scWGBS-GPT: A Foundation Model for Capturing Long-Range CpG Dependencies in Single-Cell Whole-Genome Bisulfite Sequencing to Enhance Epigenetic Analysis

Chaoqi Liang[1,9,*], Peng Ye[2,3,*], Hongliang Yan[9,*], Peng Zheng[2,*], Jianle Sun[2,4], Yanni Wang[9], Yu Li[9], Yuchen Ren[2,5], Yuanpei Jiang[9], Junjia Xiang[1], Sizhe Zhang[1], Linle Jiang[9], Weiqiang Bai[2], Xinzhu Ma[2,3], Tao Chen[6], Wangmeng Zuo[1,†], Lei Bai[2,†], Wanli Ouyang[2,†], Jia Li[7,8,9,†]

[1]Harbin Institute of Technology, [2]Shanghai Artificial Intelligence Laboratory,
[3]The Chinese University of Hong Kong, [4]Carnegie Mellon University,
[5]The University of Sydney, [6]Fudan University
[7]State Key Laboratory of Respiratory Disease, Guangzhou Institute of Cancer Research,
the Affiliated Cancer Hospital, Guangzhou Medical University,
[8]Department of Laboratory Medicine, National Clinical Research Center for Respiratory Disease,
National Center for Respiratory Medicine, Guangzhou Institute of Respiratory Health,
The First Affiliated Hospital of Guangzhou Medical University, Guangzhou Medical University,
[9]Guangzhou National Laboratory
[*]Equal contribution, [†]Corresponding author

ABSTRACT

Single-Cell Whole-Genome Bisulfite Sequencing (scWGBS) is a powerful technique for profiling DNA methylation at single-cell and single-nucleotide resolution, providing critical insights into epigenetic regulation in development, disease, and cellular heterogeneity. However, analyzing scWGBS data remains a significant challenge due to ultra-long genomic sequences with millions of CpG sites and sparse, stochastic coverage. Here, we introduce scWGBS-GPT, the first generative foundational language model specifically designed for high-resolution analysis of scWGBS data at the single-CpG level. By combining CpG special token design, Mamba backbone and Cross-Attention head, scWGBS-GPT efficiently processes ultra-long sequences while preserving both local CpG interactions and broader genomic context, enabling biologically meaningful interpretations. Pretrained on 1,000,000 single cells from 28 human and mouse tissues, scWGBS-GPT reconstructs sparse methylation landscapes, enhancing the resolution and accuracy of epigenetic analyses. It significantly outperforms existing methods in key biomedical applications—improving cell clustering to uncover previously unrecognized cellular heterogeneity, enhancing trajectory inference to map precise differentiation pathways, and automating the identification of critical epigenetic biomarkers relevant to disease progression and therapeutic targeting. These advancements establish scWGBS-GPT as a transformative tool in single-cell epigenomics, setting a new standard for DNA methylation analysis and unlocking novel insights into epigenetic mechanisms underlying health and disease. The code for scWGBS-GPT is available at https://github.com/ChaoqiLiang/scWGBS-GPT.

Highlights:

- scWGBS-GPT is the first foundation language model for scWGBS data analysis, trained on over 1,000,000 single cells from 28 human and mouse tissues, spanning 221 cell types.
- scWGBS-GPT combines CpG special token design, Mamba backbone, and Cross-Attention head for efficient processing of up to 2 million CpG sites per cell.

- scWGBS-GPT achieves State-of-the-Art Performance in Cell Clustering, Annotation, Pseudotime Inference.
- scWGBS-GPT is compatible with bulk WGBS data and excels in deconvoluting cell-free DNA methylation data, precisely inferring the tissue or organ of origin, enhancing non-invasive diagnostic capabilities.

1 INTRODUCTION

The rapid advancement of single-cell data analysis technologies has revolutionized our understanding of cellular heterogeneity, lineage differentiation, and epigenetic regulation. At the forefront of this revolution are generative language models, particularly Transformer-based architectures, which have transformed single-cell omics research. Models like scBERT (Yang et al., 2022), scGPT (Cui et al., 2024) and scFoundation (Hao et al., 2024) have demonstrated state-of-the-art performance in single-cell RNA sequencing (scRNA-seq) analysis, excelling in tasks such as cell clustering, annotation and gene perturb prediection. These models leverage powerful deep learning techniques to extract complex interrelations within single-cell data, significantly improving our ability to interpret biological systems. However, Transformer-based language models face a critical limitation: the computational complexity of their self-attention mechanism scales quadratically with sequence length, restricting their applicability to datasets with sequences longer than a few thousand words or elements. This limitation excludes many important types of biological data, such as single-cell whole-genome methylation data, from benefiting fully from these transformative approaches.

Single-Cell Whole-Genome Bisulfite Sequencing (scWGBS) (Farlik et al., 2015; Mulqueen et al., 2018) is a state-of-the-art technique that profiles DNA methylation at single-nucleotide and single-cell resolution, providing unparalleled insights into whole-genome epigenetic regulation. scWGBS reveals cell-to-cell variability, enabling a deeper understanding of developmental processes, cellular heterogeneity, and disease-associated epigenetic changes. The technique involves bisulfite treatment of DNA to distinguish methylated from unmethylated cytosines, followed by high-throughput sequencing to generate methylation data spanning hundreds of thousands to millions of CpG sites per cell. This comprehensive dataset holds immense potential for understanding epigenetic regulation. But their vast length and sparse coverage make them computationally prohibitive for Transformer-based language models.

Despite its promise, the analysis of scWGBS data presents significant computational and analytical challenges. The vast length of the sequences—often spanning millions of CpG sites—and their sparse, variable coverage render traditional machine learning approaches inadequate. Current methods (Farlik et al., 2016; Kremer et al., 2024) typically divide the genome into fixed regions (e.g., 10-kb windows) and calculate regional methylation rates. While this approach is computationally manageable, it sacrifices single-nucleotide precision and fails to capture the intricate dependencies between CpG sites. Moreover, these methods overlook the broader genomic context, leading to substantial information loss and limiting their ability to uncover subtle but critical biological mechanisms.

