13 (2012), pp. 1–14.

[20] Barturen, G., Rueda, A., Oliver, J. L., and Hackenberg, M. "MethylExtract: high-quality methylation maps and SNV calling from whole genome bisulfite sequencing data". *F1000Research* 2 (2013).

[21] Gao, S., Zou, D., Mao, L., Liu, H., Song, P., Chen, Y., Zhao, S., Gao, C., Li, X., Gao, Z., et al. "BS-SNPer: SNP calling in bisulfite-seq data". *Bioinformatics* 31.24 (2015), pp. 4006–4008.

[22] Guo, W., Zhu, P., Pellegrini, M., Zhang, M. Q., Wang, X., and Ni, Z. "CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data". *Bioinformatics* 34.3 (2018), pp. 381–387.

[23] Nunn, A., Can, S. N., Otto, C., Fasold, M., Díez Rodríguez, B., Fernández-Pozo, N., Rensing, S. A., Stadler, P. F., and Langenberger, D. "EpiDiverse Toolkit: a pipeline suite for the analysis of bisulfite sequencing data in ecological plant epigenetics". *NAR genomics and bioinformatics* 3.4 (2021), lqab106.

[24] Fan, Y., Vilgalys, T. P., Sun, S., Peng, Q., Tung, J., and Zhou, X. "IMAGE: high-powered detection of genetic effects on DNA methylation using integrated methylation QTL mapping and allele-specific analysis". *Genome biology* 20 (2019), pp. 1–18.

[25] Abante, J., Fang, Y., Feinberg, A., and Goutsias, J. "Detection of haplotype-dependent allele-specific DNA methylation in WGBS data". *Nature communications* 11.1 (2020), p. 5238.

[26] Bioinformatics, B. *Sherman*. https://github.com/FelixKrueger/Sherman/.

[27] Farrell, C., Thompson, M., Tosevska, A., Oyetunde, A., and Pellegrini, M. "BiSulfite Bolt: A bisulfite sequencing analysis platform". *GigaScience* 10.5 (2021), giab033.

[28] Xie, Q., Liu, Q., Mao, F., Cai, W., Wu, H., You, M., Wang, Z., Chen, B., Sun, Z. S., and Wu, J. “A Bayesian framework to identify methylcytosines from high-throughput bisulfite sequencing data”. *PLoS Computational Biology* 10.9 (2014), e1003853.

[29] Sun, X., Han, Y., Zhou, L., Chen, E., Lu, B., Liu, Y., Pan, X., Cowley Jr, A. W., Liang, M., Wu, Q., et al. “A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data”. *Bioinformatics* 34.16 (2018), pp. 2715–2723.

[30] Piaggeschi, G., Licheri, N., Romano, G., Pernice, S., Follia, L., and Ferrero, G. “MethylFASTQ: a tool simulating bisulfite sequencing data”. *2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. IEEE. 2019, pp. 334–339.

[31] Benjamini, Y. and Speed, T. P. “Summarizing and correcting the GC content bias in high-throughput sequencing”. *Nucleic acids research* 40.10 (2012), e72–e72.

[32] Chervova, O., Conde, L., Guerra-Assunção, J. A., Moghul, I., Webster, A. P., Berner, A., Larose Cadieux, E., Tian, Y., Voloshin, V., Jesus, T. F., et al. “The Personal Genome Project-UK, an open access resource of human multi-omics data”. *Scientific data* 6.1 (2019), p. 257.

[33] Hu, X., Yuan, J., Shi, Y., Lu, J., Liu, B., Li, Z., Chen, Y., Mu, D., Zhang, H., Li, N., et al. “pIRS: Profile-based Illumina pair-end reads simulator”. *Bioinformatics* 28.11 (2012), pp. 1533–1535.

[34] Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., and Davies, R. M. “HTSlib: C library for reading/writing high-throughput sequencing data”. *Gigascience* 10.2 (2021), giab007.

[35] Li, H. *WGSIM*. https://github.com/lh3/wgsim/.

[36] Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. “Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration”. *Briefings in bioinformatics* 14.2 (2013), pp. 178–192.

[37] Affinito, O., Palumbo, D., Fierro, A., Cuomo, M., De Riso, G., Monticelli, A., Miele, G., Chiariotti, L., and Cocozza, S. “Nucleotide distance influences co-methylation between nearby CpG sites”. *Genomics* 112.1 (2020), pp. 144–150.

[38] Teschendorff, A. E. “On epigenetic stochasticity, entropy and cancer risk”. *Philosophical Transactions of the Royal Society B* 379.1900 (2024), p. 20230054.

[39] Onuchic, V., Lurie, E., Carrero, I., Pawliczek, P., Patel, R. Y., Rozowsky, J., Galeev, T., Huang, Z., Altshuler, R. C., Zhang, Z., et al. “Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci”. *Science* 361.6409 (2018), eaar3146.

[40] Chen, J., Mu, W., Li, Y., and Li, D. “On the identifiability and interpretability of Gaussian process models”. *Advances in Neural Information Processing Systems* 36 (2023), pp. 70267–70278.

CHAPTER 3

Type-2 diabetes biomarker discovery and risk assessment through saliva DNA methylome

Abstract

The rising prevalence of type 2 diabetes (T2D) motivates innovative strategies to deepen disease understanding and enhance diagnostic capabilities. This study measures diabetes-specific epigenetic signals in saliva, establishing saliva DNA methylome as a promising medium for T2D screening and study. By integrating comprehensive whole-genome bisulfite sequencing (WGBS) and high-depth targeted bisulfite sequencing (TBS), we developed a cost-efficient two-step approach to profiling DNA methylation at regions of interest. WGBS analysis confirmed T2D-specific methylation signatures in saliva, revealing their enrichment in immune and metabolic regulation pathways. TBS enabled accurate cell type deconvolution, revealing minimal differences in cellular composition between diabetic and non-diabetic samples, suggesting intrinsic molecular changes drive the observed methylation changes. Epigenome-wide association studies further identified significant CpG sites, notably in the *ABCG1* region, with strong potential for T2D status prediction. These findings validate the saliva DNA methylome as a scalable, non-invasive resources for T2D biomarker discovery, advancing opportunities in T2D screening, risk assessment, and personalized medicine.

Key words: DNA methylation; Type-2 Diabetes; Non-invasive diagnostics; Whole-genome bisulfite sequencing; Targeted bisulfite sequencing; Epigenome-wide association study;

3.1 Introduction

Diabetes mellitus, a multifaceted metabolic disorder characterized by hyperglycemia, continues to pose a considerable and escalating global health challenge. According to the World Health Organization and the Centers for Disease Control and Prevention, the prevalence of diabetes has surged more than fourfold since 1980 [1], affecting approximately 529 million individuals worldwide [2] and 38.4 million in the United States in 2021 [3, 4]. Notably, over 90% of these cases are type 2 diabetes (T2D) [5]. This alarming rise (Figure 3.7) underscores the urgent need for deeper disease mechanism understanding, innovative diagnostic tools, as well as effective management strategies. Timely detection and intervention are crucial for managing diabetes, preventing associated complications, and reducing the economic burden on patients and healthcare systems.

Recent years have witnessed burgeoning interest in the role of epigenetics underlying diabetes [6–8], focusing on how environmental factors and lifestyle choices can induce gene expression changes without altering the DNA sequence. Among various epigenetic modifications, DNA methylation has garnered substantial attention for its robust and dynamic nature, playing important roles in gene regulation, cell differentiation, development and maintenance of homeostasis [9, 10]. Alterations in DNA methylation can contribute to disease and are often reflective of disease states, making them informative for disease mechanism research and diagnostic purposes [6, 11]. In the context of T2D, DNA methylation has been implicated in its onset [12, 13], progression [14], and complications [15–17], with emerging evidence highlighting its utility for diabetes risk prediction [18, 19]. Aberrant methylation patterns are also found in key genes associated with glucose metabolism [20], insulin secretion [21], insulin resistance [22], and inflammatory responses [23, 24]. These findings establish DNA methylation changes as valuable biomarkers for T2D, emphasizing their potential in elucidating disease mechanisms and developing novel diagnostic and treatment strategies.

Despite advancements in understanding DNA methylation changes in diabetes, most studies have focused on tissues such as blood, skeletal muscle, adipose tissue, and pancreas [6–8, 25–29], while the potential of saliva DNA methylation as a non-invasive biomarker remains underexplored. Saliva offers a particularly appealing option due to its ease of collection and high patient compliance, making it ideal for disease screening and routine monitoring. Recent studies have demonstrated a high similarity in methylation profiles between blood and saliva [30, 31], suggesting that disease-associated epigenetic signals identified in blood may also be detectable in saliva. This evidence forms the basis of our hypothesis that the saliva methylome can serve as a valuable medium for identifying T2D biomarkers. If validated, the saliva methylome profiles could facilitate T2D screening and monitoring, paving the road for future applications in T2D diagnostics and management.

A major challenge in current methylation profiling is the substantial resource demand, particularly with whole-genome bisulfite sequencing (WGBS), which remains prohibitively expensive for large-scale studies and clinical applications. While methylation microarrays offer a more affordable alternative and are widely used in DNA methylation research [32, 33], they capture only a limited, predetermined subset of CpG sites, potentially overlooking critical regions relevant to the disease of interest. Recognizing that many CpG sites exhibit minimal variation across cell types [34, 35] and non-cancer diseases [36], we identified an opportunity to reduce costs by selectively measuring the informative regions. In this study, we devised and implemented a cost-effective two-step strategy for T2D biomarker research (Figure 3.1). First, pooled WGBS of saliva DNA was conducted to identify T2D-associated signals, revealing 1,358 differentially methylated regions (DMRs) between diabetic and non-diabetic groups. Building on these findings, we designed custom probes to enrich these DMRs and other informative regions for targeted bisulfite sequencing (TBS). This integrated approach synergizes the broad genomic coverage of WGBS with the high-depth profiling of TBS, enabling precise DNA methylation measurements in genomic regions OF interest. By focusing sequencing efforts on relevant targets, this approach achieves cost-efficiency and makes large-scale study and routine screening more economically feasible.

Our study validated the presence of T2D-associated signals in the saliva methylome for the first time and provided key biological insights into the molecular basis of T2D. WGBS analysis revealed that the identified DMRs were significantly enriched in immune and metabolic pathways, consistent with the established pathophysiology of T2D [37]. TBS provided a high-depth profiling of the targeted regions and allowed for accurate cell-type deconvolution. This analysis revealed no major differences in cell type composition between diabetic and non-diabetic samples, suggesting the observed methylation changes are likely driven by intrinsic molecular alterations rather than shifts in cellular proportions. To further investigate the molecular changes underlying T2D, an epigenome-wide association study (EWAS) was conducted on TBS data and identified 12 significant CpG sites with the top hit in the *ABCG1* region, replicating and reinforcing findings from previous blood-based studies [32, 38]. Collectively, these findings establish saliva as a robust and practical medium for T2D research, enabling the precise identification of T2D-associated biomarkers. By integrating WGBS and TBS, this approach provides a cost-efficient and scalable framework for large-scale screening and monitoring. This study underscores the transformative potential of saliva-based epigenetic approaches in advancing T2D research and diagnostic applications.

3.2 Methods

3.2.1 Sample collection and preparation

This study involved saliva samples collected as part of the Parkinson's Environment and Genes (PEG) study [39–41]. While PEG is a case-control study focused on Parkinson's disease (PD), the saliva samples utilized in this study were primarily unrelated to PD. Participants were recruited from various sources across three counties in the Central Valley of California (Kern, Fresno, and Tulare) during two study waves (2000-2007 and 2009-2015). Population controls were enrolled from the same regions using Medicare lists and residential tax assessor records. Demographic data, medical history, medication use, and lifestyle information were collected through standardized interviews. Saliva collection tubes were mailed to participants, who then returned them via shipping or during in-person examinations. For this study, samples from participants with and without type 2 diabetes were randomly selected from those available in the PEG study, ensuring that the diabetic and non-diabetic groups were matched for age, sex, and ethnicity (Supplementary Data 1). Two batches of 96-well plates were prepared: the first in 2020 (Diabetes n=48, Non-diabetic n=48) for Whole Genome Bisulfite Sequencing (WGBS), probe design, and a pilot Targeted Bisulfite Sequencing (TBS) study, and the second in 2022 (Diabetes n=42, Non-diabetic n=54) for an expanded TBS study. The batch was included as a covariate in the downstream analyses. Each individual's saliva samples were sent to the UCLA Neuroscience Genomics Core (UNGC) for DNA extraction. Typically, 2.5 mL to 4 mL of saliva samples were collected using the Oragene saliva collection kit, followed by the standard manufacturer protocol of the Qiagen Puregene DNA extraction kit. After purification and extraction, the DNA concentration was measured using a NanoDrop 8000 spectrophotometer, and the extracted DNA samples were stored at -20°C before library preparation.

3.2.2 Whole genome bisulfite sequencing (WGBS)

To optimize cost efficiency, the extracted saliva DNA samples from the first batch were aggregated into four groups, matched by age and sex, as detailed in Supplementary Data 1. Each grouped DNA was pooled and subjected to whole genome bisulfite sequencing (WGBS) following established protocols [42]. Specifically, one microgram of purified DNA was sonicated using the Bioruptor Pico (Diagenode) for 15 cycles of 30 seconds ON and 90 seconds OFF, targeting a fragment size of 200-300 bp. The NEB Next Ultra II DNA kit (New England Biolabs) was used for subsequent end-repair, A-tailing, and ligation of pre-methylated unique-dual indexed adapters (Integrated DNA Technologies, custom synthesis). Bisulfite conversion was performed with the EZ DNA Methylation-Gold kit (Zymo Research). Final library amplification (12 PCR cycles) was conducted using KAPA HiFi U+ polymerase

(Roche Sequencing) and IDT xGen Primers. Library quality was assessed using the D1000 Assay on a 4200 Agilent TapeStation, and concentrations were quantified with the Qubit dsDNA BR Assay (Life Technologies). Sequencing was conducted on a NovaSeq 6000 platform (S4 lane), generating paired end reads of 150 base pairs.

3.2.3 WGBS data processing and DMR analysis

The raw sequencing reads underwent quality control using FastQC [43], followed by adapter and low-quality base trimming with fastp [44]. The trimmed reads were aligned to the reference genome (hg38) using BSBolt [45], with PCR duplicates marked with samtools [46]. Methylation levels of CpG sites were quantified for each sample, then aggregated into a methylation matrix. For downstream analysis, only sites with at least five counts in all four pooled samples were retained. The methylation matrix is available in Supplementary Data 2.

Differentially methylated region (DMR) analysis of the WGBS data was conducted using metilene [47] (version 0.2-8), with each candidate region required to contain a minimum of five CpG sites. In total, 162,833 genomic regions were analyzed. The statistical significance of methylation differences between diabetic and non-diabetic groups was evaluated using the Mann-Whitney U-test and the 2D Kolmogorov-Smirnov test for each region. Regions were considered significantly differentially methylated if they exhibited p-values below 0.01 for both tests and an absolute methylation difference exceeding 0.2 between the two groups. Supplementary Data 3 provides a comprehensive list of all candidate regions and identified DMRs, including their genomic coordinates, absolute methylation differences, and statistical significance levels.

3.2.4 Genomic region enrichment analysis and probe design

Following the identification of differentially methylated regions (DMRs) between diabetes and non-diabetes WGBS data, we conducted a genomic region enrichment analysis using the R package rGREAT [48] (version 2.4.0) in online mode. This analysis compared the DMRs against the total examined genomic regions as background, revealing significant enrichment patterns and biological relevance of the observed methylation changes. The identified DMRs were later submitted to Integrated DNA Technologies (IDT, https://www.idtdna.com/) for probe design, resulting in 937 custom probes targeting these regions. To further understand the regulatory context, we performed motif enrichment analysis using HOMER [49] (version 4.11), identifying enriched transcription factor binding sites (TFBS) for the probe-enriched regions.

In addition to the newly designed probes, we also incorporated previously designed probes targeting regions of interest from earlier studies [50, 51]. These probes were selected based on loci identified in public epigenome-wide association studies (EWAS) related to aging, cell types, and metabolic disorders. Due to an update in the probe set, there are slight differences between the probes used in batch 1 and batch 2, each containing a small set of extended probes labeled as 'batch1_extended' and 'batch2_extended.' The probes consistently used throughout the TBS study are collectively labeled as the 'Total Panel' and referred to as 'total probes' throughout the manuscript. The complete panel of probes, including both the Total Panel and extended probes, is detailed in Supplemental Data 4, with their sequences and target regions provided.

3.2.5 Targeted bisulfite sequencing (TBS)

For targeted bisulfite sequencing (TBS), 250 to 500 ng of purified gDNA from each sample was fragmented, and libraries were constructed following the same procedure as described in the WGBS protocol. Groups of 16 libraries, each with a unique dual index adapter, were pooled together, concentrated via SpeedVac, and subjected to targeted enrichment using custom 5'-biotinylated probes (IDT, xGen Custom Hybridization probe panel) (Supplemental Data 4). Enrichments were performed with the xGen Hybridization Capture kit (IDT), following the manufacturer's instructions, including overnight hybridization at 65°C. Bisulfite conversion of captured DNA was conducted using the EZ Methylation Gold kit (Zymo Research). Final

PCR amplification employed KAPA HiFi Uracil+ (Roche) with the following conditions: initial denaturation at 98°C for 2 minutes, followed by 16 cycles of 98°C for 20 seconds, 60°C for 30 seconds, and 72°C for 30 seconds, with a final extension at 72°C for 5 minutes. PCR products were purified using SPRI beads, and library quality control was conducted with the High-Sensitivity D1000 Assay on the 4200 Agilent TapeStation. Pools of 96 libraries were sequenced on a NovaSeq 6000 with paired-end 150-base reads.

3.2.6 TBS data processing and quality control

The raw sequencing reads of TBS data underwent a standardized preprocessing pipeline, including quality control, trimming, alignment, PCR duplicate marking, and methylation calling, as outlined in the WGBS data processing protocol. The methylation level of each CpG site is computed as follows:

$$m_i = \frac{\#\text{ methylated read count at site } i}{\#\text{ total read count at site } i} \tag{3.1}$$

To ensure data quality, samples with fewer than 2.5 million unique reads post-PCR deduplication or identified as outliers through PCA on the methylation level matrix were excluded from further analysis. After the data quality control, a total of 182 samples (Diabetic n=87, Non-diabetic n=95) were retained for further investigation.

We also performed quality control on the features. For epigenome-wide association studies, we focused on count data, retaining only those sites with read counts exceeding 10 in at least 80% of the samples. For analyses focused on methylation levels, such as cell deconvolution or machine learning model development, we retained only those sites that had at least 20 counts in at least 80% of the samples to ensure a reliable methylation level estimate. Missing values in the methylation level matrix were imputed using the KNN algorithm (k=5) implemented by R package impute [52] (version 1.70.0).

3.2.7 Cell type deconvolution

To ensure deconvolution accuracy, we first analyzed the cell composition in saliva using a single-cell RNA-seq dataset (GSE158055) [53] from the CELLxGENE Discover Data Portal (https://cellxgene.cziscience.com/), confirmed the predominance of epithelial and immune cells (Figure 3.12A-B). Building on this confirmation, we then compiled a comprehensive cell type methylation reference for deconvolution by integrating Whole Genome Bisulfite Sequencing (WGBS) profiles from the DNA methylation atlas (GSE186458) [54]. This reference encompasses epithelial and key immune cell types—granulocytes, monocytes, NK cells, B cells, and T cells (CD4, CD8, and naïve)—with detailed accession IDs and labels provided in Supplementary Data 5. Cell type-specific differentially methylated regions (DMRs) were identified by comparing each cell type against all others using metilene [47] (version 0.2-8), with parameters aligned to prior DMR analyses. Regions with a methylation difference exceeding 0.3 and an adjusted p-value below 0.05 (Benjamini-Hochberg correction) were extracted, and CpG sites in these regions were used to create the cell type methylation signature matrix. The resulting matrix, which serves as a robust reference for deconvolution, is available in Supplementary Data 6.

Given the limited number of CpG sites captured by TBS data, we further validated the deconvolution accuracy on TBS data using synthetic DNA methylation profiles. We generated 100 in-silico samples by mixing DNA methylation profiles of cell types with known proportions, with random gaussian noise (mean = 0, standard deviation = 0.05) added to mimic the variability in methylation levels due to sequencing noise. These synthetic datasets were then filtered to include only the CpG sites presented in the TBS methylation level matrix. Deconvolution was performed using the Houseman method [55], a well-established NNLS (non-negative least squares) approach for estimating cell-type compositions from bulk methylation data. The deconvolution accuracy was assessed by comparing the estimated and true cell proportions using key metrics, including R-squared and root mean square error (RMSE), confirming the precision and reliability of the method for TBS sites. This validated framework was then applied to deconvolve cell-type compositions in bulk saliva samples. Detailed cell proportions for each sample are provided in Supplementary Data 7.

3.2.8 Epigenome-wide association study

To prioritize the risk CpG sites associated with T2D, we conducted an epigenome-wide association study (EWAS) using the methylation read counts data using R package DSS [56]. Specifically, the DSS package utilized a beta-binomial modeling strategy to model the methylated counts Y_i based on the total counts N_i for site i by

$$Y_i \sim \text{binom}(N_i, p_i) \tag{3.2}$$

$$p_i \sim \text{beta}(\pi_i, \phi_i) \tag{3.3}$$

where π_i, ϕ_i are the mean and dispersion parameter for site i. The mean parameter was modeled as $g(\pi_i) = \sum_{j=1}^{J} X_j\beta_j$ where $g(\cdot)$ is the link function and X_j for $j = 1, \ldots, J$ are the covariates (including the variable of interest and other covariates). By testing the coefficient $\beta_j = 0$ using the F-test, the significance level of association between diabetes status and methylation of site i can be assessed. For our analysis, we used the methylation count matrix for the EWAS analysis, including age, sex, ethnicity, batch, and cell proportions as covariates, and tested whether a site is associated with diabetes. A Manhattan plot is used to visualize the testing results. Genes within 2kb window of the most prominent sites that passed the with suggestive p-value (10^{-4}) is annotated on to the plot. Detailed methylation count matrix and test result are available in Supplementary Data 2.

3.2.9 Data availability

The raw data and processed supplementary data can be accessed in [57]

3.3 Results

3.3.1 WGBS identifies DMRs associated with diabetes in saliva

To investigate DNA methylation changes associated with T2D while optimizing sequencing efficiency, we implemented a carefully designed sample pooling strategy followed by Whole Genome Bisulfite Sequencing (WGBS). In this study, we pooled 96 saliva samples with matched demographical attributes into four groups: Diabetic Male, Diabetic Female, Non-diabetic Male, and Non-diabetic Female, with a balanced sample size per group. This pooling approach ensured adequate representation of each group and enabled robust comparisons across groups at a reduced cost. The WGBS data were then processed and aligned to the human reference genome (hg38) with CpG methylation levels quantified. Downstream differential methylation region (DMR) analysis between diabetic and non-diabetic groups revealed 1358 potential DMRs out of 162833 total regions (0.8%), visualized using a volcano plot (Figure 3.2A) and a heatmap (Figure 3.2B). These findings highlight significant epigenetic variations (both hypo- and hyper-methylation) between diabetic and non-diabetic individuals.

Genomic region enrichment analysis was conducted to elucidate the biological relevance of the identified DMRs. The results exhibited substantial enrichment in genomic regions associated with metabolic regulation and immune response pathways (Figure 3.2C, Figure 3.8), underscoring their potential relevance in diabetes pathogenesis. Notably, several key pathways, such as leukocyte-mediated immunity and neutrophil activation, were significantly enriched, aligning with the current understanding of diabetes as a multifactorial disease involving intricate interactions between metabolic dysfunction and immune responses [37]. Based on the identified DMRs, we designed a set of probes (n=937) for targeted bisulfite sequencing. Motif analysis of these probe-enriched regions revealed seven significant transcription factor binding sites (p-value<0.01, adjusted p-value<0.1, Figure 3.2D), which are associated with T2D [58] and related traits, such as glycolysis [59] and immune response [60]. This association underscores the functional relevance of the identified DMRs and enriched regions with diabetes pathophysiology. Taken together, the WGBS analysis confirmed the presence of diabetes-specific methylation signals in saliva and facilitated the screening of genomic regions enriching these signals, paving the way for efficient profiling through Targeted Bisulfite Sequencing.

3.3.2 TBS enriches target regions with high sequencing depth

To enhance efficiency in large-scale epigenetic profiling, we implemented targeted bisulfite sequencing (TBS) using a curated set of probes. This set includes probes designed to enrich the identified DMRs from WGBS analysis, as well as additional probes targeting regions associated with phenotypes such as aging, cell type, BMI and metabolic disorders [29, 50, 51]. A total of 8154 probes were used throughout the targeted bisulfite sequencing study, capturing ∼1M bases of the genome. Genomic coordinate overlap analysis with the existing EWAS database [61, 62] revealed that more than 40% of the probes overlap with known EWAS sites associated with diabetes and related traits, including BMI, obesity, fasting glucose levels, insulin levels and resistance (Figure 3.3A), ensuring the capture of the diabetes-informative methylation regions.

With the curated probe set, we conducted targeted bisulfite sequencing on two cohorts (section 3.2), aiming for 10 million reads per sample. Our results confirmed that TBS can effectively capture the targeted genomic regions with high depth (Figure 3.3B). Of note, the enriched regions exhibited an average of 1300-fold higher depth than non-enriched background regions (Figure 3.3C, Figure 3.9), and over 80% of the CpG sites within the targeted regions had depth greater than 10 counts (Figure 3.3D). These findings demonstrate the remarkable efficiency of TBS in profiling targeted genomic regions with high depth while achieving cost efficiency. The successful capture of informative regions establishes TBS as a scalable solution for high-throughput epigenetic studies. Its high-depth coverage of CpG sites within targeted regions enables accurate and reliable DNA methylation quantification, ensuring robust statistical power for detecting differential signals in downstream analyses.

3.3.3 Cell type deconvolution reveals minimal T2D-related compositional changes in saliva

Both the WGBS and TBS technologies are applied to bulk saliva samples, which obscures the specific cell type abundance associated with T2D in saliva. To address this, we first assessed whether the TBS sites contain cell type information. We downloaded a WGBS dataset containing a comprehensive methylation atlas of normal human cell types [54] and identified cell type-specific regions. By overlapping with the TBS sites, we found a significant proportion of the TBS sites fell within these cell type-specific regions, sufficiently distinguishing the different cell types in saliva tissue (Figure 3.10). To further validate the utility of these sites for cell type deconvolution, we generated in-silico mixtures of DNA methylation profiles with known cell type proportions. Using these simulated datasets, we performed cell type deconvolution analysis using the Houseman method [55] (Figure 3.11A), achieving a root mean square error (RMSE) of less than 0.01 and an R-squared value approaching to 1 (Figure 3.11B). Repeated experiments consistently showed high accuracy (Figure 3.11C), confirming that the TBS sites support accurate cell-type deconvolution.

Following this validation, we applied the deconvolution method to bulk saliva TBS data to investigate cell type composition in our samples. The analysis revealed that monocytes, granulocytes, and epithelial cells were the most abundant cell types in saliva, consistent with previous literature and our reanalysis of recent single-cell RNA-seq data of human sputum tissue [53] (Figure 3.12). Comparing cell type proportions between diabetic and non-diabetic samples, we observed no significant changes in major cell types (Figure 3.4), except for a marginally significant difference in naïve T cells. However, this association was not significant after p-value adjustments. Our analysis also revealed that cell type proportions are highly correlated with the top Principal Components (PCs) of the DNA methylation matrix (Figure 3.13), emphasizing the dominant role of cell proportions in the epigenetic variability [33] and echoing the importance of including these variabilities in EWAS analysis to account for cell type heterogeneity [63].

In conclusion, our analysis demonstrated that TBS sites capture cell-type information and enable accurate cell-type deconvolution. Notably, the similar cell type proportions observed between diabetic and non-diabetic groups suggest that diabetes-related epigenetic changes in saliva are driven by intrinsic molecular alterations rather than shifts in cell composition.