In response to this challenge, we developed a generative foundational language model named scWGBS-GPT. scWGBS-GPT fuses three key components: the CpG special token design, Mamba's Selective State Space Models (SSMs) (Gu & Dao, 2023), and Cross-Attention head (Gheini et al., 2021), specifically designed to analyze scWGBS data at single-CpG resolution. The Mamba architecture, grounded in SSMs, is specifically designed to address the challenges associated with processing extremely long input sequences. The SSMs framework allows the model to selectively focus on biologically relevant portions of the input sequence, effectively filtering out less informative regions to reduce computational complexity. Additionally, the CpG special token design centers on a K-mer strategy around CG motifs to integrate the DNA background, while incorporating methylation information by multiplying the tokens with the methylation rate before inputting them into the model. The incorporation of Cross-Attention head in the final layer enables dynamic integration of nucleotide-level context, allowing the model to capture both local and long-range dependencies between CpG sites. This fusion is particularly well-suited to the biological significance of CpG methylation, where interactions between distant sites often play critical roles in gene regulation and epigenetic mechanisms.

By leveraging this hybrid architecture, scWGBS-GPT efficiently handles the ultra-long sequences characteristic of scWGBS data, where individual cells contain an average of 2 million CpG sites and the longest sequences reach up to 12 million. Pre-trained on data from over 1,000,000 single cells spanning 28 different tissues from humans and mice (Liu et al., 2021; 2022; Tian et al., 2023; Loyfer et al., 2023), the model benefits from a comprehensive and diverse representation of DNA methylation patterns across species and biological contexts. This extensive pre-training enables scWGBS-GPT to capture intricate patterns within sparse methylation data, preserving both local CpG interactions and broader genomic context. This innovative design ensures scWGBS-GPT not only overcomes the computational bottlenecks of traditional

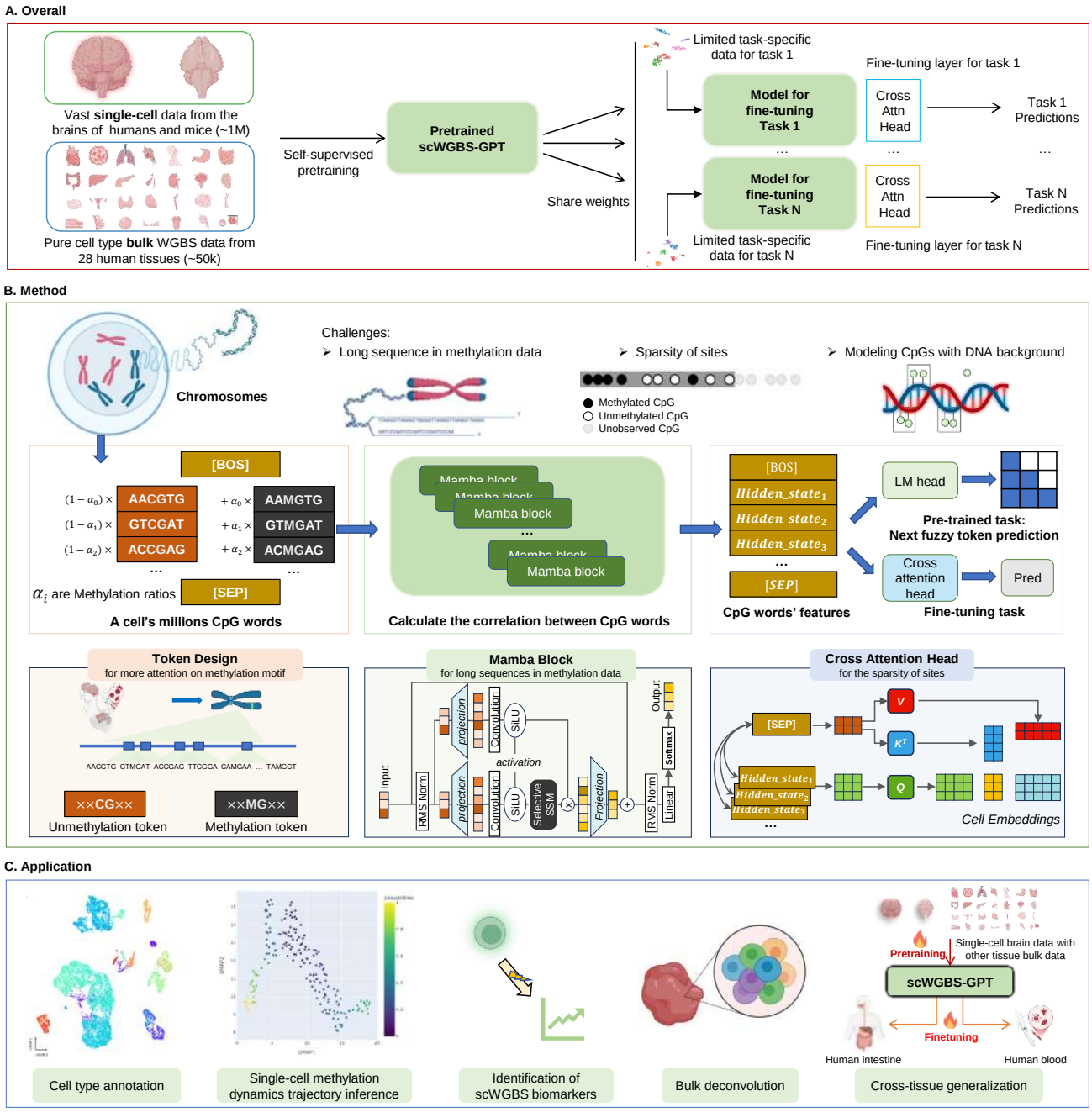

Figure 1. **Schematic of development and applications of scWGBS-GPT. a,** The overall pipeline of establishing a foundation model, including an initial self-supervised pretraining on a large-scale single-cell dataset of multiple tissues and the fine-tuning process with limited downstream data. In the fine-tuning procedure, additional layers are added to the model with shared weights to be fine-tuned on task-specific datasets for various downstream tasks. Since the model has learned strong generalization capabilities through initial large-scale pretraining, it can easily adapt to various downstream tasks with a small amount of fine-tuning data. **b,** The architecture and pretraining process of the proposed scWGBS-GPT. We highlight the main components used to deal with the challenges in building a foundation model for scWGBS, including the Mamba blocks for long sequences in methylation data, cross attention mechanism for the important CpG sites, and token design to modeling CpGs with DNA background. **c,** Main applications of scWGBS-GPT: 1) cell type annotation, 2) dynamics trajectory inference, 3) Biological interpretability analysis, and 4) bulk deconvolution. 5) cross-tissue generalization ability.

methods but also provides biologically meaningful representations of DNA methylation across the entire genome. With its ability to represent sparse, stochastic methylation patterns at unprecedented resolution, scWGBS-GPT enables precise insights into cellular heterogeneity and epigenetic regulation across a wide range of biological systems.

Our research highlights that scWGBS-GPT significantly addresses the limitations of existing scWGBS analysis methods by leveraging the power of a generative language model. In cell annotation tasks, scWGBS-GPT achieves over 94% accuracy, surpassing previous approaches by offering more distinct and biologically relevant groupings of cells, while uncovering subtle aspects of cellular heterogeneity. In trajectory inference, the model excels in accurately mapping the differentiation pathways of cells, enabling detailed and continuous representations of cell state transitions. Moreover, scWGBS-GPT automatically identifies key single-cell methylation features, providing researchers with an invaluable tool for uncovering critical epigenetic markers. Collectively, these advancements deepen our understanding of DNA methylation at the single-cell level and establish a new benchmark for analyzing ultra-long genomic sequences in epigenetic research.

2 RESULTS

2.1 The overview of scWGBS-GPT

We introduce scWGBS-GPT (as shown in 1), the first CpG language model designed to analyze single-cell methylation data with unprecedented precision and scalability. This work aims to leverage language models to decode DNA CpG site epigenetic modification information at single-cell resolution. Each CpG site, along with its surrounding nucleotides, forms a CpG motif—"XXC(M)GXX", where "X" represents any nucleotide (Zhou et al., 2018). In specific, we treat the CpG motif as a "word" and, based on this concept, construct a CpG language. We then use this framework to develop a CpG language processing model—scWGBS-GPT.

scWGBS-GPT adopts the cutting-edge Mamba architecture (Gu & Dao, 2023), specifically designed to handle the immense scale and complexity of scWGBS data, including millions of CpG words across ultra-long genomic sequences. Leveraging Selective State Space Models (SSMs) and a modular, scalable design, Mamba excels in computational efficiency through dynamic computation graphs and advanced memory optimization, minimizing memory usage while maximizing performance even on massive and sparse datasets. A key innovation is the introduction of a novel "CpG fuzzy embedding" approach, where each CpG site's embedding is a weighted sum of methylation and unmethylation motifs, reflecting their respective methylation rates to preserve probabilistic epigenetic information. To further enhance sequence structure, special tokens [BOS] and [SEP] are introduced to mark dataset boundaries, enabling effective management of variable-length sequences. The model's defining feature, Cross-Attention (Gheini et al., 2021), allows the final hidden state of the [SEP] token to dynamically attend to the hidden states of all CpG sites, integrating both local and global dependencies to generate expressive single-cell representations. This innovative design not only captures the biological complexity of DNA methylation patterns but also scales seamlessly for diverse applications, from cell annotation and differentiation trajectory inference to cross-species and cross-tissue studies, making scWGBS-GPT a groundbreaking tool in single-cell epigenomics.

In the pretraining stage, scWGBS-GPT employs a novel **next fuzzy token prediction** strategy, an extension of traditional next token prediction tasks (Radford et al., 2018) tailored for DNA methylation data. At each genomic position, the model simultaneously predicts both the methylation and unmethylation tokens, with the loss weights dynamically adjusted based on the methylation rate and its complement (1-methylation rate). This design effectively captures the probabilistic nature of CpG methylation, enabling the model to learn nuanced epigenetic patterns from sparse and complex single-cell data. During fine-tuning, the final representation of the `[SEP]` token, obtained through a Cross-Attention head across all input tokens, serves as a comprehensive single-cell embedding. This embedding integrates both local and long-range dependencies between CpG sites, facilitating downstream tasks such as cell annotation, differentiation trajectory inference, and disease state prediction with high precision and biological relevance.

To leverage the scaling law of language models, we have curated a comprehensive cross-species dataset of one million single-cell methylation profiles from humans and mice, encompassing a diverse range of tissues. Our dataset includes methylation profiles from all 28 mouse tissue types and various human tissues, representing one of the most extensive collections for single-cell methylation studies. A total of 900,000 single-cell methylation profiles were collected from publicly available data repositories, including the NCBI Gene Expression Omnibus and Sequence Read Archive (Liu et al., 2021; 2022; Tian et al., 2023). Among these, 400,000 profiles are from mouse brain regions, while 500,000 profiles are derived from human brain regions. To further enhance tissue diversity, we incorporated bulk WGBS data from over 20 different tissues and organs, focusing on purified cell types (Loyfer et al., 2023). This approach allows us to approximate