3.3.4 EWAS reveals differential DNA methylation associated with T2D status

Another distinct advantage of TBS is its ability to elucidate the epigenetic mechanisms underlying diabetes at the molecular level, providing valuable insights into disease pathways and potential therapeutic targets. To demonstrate this potential, we conducted an epigenome-wide association study (EWAS) on the TBS data. In this analysis, we accounted for key covariates such as age, sex, ethnicity, study batches, and cell-type proportions, to mitigate the influence of confounding factors and identified CpG sites associated with diabetic states. The EWAS results, visualized with a Manhattan plot (Figure 3.5A) and a QQ plot (Figure 3.14), revealed 12 CpG sites significantly associated with T2D, with 7 of these sites near genes previously implicated in diabetes pathogenesis, such as *ABCG1* [32], *LDLRAD4* [64], and *TYK2* [65]. Figure 3.5B shows methylation level differences at the top CpG sites between diabetic and non-diabetic groups after adjusting for covariates.

Notably, the strongest signal was observed in the *ABCG1* gene region, corroborating a recent meta-analysis of blood-based EWAS [32] that identified *ABCG1* as a top hit across five cohorts with over 3,000 samples. *ABCG1* plays a crucial role in regulating lipid metabolism and cholesterol efflux, which are essential for maintaining cellular lipid homeostasis [66]. The dysfunction of *ABCG1* is particularly detrimental in the context of diabetes, where impaired cholesterol efflux can exacerbate insulin resistance and promote atherosclerosis [67], a common complication of the disease. Furthermore, the accumulation of lipids can result in cellular stress and apoptosis [68], which in turn triggers an immune response and leads to chronic inflammation, further accelerating diabetes progression and increasing cardiovascular disease risk. The role of *ABCG1* in lipid regulation and its broader impact on inflammation and cell viability highlight its potential as a therapeutic target in diabetes management.

Our EWAS findings, particularly the significant signal at the *ABCG1* region, highlight the gene's critical role in diabetes. These results validate the utility of saliva-based DNA methylation analysis in diabetes research and emphasize the potential of these epigenetic markers as biomarkers for diagnosing diabetes, predicting risk, and informing the development of targeted therapeutic strategies.

3.3.5 Predictive performance of individual methylation sites for T2D status

To evaluate the potential of DNA methylation as a biomarker for diabetes diagnosis, we analyzed the predictive performance of individual methylation sites using ROC analysis. Figure 3.6 illustrates the ROC curves of all tested sites, with chr19:10380958 (*TYK2*) and chr21:42236481 (*ABCG1*) achieving AUC values of 0.683 and 0.681, respectively, indicating moderate predictive ability. The shaded region, representing the 95% quantile range of ROC curves across all sites, highlights the variability in predictive performance. These results demonstrate that while some individual sites show moderate performance, most exhibit weak signals, underscoring the importance of refining site selection. This validates the need for a targeted sequencing strategy, as it can effectively enrich informative loci, improving the signal-to-noise ratio and enabling precise and efficient methylation profiling.

Additionally, although the predictive power of individual loci is limited, combining methylation profiles within multivariate or ensemble frameworks offers a promising path forward. Future model development should focus on integrating information across multiple loci to enhance predictive accuracy and robustness. These strategies have the potential to yield reliable, clinically actionable tools for diabetes diagnosis and risk stratification, underscoring the transformative potential of the saliva DNA methylome as a scalable, non-invasive approach for advancing T2D biomarker discovery and improving disease management.

3.4 Discussion

The rising prevalence of type 2 diabetes (T2D) underscores the need for innovative approaches that extend beyond traditional diagnostics to explore the molecular mechanisms underpinning the disease. Identifying reliable biomarkers and investigating epigenetic modifications, such as DNA methylation, can deepen our understanding of T2D pathogenesis, enable early detection, and inform the development of targeted therapeutic strategies. To address the need for accessible and noninvasive approaches, this study evaluated the potential of saliva DNA methylome for T2D biomarker discovery and diagnostic applications, offering insights into the molecular and cellular dynamics underlying the disease.

One key challenge in methylation profiling is the high cost of obtaining informative and accurate measurements. Whole-genome bisulfite sequencing (WGBS) provides comprehensive coverage but requires high sequencing depth, making it prohibitively expensive for large-scale studies. Methylation arrays, while more affordable, capture only a fixed, small subset of CpG sites, potentially overlooking critical variations relevant to disease. To overcome these limitations, we developed a cost-efficient two-step strategy, combining WGBS to identify key regions with targeted bisulfite sequencing (TBS) for high-depth profiling. This approach significantly reduces costs while maintaining precision, making it suitable for broader and cohort-level applications.

Using this combined strategy, we obtained compelling evidence supporting the use of saliva DNA methylation for T2D biomarker discovery and risk assessment. Through WGBS, we identified differentially methylated regions (DMRs) associated with T2D, particularly enriched in pathways related to immune response and metabolic regulation. These results align with existing blood-based studies [37], confirming that saliva, like blood, harbors diabetes-specific epigenetic signatures. The subsequent application of targeted bisulfite sequencing (TBS) enabled precise quantification of DNA methylation in these key regions at the cohort scale.

Importantly, cell type deconvolution of the TBS data revealed no significant differences in cell proportions between diabetic and non-diabetic groups, suggesting that the observed methylation changes are primarily intrinsic rather than driven by shifts in cell composition. Further supporting these findings, an epigenome-wide association study (EWAS) conducted on the TBS data identified significant CpG sites, with the top hit in the *ABCG1* gene region, consistent with prior blood-based findings [32]. Collectively, our findings provide the first validation of T2D-specific methylation signals in saliva, establishing a novel paradigm for non-invasive diabetes screening and offering valuable insights into the epigenetic basis of this prevalent disease.

Despite these promising findings, our study has limitations that warrant further investigation. The relatively small sample size may have reduced the statistical power of our findings, potentially leading to missing important epigenetic signals. Expanding the sample size and including a more diverse population would enhance the robustness and generalizability of the results. Additionally, many diabetic participants were under good glycemic control, which may have attenuated the strength of detectable epigenetic changes. Future studies should include individuals at various stages of disease progression to capture a broader range of epigenetic variations. While our probe panel targeted diabetes-related sites, it could be further optimized by integrating prior knowledge to capture a wider range of diabetes-associated signals, particularly regions near genes involved in insulin signaling, glucose metabolism, and related pathways. Additionally, advanced machine learning approaches, such as ensemble and contrastive learning [69, 70], hold promise for enhancing diagnostic model performance by effectively integrating subtle signals linked to different disease states. Addressing these limitations through larger cohorts, refined probe designs, and advanced modeling techniques will be crucial for maximizing the potential of saliva DNA methylation in diabetes research and diagnostics.

Looking ahead, further research could greatly enhance the utility and impact of our approach. Advanced barcoding and multiplexing techniques, such as Time-Seq [71], could further reduce costs, making this method even more accessible for large-scale studies and routine clinical applications. The non-invasive nature of saliva collection, combined with cost-effective methylation profiling, offers a practical and scalable solution for diabetes screening and longitudinal monitoring. Conducting longitudinal studies will be critical to establish causal relationships between DNA methylation changes and T2D progression, providing deeper insights into disease mechanisms and enabling timely interventions. Ultimately, integrating saliva DNA methylation profiling into clinical practice has the potential to revolutionize diabetes diagnostics and monitoring, facilitating earlier detection, personalized treatment, and more effective disease management.

In conclusion, this proof-of-concept study validates diabetes-specific epigenetic signals in saliva, establishing saliva DNA methylation as a promising biomarker source for non-invasive T2D research and screening. By employing an innovative sequencing strategy that enhances precision while reducing costs, we have made epigenetic profiling feasible for large-scale studies and clinical applications. While further research with larger, more diverse cohorts is needed, this approach lays the groundwork for transforming diabetes diagnostics and monitoring, paving the way for more personalized and accessible care.

3.5 Acknowledgements

I want to thank all study participants for making this work possible. I also want to thank Dr. Marco Morselli for conducting the WGBS experiment and initiating its analysis, as well as Dr. Kimberly C. Paul and Dr. Beate Ritz for providing the saliva samples. I also appreciate the UCLA Neuroscience Genomics Core (UNGC) and the Broad Stem Cell Sequencing Core (BRSRC) for assistance in generating WGBS and TBS data. Additionally, I would like to thank Dr. David Wong from the UCLA School of Dentistry for the helpful discussions. Figure 3.1 was created with BioRender.com, thus acknowledging it by courtesy.

3.6 Tables and figures

Table 3.1: Characteristics of the study population.

Diabetes	**No (N=95)**	**Yes (N=87)**	**p value**
Age (years), mean ± SD	67.505 9.500	67.816 (8.165)	0.814
Sex (male), n (%)	54 (56.8%)	52 (59.8%)	0.689
Ethnicity (White), n (%)	73 (76.8%)	63 (72.4%)	0.492
Batch (1), n (%)	48 (50.5%)	47 (54.0%)	0.637
Parkinson Disease, n (%)	72 (75.8%)	64 (73.6%)	0.730
Smoker, n (%)	49 (52.1%)	41 (47.1%)	0.501

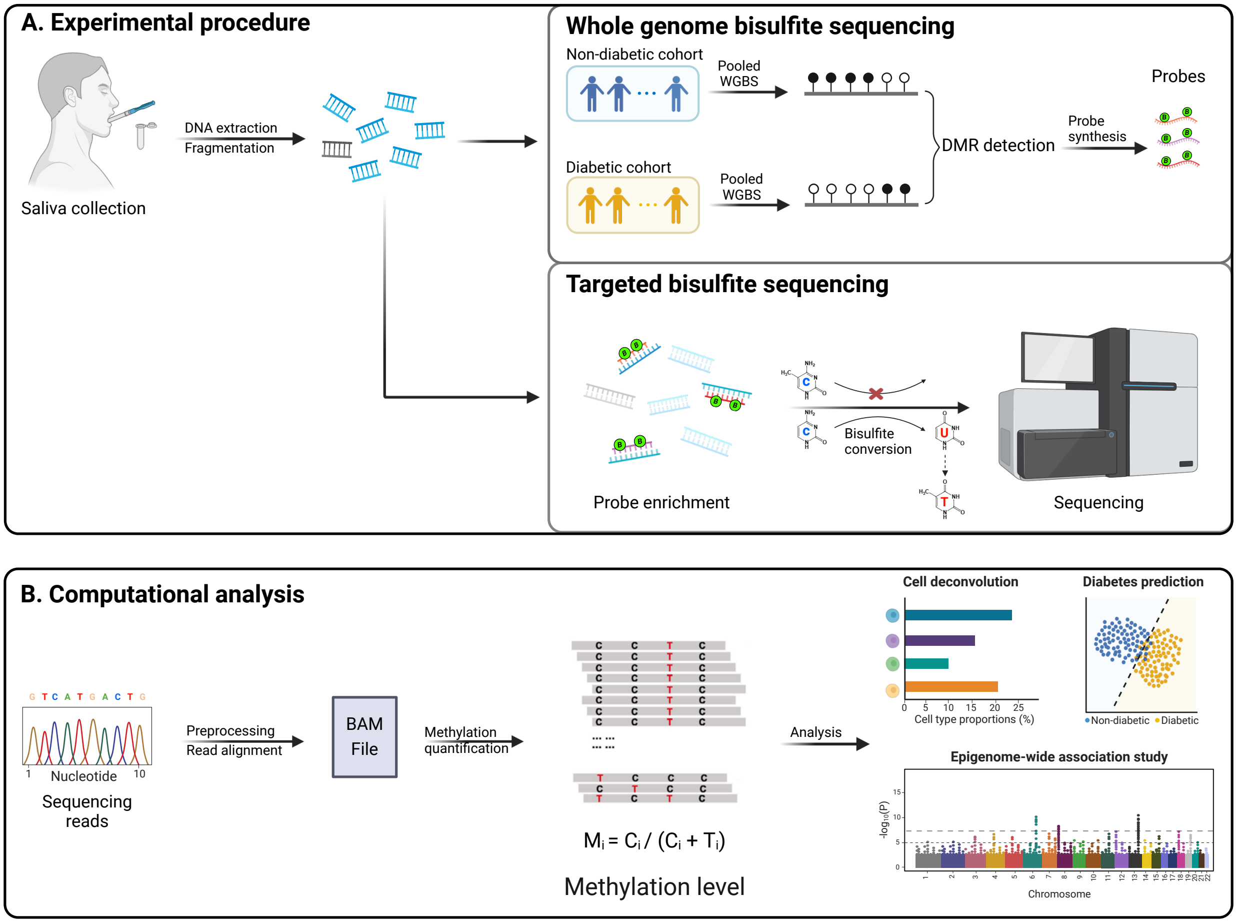
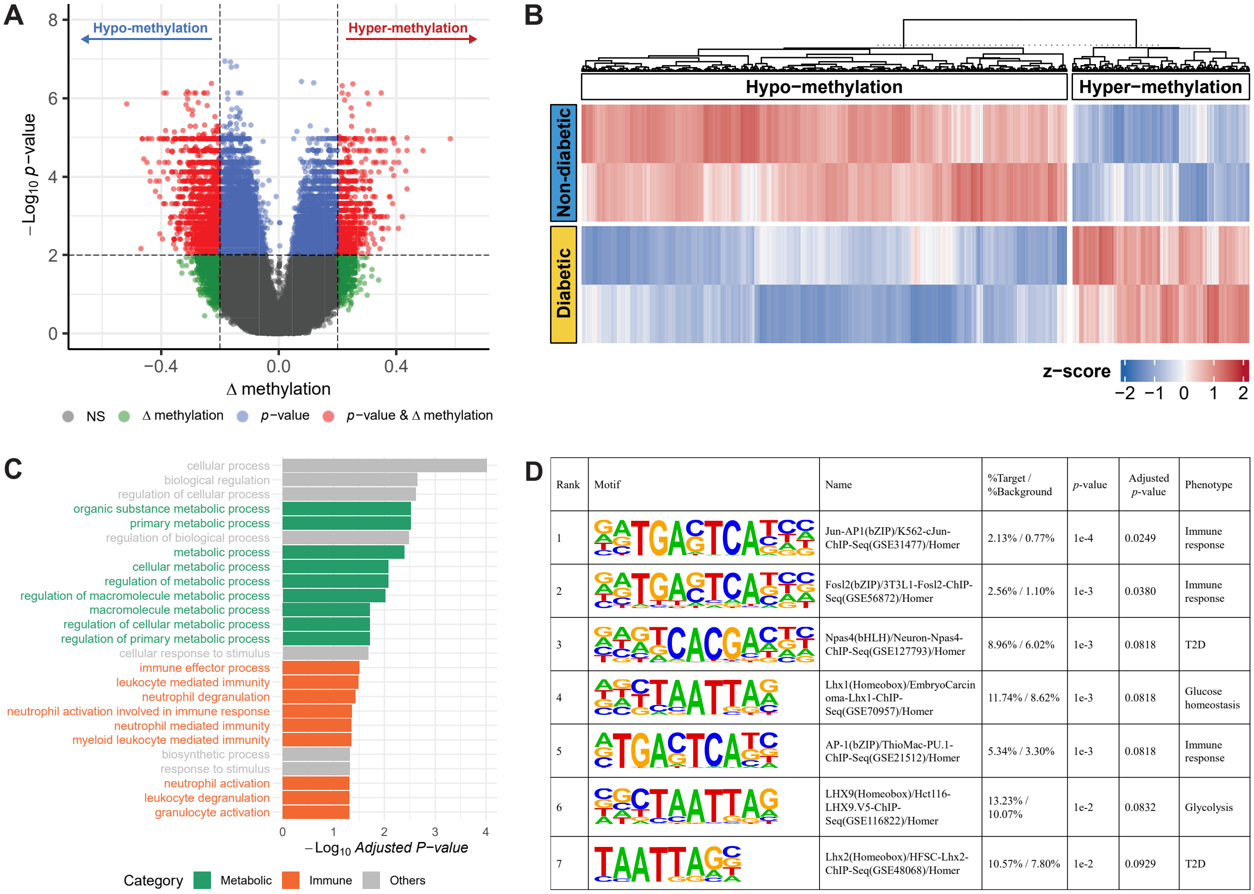


Figure 3.1: Study design for saliva DNA methylome analysis in Type 2 diabetes. (A) Experimental procedure. Participants' saliva samples were collected, followed by DNA extraction and fragmentation. Pooled samples from non-diabetic and diabetic cohorts were then subjected to whole-genome bisulfite sequencing (WGBS) to identify differentially methylated regions (DMRs) associated with T2D. Probes targeting these DMRs were synthesized and used for targeted region enrichment, followed by bisulfite conversion and sequencing in high-efficiency Targeted Bisulfite Sequencing (TBS). (B) Computational Analysis. Sequencing reads underwent preprocessing and alignment, with methylation levels quantified as the ratio of methylated cytosine (C) counts to the total counts at each CpG site. The methylation data were used for downstream analysis, including cell type deconvolution, an epigenome-wide association study, and diabetes status prediction.



Rank	Motif	Name	%Target / %Background	*p*-value	Adjusted *p*-value	Phenotype
1		Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	2.13% / 0.77%	1e-4	0.0249	Immune response
2		Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer	2.56% / 1.10%	1e-3	0.0380	Immune response
3		Npas4(bHLH)/Neuron-Npas4-ChIP-Seq(GSE127793)/Homer	8.96% / 6.02%	1e-3	0.0818	T2D
4		Lhx1(Homeobox)/EmbryoCarcinoma-Lhx1-ChIP-Seq(GSE70957)/Homer	11.74% / 8.62%	1e-3	0.0818	Glucose homeostasis
5		AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	5.34% / 3.30%	1e-3	0.0818	Immune response
6		LHX9(Homeobox)/Hct116-LHX9.V5-ChIP-Seq(GSE116822)/Homer	13.23% / 10.07%	1e-2	0.0832	Glycolysis
7		Lhx2(Homeobox)/HFSC-Lhx2-ChIP-Seq(GSE48068)/Homer	10.57% / 7.80%	1e-2	0.0929	T2D

Figure 3.2: Differential methylation region and genomic region enrichment analysis for saliva WGBS data. (A) Volcano plot showing differential methylation region (DMR) analysis results, comparing diabetic group to non-diabetic controls. The x-axis represents the difference in methylation levels (Δmethylation), while the y-axis displays the -log10 p-values. Regions where both Δmethylation and the p-value exceed their respective thresholds are highlighted in red, representing hypo-methylation (left) and hyper-methylation (right). Regions where only the Δmethylation or p-value passe their corresponding threshold are shown in green and blue, respectively. Non-significant regions are depicted in gray. (B) Hierarchical clustering heatmap of DMRs' methylation levels across diabetic and non-diabetic groups. The color scale represents z-scores, with hypo-methylated regions indicated in blue and hyper-methylated regions in red, highlighting differential methylation between the two groups. (C) Bar plot showing the genomic region enrichment analysis results of DMRs. The x-axis represents the -log10 adjusted p-value of enrichment, and the y-axis lists the enriched Gene Ontology (GO) terms of biological processes. Metabolic-related processes are highlighted in green, immune-related processes in orange, and others in gray, with notable enrichment in pathways related to cellular metabolic and immune responses. (D) Table summarizing the significantly enriched transcription factor binding sites. Each motif was ranked by significance, and the percentage of target versus background regions, p-value, adjusted p-value, and associated phenotype were provided.

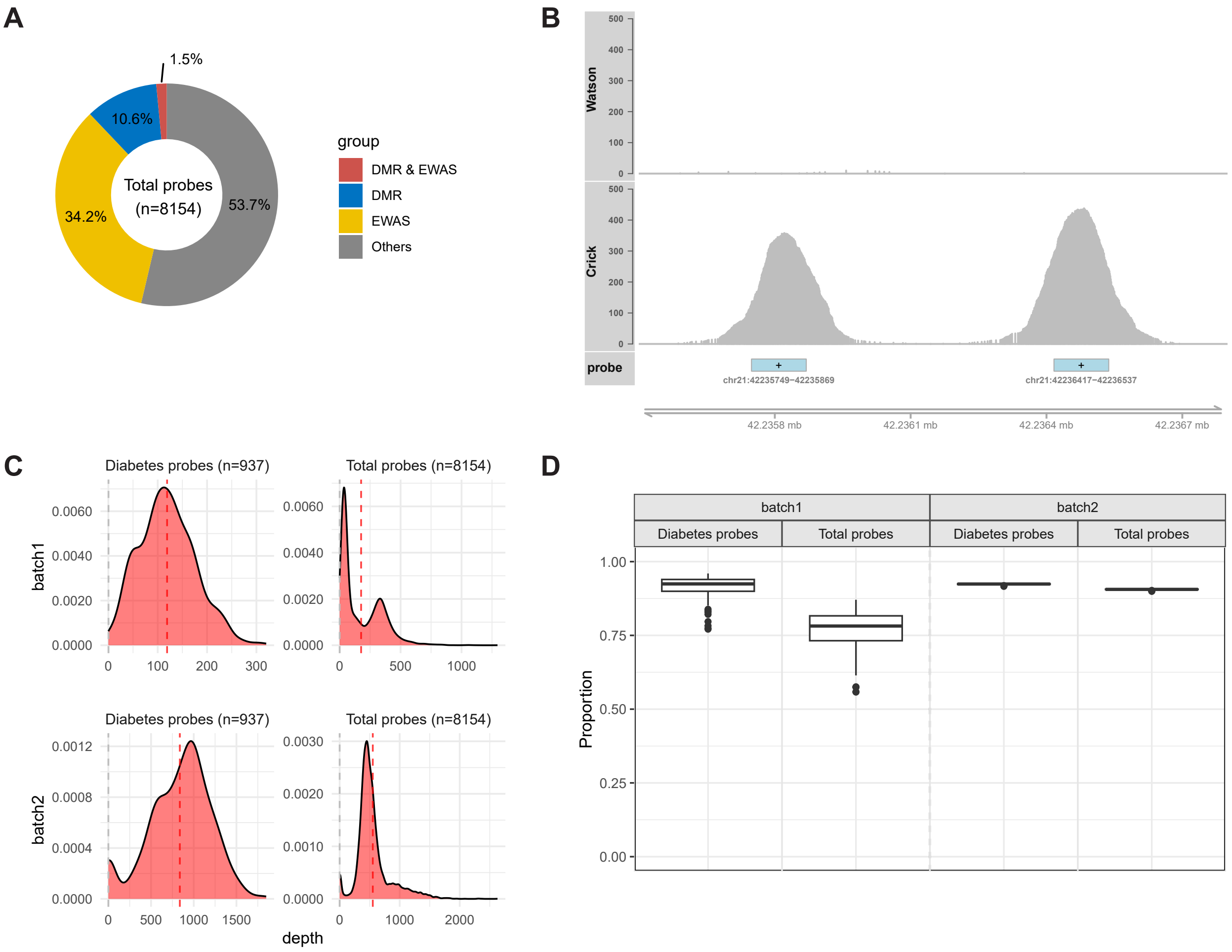


Figure 3.3: TBS captures desired region with high depth with reduced cost. (A) Pie chart illustrating the composition of the probe set (n=8154), highlighting its overlap with the differentially methylated regions (DMRs) identified in WGBS data and the public EWAS database. The probes are categorized as overlapping with DMR & EWAS (blue), DMR only (red), EWAS only (yellow), and other regions (gray). (B) Coverage plot showcasing an example of read coverage across a targeted genomic region (chr21:42,235,500-42,236,800) in a sample's TBS data. The x-axis represents the genomic coordinates, and the y-axis shows the depth at each locus. Both Watson and Crick strands are displayed, with the targeted probe region highlighted in blue. The plus signs indicate probes designed on the Watson strand to capture the Crick strand. (C) Density plots showing the depth distribution of probes targeting diabetes DMR regions (n=937) and the total probe set (n=8154) across two batch samples. The red dashed line indicates the average depth of the enriched regions, with grey dashed lines indicating the non-enriched background regions. (D) Box plots displaying the percentage of CpG sites within the probe regions that achieve a sequencing depth greater than 10x. The plots demonstrate the efficiency of TBS in achieving high sequencing depth for the targeted regions across probe sets and batches.

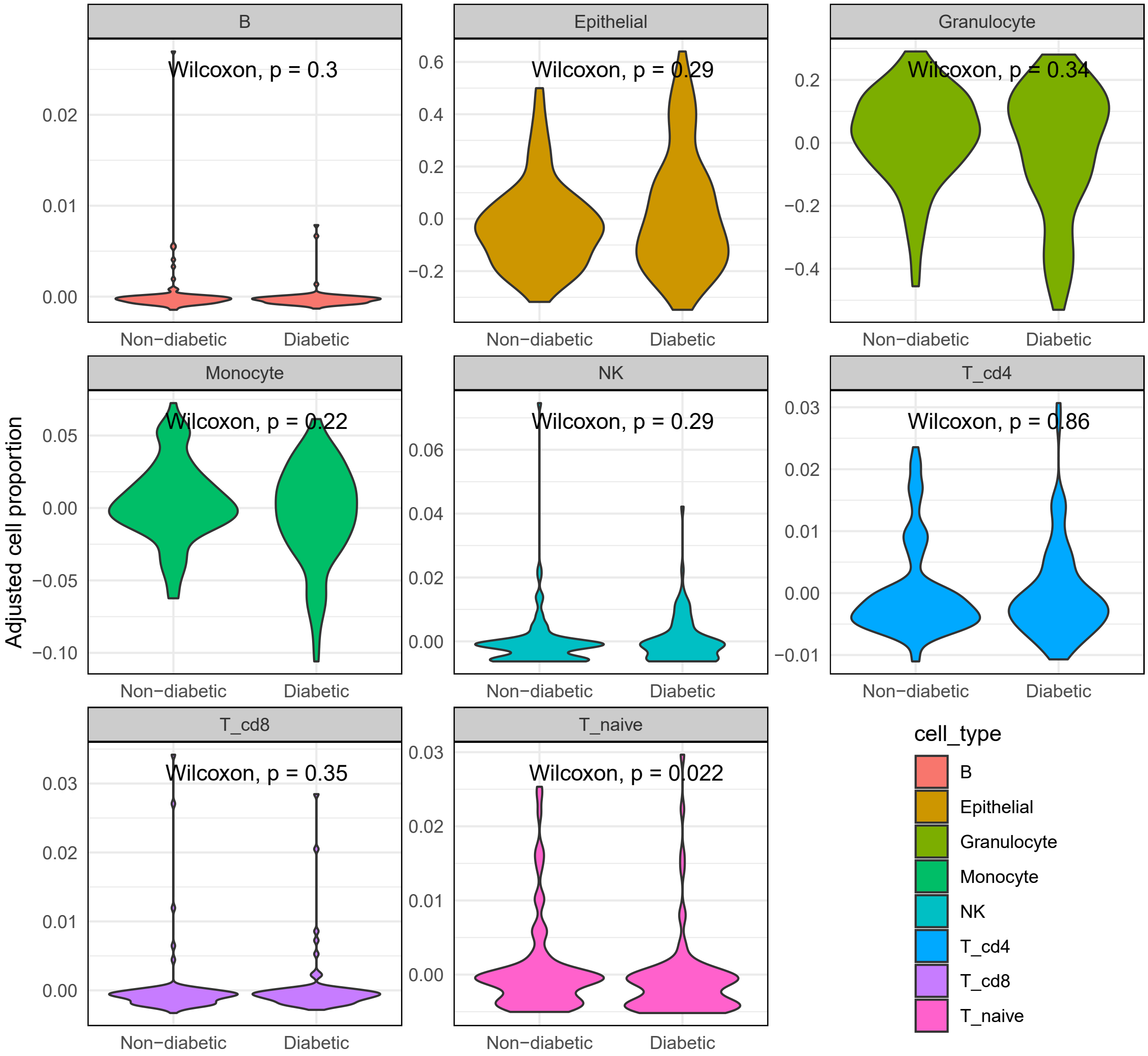


Figure 3.4: Differential cell type proportions between diabetic and non-diabetic samples. Violin plots show the difference in cell type proportions between diabetic and non-diabetic samples for each cell type after adjusting other covariates (age, sex, ethnicity, and batch). Wilcoxon p-values are annotated in each subplot, revealing no significant difference in cell proportions between the two groups, except a marginally significance for naïve T cell (p=0.022).