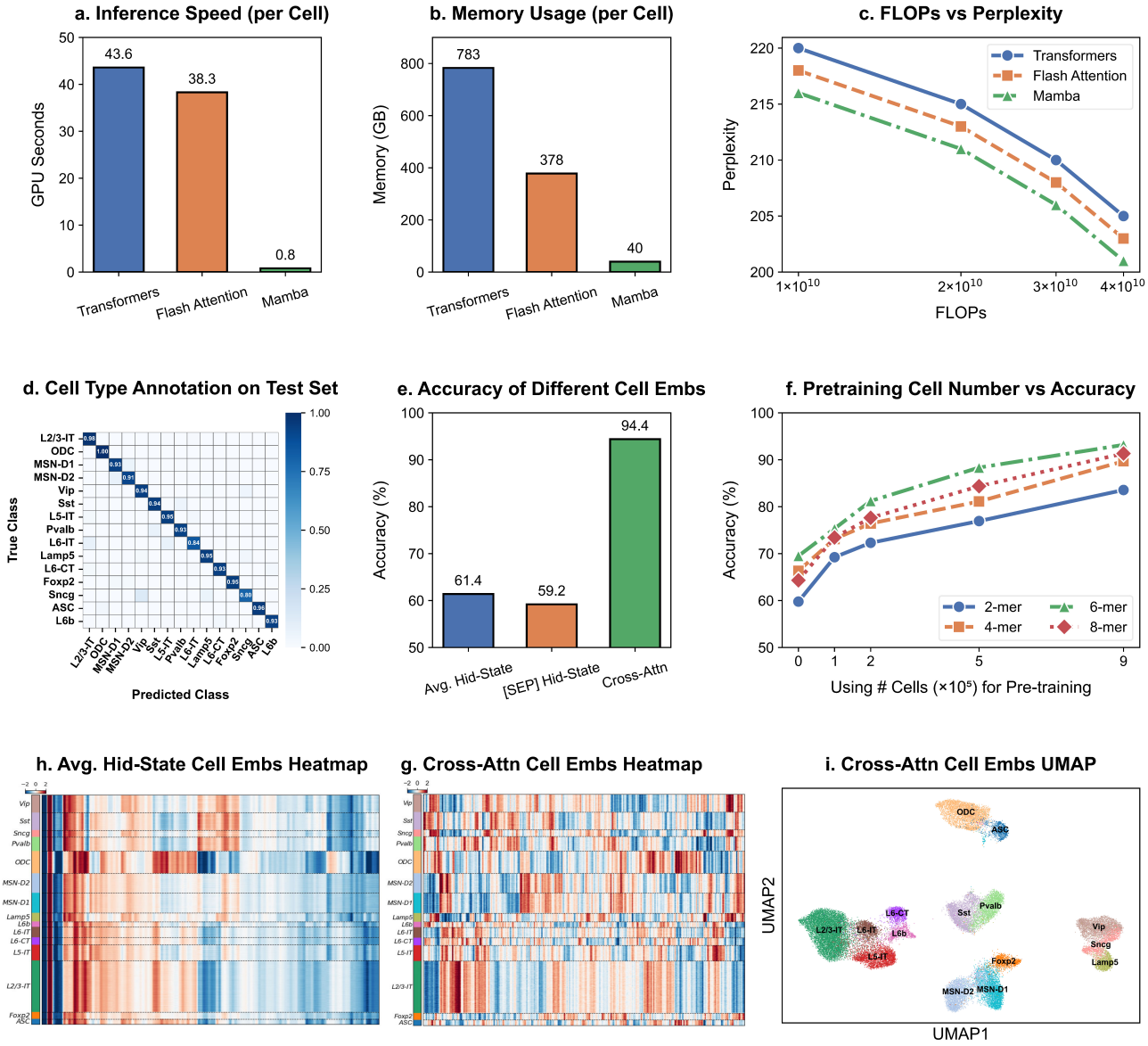

Figure 2. **Performance and embedding analysis of the scWGBS-GPT model for single-cell annotation. a,** Inference speed per cell: Mamba significantly reduces computational cost in comparison to Transformers and Flash Attention. **b,** Memory usage per cell: Mamba requires substantially less GPU memory for processing ultra-long scWGBS sequences. **c,** FLOPs vs perplexity: Mamba achieves the lowest perplexity at similar computational cost, demonstrating its efficiency in modeling scWGBS data. **d,** Cell type annotation accuracy on the test set: The confusion matrix shows high concordance between predicted and true cell types. **e,** Comparison of different cell embeddings: Cross-attention embeddings yield the highest annotation accuracy (94.4%), outperforming hidden-state-based representations. **f,** Pretraining cell number vs accuracy: Larger pretraining datasets improve annotation accuracy across different k-mer CpG motif strategies. **g-h**, Heatmaps of cell embeddings: Cross-attention embeddings (g) provide better separation of cell types compared to average hidden-state embeddings (h). **i**, UMAP visualization of cross-attention embeddings: Cells are well-clustered according to their biological identities, indicating strong representation learning.

these bulk datasets to scWGBS profiles, contributing approximately 50,000 samples. On average, each single-cell sample yields 1 million "CpG words," providing a rich resource for training. These datasets, encompassing methylation profiles from diverse human and mouse tissues, offer a comprehensive foundation for training models to learn tissue-specific methylation patterns and their regulatory roles in gene expression. By incorporating single-cell resolution and cross-species data, our dataset enables the model to capture cellular heterogeneity and identify key regulatory marks across the genome, providing valuable insights into epigenetic regulation and disease mechanisms.

2.2 Evaluation of scWGBS-GPT in Single-Cell Methylation-Based Cell Type Annotation

Before applying our model to cell-type annotation in single-cell whole-genome bisulfite sequencing (scWGBS) data, we conducted a comprehensive evaluation of three mainstream sequence modeling frameworks—Transformers, Flash Attention, and Mamba—in the context of scWGBS data (as shown in Fig.2a, b and c). Our findings reveal that Mamba significantly outperforms the other frameworks in terms of inference speed and memory utilization, making it exceptionally well-suited for handling ultra-long scWGBS sequences, which can exceed millions of tokens. Notably, in our custom-designed Next Fuzzy Token Prediction pretraining task, Mamba achieved lower perplexity with fewer floating-point operations (FLOPs), demonstrating superior scalability for large-scale epigenomic analyses. Based on these results, we selected the Mamba architecture as the backbone of scWGBS-GPT for subsequent cell-type annotation in scWGBS data. In the following, we present a detailed assessment of performance on this annotation task.

Cell annotation is a fundamental step in single-cell data analysis, aiming to classify each cell into a specific cell type or state based on its molecular features. This task serves as a primary benchmark to evaluate the performance of single-cell models. To assess our model's annotation ability, we utilized the scWGBS dataset from Tian et al. (2023), which provides cell-type labels for over 200,000 single cells and its cell type labels derived from human brain tissue. The data were split into training, validation, and test sets at a 6:2:2 ratio, leading to a final annotation accuracy of 94.4% on test set. Fig. 2d offers a detailed view of our annotation performance via a confusion matrix for a range of neuronal subtypes (e.g., L2/3-IT, MSN-D1, Vip). Each subtype's recall exceeded 90%, underscoring the robustness of our model across diverse cell populations. Notably, some subtypes (e.g., ODC, ASC) exceeded 95% accuracy, indicating that the model effectively captures subtype-specific methylation signatures. These findings highlight the utility of our single-cell annotation framework for large-scale scWGBS data and confirm its suitability for high-resolution cellular analyses in the brain.

We explored multiple embedding strategies to distill cell-type–specific features, focusing on both the representation of individual CpGs and the integration of these representations at the whole-genome level. To model each CpG, we drew on the findings of Zhou et al. (2018), who reported that a 4-mer centered on "CG" is biologically informative. Building on this, we introduced extended k-mers around each CpG site to capture local DNA context. Specifically, our model represents each CpG by incorporating flanking nucleotides on both sides, yielding k-mers ranging from 2-mer (bare CpG) to 8-mers (i.e., ± 3 nucleotides on each side). As shown in Fig. 2f, we compared these k-mer representations under varying numbers of pretraining cells and observed a clear scaling law: while performance improves substantially all k-mers, 6-mers offer the most balanced and robust representation. To obtain a single embedding for each cell, we evaluated three strategies: (i) taking the hidden-state average across all CpG tokens, (ii) extracting the final hidden state of a special `[SEP]` token, and (iii) employing a cross-attention head to automatically weight the most informative CpG k-mers. As shown in Fig. 2e, our results show that the conventional averaging or `[SEP]`-only approaches yield classification accuracies of 59.2% and 61.4%, respectively—substantially lower than the 94.4% achieved with cross-attention–based embeddings.