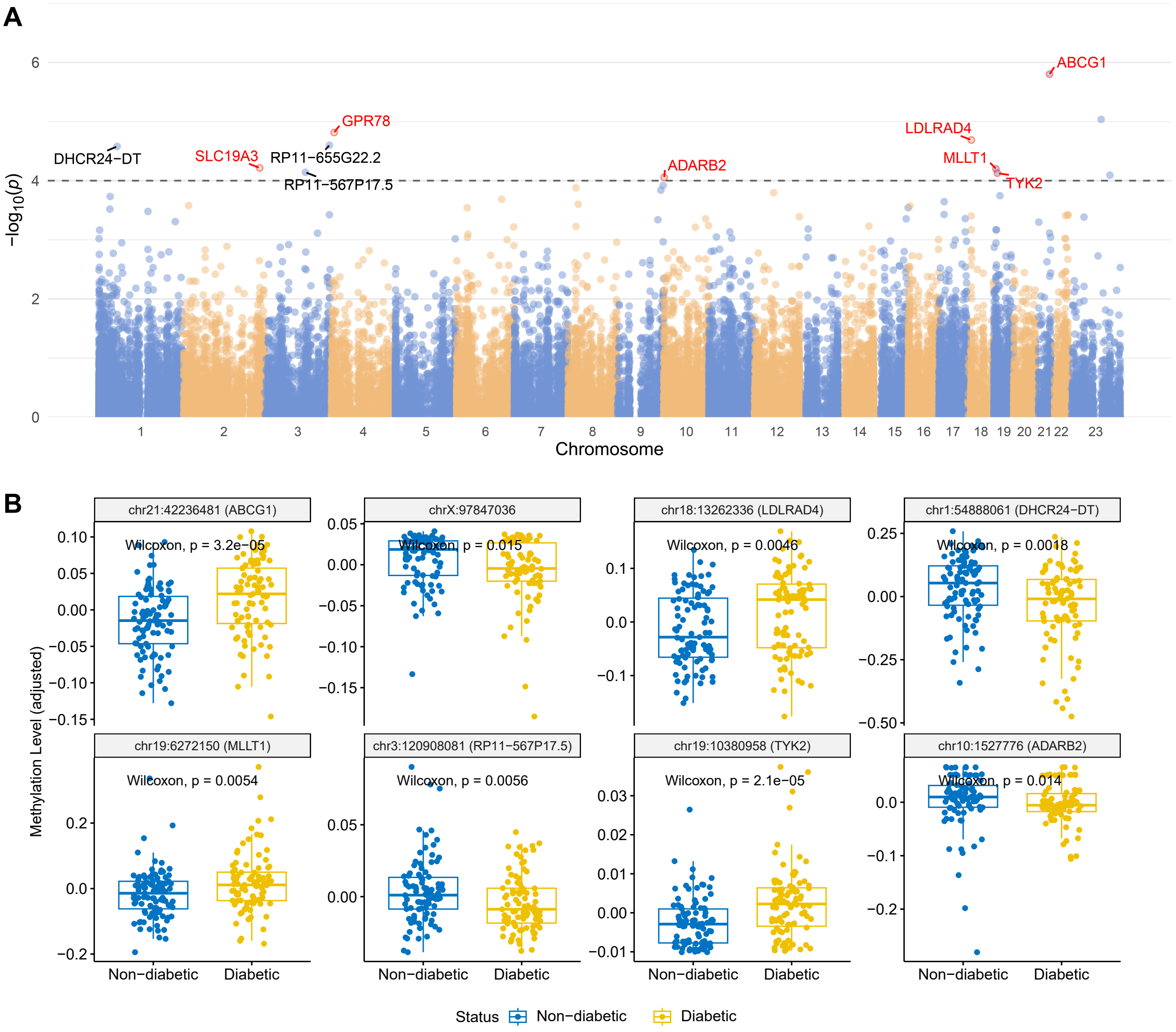


Figure 3.5: EWAS analysis identifies methylation sites associated with diabetes status. (A) Manhattan plot depicting the epigenome-wide association between DNA methylation levels and T2D status. Each dot represents a CpG site, with the -log10(p-value) plotted against its chromosomal position. The horizontal dashed line indicates the suggestive significance threshold (10^{-4}). Genes located within a 2kb window of the top CpG sites are annotated, with established diabetes-related genes highlighted in red, such as *ABCG1*, *LDLRAD4*, *TYK2*, etc. (B) Boxplots illustrating the methylation levels (adjusted for covariates) at selected top CpG sites. Each plot panel compares the methylation levels between diabetic (yellow) and non-diabetic (blue) samples at the specific CpG site, highlighting their potential role in diabetes pathogenesis.

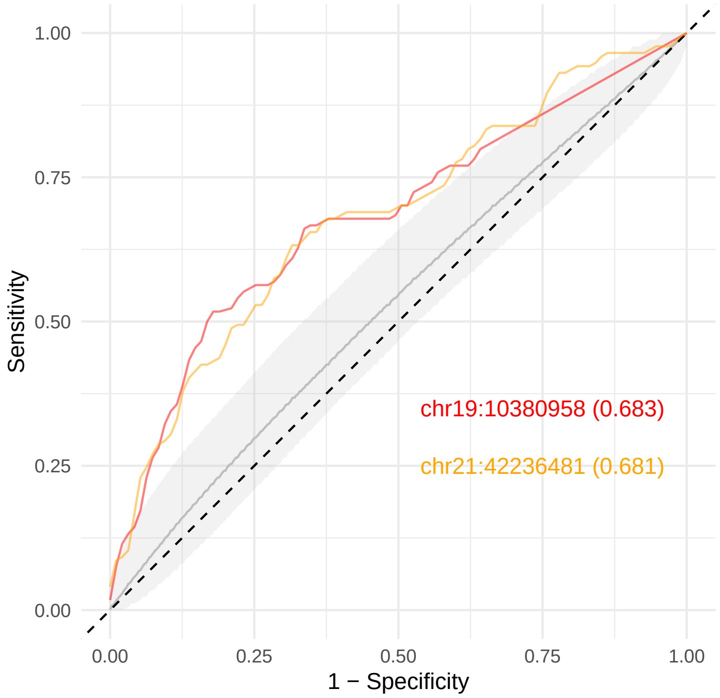


Figure 3.6: ROC curve for diabetes status classification using individual methylation sites. This ROC curve highlights the classification performance of two key methylation sites, chr19:10380958 and chr21:42236481, in predicting T2D status, with respective AUC values of 0.683 and 0.681. The shaded region denotes the 95% range of predictive performance across all other analyzed methylation sites, providing context for the highlighted sites' relative performance, with the dashed diagonal line representing AUC 0.5 as a reference

3.7 Supplementary materials

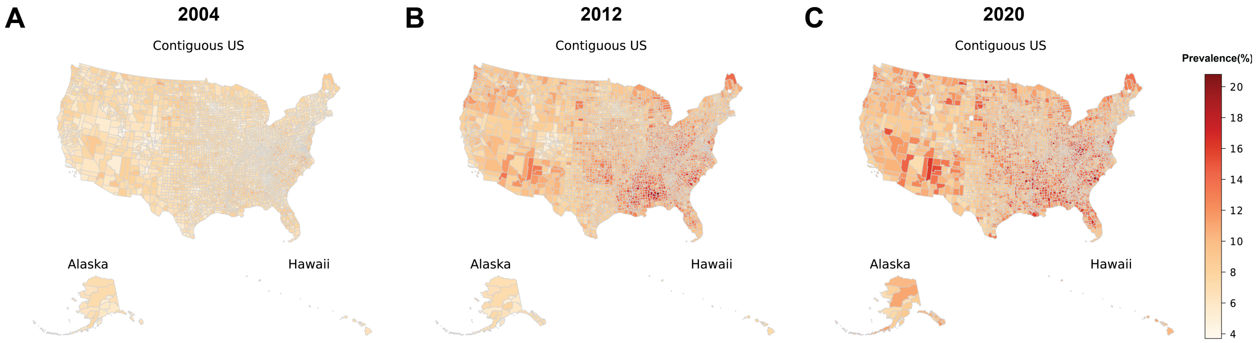


Figure 3.7: The increasing prevalence of diabetes across U.S. counties from 2004 to 2020. Choropleth map displaying the escalating diabetes prevalence in U.S. counties from 2004 (A), through 2012 (B), to 2020 (C), which underscores the growing public health challenge and the need for targeted interventions. The color gradient indicates the percentage of the population with diabetes, with darker colors representing increasing prevalence, as shown in the accompanying legend (4% to 20%). County-level diabetes prevalence data was obtained from the United States Diabetes Surveillance System (https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html).

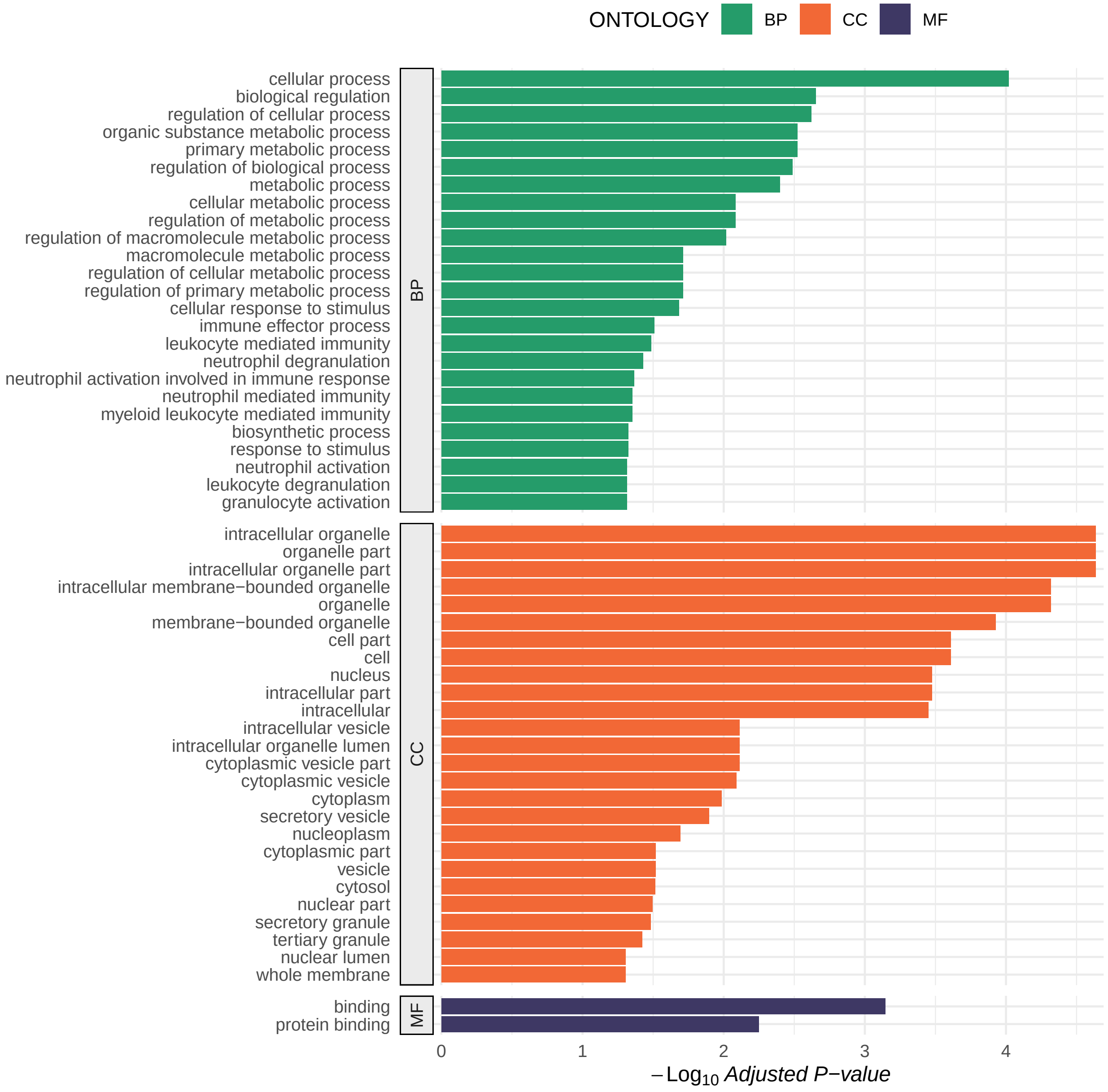


Figure 3.8: GO pathway enrichment for DMR regions in WGBS analysis. Genomic region enrichment analysis for differentially methylated regions (DMRs) identified in the Whole Genome Bisulfite Sequencing (WGBS) data. The bar plot presents enriched GO terms categorized by Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) ontologies, with the x-axis showing the -log10 of the adjusted p-values. Only GO terms with an adjusted p-value below 0.05 are displayed. The analysis highlights significant associations of DMRs with various biological processes, particularly those related to metabolic functions and immune responses.

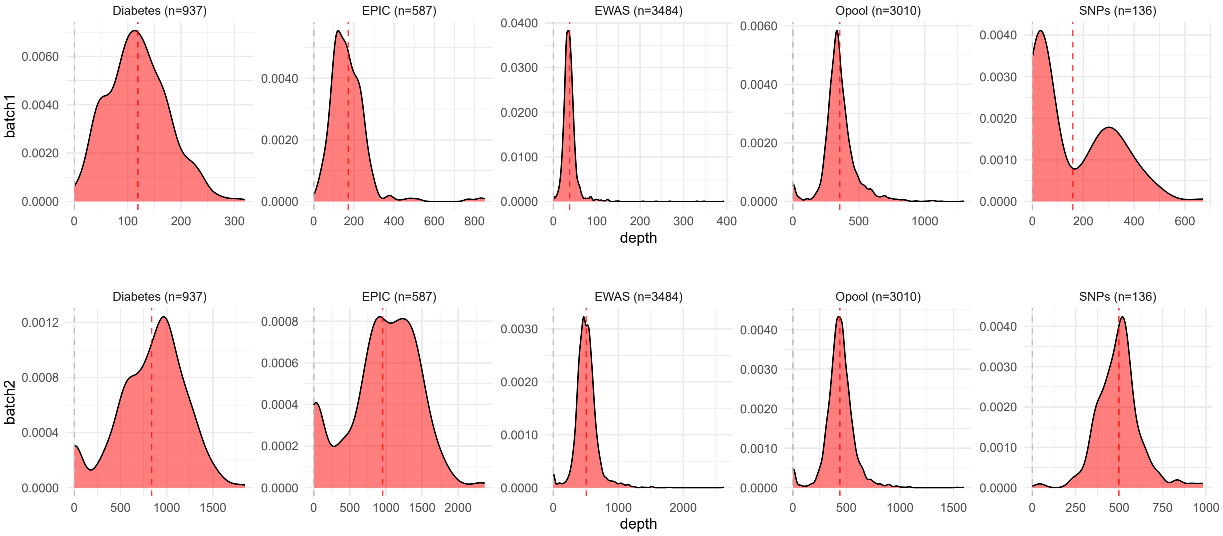


Figure 3.9: Depth distribution of targeted regions across probe sets in TBS. Density plots illustrate the depth distributions of different probe sets across two sample batches (upper panel: batch1, lower panel, batch2). Each subplot corresponds to a specific probe group—Diabetes, EPIC, EWAS, Opool, and SNPs—with the number of probes indicated in parentheses. Red dashed lines indicate the average depth for each probe set, with grey lines showing non-enriched background regions, underscoring the high efficiency of target enrichment achieved by targeted bisulfite sequencing (TBS).

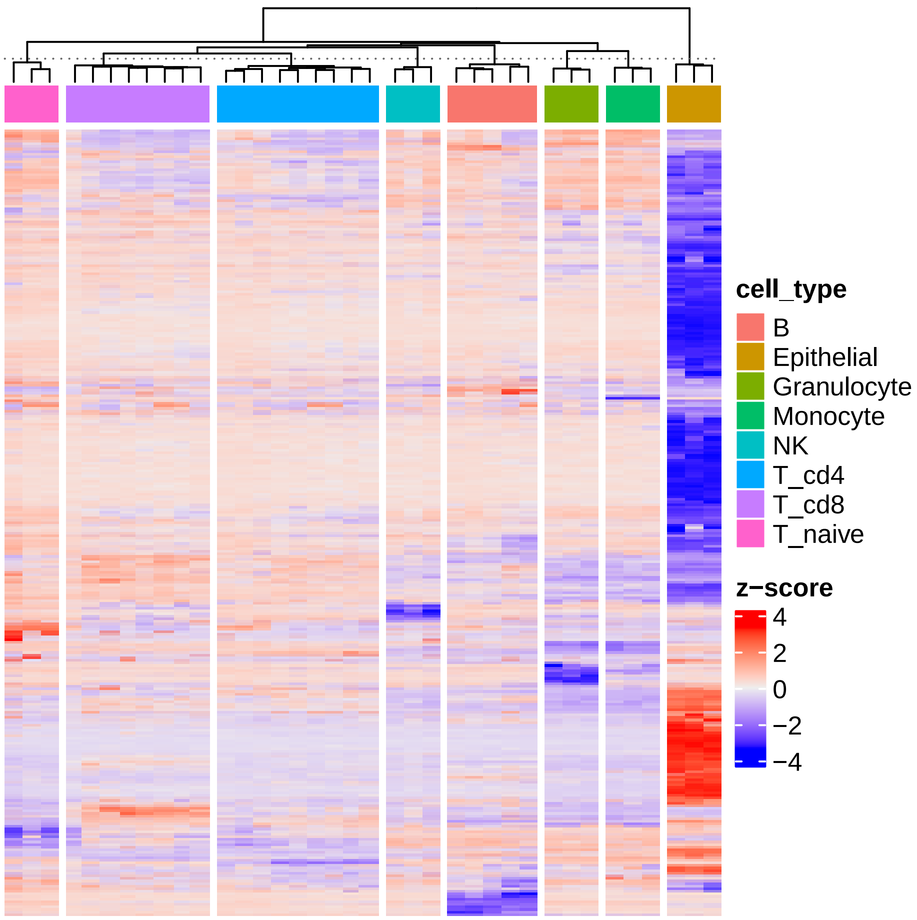


Figure 3.10: Cell-type specific methylation signatures at TBS sites. Heatmap showing the methylation profiles of various cell types, restricted to CpG sites that overlap with targeted bisulfite sequencing (TBS) data. Each column represents a sample from a specific cell type, with cell types indicated by the color bar at the top: B cells, epithelial cells, granulocytes, monocytes, NK cells, cd4+, cd8+, and naive T cells. The z-scores reflect relative methylation levels, with red indicating hypermethylation and blue indicating hypomethylation. The distinct clustering patterns in the heatmap confirm that TBS sites retain sufficient cell type identity information, allowing for a clear distinction between cell types. This demonstrates the efficacy of TBS in capturing cell type-specific epigenetic signatures, reinforcing its utility for studying cellular heterogeneity.

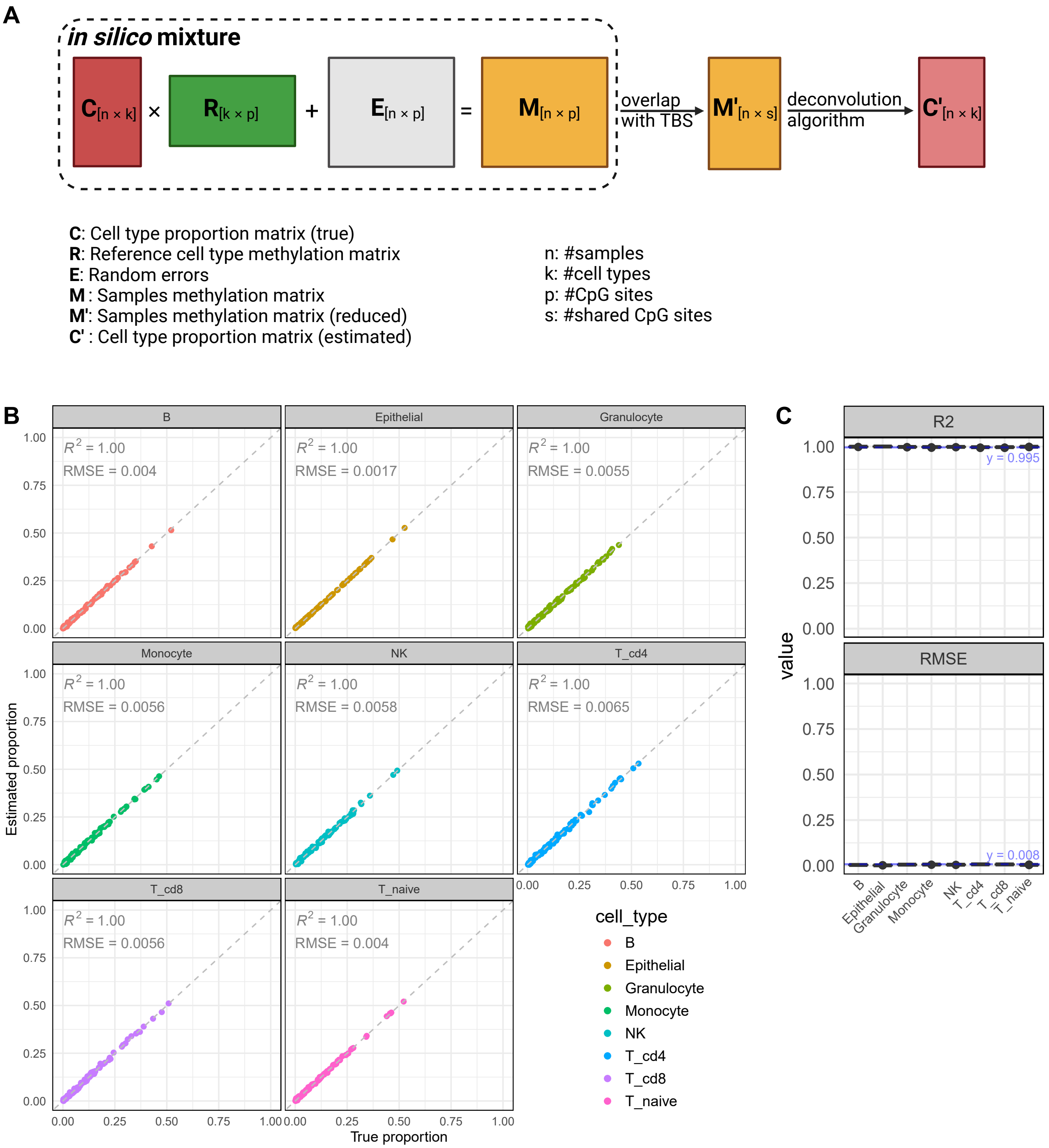


Figure 3.11: Simulated validation of TBS sites for accurate cell deconvolution. Simulation of cell deconvolution using targeted bisulfite sequencing (TBS) sites to confirm their support for accurate cell type estimation. (A) Schematic of the simulation workflow for cell deconvolution. The process begins with a true cell type proportion matrix, C, and a reference cell type methylation matrix, R, which are combined with random errors, E, to generate a simulated methylation matrix, M. This matrix is then refined to include only the sites overlapping with TBS data, creating a reduced methylation matrix, M'. A deconvolution algorithm is subsequently applied to estimate cell type proportions, C', from the reduced methylation matrix. The accuracy of the deconvolution is then evaluated by comparing these estimated proportions with the true proportions.

Figure 3.11: (B) Scatter plots showing the deconvolution accuracy across different cell types in a single simulated exaperiment, demonstrating high deconvolution accuracy. (C) Results from repeating the simulation 100 times, consistently showing high R^2 values and low RMSE across all cell types, confirming that TBS sites robustly support accurate cell type deconvolution. (R^2: coefficient of determination; RMSE: Root Mean Squared Error)

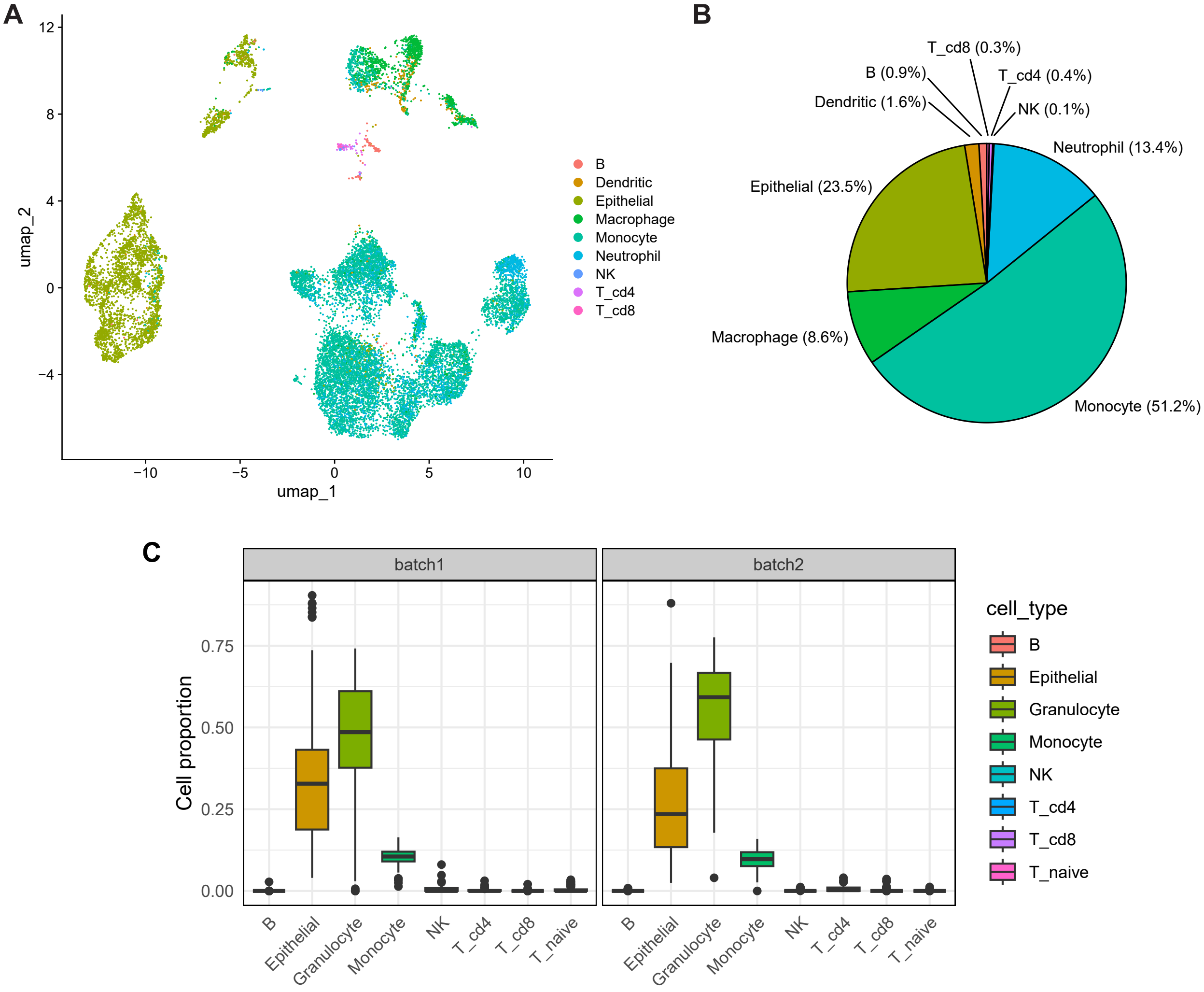


Figure 3.12: Cell type compositions in saliva: reanalysis of existing scRNA-seq dataset and TBS deconvolution results. (A-B) Cell type composition in saliva was revealed by the reanalysis of a previous single-cell RNA sequencing dataset from human sputum [53]. (A) UMAP plot displaying distinct clusters of cells, each colored according to its identified cell type. (B) Pie chart showing the abundance of cell type proportions, with Monocytes, Epithelial cells, and Neutrophils being the most abundant, followed by smaller populations of other immune cells. The reanalysis results validate the major cell types, confirming that immune cells and Epithelial cells are predominant in saliva samples. (C) Boxplots illustrating the cell type proportions in two batches (batch 1 and batch 2) derived from deconvolution analysis of bulk TBS data, highlighting the reproducibility across different batches. The deconvolution results show a similar pattern to the scRNA-seq findings, with Granulocytes, Monocytes, and Epithelial cells constituting most of the cell population. In contrast, other immune cells are present in lower proportions. The alignment between the scRNA-seq reanalysis and TBS data deconvolution results supports the reliability of deconvolution analysis. The differences in quantitative proportions may be attributed to inherent sample variation and technology biases, such as the scRNA-seq conducted on sputum from COVID-19 patients, which could have altered cell proportions and capture preferences.

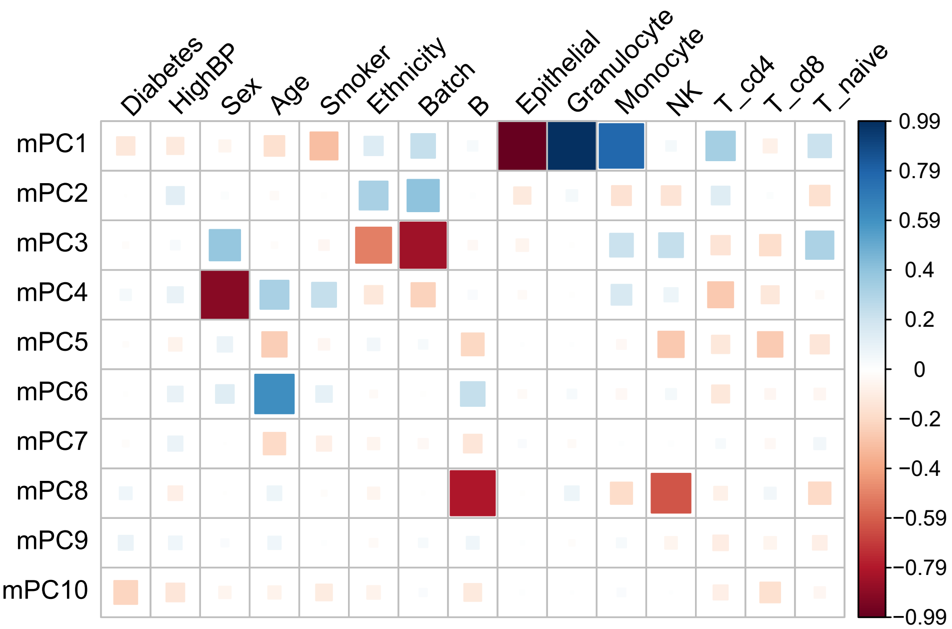


Figure 3.13: Correlation heatmap between methylation principal components (mPCs) and demographical and cellular variables. Heatmap illustrating the correlations between the top 10 methylation principal components (mPC1 to mPC10) and various demographical and cellular proportions. The color intensity and size of the squares represent the strength of the correlation, with blue indicating positive correlations and red indicating negative correlations, as shown by the color scale on the right. These correlations suggest that sex, age, ethnicity, and cell proportions are dominant factors of DNA methylation variations in the TBS data.

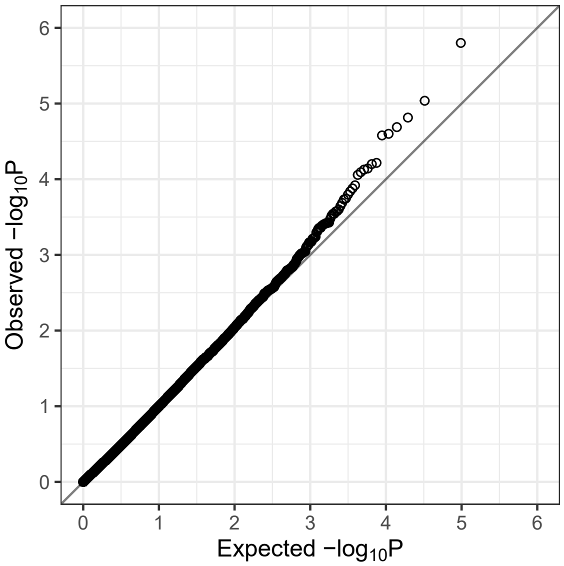


Figure 3.14: Quantile-Quantile (Q-Q) plot for EWAS analysis. The Q-Q plot compares observed -log10(p-values) from the EWAS with expected values under the null hypothesis. Points along the diagonal indicate concordance between observed and expected p-values, while deviations from the diagonal, particularly at the upper tail, suggest the presence of CpG sites with significant associations that exceed what would be expected by chance. The plot shows a slight deviation from the diagonal in the higher -log10(p-value) range, indicating the presence of true associations in the dataset.

3.8 References

[1] Zhou, B. et al. “Worldwide Trends in Diabetes since 1980: A Pooled Analysis of 751 Population-Based Studies with 4·4 Million Participants”. *The Lancet* 387.10027 (2016), pp. 1513–1530.

[2] Ong, K. L. et al. “Global, Regional, and National Burden of Diabetes from 1990 to 2021, with Projections of Prevalence to 2050: A Systematic Analysis for the Global Burden of Disease Study 2021”. *The Lancet* 402.10397 (2023), pp. 203–234.

[3] Centers for Disease Control and Prevention. *National Diabetes Statistics Report*. (accessed on August, 2024). URL: https://www.cdc.gov/diabetes/php/data-research/index.html.

[4] *Diabetes*. (accessed on August, 2024). URL: https://www.niddk.nih.gov/health-information/diabetes.

[5] Ahmad, E., Lim, S., Lamptey, R., Webb, D. R., and Davies, M. J. “Type 2 Diabetes”. *The Lancet* 400.10365 (2022), pp. 1803–1820.

[6] Wu, Y.-L., Lin, Z.-J., Li, C.-C., Lin, X., Shan, S.-K., Guo, B., Zheng, M.-H., Li, F., Yuan, L.-Q., and Li, Z.-h. “Epigenetic Regulation in Metabolic Diseases: Mechanisms and Advances in Clinical Study”. *Signal Transduction and Targeted Therapy* 8.1 (2023), pp. 1–27.

[7] Ling, C., Bacos, K., and Rönn, T. “Epigenetics of Type 2 Diabetes Mellitus and Weight Change — a Tool for Precision Medicine?” *Nature Reviews Endocrinology* 18.7 (2022), pp. 433–448.

[8] Ling, C. and Rönn, T. “Epigenetics in Human Obesity and Type 2 Diabetes”. *Cell Metabolism* 29.5 (2019), pp. 1028–1044.

[9] Greenberg, M. V. C. and Bourc'his, D. "The Diverse Roles of DNA Methylation in Mammalian Development and Disease". *Nature Reviews Molecular Cell Biology* 20.10 (2019), pp. 590–607.

[10] Smith, Z. D. and Meissner, A. "DNA Methylation: Roles in Mammalian Development". *Nature Reviews Genetics* 14.3 (2013), pp. 204–220.

[11] Stackpole, M. L. et al. "Cost-Effective Methylome Sequencing of Cell-Free DNA for Accurately Detecting and Locating Cancer". *Nature Communications* 13.1 (2022), p. 5566.

[12] Chambers, J. C. et al. "Epigenome-Wide Association of DNA Methylation Markers in Peripheral Blood from Indian Asians and Europeans with Incident Type 2 Diabetes: A Nested Case-Control Study". *The Lancet Diabetes & Endocrinology* 3.7 (2015), pp. 526–534.

[13] Toperoff, G. et al. "Genome-Wide Survey Reveals Predisposing Diabetes Type 2-Related DNA Methylation Variations in Human Peripheral Blood". *Human Molecular Genetics* 21.2 (2012), pp. 371–383.

[14] Gillberg, L. and Ling, C. "The Potential Use of DNA Methylation Biomarkers to Identify Risk and Progression of Type 2 Diabetes". *Frontiers in Endocrinology* 6 (2015), p. 43.

[15] Li, K. Y. et al. "DNA Methylation Markers for Kidney Function and Progression of Diabetic Kidney Disease". *Nature Communications* 14.1 (2023), p. 2543.

[16] Chen, Z. et al. "DNA Methylation Mediates Development of HbA1c-associated Complications in Type 1 Diabetes". *Nature Metabolism* 2.8 (2020), pp. 744–762.

[17] Christiansen, C. et al. "Enhanced Resolution Profiling in Twins Reveals Differential Methylation Signatures of Type 2 Diabetes with Links to Its Complications". *eBioMedicine* 103 (2024).

[18] Dhawan, S. and Natarajan, R. “Epigenetics and Type 2 Diabetes Risk”. *Current Diabetes Reports* 19.8 (2019), p. 47.

[19] Cheng, Y. et al. “Development and Validation of DNA Methylation Scores in Two European Cohorts Augment 10-Year Risk Prediction of Type 2 Diabetes”. *Nature Aging* 3.4 (2023), pp. 450–458.

[20] Baca, P. et al. “DNA Methylation and Gene Expression Analysis in Adipose Tissue to Identify New Loci Associated with T2D Development in Obesity”. *Nutrition & Diabetes* 12.1 (2022), pp. 1–7.

[21] Bacos, K. et al. “Blood-Based Biomarkers of Age-Associated Epigenetic Changes in Human Islets Associate with Insulin Secretion and Diabetes”. *Nature Communications* 7.1 (2016), p. 11089.

[22] Maude, H., Sanchez-Cabanillas, C., and Cebola, I. “Epigenetics of Hepatic Insulin Resistance”. *Frontiers in Endocrinology* 12 (2021).

[23] Ding, Q., Gao, Z., Chen, K., Zhang, Q., Hu, S., and Zhao, L. “Inflammation-Related Epigenetic Modification: The Bridge Between Immune and Metabolism in Type 2 Diabetes”. *Frontiers in Immunology* 13 (2022).

[24] Shao, B.-Y., Zhang, S.-F., Li, H.-D., Meng, X.-M., and Chen, H.-Y. “Epigenetics and Inflammation in Diabetic Nephropathy”. *Frontiers in Physiology* 12 (2021).

[25] Bansal, A. and Pinney, S. E. “DNA Methylation and Its Role in the Pathogenesis of Diabetes”. *Pediatric Diabetes* 18.3 (2017), pp. 167–177.

[26] Davegårdh, C., García-Calzón, S., Bacos, K., and Ling, C. “DNA Methylation in the Pathogenesis of Type 2 Diabetes in Humans”. *Molecular Metabolism* 14 (2018), pp. 12–25.

[27] Ahmed, S. A. H., Ansari, S. A., Mensah-Brown, E. P. K., and Emerald, B. S. “The Role of DNA Methylation in the Pathogenesis of Type 2 Diabetes Mellitus”. *Clinical Epigenetics* 12.1 (2020), p. 104.

[28] Nadiger, N., Veed, J. K., Chinya Nataraj, P., and Mukhopadhyay, A. “DNA Methylation and Type 2 Diabetes: A Systematic Review”. *Clinical Epigenetics* 16.1 (2024), p. 67.

[29] Orozco, L. D. et al. “Epigenome-Wide Association in Adipose Tissue from the METSIM Cohort”. *Human Molecular Genetics* 27.10 (2018), pp. 1830–1846.

[30] Braun, P. R. et al. “Genome-Wide DNA Methylation Comparison between Live Human Brain and Peripheral Tissues within Individuals”. *Translational Psychiatry* 9.1 (2019), p. 47.

[31] Nishitani, S. et al. “Cross-Tissue Correlations of Genome-Wide DNA Methylation in Japanese Live Human Brain and Blood, Saliva, and Buccal Epithelial Tissues”. *Translational Psychiatry* 13.1 (2023), p. 72.

[32] Fraszczyk, E. et al. “Epigenome-Wide Association Study of Incident Type 2 Diabetes: A Meta-Analysis of Five Prospective European Cohorts”. *Diabetologia* 65.5 (2022), pp. 763–776.

[33] Sehl, M. E., Guo, W., Farrell, C., Marino, N., Henry, J. E., Storniolo, A. M., Papp, J., Li, J. J., Horvath, S., Pellegrini, M., et al. “Systematic dissection of epigenetic age acceleration in normal breast tissue reveals its link to estrogen signaling and cancer risk”. *bioRxiv* (2024).

[34] Ziller, M. J. et al. “Charting a Dynamic DNA Methylation Landscape of the Human Genome”. *Nature* 500.7463 (2013), pp. 477–481.

[35] Moss, J. et al. “Comprehensive Human Cell-Type Methylation Atlas Reveals Origins of Circulating Cell-Free DNA in Health and Disease”. *Nature Communications* 9.1 (2018), p. 5068.

[36] Dayeh, T. et al. “Genome-Wide DNA Methylation Analysis of Human Pancreatic Islets from Type 2 Diabetic and Non-Diabetic Donors Identifies Candidate Genes That Influence Insulin Secretion”. *PLOS Genetics* 10.3 (2014), e1004160.

[37] DeFronzo, R. A. et al. “Type 2 Diabetes Mellitus”. *Nature Reviews Disease Primers* 1.1 (2015), pp. 1–22.

[38] Dayeh, T. et al. “DNA Methylation of Loci within ABCG1 and PHOSPHO1 in Blood DNA Is Associated with Future Type 2 Diabetes Risk”. *Epigenetics* 11.7 (2016), pp. 482–488.

[39] Chuang, Y.-H., Paul, K. C., Bronstein, J. M., Bordelon, Y., Horvath, S., and Ritz, B. “Parkinson’s Disease Is Associated with DNA Methylation Levels in Human Blood and Saliva”. *Genome Medicine* 9.1 (2017), p. 76.

[40] Duarte Folle, A., Paul, K. C., Bronstein, J. M., Keener, A. M., and Ritz, B. “Clinical Progression in Parkinson’s Disease with Features of REM Sleep Behavior Disorder: A Population-Based Longitudinal Study”. *Parkinsonism & Related Disorders* 62 (2019), pp. 105–111.

[41] Paul, K. C. et al. “A Pesticide and iPSC Dopaminergic Neuron Screen Identifies and Classifies Parkinson-relevant Pesticides”. *Nature Communications* 14.1 (2023), p. 2803.

[42] Morselli, M. et al. “In Vivo Targeting of de Novo DNA Methylation by Histone Modifications in Yeast and Mouse”. *eLife* 4 (2015). Ed. by Ren, B., e06205.

[43] *FastQC: A Quality Control Tool for High Throughput Sequence Data*. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[44] Chen, S., Zhou, Y., Chen, Y., and Gu, J. “Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor”. *Bioinformatics* 34.17 (2018), pp. i884–i890.

[45] Farrell, C., Thompson, M., Tosevska, A., Oyetunde, A., and Pellegrini, M. “BiSulfite Bolt: A bisulfite sequencing analysis platform”. *GigaScience* 10.5 (2021), giab033.

[46] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. “The Sequence Alignment/Map Format and SAMtools”. *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[47] Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., and Hoffmann, S. “Metilene: Fast and Sensitive Calling of Differentially Methylated Regions from Bisulfite Sequencing Data”. *Genome Research* 26.2 (2016), pp. 256–262.

[48] Gu, Z. and Hübschmann, D. “rGREAT: An R/Bioconductor Package for Functional Enrichment on Genomic Regions”. *Bioinformatics* 39.1 (2023), btac745.

[49] Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. “Simple Combinations of Lineage-Determining Transcription Factors Prime *Cis*-Regulatory Elements Required for Macrophage and B Cell Identities”. *Molecular Cell* 38.4 (2010), pp. 576–589.

[50] Morselli, M., Farrell, C., Montoya, D., Gören, T., Sabırlı, R., Türkçüer, İ., Kurt, Ö., Köseler, A., and Pellegrini, M. “DNA Methylation Profiles in Pneumonia Patients Reflect Changes in Cell Types and Pneumonia Severity”. *Epigenetics* 17.12 (2022), pp. 1646–1660.

[51] Protti, G., Rubbi, L., Gören, T., Sabirli, R., Civlan, S., Kurt, Ö., Türkçüer, İ., Köseler, A., and Pellegrini, M. “The Methylome of Buccal Epithelial Cells Is Influenced by Age, Sex, and Physiological Properties”. *Physiological Genomics* 55.12 (2023), pp. 618–633.

[52] *Impute*. URL: http://bioconductor.org/packages/impute/.

[53] Ren, X. et al. “COVID-19 Immune Features Revealed by a Large-Scale Single-Cell Transcriptome Atlas”. *Cell* 184.7 (2021), 1895–1913.e19.

[54] Loyfer, N. et al. “A DNA Methylation Atlas of Normal Human Cell Types”. *Nature* 613.7943 (2023), pp. 355–364.

[55] Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. “DNA Methylation Arrays as Surrogate Measures of Cell Mixture Distribution”. *BMC Bioinformatics* 13.1 (2012), p. 86.

[56] Park, Y. and Wu, H. "Differential Methylation Analysis for BS-seq Data under General Experimental Design". *Bioinformatics* 32.10 (2016), pp. 1446–1453.

[57] Guo, W., Morselli, M., Paul, K. C., Thompson, M., Ritz, B., and Pellegrini, M. "Type-2 diabetes biomarker discovery and risk assessment through saliva DNA methylome". *medRxiv* (2024).

[58] Sabatini, P. V. et al. "Neuronal PAS Domain Protein 4 Suppression of Oxygen Sensing Optimizes Metabolism during Excitation of Neuroendocrine Cells". *Cell Reports* 22.1 (2018), pp. 163–174.

[59] Luo, X., Ge, J., Chen, T., Liu, J., Liu, Z., Bi, C., and Lan, S. "LHX9, a P53-Binding Protein, Inhibits the Progression of Glioma by Suppressing Glycolysis". *Aging (Albany NY)* 13.18 (2021), pp. 22109–22119.

[60] Renoux, F. et al. "The AP1 Transcription Factor Fosl2 Promotes Systemic Autoimmunity and Inflammation by Repressing Treg Development". *Cell Reports* 31.13 (2020), p. 107826.

[61] Battram, T. et al. "The EWAS Catalog: A Database of Epigenome-Wide Association Studies". *Wellcome Open Research* 7 (2022), p. 41.

[62] Li, M. et al. "EWAS Atlas: A Curated Knowledgebase of Epigenome-Wide Association Studies". *Nucleic Acids Research* 47.D1 (2019), pp. D983–D988.

[63] Jaffe, A. E. and Irizarry, R. A. "Accounting for Cellular Heterogeneity Is Critical in Epigenome-Wide Association Studies". *Genome Biology* 15.2 (2014), R31.

[64] Vujkovic, M. et al. "Discovery of 318 New Risk Loci for Type 2 Diabetes and Related Vascular Outcomes among 1.4 Million Participants in a Multi-Ancestry Meta-Analysis". *Nature Genetics* 52.7 (2020), pp. 680–691.

[65] Chandra, V. et al. "The Type 1 Diabetes Gene TYK2 Regulates β-Cell Development and Its Responses to Interferon-α". *Nature Communications* 13.1 (2022), p. 6363.

[66] Kennedy, M. A., Barrera, G. C., Nakamura, K., Baldán, Á., Tarr, P., Fishbein, M. C., Frank, J., Francone, O. L., and Edwards, P. A. “ABCG1 Has a Critical Role in Mediating Cholesterol Efflux to HDL and Preventing Cellular Lipid Accumulation”. *Cell Metabolism* 1.2 (2005), pp. 121–131.

[67] Bornfeldt, K. E. and Tabas, I. “Insulin Resistance, Hyperglycemia, and Atherosclerosis”. *Cell metabolism* 14.5 (2011), pp. 575–585.

[68] Jenkins, A. J., Grant, M. B., and Busik, J. V. “Lipids, Hyperreflective Crystalline Deposits and Diabetic Retinopathy: Potential Systemic and Retinal-Specific Effect of Lipid-Lowering Therapies”. *Diabetologia* 65.4 (2022), pp. 587–603.

[69] Tu, X., Hutter, J.-C., Wang, Z. J., Kudo, T., Regev, A., and Lopez, R. “A Supervised Contrastive Framework for Learning Disentangled Representations of Cell Perturbation Data”. *bioRxiv* (2024), pp. 2024–01.

[70] Abid, A. and Zou, J. “Contrastive variational autoencoder enhances salient features”. *arXiv preprint arXiv:1902.04601* (2019).

[71] Griffin, P. T. et al. “TIME-seq Reduces Time and Cost of DNA Methylation Measurement for Epigenetic Clock Construction”. *Nature Aging* 4.2 (2024), pp. 261–274.

CHAPTER 4

Systematic dissection of epigenetic age acceleration in normal breast tissue reveals its link to estrogen signaling and cancer risk

Abstract

Breast aging encompasses intricate molecular and cellular changes that elevate cancer risk. Our study profiled DNA methylation and gene expression of 181 normal breast samples and systematically evaluated eight epigenetic clocks. We found that clocks trained using breast tissues demonstrate improved age prediction in normal breast tissue, and bias universally exists in epigenetic clocks, necessitating a proper definition of age acceleration. Cell composition analysis revealed significant age-related alterations and highlighted its distinct associations with age acceleration, including increased luminal epithelial and myoepithelial cells and reduced adipocytes and immune cells, connecting age acceleration to carcinogenesis from a cell compositional perspective. Additionally, CpG sites associated with age acceleration were enriched for estrogen receptor binding sites, providing a mechanistic link between estrogen exposure, accelerated aging, and cancer. These findings highlight the importance of cellular heterogeneity in epigenetic age estimates and the potential of age acceleration to guide risk stratification and prevention strategies.

Key words: Breast aging; epigenetic clocks; age acceleration; DNA methylation; tissue heterogeneity; cell composition; estrogen receptor; breast cancer risk.

4.1 Introduction

Aging is an intricate biological process accompanied by progressive molecular, cellular, and tissue-level changes, leading to functional decline and increased disease susceptibility [1]. In breast tissue, aging induces a cascade of molecular alterations such as DNA damage accumulation [2], telomere shortening [3] and epigenetic modifications [4]. At the cellular level, these changes manifest as senescence [5], altered cell composition and disrupted tissue architecture [6]. Specifically, breast aging is associated with the accumulation of dysfunctional luminal epithelial cells [7], a reduction in hormone-sensitive cells [8], and an increase in immune cell infiltration along with a decline in adaptive immune cells [9]. These alternations contribute to genome instability and a pro-inflammatory environment with compromised immune surveillance, heightening the risk of malignant transformation and breast cancer development. Understanding these aging-related changes is essential for developing strategies to alleviate the adverse effects of aging on breast health and prevent breast cancer.

The concept of epigenetic clock has emerged as a powerful tool for studying aging and elucidating its molecular mechanisms. Various epigenetic clocks have been developed over the years [10–16], each leveraging DNA methylation patterns at specific genomic loci to accurately measure biological age. Previous research has shown significant associations between epigenetic age and a wide range of age-related diseases, including breast cancer [17]. Beyond reflecting biological age, deviations in epigenetic age from a normal trajectory—known as age acceleration or deceleration—have garnered considerable attention due to their profound implications for diverse health outcomes [18]. In the breast aging field, pioneering work has established the link between accelerated epigenetic aging with lifetime estrogen exposure [19] and increased breast cancer risk [20]. Despite these advancements, a comprehensive understanding of breast age acceleration, particularly its mechanistic link with estrogen exposure and cancer risk from molecular and cellular perspectives, remains elusive.

A critical gap in current research is the lack of consideration for tissue heterogeneity. Breast tissue consists of diverse cell types, each potentially aging at different rates and in distinct ways, while most existing epigenetic clocks are built with bulk DNA methylation mainly using blood samples [21], failing to account for the complexity and variability of breast tissue. This raises questions about the accuracy of these clocks in predicting breast biological age and poses additional challenges on understanding breast-specific aging processes. Additionally, most epigenetic aging studies overlook the complex cell compositional changes with age [22], focusing instead on bulk tissue analysis, which can obscure insights into the dynamics of specific cell types, their distinct relationships with epigenetic age and age acceleration, and their implications for cancer risks. Detailed investigations and characterizations of breast age acceleration from cellular, epigenetic, and transcriptomic perspectives are needed to elucidate breast aging processes, providing insights into potential intervention targets to promote breast health.

To address these gaps, we profiled and analyzed the DNA methylation and gene expression of 181 normal breast samples aged 19 to 90 (Figure 4.1A-B). By deconvolving the cell type abundance of breast tissue, we comprehensively studied the epigenetic aging at both the molecular and cellular levels. First, we systematically evaluated the predictive accuracy of eight epigenetic clocks in normal breast tissue, including two pan-tissue, two breast-specific, two second-generation, and two first-generation clocks. Our findings show that pan-tissue and breast-specific clocks accurately predicted age, while blood-based clocks underperformed due to tissue heterogeneity. We also revealed systematic biases in epigenetic clocks for age prediction and advocate for a proper definition of age acceleration to avoid confounding by chronological age. Cell type abundance analysis indicated significant age-related changes in breast tissue composition, including increased adipocytes and vascular endothelial cells, and decreased luminal epithelial and basal myoepithelial cells. Notably, epigenetic age acceleration was associated with distinct cellular changes, suggesting a higher risk of breast carcinogenesis from a cellular population perspective. We also identified CpG sites associated with age acceleration, enriched for estrogen receptor binding sites (ESR1), linking estrogen exposure to accelerated breast aging and increased cancer risk from the molecular level. Transcriptome analysis revealed differentially expressed genes associated with age acceleration for each clock, while the overlap among different clocks is minimal and the pathway enrichments reflect unique biological signals captured by different clocks. Overall, our findings emphasize the need for considering cellular and tissue heterogeneity in epigenetic aging studies, providing valuable insights into the molecular mechanisms linking estrogen exposure, epigenetic aging, and breast cancer risk. Future research should focus on refining epigenetic clocks and exploring their clinical applications in breast aging and cancer risk assessment.

4.2 Methods

4.2.1 Study samples and specimens

We utilized breast tissue specimens from the Susan G. Komen Tissue Bank (KTB) at the Indiana University Simon Comprehensive Cancer Center, a unique repository of samples from healthy female donors. Each tissue sample is well annotated with the donor's race and ethnicity, height, weight, family history, reproductive history, and medication use. All participants in the study have provided informed consent, and the study received approval from the UCLA Institutional Review Board. Data for this study were collected as part of a cross-sectional study designed to investigate the associations between breast epigenetic age and hormonal factors. We initially recruited 200 female participants, categorized into four groups: (i) premenopausal and nulliparous, (ii) premenopausal and with at least 1live birth, (iii) postmenopausal and nulliparous, and (iv) postmenopausal and with at least one live birth. Each donor underwent six core biopsies from the upper outer quadrant of one breast under local anesthesia. Within five minutes of collection, one biopsy was immediately placed into an embedding cassette and stored in 10% buffered formalin at room temperature before being embedded in paraffin. The remaining five biopsies were flash frozen in liquid nitrogen, then placed in labeled, chilled cryovials, and stored in liquid nitrogen, as described in our previous study [3, 19]

4.2.2 DNA and RNA extraction

Breast tissue samples, each weighing 50 milligrams, were shipped to the Neurogenetics Core Sequencing Laboratory (UNGC) at UCLA for DNA methylation and transcriptome profiling. Frozen tissue (30 mg) was lysed in 600 μL of guanidine-isothiocyanate-containing Buffer RLT Plus in a 2.0 mL microcentrifuge tube and homogenized using TissueLyser II (Qiagen) with 5 mm stainless steel beads. The tissue lysate was then processed following the AllPrep protocol (Qiagen, catalog no. 80224) to simultaneously extract genomic DNA and total RNA, utilizing RNeasy Mini spin column technology. Extracted DNA underwent bisulfite conversion for methylation quantification, while RNA was used for transcriptome quantification and analyses.

4.2.3 DNA methylation quantification and processing

DNA Methylation for each sample was measured using the Human Methylation EPIC (850K) array BeadChip (Illumina). 500 nanograms of DNA was bisulfite-converted using the EZ Methylation Kit (Zymo Research). Following bisulfite conversion, the DNA was hybridized to the EPIC array probes. Fluorescence data from the hybridized chip were scanned on an iScan (Illumina), where the methylated intensity (M_i) and unmethylated intensity (U_i) for each CpG site i were measured. Probe quality control and data processing were conducted using R package minfi (version 1.48.0). Specifically, we employed the `processIllumina()` function to perform background subtraction and control normalization, and calculated DNA methylation level (beta-value) for each CpG site based on the intensity ratio between methylated and an unmethylated signal using the following formula

$$\beta_i = \frac{\max(M_i, 0)}{\max(M_i, 0) + \max(U_i, 0) + \alpha} \tag{4.1}$$

where α (default 100) is an offset to regularize beta value when both methylated and unmethylated probe intensities are low. By definition, the beta values range between 0 and 1, with 0 indicating completely unmethylated, and values approaching 1 indicating completely methylated.

4.2.4 Bulk RNA sequencing and processing

Transcriptome profiling was performed using the Lexogen QuantSeq 3' mRNA-Seq FWD kit to generate RNA sequencing libraries. Sequencing was conducted with 65 bp single-end reads on an Illumina HiSeq 4000. The raw sequencing data underwent quality control using FastQC [23] (version 0.11.9). Adapters and low-quality bases were trimmed using fastp [24] (version 0.23.2). Trimmed reads were subsequently aligned to the human reference genome (GRCh38) with Ensembl annotation file (v84) using STAR [25] (version 2.7.9a). Gene expression counts for each sample were obtained using HTSeq [26] (version 1.99.2) and merged into a gene expression count matrix using in-house scripts. Genes with no more than 10 counts in less than 10% of samples were excluded from further analysis. Finally, the gene expression count per million (CPM) values were calculated using the `cpm()` function implemented in the edgeR package [27] (version 3.33.5).