To further elucidate this gap in performance, Figs. 2g and 2h compare the heatmaps generated from averaging-based versus cross-attention–based embeddings, respectively. Unlike the averaging-based heatmap, which concentrates most variation in the early feature dimensions and struggles to distinguish certain cell types, the cross-attention embedding clearly separates different types, reflecting the model's ability to focus on biologically relevant CpG sites. Moreover, we visualized the cross-attention–based embeddings using UMAP, as shown in Fig. 2i. The resulting clusters are well-defined, reflecting the richness of the learned embeddings. Cell types with closer biological relationships (e.g., L2/3, L5, and L6 excitatory neurons) lie near one another in the UMAP space, whereas more divergent cell populations are positioned farther apart. The clear separation underscores both the high annotation accuracy of our approach and its ability to capture key biological distinctions among single cells. These all results underscore the synergy between Mamba's scalability for ultra-long scWGBS CpG words and the capacity of cross-attention to highlight crucial CpG words, capturing biologically meaningful patterns essential for accurate single-cell annotation.

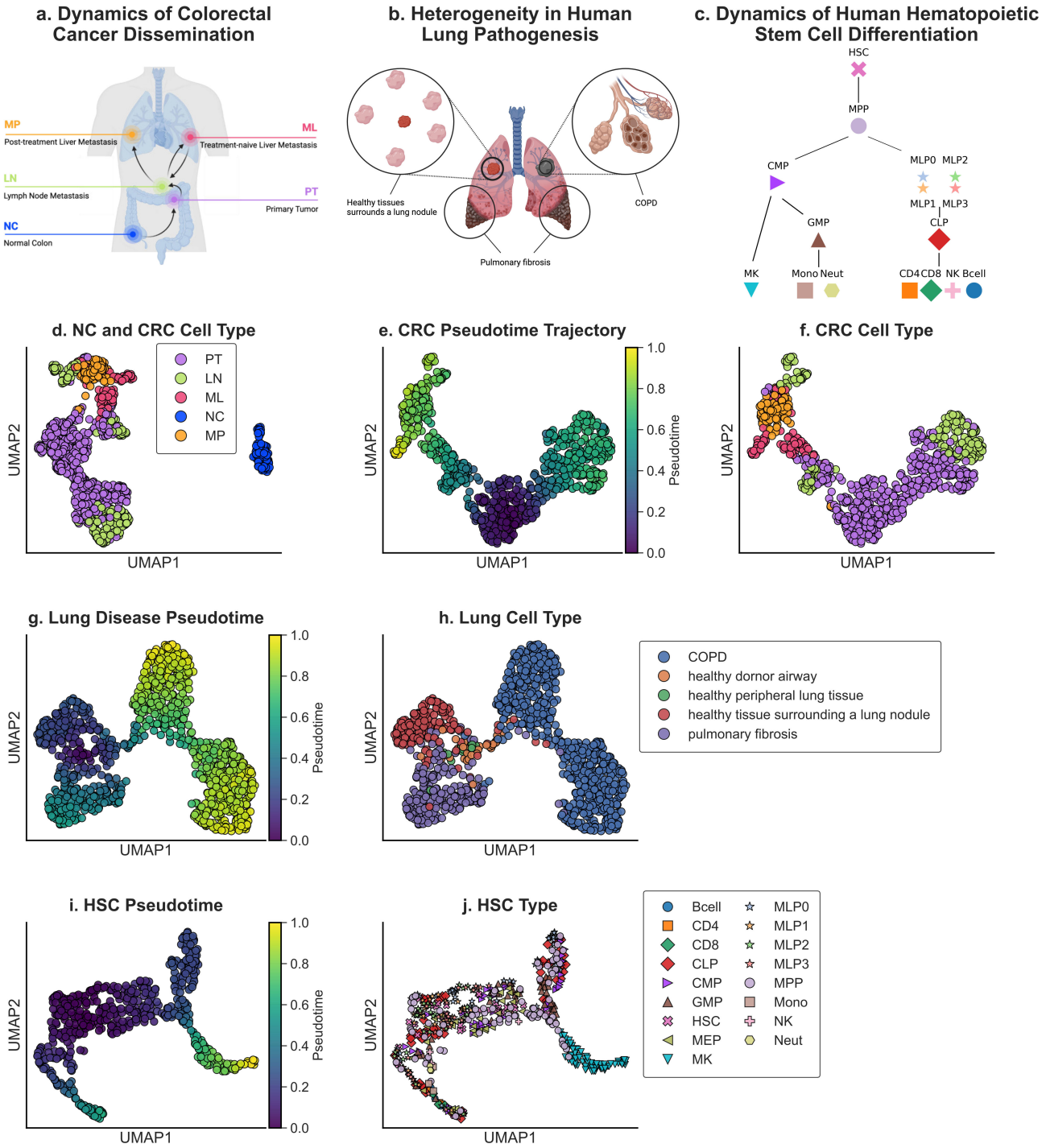

Figure 3. **Single-cell methylation lineage dynamics revealed by scWGBS-GPT. a,** Schematic representation of colorectal cancer (CRC) dissemination, depicting primary tumors (PT), normal colon (NC), lymph node metastasis (LN), post-treatment liver metastasis (MP), and treatment-naïve liver metastasis (ML). **b,** Illustration of cellular heterogeneity in human lung pathogenesis, highlighting COPD and pulmonary fibrosis in different lung microenvironments. **c,** Diagram of human hematopoietic stem cell (HSC) differentiation, showing the hierarchical progression from HSC to multiple progenitor and immune cell lineages. **d,** UMAP visualization of normal colon (NC) and CRC cell types, demonstrating scWGBS-GPT's ability to distinguish between different metastatic and primary tumor states. **e,** CRC pseudotime trajectory inferred using scWGBS-GPT, capturing the progression of CRC cells along a developmental path. **f,** UMAP representation of CRC cell types, showing clear separation of metastatic and tumor-originating populations. **g,** Lung disease pseudotime trajectory, illustrating the progression of cells along a disease continuum modeled by scWGBS-GPT. **h,** UMAP of lung cell types, highlighting distinct clustering of COPD, healthy airway, and pulmonary fibrosis cells. **i,** Pseudotime trajectory of hematopoietic stem cells (HSC), capturing the differentiation process as learned by scWGBS-GPT. **j,** UMAP of HSC differentiation, showing distinct immune and progenitor cell types along the differentiation spectrum, further validating the model's ability to capture biologically meaningful cell transitions.

2.3 Cross-tissue Generalization and Single-cell Analysis Using scWGBS-GPT

Large-scale pretrained language models have shown strong few-shot learning abilities. OpenAI's GPT-3 study (Brown et al., 2020) demonstrated that models trained on large datasets can adapt to new tasks with minimal fine-tuning. Similarly, scWGBS-GPT, after being extensively pretrained on large-scale single-cell methylation data, can be adapted to diverse tissue types using only a small number of fine-tuning samples—typically ranging from a few hundred to a thousand single cells per dataset. To evaluate its generalizability, we applied scWGBS-GPT to single-cell methylation data from three distinct biological systems that were absent from the pretraining dataset: colorectal cancer (CRC), lung disease, and hematopoietic stem cell (HSC) differentiation. We fine-tuned the model using 1,186 human CRC scWGBS samples, 1,094 lung scWGBS samples, and 639 HSC scWGBS samples, splitting each dataset into a 7:3 ratio for training and testing, respectively. This experimental design allowed us to assess the model's cross-tissue transferability and its ability to perform single-cell annotation across vastly different biological contexts (Fig. 3).

Colorectal cancer progression: We first evaluated scWGBS-GPT on single-cell methylation profiles from normal colon (NC) and colorectal cancer (CRC) samples across different metastatic stages, using publicly available scWGBS data from Bian et al. (2018). As shown in Fig. 3d, our model successfully distinguished primary tumors (PT), lymph node metastases (LN), treatment-naïve liver metastases (ML), and post-treatment liver metastases (MP), revealing distinct single-cell clusters. To further explore the progression of CRC, we performed pseudotime trajectory inference (Fig. 3e), where scWGBS-GPT captured a clear developmental path of tumor evolution, reflecting the transition from primary CRC to metastatic states. This finding was further supported by cell-type classification results (Fig. 3f), which demonstrated our model's ability to recover CRC subtype-specific methylation signatures.

Lung disease heterogeneity: To assess the model's effectiveness in capturing epigenetic heterogeneity in human lung diseases, we analyzed in-house generated scWGBS data from chronic obstructive pulmonary disease (COPD), pulmonary fibrosis, and healthy lung tissue. The pseudotime trajectory inferred by scWGBS-GPT (Fig. 3g) suggests a continuum of cell states from normal lung cells to diseased states, highlighting an epigenetic trajectory associated with disease progression. The UMAP projection of lung cell types (Fig. 3h) further confirms the model's ability to resolve distinct disease-related cell populations, with clear separation between healthy lung cells, cells surrounding lung nodules, and diseased cells. This indicates that scWGBS-GPT effectively extracts biologically meaningful embeddings that align with known disease phenotypes.

Hematopoietic stem cell differentiation: We applied scWGBS-GPT to human hematopoietic stem cells (HSCs) and their differentiated progeny, using scWGBS data from Farlik et al. (2016), to evaluate its ability to infer cell differentiation trajectories. The pseudotime trajectory inferred from scWGBS-GPT embeddings (Fig. 3i) captures a well-defined lineage hierarchy, from HSCs to committed progenitors and fully differentiated immune cell types. The corresponding UMAP visualization (Fig. 3j) demonstrates that cells are organized according to their differentiation states, with closely related immune cell types clustering together. Notably, B cells, T cells (CD4/CD8), and natural killer (NK) cells are positioned in adjacent regions, reflecting their shared developmental origin, whereas monocytes (Mono) and granulocyte-macrophage progenitors (GMP) are located further apart, consistent with known hematopoietic differentiation trajectories.

These results highlight the ability of scWGBS-GPT to accurately annotate cell types, infer cell state transitions, and reconstruct differentiation trajectories from single-cell methylation data. The model's strong few-shot transferability allows it to generalize across tissues with minimal fine-tuning, making it a powerful tool for cross-tissue single-cell epigenomic analysis.

2.4 Interpretable Analysis of scWGBS-GPT and Identification of Tumor-Specific scWGBS Biomarkers

Lager language model for single-cell methylation analysis must not only achieve high predictive performance but also provide biologically meaningful insights. To evaluate the interpretability of scWGBS-GPT, we analyzed its attention weight distribution across different colorectal cancer (CRC) progression stages. Our results demonstrate that scWGBS-GPT effectively identifies tumor-specific hypomethylation biomarkers, highlighting its ability to extract biologically relevant CpG sites. Furthermore, we observed distinct attention weight patterns between primary tumors and metastatic tumors, suggesting that scWGBS-GPT captures key epigenetic differences associated with tumor progression.

To understand how scWGBS-GPT prioritizes CpG sites, we examined the relationship between methylation ratios and model attention weights in normal colon (NC) and primary tumor (PT) samples (Fig.4a,b). In normal colon samples, attention weights remain uniformly low across different methylation levels, indicating that the model does not assign high