4.2.5 Sample quality control

Principal component analysis (PCA) was conducted on both the DNA methylation matrix (beta values) and the gene expression matrix (log-transformed CPM values) to identify potential sample outliers. After removing outliers and excluding samples from participants with breast cancer, we retained 181 normal breast tissue samples with paired DNA methylation and gene expression data. All downstream analyses were based on these 181 samples. The demographic and clinical characteristics of the finalized study cohort are detailed in Table 4.1 and Supplementary Data 1.

4.2.6 Epigenetic age and age acceleration calculation

To comprehensively assess the age prediction accuracy of DNA methylation (DNAm) epigenetic clocks in normal breast tissue, we included two pan-tissue clocks: Horvath's pan-tissue clock [10] and AltumAge [16]. These clocks are designed to operate across various tissue and cell types, including breast. Additionally, we examined two second-generation clocks, GrimAge [14], and Phenotypic Age [12], tailored to predict overall health span and lifespan more effectively, particularly in blood samples. GrimAge is of particular interest due to its potential to predict cancer onset and its association with menopausal age. Our analysis also included two first-generation clocks, the Hannum clock [11] and the Skin&Blood Clock [13], representing previous efforts for biological age prediction. The methodological details and applications of these clocks are further illustrated in Figure 4.1C, providing a clear summary of their distinct characteristics and diverse biological contexts. DNA methylation beta values were used to calculate the epigenetic age. Epigenetic age for Horvath, Hannum, GrimAge, Phenotypic Age, Skin&Blood was calculated using DNA methylation Age Calculator (https://dnamage.genetics.ucla.edu/home). AltumAge was calculated according to its official tutorial (https://github.com/rsinghlab/AltumAge).

Additionally, we trained two breast-specific clocks using the KTB DNA methylation data. One employs the Elastic Net algorithm [28] to predict age based on methylation levels, and the other utilizes Epigenetic Pacemaker (EPM) model [15], which uses inverse regression to derive age estimates from DNA methylation patterns. To avoid data leakage, we implemented a nested cross-validation strategy: The outer loop used a leave-one-out approach to split the data into training and testing sets, while the inner loop employed a 5-fold cross-validation to tune the hyperparameters and train the epigenetic clocks. This strategy ensured that test data was never seen during model training. The trained model was then used to predict the age of each test data point, iterating through each sample to collect age predictions for all samples.

The relationship between chronological age and epigenetic age of different clocks was assessed using a generalized additive model with a cubic spline; we confirmed that the trend is predominantly linear on our data (Figure 4.8A) and thus computed age acceleration by obtaining the residuals from a linear regression of epigenetic age on chronological age. This residual measure, designed to be age-adjusted, showed no correlation with chronological age (Figure 4.9B). Despite concerns about sampling uncertainty affecting the residual calculations, we used bootstrapping to confirm that age acceleration of each subject remains stable against variations in sample composition for each clock (Figure 4.9C). The robust measurement underscores the reliability of age acceleration and our findings.

4.2.7 Cell type deconvolution and immune enrichment score calculation

To quantify the cell type abundance in normal breast tissue samples, we utilized the Genotype-Tissue Expression (GTEx) v8 single-nucleus RNA-seq (snRNA-seq) data [**eraslan2022single**] and extracted gene expression data specific to normal breast tissue. The snRNA-seq breast dataset identified eight major cell types: adipocytes, luminal epithelial cells, basal myoepithelial cells, vascular endothelial cells, lymphatic endothelial cells, immune cells (dendritic cells/macrophages), pericytes & smooth muscle cells, and fibroblasts (Figure 4.10A-B). Cell doublets and genes expressed in fewer than 10 cells were excluded from the snRNA-seq data. The processed expression count matrix was converted into counts per million (cpm) values. Using this data as a reference, we constructed a cell type signature matrix and deconvolved the bulk RNA-seq data to determine cell type abundance proportions for each sample using CIBERSORTx [29] with batch correction mode (S-mode).

To achieve higher resolution of immune cell composition in normal breast tissue, we used ImmuneCellAI [30](`http://bioinfo.life.hust.edu.cn/ImmuCellAI/`) to calculate enrichment scores for 24 immune cell types and immune infiltration scores for each sample based on gene expression data. All software parameters were set as default unless otherwise specified.

4.2.8 Cancer risk score calculation

In addition to the Gail and Tyrer-Cuzick breast cancer risk measurements, which were computed in previous work using demographic, reproductive and family history data [3], we expanded our analysis to include cancer risk scores derived from molecular data in this analysis. Specifically, we utilized the code in epiTOC2 [31] to compute cancer risk scores for each sample based on the DNA methylation matrix. We also calculated cell senescence scores using gene expression data by applying single-sample Gene Set Enrichment Analysis (ssGSEA) with senescence signature genes from the CellAge database [32], implemented through the GSVA package [33] (version 1.50.5). By integrating these molecular-based risk scores with traditional risk assessments, we aimed to provide a more holistic view of breast cancer risk. The processed data, including epigenetic age estimates, age acceleration of each clock, cell proportions, immune enrichment scores, and cancer risk scores can be found in Supplementary Data 2.

4.2.9 Correlation and mediation analysis

We computed the pairwise Pearson correlation coefficients among various variables of interest. These include demographic variables, reproductive history, breast cancer risk and senescence scores, epigenetic age from eight different clocks, age acceleration (both raw and adjusted by cell proportions), the top 10 principal components from DNA methylation and gene expression, eight cell type proportions, 24 immune cell scores, and immune infiltration scores. The comprehensive correlation matrix is provided in Supplementary Data 3 and visualized in Figure 4.12.

To examine how changes in cell type proportions mediate the relationship between chronological age and epigenetic age, we conducted a mediation analysis. In this analysis, chronological age served as the independent variable (IV), epigenetic age as the dependent variable (DV), and each cell type abundance as mediators. The structural equation model used for this analysis is detailed in the Supplementary Data 4. We utilized the R package lavaan [34] (version 0.6-18) to perform the mediation analysis for each clock, estimating both direct and indirect effects and the contribution of each mediator.

4.2.10 Epigenome-wide association study and genomic region enrichment analysis

We conducted an Epigenome-Wide Association Study (EWAS) to identify CpG sites associated with age acceleration, controlling for chronological age and cell proportions. The EWAS analysis was performed using GEMMA [35] (version 0.98.6), where the age acceleration for each clock is the quantitative trait, and the CpG sites are the markers, with age and cell proportions as covariates. Since the cell proportions sum up to 1, we excluded Pericyte/SMC from the covariates to avoid the model identifiability issue. The association between each CpG site and age acceleration was then tested using a linear mixed model framework in GEMMA, with p-values calculated via likelihood ratio tests. CpG sites with q-values smaller than 0.05 were considered significantly associated (Supplementary Data 5) and were subsequently subjected to genomic region enrichment analysis using LOLA [36] (version 1.32.0), with all CpG sites serving as the background (Supplementary Data 6). The top 10 most significantly enriched Transcription Factor Binding Sites (TFBS) were extracted and visualized in Figure 4.5.

4.2.11 Differential expression and gene-set enrichment analysis

To further explore the relationship between gene expression and age acceleration for each clock, we conducted a differential expression (DE) analysis. First, we converted the gene expression matrix (CPM) for each gene by applying an inverse Gaussian transformation, ensuring that each gene's expression value Y_i follows a Gaussian distribution. Next, we performed DE analysis using a linear regression framework with age and age acceleration (AA) as covariates, both before (1) and after (2) adjusting for cell proportions:

$$Y_i \sim AA + Age \tag{4.2}$$

$$Y_i \sim AA + Age + Cell_1 + \cdots + Cell_{(k-1)} \tag{4.3}$$

$$\tag{4.4}$$

Where AA is the age acceleration of a specific clock, and k=8 is the total number of major cell types. Since the cell proportions sum up to 1, we excluded Pericyte/SMC when fitting the second model to avoid model identifiability issue. We implemented these models using `lm()` function in R. For both models, we tested whether AA is associated with the response Y_i for each gene. The association p-values were obtained using t-tests (Supplementary Data 7).

Genes with an adjusted p-value less than 0.05 in model (2) were defined as differentially expressed for each clock and subjected to Gene Ontology (GO) overrepresentation enrichment analysis. Additionally, we performed Gene Set Enrichment Analysis (GSEA) using the differential analysis results. Both GO overrepresentation and GSEA analyses were conducted and visualized using clusterProfiler [37] (version 4.10.1). The gene expression enrichment analysis results are provided in Supplementary Data 8.

4.2.12 Statistical analysis

Besides the statistical analysis described above, the pairwise Pearson correlation coefficient and significance (p-values) are calculated using R package Hmisc [38] (version 5.1-2). Correlation heatmaps displayed the Pearson correlation coefficients using the R package corrplot [39] (version 0.92). To account for multiple hypothesis testing, we reported the adjusted p-values using the Benjamini-Hochberg [40] procedure implemented by the `p.adjust()` function in R. All statistical tests were conducted in R (version 4.3).

4.2.13 Data availability

Raw and supplementary data can be obtained from [41].

4.3 Results

4.3.1 Age prediction accuracy of epigenetic clocks in normal breast tissue

Evaluating the age prediction accuracy in healthy breast tissue using eight epigenetic clocks (Figure 4.1C, Methods), we observed strong correlations between epigenetic age and chronological age across most clocks, except for Phenotypic Age (Figure 4.2, Table 4.2). Significant correlations among different epigenetic age estimates were also noted (Figure 4.9A). Notably, the pan-tissue and breast-specific clocks demonstrated high accuracy in predicting age in breast tissue, with the ElasticNet clock trained on Komen Tissue Bank (KTB) breast tissue showing the highest accuracy (Pearson correlation coefficient (Corr.) = 0.94, mean absolute error (MAE) = 3.33 years). Other clocks such as HO (Corr. = 0.90, MAE = 7.92 years), Altum (Corr. = 0.88, MAE = 5.24 years), and EPM (Corr. = 0.78, MAE = 6.56 years) also performed well.

Conversely, blood-based (first- and second-generation) clocks that did not include breast tissue in their original training set showed significantly lower correlations or higher MAEs, indicating reduced predictive accuracy. An exception was GrimAge, which includes chronological age as a predictor and thus showed a high correlation of Corr.= 0.94, similar to the best-performing breast-specific clocks, but with a much higher MAE (11.96 years). These findings suggest that blood-based clocks tend to underperform for predicting age in breast tissue, likely due to tissue heterogeneity. This underscores the importance of using tissue-specific or pan-tissue clocks to minimize cross-tissue biases and improve age prediction accuracy.

Our analysis also revealed that the regression line between predicted (epigenetic) age and chronological age consistently exhibited a positive intercept and a slope less than 1, regardless of the clock used. The large positive intercepts of first- and second-generation clocks, indicating predicted age in breast tissue at chronological age zero, might suggest that breast tissue appears "older" compared to other tissues such as blood [22]. However, the slope being less than 1 raises concerns about underestimating age in older subjects, highlighting a systematic bias in the epigenetic clock models. This issue will be further discussed in the following section.

4.3.2 Systematic biases in epigenetic clocks and justification of age acceleration

To elucidate the observed systematic biases in age estimation stemming from model bias, we simulated a scenario where DNA methylation can fully explain the variance in chronological age. Using penalized regression techniques (LASSO, Ridge, and Elastic Net regression) to construct the clocks, we consistently found that the regression slope of predicted age on chronological age was below 1 (Figure 4.7A). This occurs because penalized regression methods, used to manage the high-dimensional challenge where features (CpG sites) outnumber samples, shrink regression coefficients towards zero. As a result, predicted age (epigenetic age) rarely matches chronological age exactly and tends to have reduced variance compared to chronological age. Consequently, the prediction error (predicted age - chronological age) correlates with chronological age (Figure 4.7C).

These biases complicate the definition of age acceleration. Throughout the literature, we recognized various studies have used different definitions, with no consensus reached. Some studies define the age acceleration as the difference between epigenetic age and chronological age (difference definition), while others define it as the deviation of epigenetic age from its expected value given chronological age (residual definition). A slope smaller than 1 indicates that age difference is inversely associated with chronological age (Figure 4.7C, Figure 4.8C), leading to spurious associations between health outcomes and age differences due to confounding by chronological age. In contrast, the residual definition represents the part of epigenetic age unexplained by chronological age and is orthogonal to chronological age (Figure 4.7B, Figure 4.8B), eliminating age confounding concerns. In this study, we advocate against the use of age difference definition since it can easily lead to false discoveries when not calibrated with chronological age; And we defined the age acceleration as the regression residual between epigenetic age and chronological age.

4.3.3 Age and other demographic related changes in breast cell composition

We next investigated how cell type abundance changes with advancing age in breast tissue. Using CIBERSORTx [29] with GTEx normal breast snRNA-seq data [42] as a reference, we deconvolved the cell type abundance from bulk gene expression data for each sample and observed significant age-related changes in breast tissue composition (Figure 4.3A-B). Specifically, advancing chronological age is associated with a notable increase in the proportion of adipocytes ($p<4.12e-8$, adjusted $p<1.32e-06$) and vascular endothelial cells ($p<1.02e-4$, adjusted $p<1.21e-03$), and a significant decrease in proportions of luminal epithelial cells ($p<9.1e-11$, adjusted $p<5.83e-09$) and basal myoepithelial cells ($p<1.13e-4$, adjusted $p<1.21e-03$). Additionally, the proportion of immune dendritic cells/macrophages significantly increased ($p<1.93e-6$, adjusted $p<4.11e-05$), indicating an elevated inflammatory landscape with aging.

We also examined the associations between cell type abundance and various demographic and reproductive variables (Supplementary Data 3). Hispanic ethnicity was associated with a lower proportion of fibroblasts ($p<0.017$). Higher body mass index (BMI) correlated with increasing proportions of immune dendritic cells/macrophages ($p<0.014$) and decreased luminal epithelial cells ($p<0.033$). Tobacco smoking history was linked to a reduced proportion of pericytes/smooth muscle cells ($p<0.034$), though these associations did not remain significant after p-value adjustment. Interestingly, adipocyte proportion did not correlate with BMI (Figure 4.12), suggesting that an individual BMI is primarily related to overall adiposity rather than the proportion of adipocytes in breast tissue. There were no significant associations between cell proportions and age at menarche, parity, or history of breastfeeding.

4.3.4 Changes in breast cell composition with breast epigenetic age

Given that current epigenetic clocks primarily focus on bulk tissues with surrogate measures of DNA methylation, it is intuitive that changes in cell composition could influence epigenetic age. We investigated how cell type proportions correlate with epigenetic age for each clock, using chronological age as a reference (Figure 4.3C). Our analysis revealed significant associations between epigenetic age estimates and at least one cell type proportion for almost all clocks. The correlation patterns generally mirrored those observed with chronological age, except for the Phenotypic Age, Hannum clock, and Skin&Blood clock, which did not predict age accurately in breast tissue.

Specifically, adipocyte proportions displayed negative correlations with epigenetic ages across multiple clocks. Vascular endothelial cells consistently correlated positively with epigenetic age, indicating vascular remodeling. Luminal epithelial and basal myoepithelial cells exhibited strong negative correlations, reflecting a decline in breast density with epigenetic aging. The positive correlation between immune dendritic cells/macrophages and epigenetic age highlights an increased inflammatory environment. These findings underscore the complexity of epigenetic aging, validating the biological relevance of epigenetic clocks in capturing aging-related cellular changes, and emphasize the importance of considering these changes when interpreting epigenetic age estimates.

Beyond the major cell types, we also examined the correlation between epigenetic age and immune cell enrichment scores calculated by immuneCellAI [30] using gene expression data. We found that macrophage scores positively correlated with both chronological and epigenetic age, while gamma-delta T ($\gamma\delta$T) cell scores displayed a negative correlation (Figure 4.11A), both well aligning with previous observations [9, 43]. The increased macrophage enrichment scores highlight an elevated inflammatory status, while decreased $\gamma\delta$T cell enrichment scores potentially reflect a decline in maintenance of tissue homeostasis [43]. Together, these findings indicate altered immune surveillance dynamics during aging in the normal breast microenvironment.

4.3.5 Changes in breast cell composition with breast age acceleration

We next investigated whether epigenetic age accelerations are associated with cell compositional changes in breast tissue. Figure 4.3D illustrates the correlations between epigenetic age acceleration and cell abundance for each clock measure and cell type. We first noted the correlation variability among different clocks: Breast-specific clocks, optimized for predicting chronological age, showed the least correlation with cell abundance. In contrast, other clocks demonstrated significant associations with cell compositional changes. Specifically, blood-based clocks (both first- and second-generation) showed significant correlations with immune cell enrichment scores after p-value adjustment (Figure 4.11B). Notably, the pan-tissue clock AltumAge exhibited greater sensitivity to cellular changes compared to the Horvath clock, likely due to it used higher number (70 times more) of CpG sites for age prediction.

At the cell-type level, interestingly, we observed that adipocyte and vascular endothelial cell proportions negatively correlated with epigenetic age acceleration across pan-tissue, first-, and second-generation clocks ($p<0.01$, adjusted p-value< 0.05). This suggests a decrease in these cells with advancing epigenetic age, contrasting with their positive correlation with chronological age. Conversely, Luminal epithelial cells, basal myoepithelial cells lymphatic endothelial cells showed significant positive correlations with age acceleration across clocks. Additionally, the proportion of immune dendritic cells/macrophages was significantly lower in accelerated breast aging tissue, reflecting compromised immune surveillance. These findings underscore the distinct characteristics of accelerated breast aging compared to normal aging. The increase in luminal epithelial and basal myoepithelial cells, along with the reduction in immune cells in accelerated aging breast, suggests a potential higher risk of breast carcinogenesis from a cell compositional perspective.

To further explore whether cell composition mediates the increase in epigenetic age during aging process, we conducted a mediation analysis. Consistent with the correlation analysis findings, we found significant mediation effects for blood-based clocks (first- and second-generation), with GrimAge being an exception due to its inclusion of chronological age in the predictors. Our analysis also revealed that adipocytes negatively contribute to epigenetic age, while luminal epithelial and basal myoepithelial cells positively contribute, although the strength of these mediating effects were marginally significant.

4.3.6 Association between breast age acceleration and breast cancer risk measurement

We further examined the association between breast epigenetic age acceleration and breast cancer risk using the Breast Cancer Risk Assessment Tool (Gail Model) and the Tyrer-Cuzick Risk Calculator (IBISv8, for 10-year and lifetime breast cancer risk). Figure 4.13A presents the correlations between epigenetic age acceleration and breast cancer risk estimates for each risk score and clock. Among them, only Horvath clock demonstrated a marginally significant association (corr.=0.14, p=0.054) and lifetime Tyrer-Cuzick score (corr.=0.14, p=0.068). Interestingly, after adjusting for the cell proportions, we observed an increase in correlation between Horvath age acceleration and 10-year Tyrer-Cuzick score (corr.=0.17, p=0.022) and lifetime Tyrer-Cuzick score (corr.=0.17, p=0.022) (Figure 4.4A), while the correlation for other clocks remains insignificant (Figure 4.13B).

Beyond risk estimations derived from demographic and reproductive history, we also assessed DNA methylation-based risk using epiTOC2 [31], a breast cancer risk calculator that quantifies the total number of stem cell divisions. All but breast-specific clocks displayed a significant positive association between epiTOC2 score and age acceleration adjusted by cell proportions. These results showed that the link between age acceleration and cancer risk is not purely due to the cell compositional changes, although different clocks showed varying association strengths with cancer risk scores.

4.3.7 Identification of CpG sites associated with breast age acceleration and its link to estrogen receptor

To characterize the CpG sites associated with age acceleration after adjusting for cell proportions, we performed an epigenome-wide association study (EWAS). Figure 4.14 presents the QQ and Manhattan plot of the EWAS analysis results for each clock. CpG sites with q-value smaller than 0.05 are defined as significant and summarized in Supplementary Data 5 for each clock, along with their nearby genes annotated. Notably, blood-based clocks identified more significant CpG sites associated with age acceleration compared to both pan-tissue clocks and breast-specific clocks.

To better understand the biological mechanisms underlying these CpG sites, we performed genomic region set enrichment analysis on the significant CpG sites, using the total CpG sites as the background. The top 10 most significantly enriched Transcription Factor Binding Sites (TFBS) of each clock were collected and visualized in Figure 4.5. Our analysis revealed that estrogen receptor 1 (ESR1) binding sites were among the top 10 enriched TFBS for most clocks examined, providing a potential link between age acceleration and estrogen exposure. Although the association between age acceleration and estrogen exposure has been reported in our previous study [19], this analysis provides direct molecular evidence supporting the association for the first time.

Estrogen is known to stimulate the division and proliferation of breast tissue, leading to increased cellular turnover and DNA double-strand breaks [44]. This heightened cellular activity, coupled with the accumulation of DNA damage, promotes genomic instability—a hallmark of cancer development. Recent research also found that DNA double-strand breaks erode the epigenetic landscape, contributing to mammalian aging [45]. Thus, the enrichment of ESR1 binding sites for age acceleration-associated CpG sites provides a plausible mechanism linking age acceleration and increased cancer risk: estrogen exposure accelerates aging through DNA double-strand breaks. These findings suggest that estrogen exposure may have a significant role in the epigenetic changes observed in accelerated breast aging, potentially heightening breast cancer risk.

4.3.8 Transcriptomic alternations associated with breast age acceleration

To identify genes whose expression is associated with age acceleration, we conducted a differential gene expression (DE) analysis using a regression framework, testing the associations between each gene's expression and age acceleration, with chronological age included as a covariate. This analysis was conducted for each epigenetic clock, both before and after adjusting for cell proportions. We found that while thousands of genes were identified as DE genes before accounting for cell compositional changes, this number dramatically decreased after adjusting for cell proportions (Figure 4.15A), indicating that many observed differences in gene expression were primarily due to cell population dynamics. Focusing on the DE genes after adjusting cell proportions, we noticed there was minimal overlap among different clocks (Figure 4.6A), suggesting that each clock captures distinct biological signals.

We also performed Gene Ontology (GO) enrichment analysis with the DE genes and fount the enrichment results revealed distinct patterns in biological pathways for each epigenetic clock (Figure 4.6B). Specifically, there were no enriched biological processes for the Horvath and ElasticNet clocks. While the Altum clock showed enrichment for processes related to epithelial migration, and the EPM clock highlighted pathways involved in extracellular matrix organization. Blood-based clocks demonstrated greater overlap in enriched biological pathways, emphasizing processes such as kinas activity regulation, steroid hormone signaling and cell polarity. Furthermore, Gene Set Enrichment Analysis (GSEA) based on the DE analysis shows a similar pattern where different clocks enrich for diverse pathways, with the blood-based clocks to have bigger overlap, featuring immune and metabolic processes. (Figure 4.15B).

Collectively, these findings in the transcriptome analysis underscore the critical importance of accounting for cell composition in epigenetic studies [46]. The pathway enrichment analysis also highlight that different epigenetic clocks capture unique and complementary aspects of biological aging in breast tissue, providing a nuanced understanding of the underlying molecular mechanisms. This insight highlights the complexity of breast aging and interplay between epigenetic dynamics and transcriptomic variations in breast aging processes, and further investigation are needed to advance the precision and applicability of epigenetic clocks in breast aging research and clinical interventions.

4.4 Discussion

In this study, we systematically evaluated the predictive accuracy of eight epigenetic clocks in normal breast tissue, the associations between epigenetic age estimates and cell composition, and their potential implications for breast cancer risk. Our findings demonstrate that pan-tissue and breast-specific epigenetic clocks exhibit superior accuracy in age prediction compared to clocks developed for other tissues. This discrepancy underscores the limitations of cross-tissue applications of epigenetic clocks and emphasizes the importance of considering tissue specificity in epigenetic age estimation. Our analysis also revealed systematic biases in age prediction, with regression lines consistently showing positive intercepts and slopes less than one. Simulations indicated that these biases arise from the shrinkage of regression coefficients in penalized regression, which is common in high-dimensional data settings (e.g., epigenetic clock). While previous studies have reported the underestimate bias in the old subjects of Horvath clock [47], we are, to our best knowledge, the first to illustrate that the bias is a universal problem for all epigenetic clocks examined, and that it is rooted in the statistical models used to construct the clocks. These biases complicate the definition of age acceleration. To address this, we advocate for the residual definition of age acceleration, which eliminates confounding by chronological age and provides a more robust measure for studying age-acceleration-related health outcomes. Using cell type abundance estimated from transcriptomic data, we confirmed several cell compositional changes associated with advancing chronologic age described in previous studies, including increased proportions of adipocytes and vascular endothelial cells [48], and decreases in luminal epithelial and basal myoepithelial cells. We also found that immune dendritic cells/macrophages increased with advancing chronological age in healthy breast tissue. These findings indicate an age-associated remodeling of breast tissue, characterized by a decline of breast density and an elevated inflammatory landscape. The observed correlations between cell composition and epigenetic age mirrored those with chronological age, validating the biological relevance of epigenetic clocks in capturing aging-related cellular alterations. However, the variability of correlations also highlights the influence of cell composition on epigenetic/biological age estimates, emphasizing the necessity of considering cell composition heterogeneity when interpreting epigenetic aging data.

While previous studies have examined cell compositional changes in breast tissue with chronologic age, this is the first study to characterize the association between cell composition and epigenetic age accelerations in normal breast tissue. Interestingly, we found that epigenetic age acceleration of multiple clocks is associated with distinct changes of cell composition, with a significant increase in both luminal epithelial and basal myoepithelial cell proportions, and a decrease in adipocytes, vascular endothelial cells and immune dendritic cell/macrophages. The rise in luminal and basal myoepithelial cells may heighten breast cancer risk, as these cells are progenitors of common breast cancer subtypes [49]. Reduced adipocytes and vascular endothelial cells can disrupt hormonal balance [50] and angiogenesis [51] respectively, potentially affecting tissue homeostasis and contributing to an environment that favors tumor development. Lastly, the decrease in immune dendritic cells/macrophages suggests a compromised immune surveillance [9], allowing abnormal cells to proliferate unchecked. These changes jointly highlight the unique cellular dynamics of accelerated breast aging and implies a higher potential for breast carcinogenesis from a cellular population perspective.

Our study also explored the associations between breast epigenetic age acceleration and breast cancer risk. We assessed the association between epigenetic age acceleration and Gail or Tyrer-Cuzick scores and did not find a significant association. Interestingly, age acceleration in the Horvath clock is significantly associated with the Tyrer-Cuzick score after adjusting for cell proportions, suggesting that cellular heterogeneity plays a role in modulating cancer risk. The positive correlation between age acceleration and the total number of stem cell divisions, as measured by epiTOC2, supports the link between accelerated epigenetic aging and increased cancer risk. These findings highlight the potential utility of epigenetic clocks in assessing breast cancer risk and underscore the importance of considering cell composition in these cancer risk analyses.

Epigenome-wide association studies (EWAS) have identified a handful of CpG sites associated with age acceleration after adjusting for cell proportions, with significant enrichment for estrogen receptor 1 (ESR1) binding sites across multiple epigenetic clocks. Estrogen exposure has been found to promote DNA double-strand breaks [52], which contribute to mammalian aging [45], suggesting a mechanistic link between estrogen exposure and accelerated epigenetic aging. Estrogen-driven cellular proliferation and DNA damage induce genomic instability, a hallmark of both aging and cancer development [53, 54], linking the age acceleration with cancer risk. Although previous studies have reported age acceleration is potentially associated with estrogen exposure [19] and breast cancer risk [55], respectively, our findings provide direct molecular evidence that estrogen exposure accelerates breast tissue aging, potentially heightening the risk of breast cancer. Complementing these epigenetic insights, differential gene expression analysis has revealed significant transcriptomic alterations linked to age acceleration, many of which are driven by underlying shifts in cell composition. Gene ontology and gene set enrichment analyses have highlighted distinct biological processes associated with different epigenetic clocks, reflecting their unique biological signals. These insights emphasize the complexity of breast aging, driven by the interplay between estrogen exposure, epigenetic dynamics, and transcriptomic variations during aging. Understanding these interactions is crucial for refining epigenetic clocks and enhancing their applicability in clinical settings to assess aging and cancer risk.

There are some limitations to our work. Firstly, the cross-sectional nature of the data, similar to most epigenetic aging studies, makes it difficult to disentangle subject-specific effects and aging effects on the epigenetic landscape. We believe longitudinal data will benefit us and help us profile the aging rates and trajectories more accurately. Secondly, the relatively short duration after sample collection did not allow us to assess breast cancer incidence among the study population; we relied on breast cancer risk scores like the Gail or Tyrer-Cuzick scores, which may not accurately reflect the future likelihood of developing cancer [56]. This reliance limits our ability to assess the relationship between age acceleration and the true potential of developing breast cancer.

Additionally, our analysis was based on bulk DNA methylation and gene expression data, which lacks single-cell level resolution. While we recovered some cell type-level dynamics via deconvolution, it can be further improved through the lens of single-cell techniques. We envision that single-cell DNA methylation, coupled with single-cell transcriptomic profiling technologies, would provide a more detailed understanding of how epigenomic changes interplay with transcriptomics during normal aging. Future research should focus on longitudinal studies and single-cell analyses to refine these clocks and explore their clinical applications in breast aging and cancer risk assessment.

In summary, our study underscores the critical need for tissue-specific epigenetic clocks to accurately predict age and assess age acceleration. The observed biases in age prediction models highlight the importance of using appropriate definitions for age acceleration. Age-related changes in breast cell composition and their impact on epigenetic age emphasize the need to consider cellular heterogeneity in aging studies, and age acceleration shows a distinct association with cell composition, suggesting its link with cancer risk from the cell compositional perspective. Our findings also provide valuable insights into the molecular mechanisms linking estrogen exposure, epigenetic aging, and breast cancer risk, paving the way for future research and clinical interventions in breast aging. Future studies should focus on refining epigenetic clocks for broader applicability and investigating their potential for early detection and prevention of age-related diseases, including breast cancer.

4.5 Acknowledgements

I thank the study donors and contributors to the Susan G. Komen Tissue Bank at the Indiana University Simon Cancer Center, whose participation, data collection, and support made this work possible. I also want to thank Dr. Mary Sehl for inviting me to work on this project. Figure 4.1A was created with BioRender.com, thus acknowledging it by courtesy.

4.6 Tables and figures

Table 4.1: Characteristics of the study population.

Variable	Age < 50 years (n=89)	Age ≥ 50 years (n=92)	p-value
Demographics			
Age (years), mean ± SD	40.8 ± 7.7	59.0 ± 6.9	<2.2e-16
Ethnicity Hispanic, n (%)	3 (3.4)	6 (6.5)	0.53
Body mass index, mean ± SD	28.2 ± 7.4	28.3 ± 6.2	0.93
Tobacco smoking, ever, n (%)	21 (23.6)	34 (37)	0.07
Alcohol use, current, n (%)	61 (68.5)	64 (69.6)	1
Gynecologic history			
Age at menarche, mean ± SD	12.8 ± 1.6	12.7 ± 1.5	0.77
Premenopausal, n (%)	81 (91)	14 (15.2)	<2.2e-16
Age at menopause, mean ± SD	36.7 ± 7.4	47.5 ± 6.6	0.0077

Continued on next page

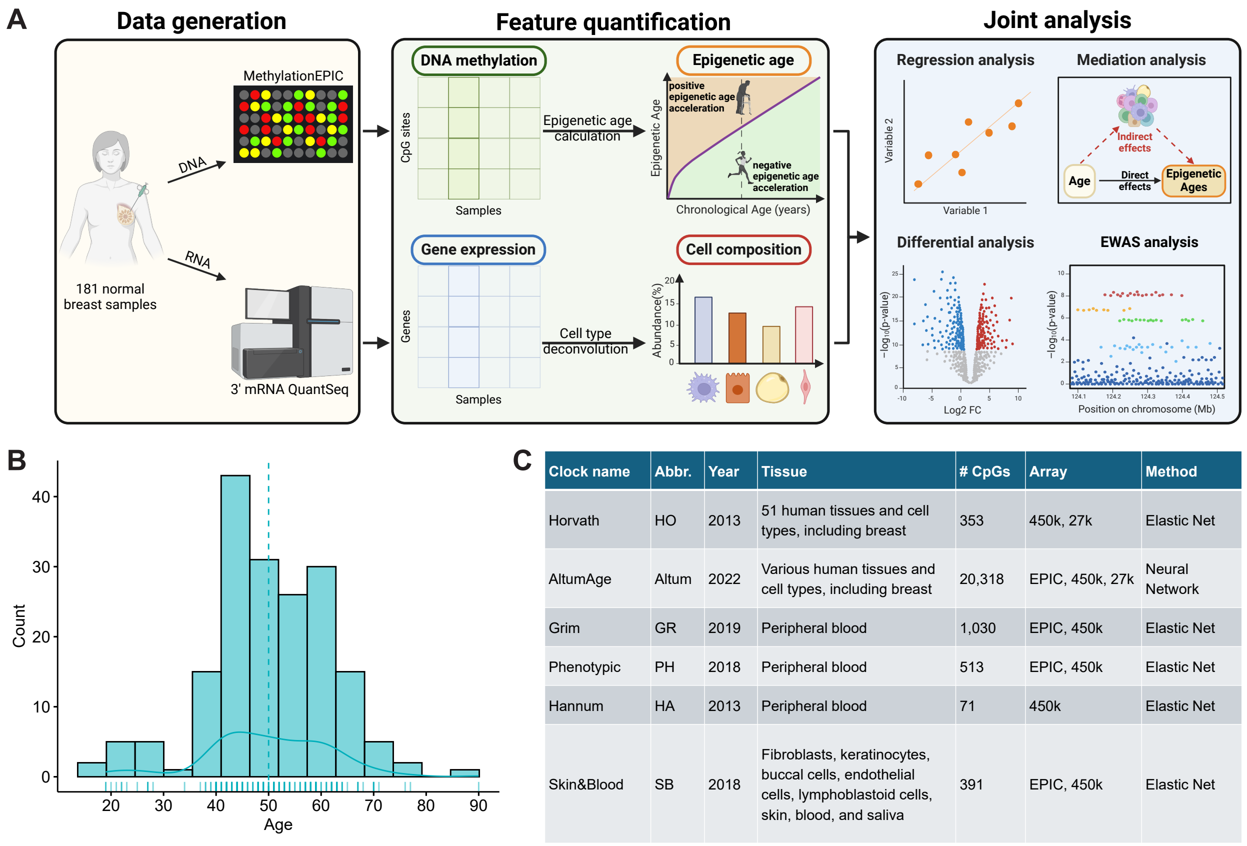
Variable	Age $<$ 50 years (n=89)	Age $\geq$ 50 years (n=92)	p-value
Total menstrual years, mean ± SD	27.5 ± 8.1	35.6 ± 6.5	8.9e-12
Reproductive history			
Nulliparous, n (%)	43 (48.3)	46 (50)	0.94
Age at first full-term birth, mean ± SD	27.3 ± 5.3	26.0 ± 4.8	0.23
Hormonal replacement therapy, ever, n (%)	1 (1.1)	37 (40.2)	<2.2e-16
Estimated breast cancer risk scores			
Gail score, mean ± SD	13.3 ± 5.7	10.7 ± 5.3	0.0027
Tyrer-Cuzick, lifetime, mean ± SD	2.6 ± 1.9	4.4 ± 3.0	2.7e-06
Tyrer-Cuzick, 10-year, mean ± SD	15.6 ± 7.1	11.2 ± 6.4	2.5e-05

SD: Standard deviation; n: number of samples; %: percentage; p-value: Two-tailed Student's t-test for continuous variables and chi-square test for categorical variables.

Table 4.2: Regression summary for eight epigenetic clocks in KTB data.

Clock	**Intercept**	**Slope**	**MAE**	**RMSE**	**Corr.**	**p-value**	**adjusted p-value**
Horvath	28.09	0.58	7.92	9.25	0.90	9.25e-68	2.47e-67
AltumAge	16.24	0.74	5.24	6.42	0.88	1.16e-59	2.32e-59
ElasticNet	7.48	0.85	**3.33**	**4.11**	**0.94**	3.61e-83	1.44e-82
Epigenetic Pacemaker	**4.10**	**0.92**	6.56	8.55	0.78	1.05e-38	1.68e-38
GrimAge	29.01	0.66	11.96	12.85	**0.94**	6.76e-88	5.41e-87
Phenotypic age	14.91	0.23	23.83	27.20	0.25	6.29e-04	6.29e-04
Hannum clock	12.64	0.34	20.39	22.03	0.60	6.95e-26	6.96e-26
Skin&Blood clock	24.19	0.63	8.73	10.72	0.68	1.43e-25	1.63e-25

Intercept and Slope: the regression coefficients of epigenetic age against chronological age for each clock; MAE: Mean Absolute Error; RMSE: Root Mean Square Error; Corr.: Pearson Correlation coefficient; p-value: significance level of correlation via t-test; adjusted p-value: significance level of correlation adjusted by Benjamini-Hochberg procedure. The best prediction performance metrics (Intercept closest to 0, Slope closest to 1, smallest MAE/RMSE, and biggest Corr.) are highlighted in bold font.



Clock name	Abbr.	Year	Tissue	# CpGs	Array	Method
Horvath	HO	2013	51 human tissues and cell types, including breast	353	450k, 27k	Elastic Net
AltumAge	Altum	2022	Various human tissues and cell types, including breast	20,318	EPIC, 450k, 27k	Neural Network
Grim	GR	2019	Peripheral blood	1,030	EPIC, 450k	Elastic Net
Phenotypic	PH	2018	Peripheral blood	513	EPIC, 450k	Elastic Net
Hannum	HA	2013	Peripheral blood	71	450k	Elastic Net
Skin&Blood	SB	2018	Fibroblasts, keratinocytes, buccal cells, endothelial cells, lymphoblastoid cells, skin, blood, and saliva	391	EPIC, 450k	Elastic Net

Figure 4.1: Study design and overview. (A) In this study, 181 normal breast tissue samples from Komen Tissue Bank (KTB) with paired DNA methylation and gene expression data were analyzed. DNA methylation data were used to calculate the epigenetic age for different clocks. Gene expression data were used to estimate cell type abundances. The DNA methylation, gene expression, epigenetic age, and cell composition were then jointly analyzed to comprehensively characterize the epigenetic aging and cell compositional landscape of normal breast tissue during aging. (B) Histogram and density plot showing the chronological age distribution of the 181 study samples. The vertical dashed line shows the median age (50 years old). The solid line represents the distribution density. (C) Table summarizing the features of six popular clocks, including two pan-tissue clocks, Horvath clock and AltumAge, two second-generation clocks, GrimAge and Phenotypic Age, two first-generation clocks, Hannum clock and Skin&Blood clock. Besides the six published clocks, we also trained two breast-specific clocks in KTB samples using Elastic Net algorithms and Epigenetic Pacemaker and included them in the following evaluations.

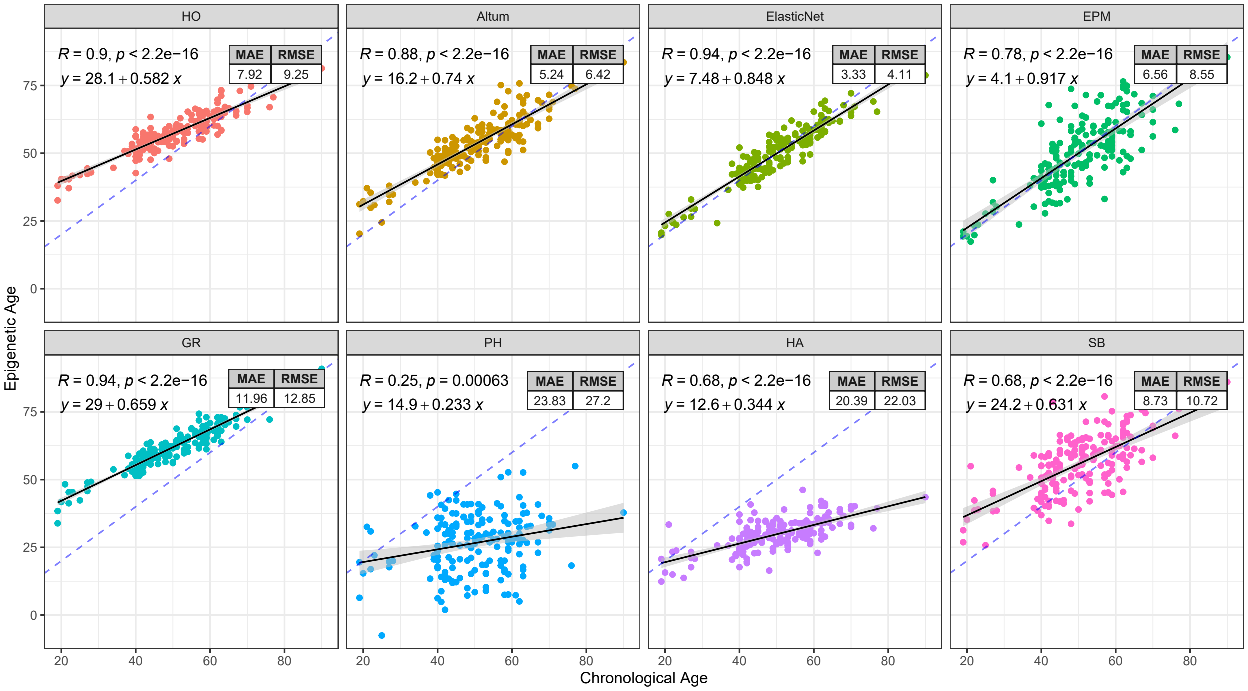


Figure 4.2: Age prediction accuracy across eight epigenetic clocks. The top panel shows the epigenetic clocks for pan-tissue (HO: Horvath clock, Altum: AltumAge) and breast-specific clocks (ElasticNet: clock trained with Elastic Net algorithm, EPM: clock trained with Epigenetic Pacemaker). The bottom panel shows blood-based epigenetic clocks, including second-generation clocks (GR: GrimAge, PH: Phenotypic Age) and first-generation clocks (HA: Hannum clock, SB: Skin&Blood clock). The dashed blue line shows a y=x diagonal line, and the black line represents the regression line between epigenetic age and chronological age for each clock, with the shadowed region indicating a 95% Confidence Interval for regression. Pearson correlation coefficient statistics, regression line equation, and prediction error (MAE: mean square error, RMSE: root mean square error) are annotated at the top of each panel on the plot.

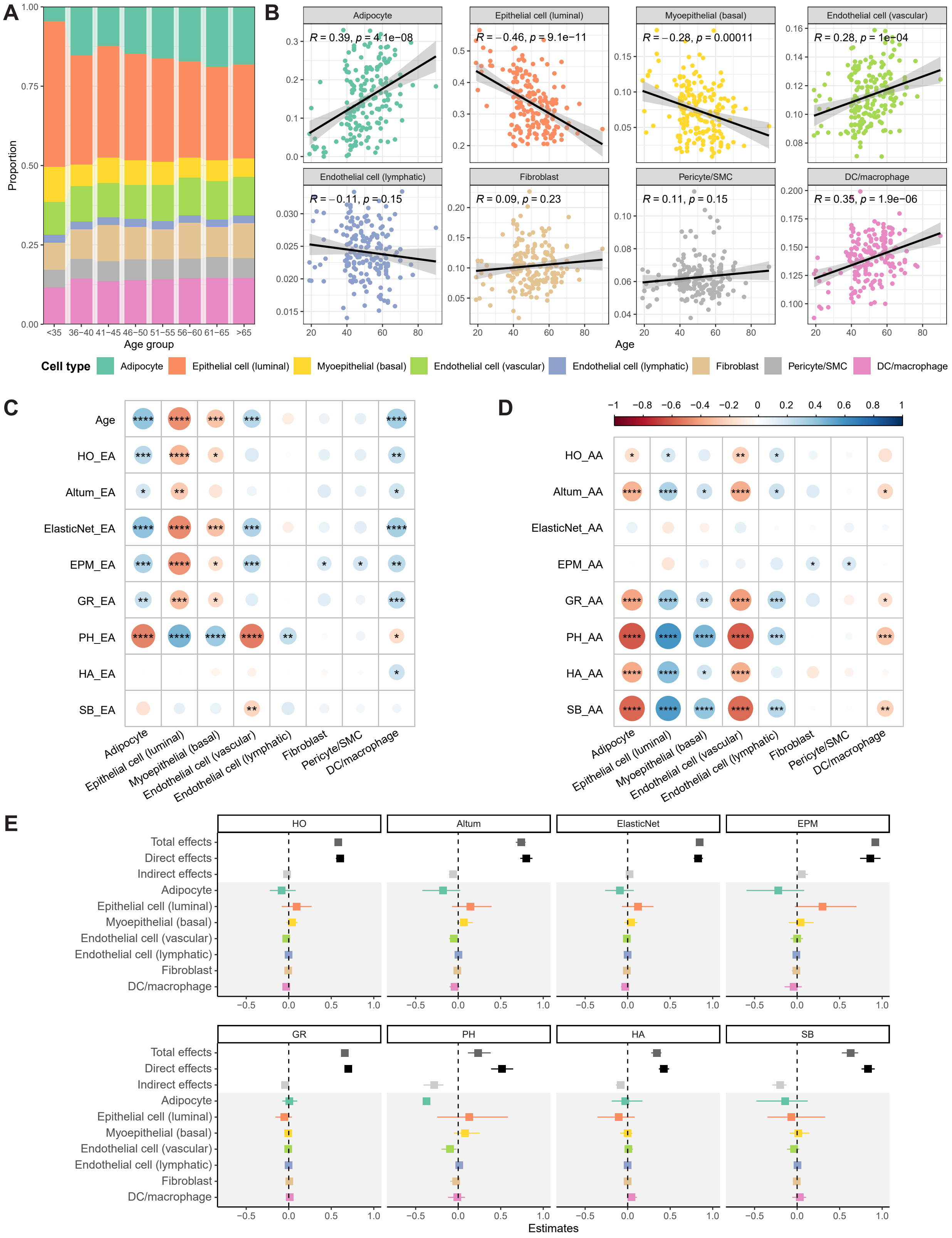


Figure 4.3: Cell composition's correlation with epigenetic age and age acceleration.

Figure 4.3: (A-B) Cell composition's correlation with advancing chronological age. (A) Bar plot showing the average cell proportion dynamics across age groups. (B) A scatterplot shows the trend and strength of proportion change with chronological age for each cell type. The black line indicates the regression line between proportion and chronological age, with the shadow region indicating 95% regression Confidence Intervals. The Pearson correlation coefficients and significance level are annotated at the top of each panel on the plot. (C-D) Heatmap showing the correlation between cell composition and epigenetic age (C) and age acceleration (D). Chronological age was added to the top row as a reference. Circle size and color scale represent the strength of correlation, with blue color showing positive correlation and red color indicating negative correlation. The asterisk indicates the significance level after the p-value adjustment. *: adjusted p-value < 0.05; **: adjusted p-value < 0.01; ***: adjusted p-value < 0.001; ****: adjusted p-value < 0.0001. (E) Forest plot displaying the cell composition's mediation effects on the relationship from chronological age to epigenetic age for each clock. The total effects, direct effects, and indirect effects are annotated at the top. Grey shadowed region shows the individual contribution of each cell type to epigenetic age. The dot represents the effects of mediation analysis, with the line width representing the 95% Confidence Interval.

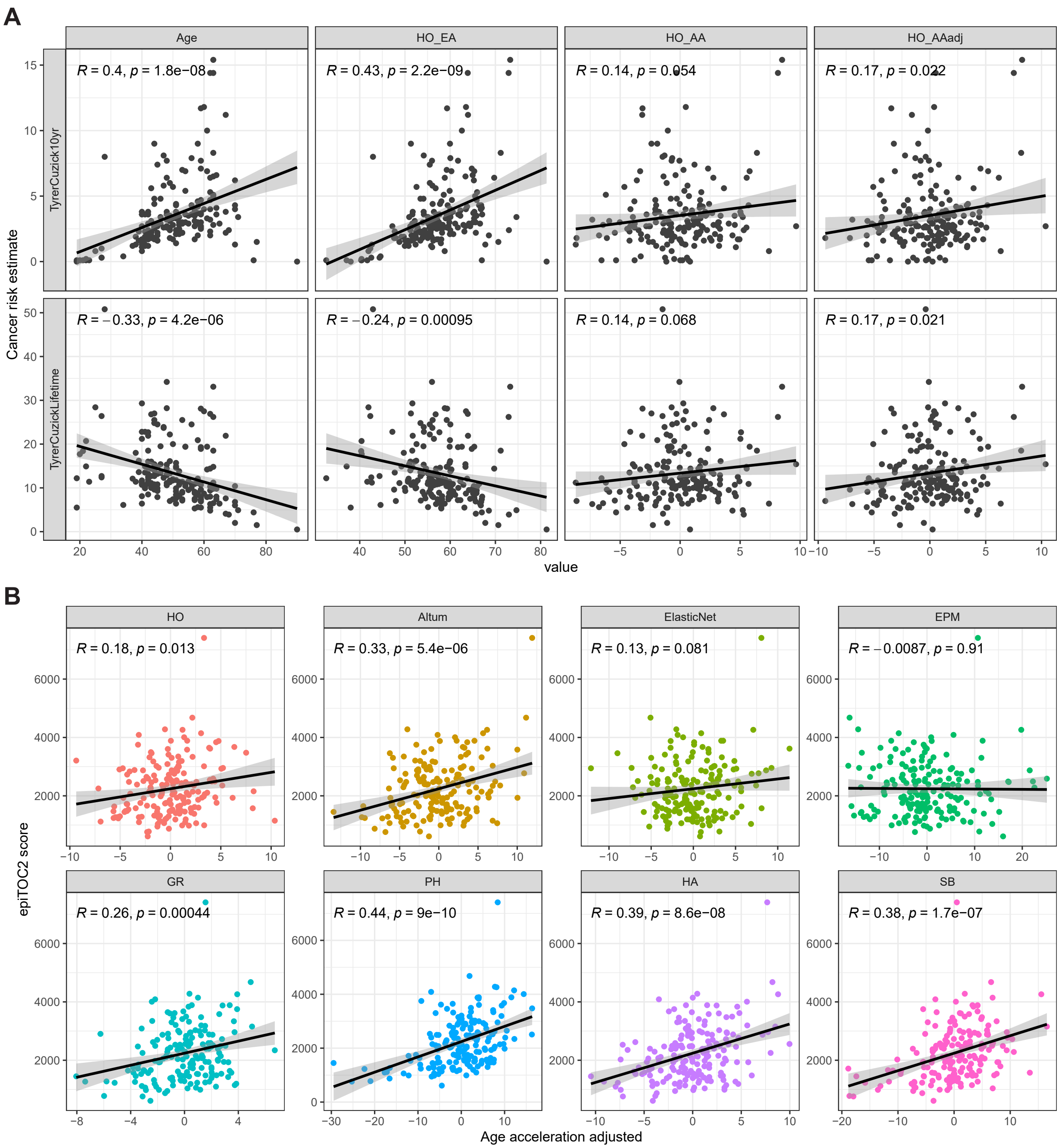


Figure 4.4: Age acceleration's association with cancer risk estimates. (A) Scatter plot showing the association between Tyrer-Cuzick score and chronological age, Horvath clock's epigenetic age, age acceleration, and age acceleration adjusted by cell proportions (top panel: 10-year risk, bottom panel: lifetime risk). (B) Scatter plot showing the association between DNAm-based cancer risk (epiTOC2 score) and age acceleration adjusted by cell proportions for each clock. The black line represents the regression line with a shadowed region indicating 95% Confidence Intervals. Pearson correlation coefficients and significance levels are annotated at the top of each panel on the plot.

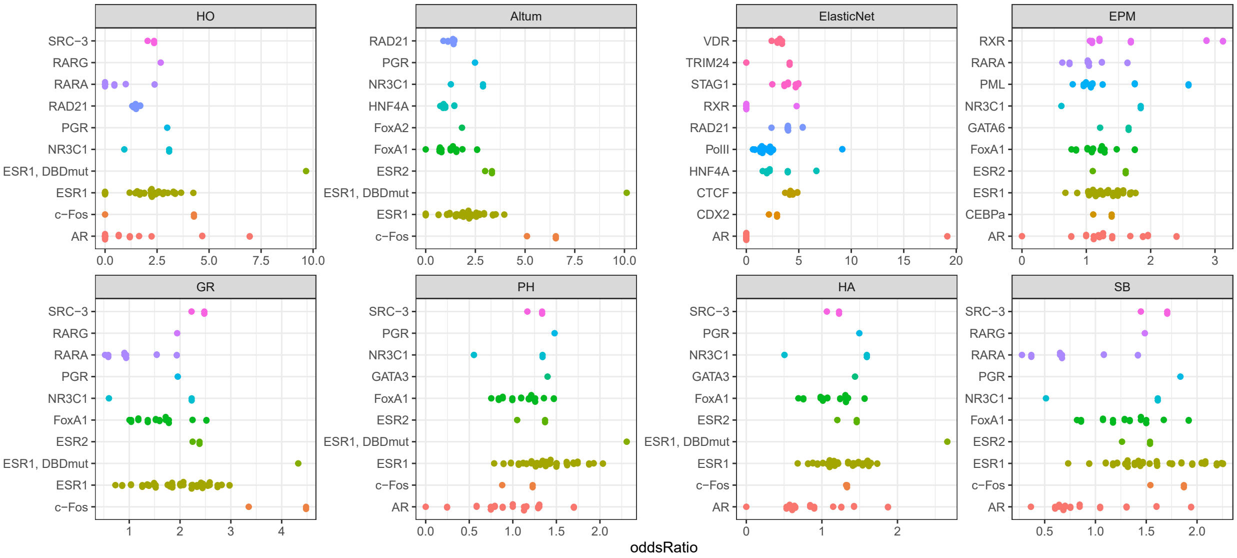


Figure 4.5: Age-acceleration-related CpG sites enriching for estrogen receptor binding sites. EWAS analysis was performed to identify CpG sites associated with age acceleration after adjusting cell proportions for each clock, followed by genomic region enrichment analysis. The top 10 most significantly enriched Transcription Factor Binding Sites (TFBS) for each clock were visualized, with dots indicating the enrichment odds ratio of genomic regions for a particular TFBS. Estrogen receptor 1 (ESR1) binding sites were found among the top 10 most significantly enriched TFBS for most clocks examined.

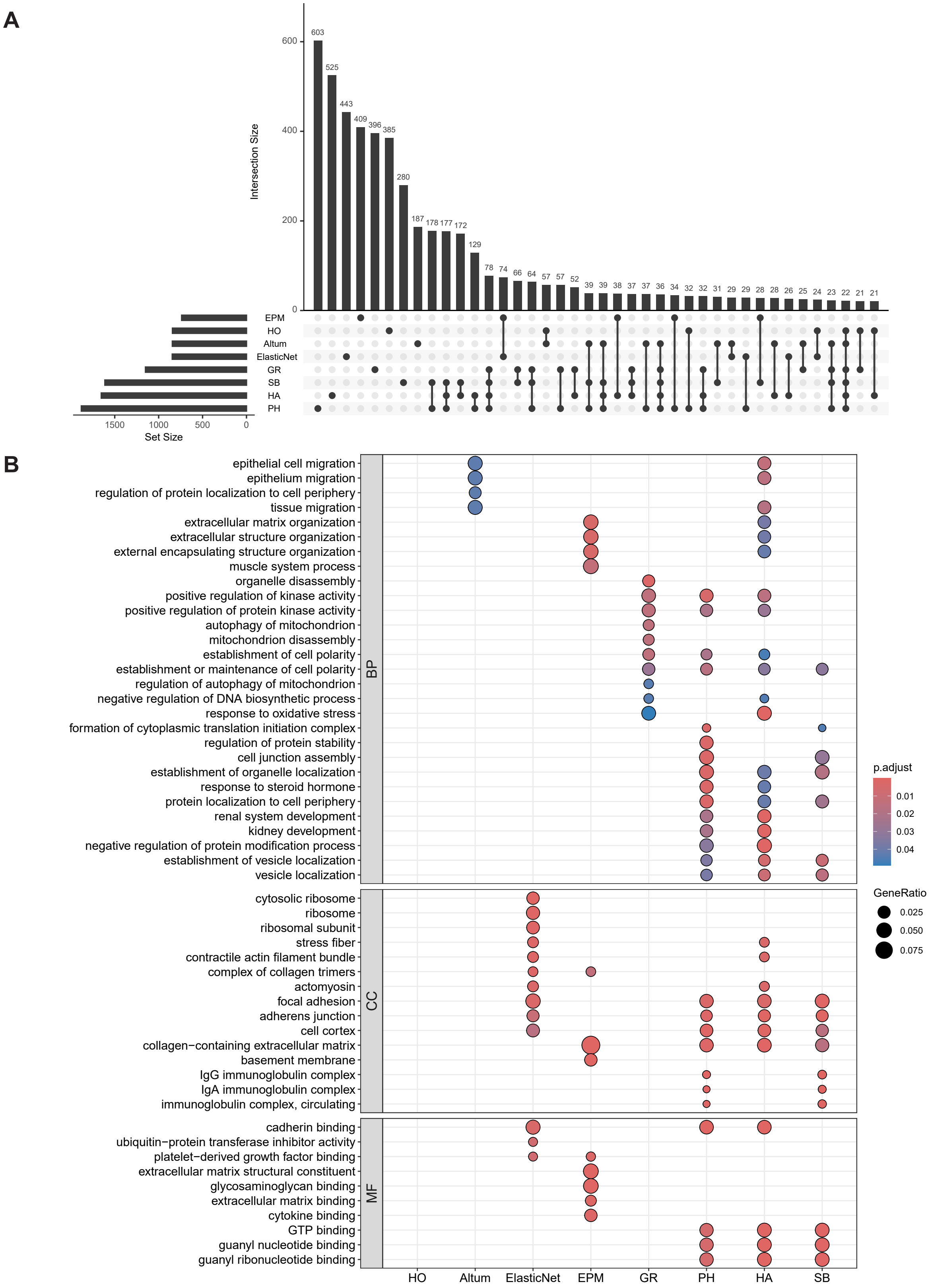


Figure 4.6: Transcriptome analysis characterizing genes and pathways associated with age acceleration.