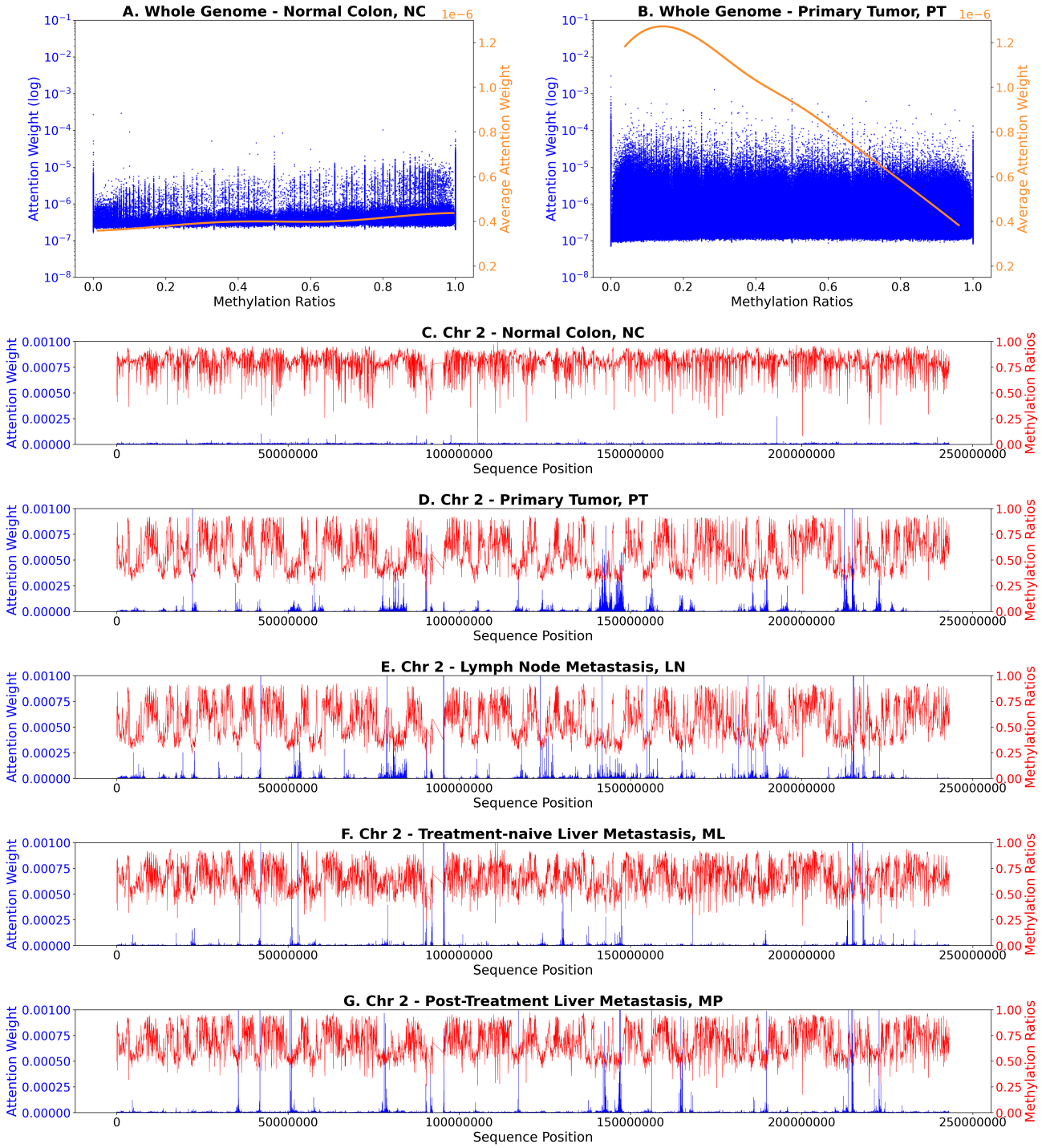

Figure 4. **Attention weight analysis reveals tumor-specific hypomethylation biomarkers identified by scWGBS-GPT. a,** Scatter plot showing the relationship between methylation ratios and attention weights (log scale) in normal colon (NC) samples. Each blue point represents a single CpG site, with its methylation ratio on the x-axis and the corresponding model attention weight on the y-axis. The yellow curve represents the binned average attention weight, computed by dividing the x-axis into bins of length 0.1 and averaging attention weights within each bin. The model exhibits uniformly low attention weights across all CpG sites, suggesting no specific prioritization in normal tissues. **b,** Scatter plot of methylation ratios and attention weights in primary tumor (PT) samples, with the same representation as panel a. The model assigns significantly higher attention weights to hypomethylated CpGs, particularly in unmethylated regions, highlighting tumor-associated hypomethylation sites. **c,** Genomic distribution of methylation ratios (red, right axis) and scWGBS-GPT attention weights (blue, left axis) along chromosome 2 in normal colon (NC) samples. The model does not exhibit preferential focus on any specific CpG sites. **d,** Distribution of methylation ratios and attention weights in primary tumor (PT) samples. The model highlights distinct hypomethylated CpG sites, but attention patterns differ from those observed in metastatic tumors. **e,** Chromosome 2 methylation and attention profiles in lymph node metastasis (LN) samples. The model shows a strong focus on specific hypomethylated regions, particularly near position 50,000,000. **f,** Genomic distribution in treatment-naïve liver metastasis (ML) samples, showing a similar attention pattern to other metastatic states, with high attention around position 50,000,000. **g,** Post-treatment liver metastasis (MP) samples, where scWGBS-GPT maintains a highly consistent attention profile with other metastatic tumors, further supporting the presence of metastasis-associated hypomethylation biomarkers.

importance to any specific CpG sites (Fig.4a). However, in tumor samples, the model selectively assigns high attention weights to hypomethylated CpGs (Fig. 4b), suggesting that these CpG regions may play key roles in tumorigenesis.

To further investigate these attention mechanisms, we analyzed scWGBS-GPT's attention weights across chromosome 2 in different CRC stages (Fig.4c–g). In normal colon samples (Fig.4c), the model assigns low and uniform attention across the chromosome, consistent with its behavior in normal tissues. In primary tumor samples (Fig. 4d), the model assigns high attention weights to certain tumor-specific hypomethylated regions, but these patterns are distinct from those observed in metastatic tumors.

This trend persists in metastatic CRC samples, including lymph node metastases (LN, Fig. 4e), treatment-naïve liver metastases (ML, Fig. 4f), and post-treatment liver metastases (MP, Fig. 4g). Across all metastatic samples, scWGBS-GPT systematically assigns high attention scores to hypomethylated CpGs, highlighting their potential roles as epigenetic biomarkers of CRC progression. Importantly, these high-attention CpGs are not randomly distributed but instead cluster around genomic loci enriched for known tumor suppressor genes, further supporting their biological relevance. Notably, among metastatic samples (Fig. 4e–g), the model exhibits a highly consistent attention pattern, particularly near position 50,000,000 on chromosome 2, where a strong attention signal is observed across all metastatic subtypes (LN, ML, and MP). This striking pattern is absent in both normal and primary tumor samples, suggesting that these genomic loci may be metastasis-associated epigenetic markers. The model's ability to distinguish primary tumors from metastatic tumors based on methylation patterns further validates its capacity to capture biologically meaningful features relevant to cancer progression.

These findings demonstrate that scWGBS-GPT not only recognizes tumor-specific hypomethylation patterns but also differentiates between primary and metastatic tumors based on distinct epigenetic signatures. This suggests that attention-based deep learning approaches can be leveraged to identify metastasis-specific biomarkers from single-cell methylation data. The identification of highly attended regions unique to metastatic states highlights the model's potential utility for epigenetic biomarker discovery and cancer progression modeling.