Figure 4.6: (A) Differential expression (DE) analysis was performed to identify genes significantly associated with age acceleration after adjusting cell proportions. The bar plot shows the set size of DE genes of each clock and their intersections. Little overlap was found in the DE genes among clocks. (B) Gene ontology (GO) enrichment analysis was performed on the DE genes of each clock to characterize the enriched pathways for each clock. The top 10 most significantly enriched pathways were shown for each clock and arranged into three categories (BP: Biological Pathways, CC: Cellular Component, MF: Molecular Function). The dot size represents the gene ratio of a specific pathway overlapping with DE genes, and the color scale represents the adjusted significance level of enrichment.

4.7 Supplementary materials

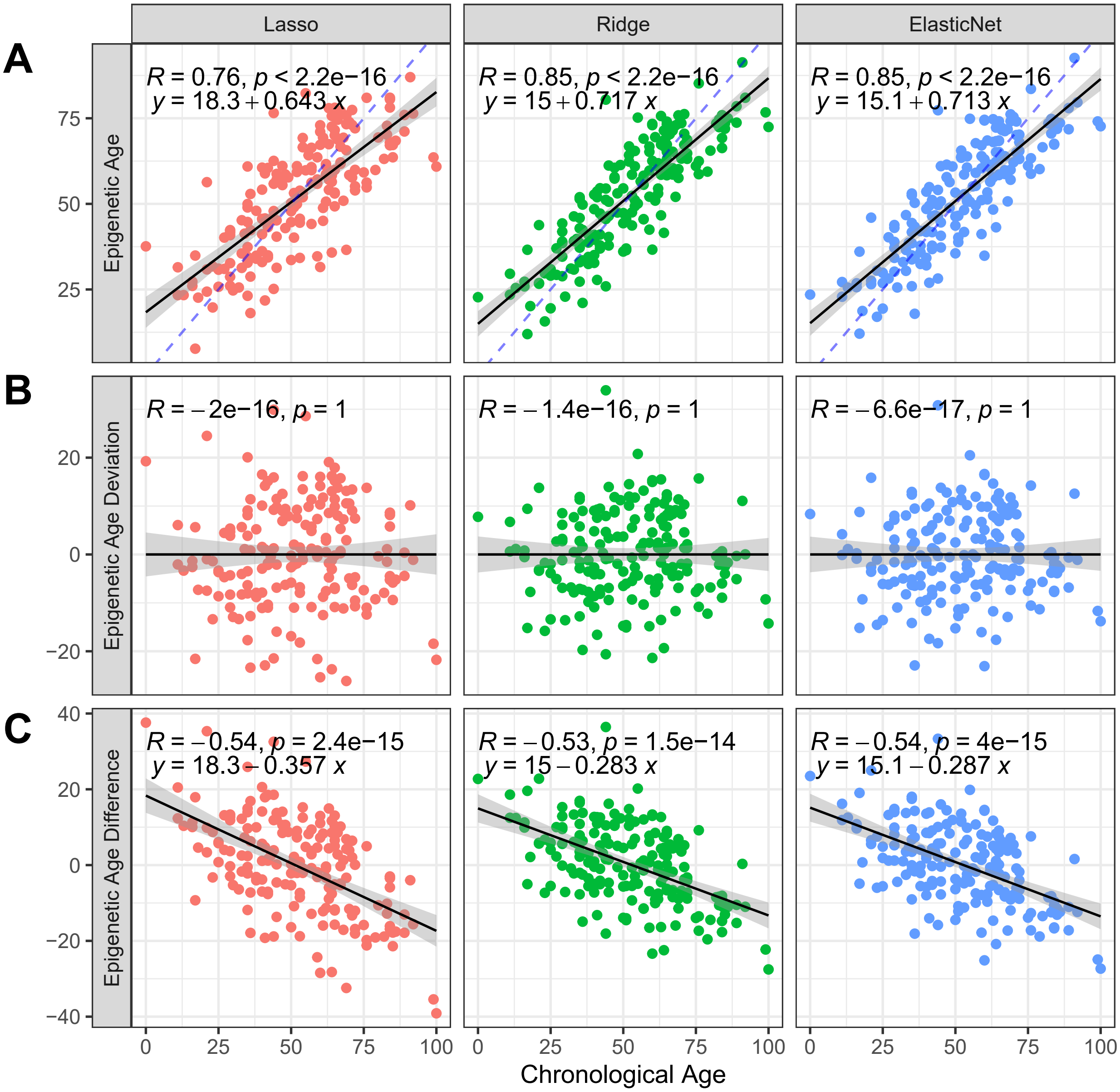


Figure 4.7: Age deviation and Age difference comparison across different penalized regression methods in simulation. The scatter plot shows chronological age is strongly correlated with predicted/epigenetic age (A), not correlated with Age deviation (B), and inversely correlated with Age difference (C) across penalization methods in simulation, indicating the confounding issue of using difference as age acceleration measurement. Columns represent the three different penalization techniques (Lasso: least absolute shrinkage and selection operator, Ridge: Ridge regression, ElasticNet: Elastic Net regression). The black line represents the regression line with a shadowed region indicating a 95% Confidence Interval.

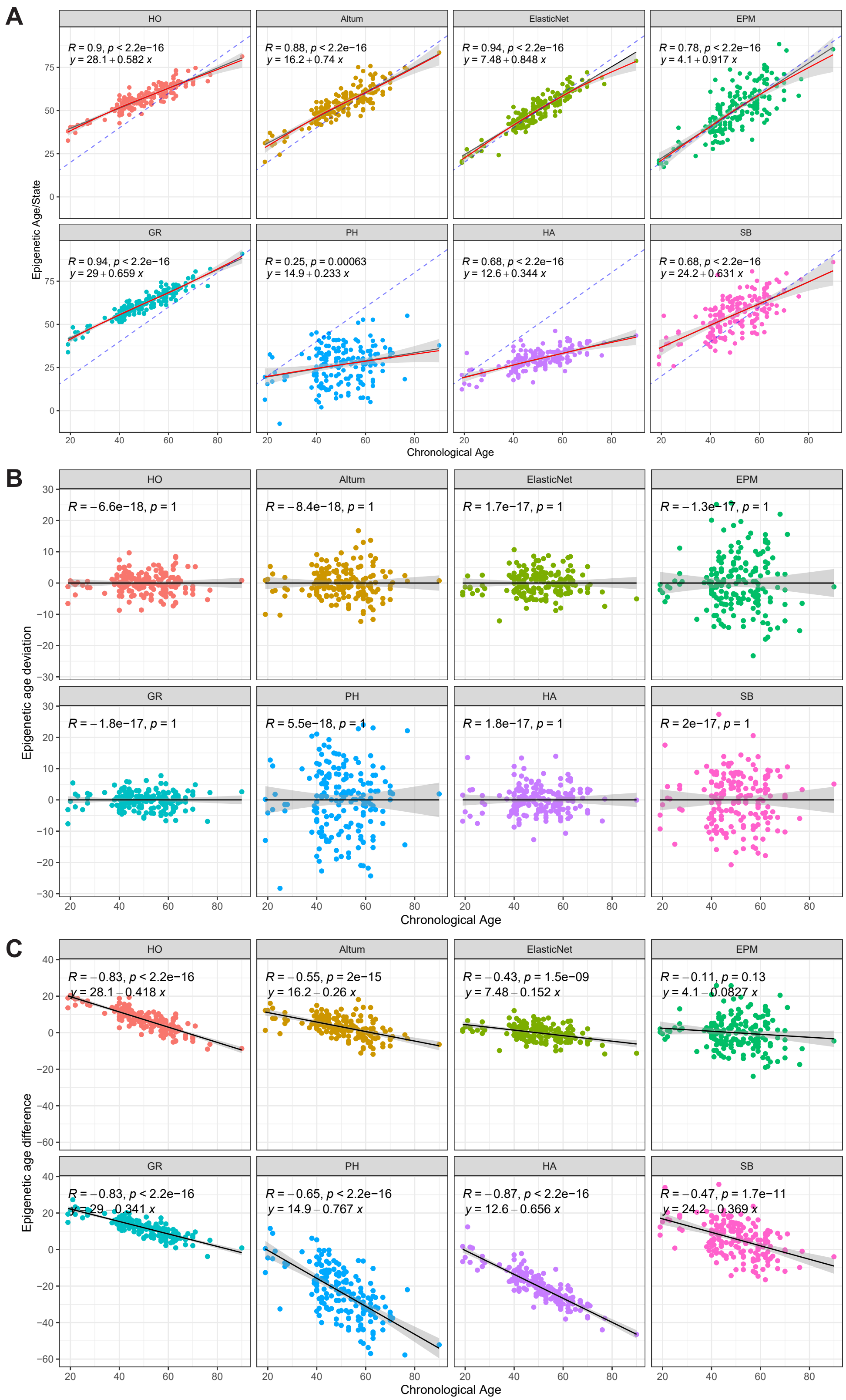
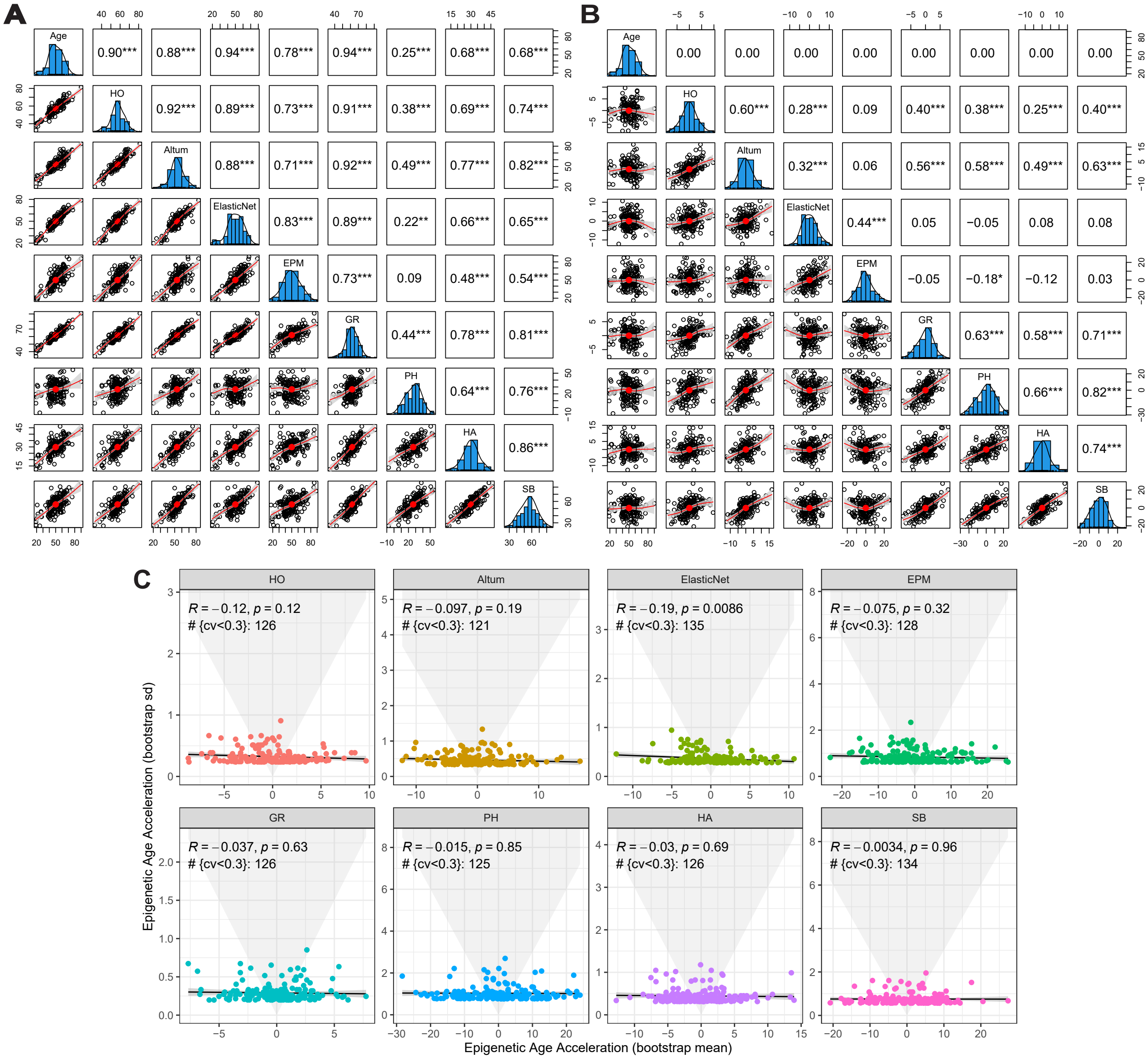


Figure 4.8: Age deviation and Age difference comparison across epigenetic clocks in KTB data.

Figure 4.8: The scatter plot shows chronological age is strongly correlated with predicted/epigenetic age (A) in KTB data, with the redline representing the cubic spline regression and confirming the change tendency between epigenetic age and chronological age is predominantly linear across epigenetic clocks. Chronological age is not correlated with Age deviation (B), and inversely correlated with Age difference (C) in real data, indicating the confounding issue of using difference as age acceleration measurement. The black line represents the regression line with a shadowed region indicating a 95% Confidence Interval.



Shadow region indicates coefficient of variation > 30%

Figure 4.9: Correlations among epigenetic age, age acceleration, and the robustness of age acceleration to sampling in KTB data. (A-B) Scatterplot matrices for epigenetic age (A) and age acceleration (B), with chronological age added as a reference. The lower panels show the scatter plot of a variable pair. The upper panels display the pairwise Pearson correlation coefficient and significance level (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). (C) Robustness of age acceleration to variations of sample composition. Bootstrap sampling was conducted for 1000 times with bootstrap mean (x-axis) and standard deviation (y-axis) of age acceleration calculated for each sample, shadowed region represents coefficients of variation (cv) greater than 0.3, showing most samples have stable age acceleration across clocks.

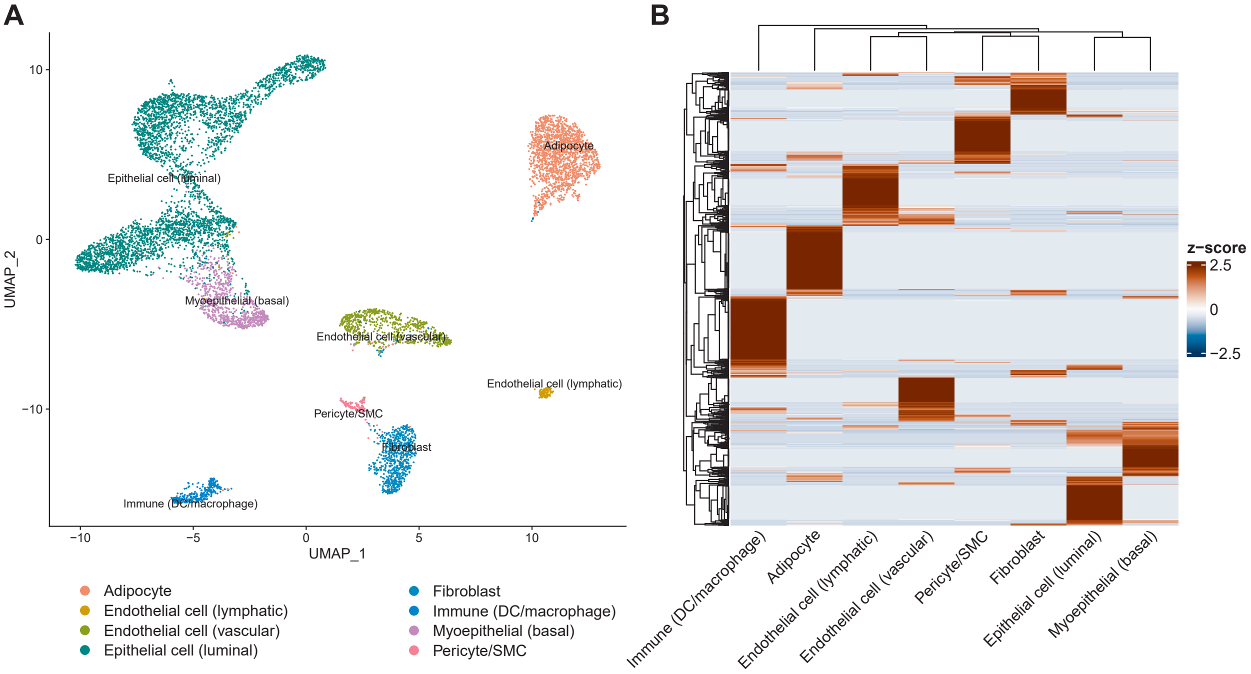


Figure 4.10: GTEx normal breast snRNA data for cell deconvolution. (A) UMAP showing eight major cell types in GTEx normal breast tissue snRNA-seq data. (B) Heatmap showing the single cell gene signature matrix constructed by CIBERSORTx and used for cell deconvolution, the standardized gene expression value (z-score) is shown for each cell type.

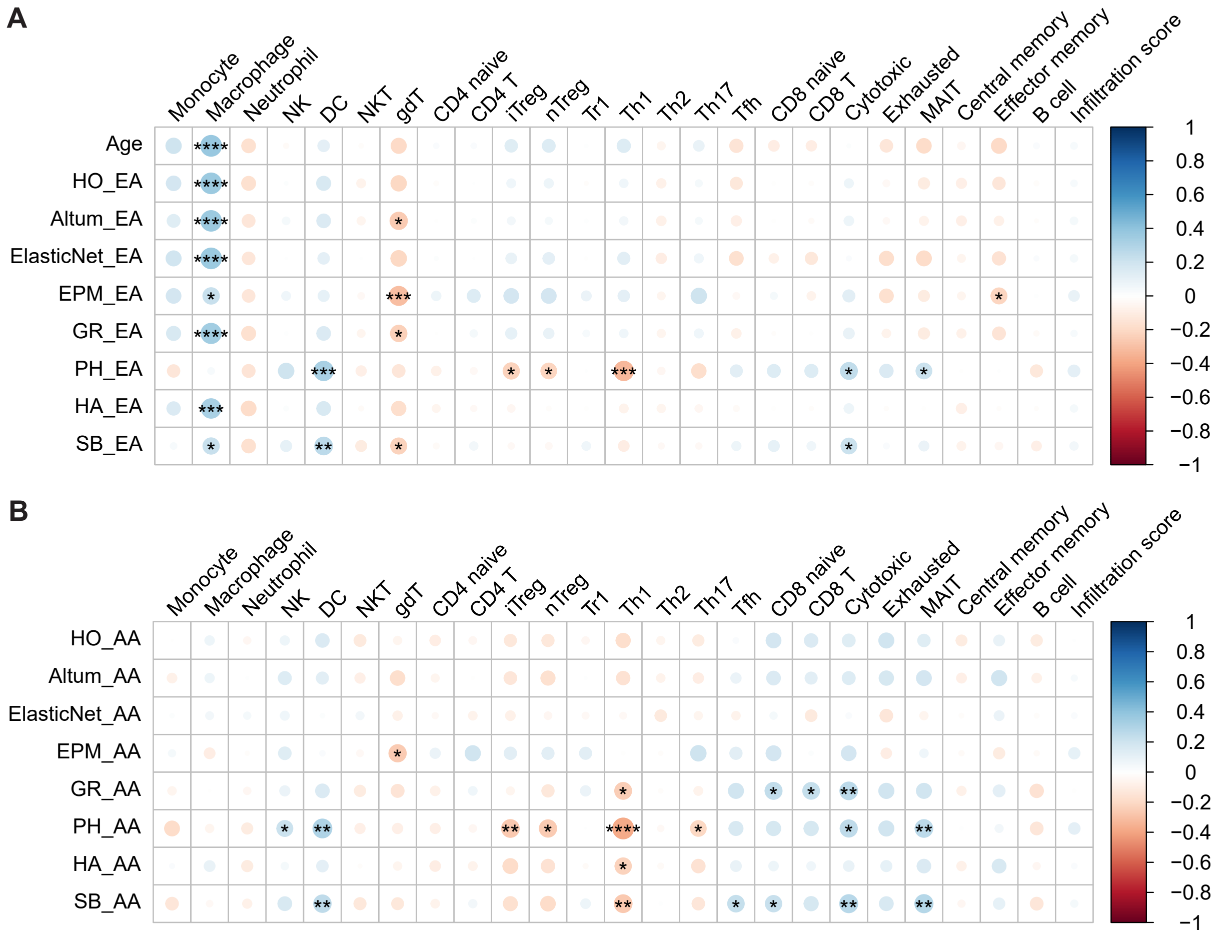


Figure 4.11: Association of immune cell scores with epigenetic age and age acceleration. Correlation heatmap showing the relationship between immune cell scores and epigenetic age (A) and age acceleration (B) across epigenetic clocks. Chronological age was added to the top row as a reference. Circle size and color scale represent the strength of correlation, with blue color showing positive correlation and red color indicating negative correlation. The asterisk indicates the significance level after the p-value adjustment. *: adjusted p-value $<$ 0.05; **: adjusted p-value $<$ 0.01; ***: adjusted p-value $<$ 0.001; ****: adjusted p-value $<$ 0.0001.

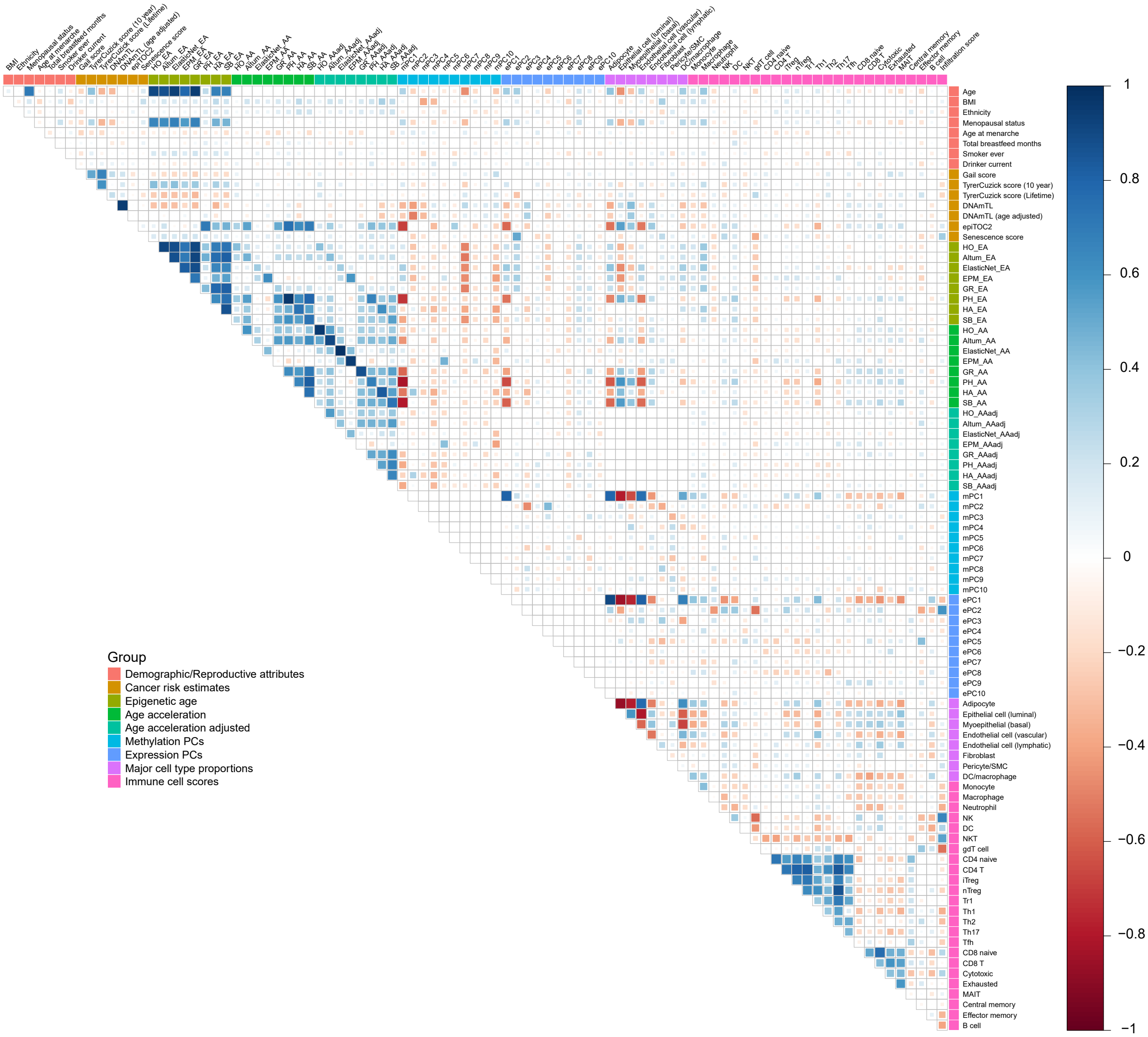


Figure 4.12: Pairwise correlation heatmap of examined variables. Correlation heatmap showing the pairwise correlation among demographical variables (Age, Ethnicity, BMI, Smoker ever, Acholic drinks per week), reproductive history (Menopausal status, Age at menarche, Total breastfeeding months), breast cancer risk estimates, eight epigenetic clocks' epigenetic age, age acceleration, age acceleration adjusted by cell proportions, top 10 principal components from DNA methylation (mPC1-mPC10) and gene expression (ePC1-ePC10), eight cell proportions, 24 immune cell scores and immune infiltration score.

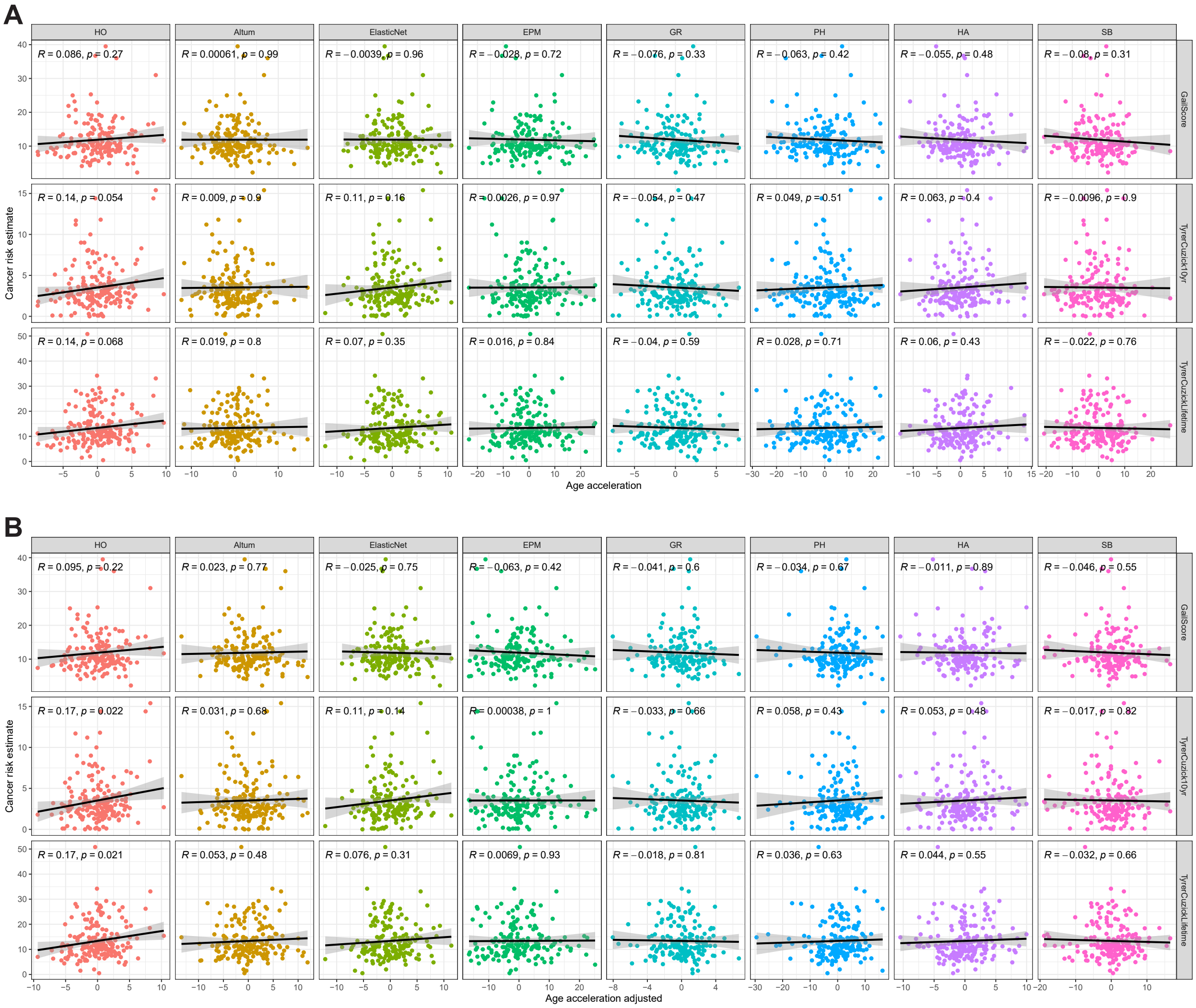


Figure 4.13: Association between cancer risk estimates and age acceleration. Scatterplot showing the association between cancer risk measurement and age acceleration before (A) and after (B) adjusting cell proportions. Columns show the eight epigenetic clocks, and rows represent the three breast cancer risk measurements. The black line represents the regression line with a shadowed region indicating a 95% Confidence Interval. Pearson correlation coefficient statistics are annotated on the top of each panel.

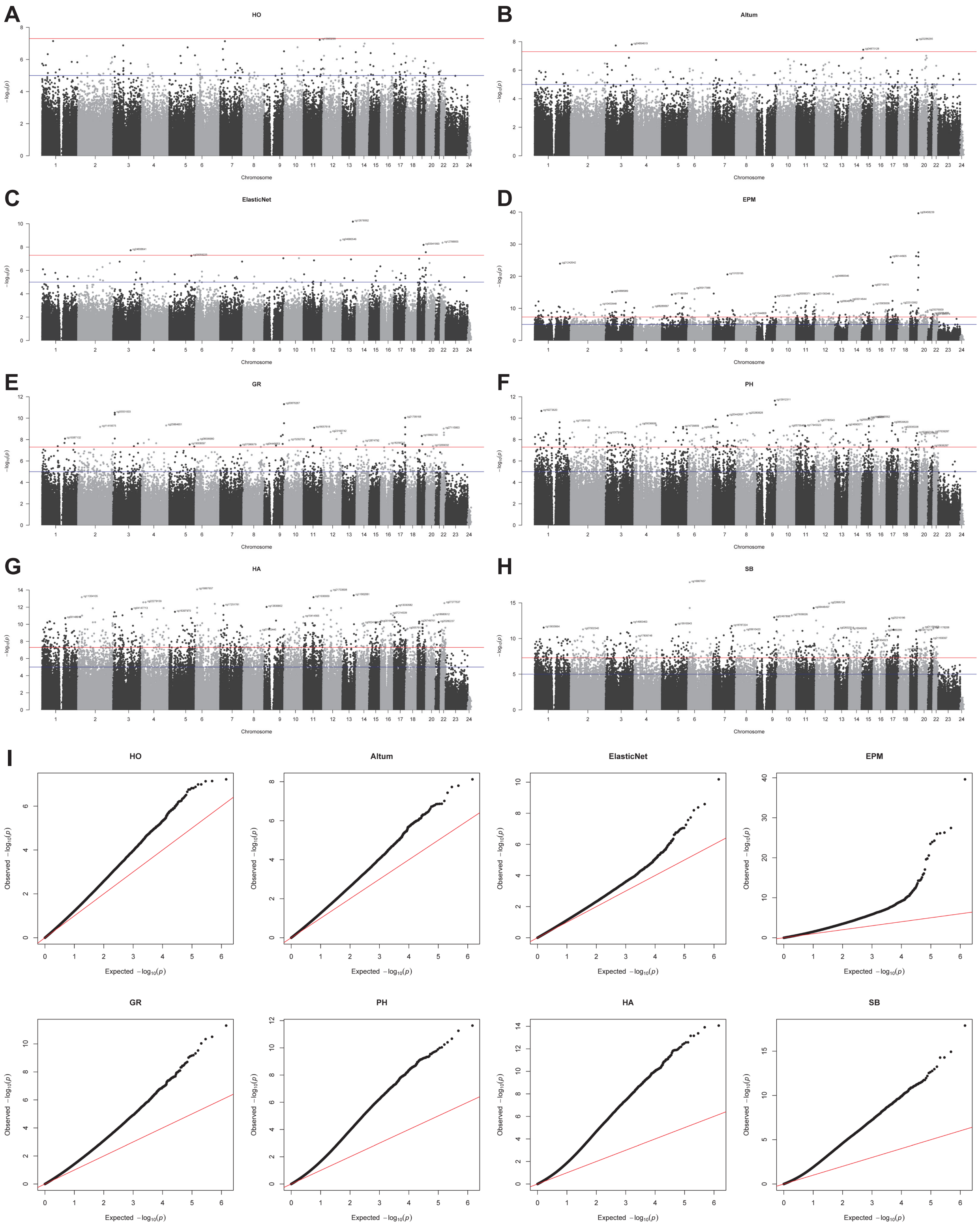


Figure 4.14: Manhattan and QQ plot for EWAS analysis.

Figure 4.14: (A-H) Manhattan plot of EWAS analysis for pan-tissue clocks: Horvath clocks (A) and AltumAge (B), breast-specific clocks: clock trained using ElasticNet algorithm (C) and Epigenetic pacemaker (D), second generation clocks: GrimAge (E) and Phenotypic age (F), first-generation clocks Hannum clock (G) and Skin&Blood clock (H); Blue and red horizontal line on the Manhattan plot shows 10^{-5} and $5*10^{-8}$ p-value threshold, respectively. The top sites on each chromosome whose p-value passed the Bonferroni correction threshold are labeled on the plot. (I): Quantile-Quantile (QQ) plot for all eight clocks examined.

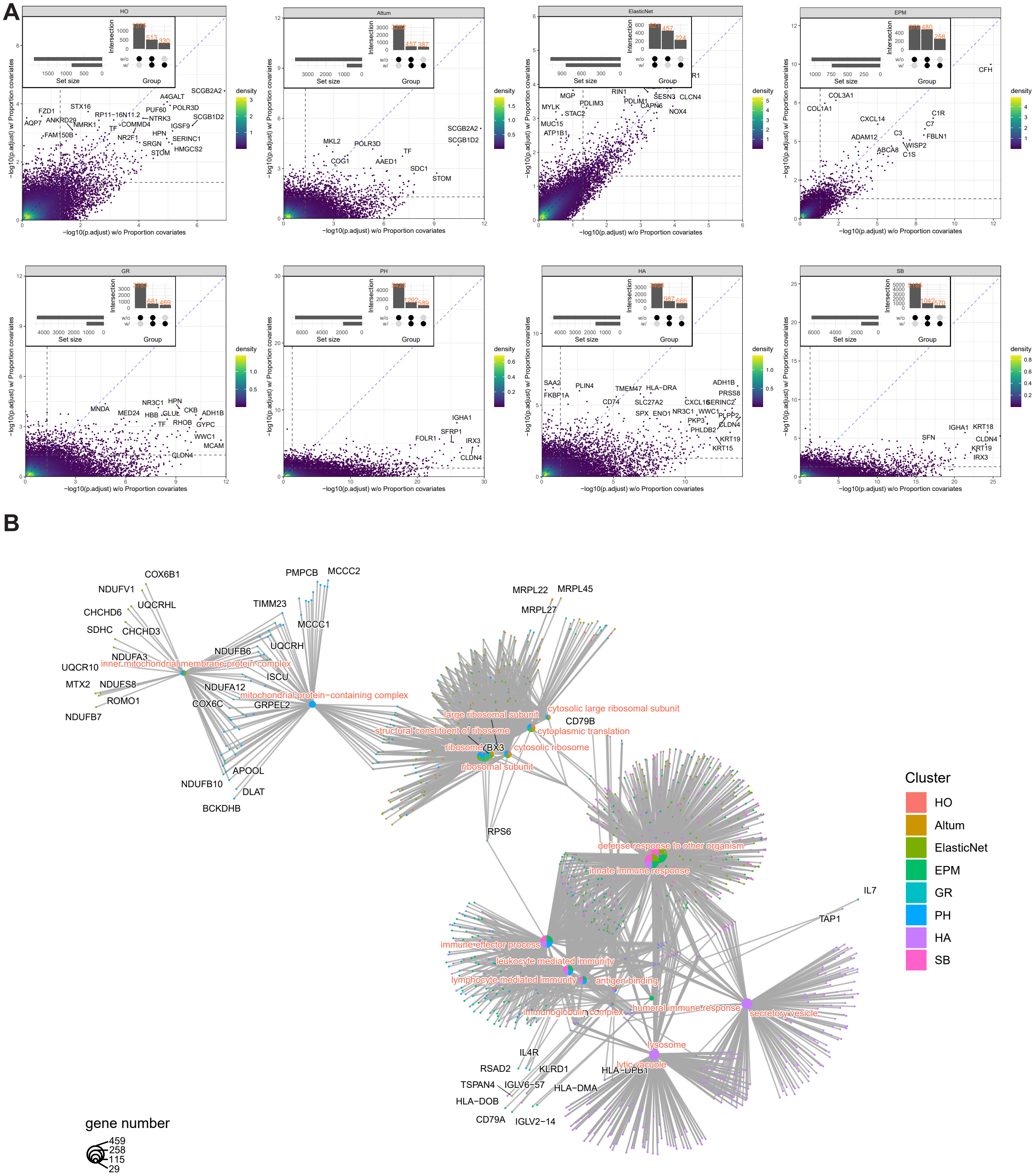


Figure 4.15: Differential gene expression and gene set enrichment analysis. Differential expression (DE) analysis was performed to identify genes associated with age acceleration for each clock, followed by gene set enrichment analysis.

Figure 4.15: (A) scatterplot showing the DE analysis before and after adjusting cell proportions. Each dot is a gene with -log10 of adjusted p-value before (x-axis) and after (y-axis) adjusting cell proportions. Vertical and horizontal dashed line indicates a 0.05 threshold. Genes with adjusted p-values smaller than 0.05 are defined as differentially expressed (DE) genes, and the number of DE pre/post adjusting cell composition, as well as their overlap, were summarized using the bar plot annotated to the top-left of the scatter plot. The genes accumulating below the diagonal dashed line indicate many of the differential genes are due to cell compositional changes. (B) Cnetplot displaying the GSEA pathway enrichment analysis results. Large nodes with pinkish font represent pathways, while small nodes with black font represent genes connected to the pathways by edges. Node size represents the number of genes of a pathway, and the colors of large nodes indicate whether the given pathway is enriched in the corresponding clocks.

4.8 References

[1] Horvath, S. and Raj, K. “DNA Methylation-Based Biomarkers and the Epigenetic Clock Theory of Ageing”. *Nature Reviews Genetics* 19.6 (2018), pp. 371–384.

[2] Davis, J. D. and Lin, S.-Y. “DNA Damage and Breast Cancer”. *World Journal of Clinical Oncology* 2.9 (2011), pp. 329–338.

[3] Sehl, M. E., Henry, J. E., Storniolo, A. M., Horvath, S., and Ganz, P. A. “The Impact of Reproductive Factors on DNA Methylation-Based Telomere Length in Healthy Breast Tissue”. *npj Breast Cancer* 8.1 (2022), pp. 1–4.

[4] Johnson, K. C., Koestler, D. C., Cheng, C., and Christensen, B. C. “Age-Related DNA Methylation in Normal Breast Tissue and Its Relationship with Invasive Breast Tumor Methylation”. *Epigenetics* 9.2 (2014), pp. 268–275.

[5] Bai, H., Liu, X., Lin, M., Meng, Y., Tang, R., Guo, Y., Li, N., Clarke, M. F., and Cai, S. “Progressive Senescence Programs Induce Intrinsic Vulnerability to Aging-Related Female Breast Cancer”. *Nature Communications* 15.1 (2024), p. 5154.

[6] Benz, C. C. “Impact of Aging on the Biology of Breast Cancer”. *Critical Reviews in Oncology/Hematology* 66.1 (2008), pp. 65–74.

[7] Pelissier, F. A., Garbe, J. C., Ananthanarayanan, B., Miyano, M., Lin, C., Jokela, T., Kumar, S., Stampfer, M. R., Lorens, J. B., and LaBarge, M. A. “Age-Related Dysfunction in Mechanotransduction Impairs Differentiation of Human Mammary Epithelial Progenitors”. *Cell Reports* 7.6 (2014), pp. 1926–1939.

[8] Gray, G. K. et al. “A Human Breast Atlas Integrating Single-Cell Proteomics and Transcriptomics”. *Developmental Cell* 57.11 (2022), 1400–1420.e7.

[9] Zirbes, A., Joseph, J., Lopez, J. C., Sayaman, R. W., Basam, M., Seewaldt, V. L., and LaBarge, M. A. “Changes in Immune Cell Types with Age in Breast Are Consistent with a Decline in Immune Surveillance and Increased Immunosuppression”. *Journal of Mammary Gland Biology and Neoplasia* 26.3 (2021), pp. 247–261.

[10] Horvath, S. “DNA Methylation Age of Human Tissues and Cell Types”. *Genome Biology* 14.10 (2013), R115.

[11] Hannum, G. et al. “Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates”. *Molecular Cell* 49.2 (2013), pp. 359–367.

[12] Levine, M. E. et al. “An Epigenetic Biomarker of Aging for Lifespan and Healthspan”. *Aging (Albany NY)* 10.4 (2018), pp. 573–591.

[13] Horvath, S. et al. “Epigenetic Clock for Skin and Blood Cells Applied to Hutchinson Gilford Progeria Syndrome and Ex Vivo Studies”. *Aging* 10.7 (2018), pp. 1758–1775.

[14] Lu, A. T. et al. “DNA Methylation GrimAge Strongly Predicts Lifespan and Healthspan”. *Aging* 11.2 (2019), pp. 303–327.

[15] Farrell, C., Snir, S., and Pellegrini, M. “The Epigenetic Pacemaker: Modeling Epigenetic States under an Evolutionary Framework”. *Bioinformatics* 36.17 (2020). Ed. by Martelli, P. L., pp. 4662–4663.

[16] de Lima Camillo, L. P., Lapierre, L. R., and Singh, R. “A Pan-Tissue DNA-methylation Epigenetic Clock Based on Deep Learning”. *npj Aging* 8.1 (2022), p. 4.

[17] Fransquet, P. D., Wrigglesworth, J., Woods, R. L., Ernst, M. E., and Ryan, J. “The Epigenetic Clock as a Predictor of Disease and Mortality Risk: A Systematic Review and Meta-Analysis”. *Clinical Epigenetics* 11.1 (2019), p. 62.

[18] Xiao, F.-H., Wang, H.-T., and Kong, Q.-P. “Dynamic DNA Methylation During Aging: A “Prophet” of Age-Related Outcomes”. *Frontiers in Genetics* 10 (2019).

[19] Sehl, M. E., Henry, J. E., Storniolo, A. M., Horvath, S., and Ganz, P. A. "The Effects of Lifetime Estrogen Exposure on Breast Epigenetic Age". *Cancer Epidemiology, Biomarkers & Prevention* 30.6 (2021), pp. 1241–1249.

[20] Ambatipudi, S. et al. "DNA Methylome Analysis Identifies Accelerated Epigenetic Ageing Associated with Postmenopausal Breast Cancer Susceptibility". *European Journal of Cancer* 75 (2017), pp. 299–307.

[21] Bell, C. G. et al. "DNA Methylation Aging Clocks: Challenges and Recommendations". *Genome Biology* 20.1 (2019), p. 249.

[22] Sehl, M. E., Henry, J. E., Storniolo, A. M., Ganz, P. A., and Horvath, S. "DNA Methylation Age Is Elevated in Breast Tissue of Healthy Women". *Breast Cancer Research and Treatment* 164.1 (2017), pp. 209–219.

[23] *FastQC: A Quality Control Tool for High Throughput Sequence Data.* URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[24] Chen, S., Zhou, Y., Chen, Y., and Gu, J. "Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor". *Bioinformatics* 34.17 (2018), pp. i884–i890.

[25] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. "STAR: Ultrafast Universal RNA-seq Aligner". *Bioinformatics* 29.1 (2013), pp. 15–21.

[26] Anders, S., Pyl, P. T., and Huber, W. "HTSeq–a Python Framework to Work with High-Throughput Sequencing Data". *Bioinformatics* 31.2 (2015), pp. 166–169.

[27] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data". *Bioinformatics* 26.1 (2010), pp. 139–140.

[28] Zou, H. and Hastie, T. "Regularization and Variable Selection Via the Elastic Net". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320.

[29] Steen, C. B., Liu, C. L., Alizadeh, A. A., and Newman, A. M. "Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx". *Stem Cell Transcriptional Networks.* Ed. by Kidder, B. L. Vol. 2117. New York, NY: Springer US, 2020, pp. 135–157.

[30] Miao, Y.-R., Zhang, Q., Lei, Q., Luo, M., Xie, G.-Y., Wang, H., and Guo, A.-Y. "ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and Its Application in Cancer Immunotherapy". *Advanced Science* 7.7 (2020), p. 1902880.

[31] Teschendorff, A. E. "A Comparison of Epigenetic Mitotic-like Clocks for Cancer Risk Prediction". *Genome Medicine* 12.1 (2020), p. 56.

[32] Avelar, R. A. et al. "A Multidimensional Systems Biology Analysis of Cellular Senescence in Aging and Disease". *Genome Biology* 21.1 (2020), p. 91.

[33] Hänzelmann, S., Castelo, R., and Guinney, J. "GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data". *BMC Bioinformatics* 14.1 (2013), p. 7.

[34] Rosseel, Y. "**Lavaan** : An *R* Package for Structural Equation Modeling". *Journal of Statistical Software* 48.2 (2012).

[35] Zhou, X. and Stephens, M. "Genome-Wide Efficient Mixed-Model Analysis for Association Studies". *Nature Genetics* 44.7 (2012), pp. 821–824.

[36] Sheffield, N. C. and Bock, C. "LOLA: Enrichment Analysis for Genomic Region Sets and Regulatory Elements in R and Bioconductor". *Bioinformatics* 32.4 (2016), pp. 587–589.

[37] Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. "clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters". *OMICS: A Journal of Integrative Biology* 16.5 (2012), pp. 284–287.

[38] Harrell Jr, F. E. *Hmisc: Harrell Miscellaneous.* 2003.

[39] Wei, T. and Simko, V. *Corrplot: Visualization of a Correlation Matrix.* 2010.

[40] Benjamini, Y. and Hochberg, Y. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300. JSTOR: 2346101. URL: https://www.jstor.org/stable/2346101.

[41] Sehl, M. E., Guo, W., Farrell, C., Marino, N., Henry, J. E., Storniolo, A. M., Papp, J., Li, J. J., Horvath, S., Pellegrini, M., et al. “Systematic dissection of epigenetic age acceleration in normal breast tissue reveals its link to estrogen signaling and cancer risk”. *bioRxiv* (2024).

[42] Eraslan, G. et al. “Single-Nucleus Cross-Tissue Molecular Reference Maps toward Understanding Disease Gene Function”. *Science* 376.6594 (2022), eabl4290.

[43] Hu, Y., Hu, Q., Li, Y., Lu, L., Xiang, Z., Yin, Z., Kabelitz, D., and Wu, Y. “Γδ T Cells: Origin and Fate, Subsets, Diseases and Immunotherapy”. *Signal Transduction and Targeted Therapy* 8.1 (2023), pp. 1–38.

[44] Stork, C. T., Bocek, M., Crossley, M. P., Sollier, J., Sanz, L. A., Chédin, F., Swigut, T., and Cimprich, K. A. “Co-Transcriptional R-loops Are the Main Cause of Estrogen-Induced DNA Damage”. *eLife* 5 (2016). Ed. by Aguilera, A., e17548.

[45] Yang, J.-H. et al. “Loss of Epigenetic Information as a Cause of Mammalian Aging”. *Cell* 186.2 (2023), 305–326.e27.

[46] Jaffe, A. E. and Irizarry, R. A. “Accounting for Cellular Heterogeneity Is Critical in Epigenome-Wide Association Studies”. *Genome Biology* 15.2 (2014), R31.

[47] El Khoury, L. Y. et al. “Systematic Underestimation of the Epigenetic Clock and Age Acceleration in Older Subjects”. *Genome Biology* 20.1 (2019), p. 283.

[48] Lin, J., Ye, S., Ke, H., Lin, L., Wu, X., Guo, M., Jiao, B., Chen, C., and Zhao, L. “Changes in the Mammary Gland during Aging and Its Links with Breast Diseases”. *Acta Biochimica et Biophysica Sinica* 55.6 (2023), pp. 1001–1019.

[49] Shams, A. “Re-Evaluation of the Myoepithelial Cells Roles in the Breast Cancer Progression”. *Cancer Cell International* 22.1 (2022), p. 403.

[50] Gérard, C. and Brown, K. A. “Obesity and Breast Cancer – Role of Estrogens and the Molecular Underpinnings of Aromatase Regulation in Breast Adipose Tissue”. *Molecular and Cellular Endocrinology* 466 (2018), pp. 15–30.

[51] Liu, Z.-L., Chen, H.-H., Zheng, L.-L., Sun, L.-P., and Shi, L. “Angiogenic Signaling Pathways and Anti-Angiogenic Therapy for Cancer”. *Signal Transduction and Targeted Therapy* 8.1 (2023), pp. 1–39.

[52] Yager, J. D. and Davidson, N. E. “Estrogen Carcinogenesis in Breast Cancer”. *New England Journal of Medicine* 354.3 (2006), pp. 270–282.

[53] López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. “Hallmarks of Aging: An Expanding Universe”. *Cell* 186.2 (2023), pp. 243–278.

[54] Negrini, S., Gorgoulis, V. G., and Halazonetis, T. D. “Genomic Instability — an Evolving Hallmark of Cancer”. *Nature Reviews Molecular Cell Biology* 11.3 (2010), pp. 220–228.

[55] Valencia, C. I., Saunders, D., Daw, J., and Vasquez, A. “DNA Methylation Accelerated Age as Captured by Epigenetic Clocks Influences Breast Cancer Risk”. *Frontiers in Oncology* 13 (2023), p. 1150731.

[56] Sevdalis, A., Deng, X., Bandyopadhyay, D., and McGuire, K. P. “The Value of Tyrer-Cuzick Versus Gail Risk Modeling in Predicting Benefit from Screening MRI in Breast Cancer”. *European Journal of Breast Health* 18.1 (2022), pp. 79–84.

CHAPTER 5

Conclusion and future directions

5.1 Conclusion

Data simulation and data analysis are complementary aspects of research that together provide a comprehensive framework for understanding complex systems and processes. Simulation focuses on generating realistic data based on predefined truths (generative), while analysis seeks to uncover the underlying truths from observed data (inferential). By addressing key challenges in simulating bisulfite sequencing data and analyzing DNA methylome changes associated with disease and aging, this dissertation aims to advance the understanding of DNA methylation, enabling the development of more reliable computational tools, uncovering novel biological insights, and paving the way for scientific research and clinical applications.

In chapter 2, the development of BSReadSim addressed key limitations in existing bisulfite sequencing simulators by integrating genetic variants and methylation profiles while accurately modeling biological and technical variations. BSReadSim demonstrated high fidelity in replicating reference genetic and methylation profiles, establishing itself as a robust platform for generating realistic bisulfite sequencing data. Its versatility enhances its utility for benchmarking bioinformatics tools and optimizing experimental designs, ultimately improving the reliability and rigor of computational methods for DNA methylation analysis.

chapter 3 introduced a novel approach for type 2 diabetes (T2D) biomarker study by validating saliva DNA methylation as a non-invasive indicator. The study combined Whole Genome Bisulfite Sequencing (WGBS) and Targeted Bisulfite Sequencing (TBS) to identify and profile diabetes-specific epigenetic signals in saliva, uncovering molecular alterations and cellular dynamics associated with T2D. These findings demonstrate, for the first time, the potential of saliva DNA methylation as a cost-effective and accessible biomarker, paving the way for personalized and non-invasive diagnostics and monitoring.

In chapter 4, the study of epigenetic clocks in normal breast tissue offered critical insights into the interplay between epigenetic aging, cell composition, and breast cancer risk. This research underscored the importance of tissue-specific clocks for accurate age prediction and revealed systematic biases in existing epigenetic clocks. The findings demonstrated a link between epigenetic age acceleration and changes in cell composition, particularly those associated with increased breast cancer risk, highlighting the potential of epigenetic clocks for cancer risk assessment. Furthermore, the study provided molecular evidence connecting estrogen exposure, accelerated epigenetic aging, and heightened cancer susceptibility.

5.2 Future work

While this dissertation presents several advancements, some future work can further enhance the impact and applicability of the findings.

1. Application of BSReadSim to benchmark computational tools: Future work will focus on using BSReadSim to benchmark aligners and SNP callers under realistic scenarios. This approach will enable a detailed evaluation of these tools' performance in handling bisulfite sequencing data. Building on these benchmarks, BSReadSim will be extended to assess tools for detecting allele-specific methylation (ASM), leveraging its capability to simulate ASM scenarios. These efforts will provide critical insights into the strengths and limitations of existing tools, advancing their development and application in epigenetics research.

2. Expansion of Saliva DNA Methylation Research Future studies should fully unlock saliva DNA methylation's potential as a biomarker for T2D. They should also include larger, more diverse cohorts and refine probe designs to target a broader spectrum of relevant methylation sites. Additionally, further cost reductions through advanced sequencing techniques, such as barcoding and multiplexing, would make this approach more feasible for large-scale epidemiological studies and routine clinical diagnostics.

3. Refinement and Clinical Application of Epigenetic Clocks The insights gained from studying epigenetic clocks in breast tissue suggest several directions for future research. Longitudinal studies are needed to better understand the dynamics of epigenetic aging and its relationship with cancer risk over time. Additionally, single-cell DNA methylation and transcriptomic profiling could provide a more detailed understanding of the interplay between epigenomic changes and cellular composition during aging. These advancements could refine epigenetic clocks, enhancing their clinical utility for early detection and prevention of age-related diseases, including breast cancer.

4. Integration of Multi-Omics Approaches Future research should explore integrating DNA methylation data with other omics layers, such as transcriptomics, proteomics, and metabolomics, to provide a more comprehensive understanding of the molecular mechanisms underlying aging and disease. Multi-omics approaches could reveal new biomarkers, therapeutic targets, and insights into the complex regulatory networks that govern cellular processes in health and disease.

The research presented in this dissertation highlights the pivotal role of DNA methylation in unraveling complex biological phenomena, including disease and aging. By developing innovative tools and exploring noninvasive biomarkers, this work establishes a solid foundation for future advancements in both research and clinical applications. As the field of epigenomics continues to advance, the methodologies and insights provided by this dissertation will serve as valuable resources for ongoing efforts to uncover the epigenetic mechanisms underlying health and disease.