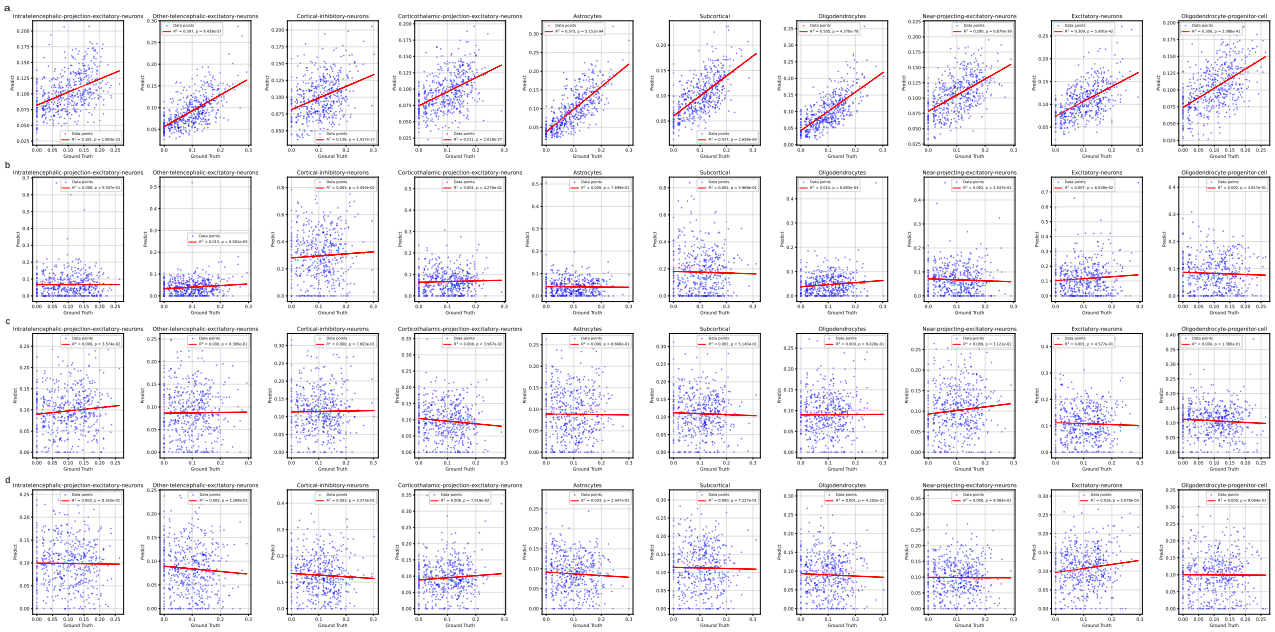

Figure 5. Performance of different methods in cell type deconvolution. We report the R^2 between predicted proportion reported by scWGBS-GPT (**a**), RefFreeCellMix (**b**), EDec (**c**), and MeDeCom (**d**) and true proportion of each cell type on the test data.

2.5 scWGBS-GPT achieves an outstanding performance in reference-free decovnlution on highly heterogeneous bulk methylation profiles

Inferring the cellular composition of a bulk sample with methylation signals, known as cell type deconvolution, is crucial for understanding physiological states and identifying diseases. Traditionally, deconvolution methods rely on reference profiles of several specific cell types, while the high quality references are often hard to obtain and also fail to capture the intra-cell-type variability (Garmire et al., 2024). We leverage the pre-trained scWGBS-GPT model, finetuned on bulk data mixed with different cell types, to perform an end-to-end, reference-free deconvolution.

We use real single-cell methylation data from 10 types of neuronal cells (Tian et al., 2023) to simulate bulk samples. For this, we randomly select 0 to 10 cells from each cell type and mix them together, aiming to preserve both inter-cell-type and intra-cell-type heterogeneity. To better mimic real bulk methylation sequencing, we preserve the variations in sequencing coverage and depth across the mixed single-cell samples, which reflects random noise in bulk sequencing due to the differences in capture efficiency among different cells and the unequal sequencing depth across loci.

We divide 12,000 simulated bulk methylation sequencing samples into training, validation, and test sets as 6:2:2. The pre-trained scWGBS-GPT model was fine-tuned on the training set by minimizing the KL divergence with the true cell type proportion distribution, and the prediction accuracy for each cell type was evaluated on the test set. For comparison, we used three classic reference-free deconvolution methods: RefFreeCellMix (Houseman et al., 2016), EDec (Onuchic et al., 2016), and MeDeCom (Lutsik et al., 2017), implemented with R package DecompPipeline (Scherer et al., 2020). For compared methods, we assigned cell type labels to the predicted results by comparing the order of the average predicted cell proportions for the 10 cell types obtained from each method on the test set, and the order of the true cell proportions. Fig.5 shows that classical methods fail to handle the tricky intra-cell type heterogeneity due to sampling strategy and sequencing noise, while scWGBS-GPT exhibits a quite robust performance.

3 Discussion

We introduce scWGBS-GPT, the first large language model specifically designed for single-cell DNA methylation analysis. By leveraging the innovative Mamba architecture and pretraining on over one million single-cell DNA methylation profiles with one million parameters, our model captures complex interactions among CpG sites and their sequence contexts at the single-cell level. This design enables scWGBS-GPTto generate biologically meaningful embeddings that robustly represent epigenetic states, ultimately providing a versatile tool for downstream analyses. A notable feature of our framework is the integration of cross attention mechanisms. Although our model is built upon the Mamba architecture—which is highly efficient in processing high-dimensional, sparse data—it still employs cross attention to effectively fuse information across different representations. This mechanism plays a critical role in aligning and integrating various aspects of the input data (e.g., local CpG site features and broader sequence context), thereby enhancing the model's capability to capture long-range dependencies and subtle epigenetic signals.

Due to the sparsity and high dimensionality of scWGBS data, clustering and cell type prediction remain significant challenges. Traditional methods mitigate sparsity by averaging DNA methylation ratios over 100,000 bp bins; however, this approach also obscures informative DNA methylation signals and leads to inconsistent clustering results. In contrast, scWGBS-GPT achieves a clustering accuracy of 94.4%, effectively generates meaningful clustering results that align with cell types, even for unseen data. Notably, regions with higher attention weights correspond to low DNA methylation ratios and are marked by active histone modifications, underscoring the interpretability of our model. The sizes of regions with higher attention scores range from hundreds to tens of thousands of base pairs, demonstrating the model's superior resolution in identifying biologically meaningful DNA methylation regions.

Beyond clustering and cell type prediction, scWGBS-GPT demonstrates promising capabilities in reconstructing cell differentiation trajectories. This feature provides novel insights into epigenetic dynamics during development and lineage commitment, positioning our model as a powerful tool for investigating cell fate decisions and the underlying mechanisms of epigenetic regulation. Additionally, preliminary experiments suggest that scWGBS-GPTmay aid in deconvolving cell type compositions from cell-free DNA methylation data—a challenging task when relying on traditional reference-based methods.

Deconvolution using cfDNA methylation data to infer cell type composition, when based on reference data, poses significant limitations. scWGBS-GPT has certain limitations. First, its pretraining data primarily originate from brain samples, which may not sufficiently capture the features of all organs and diseases. Expanding the training dataset to include

scWGBS data from diverse organs, diseases, and cell types will be essential for improving its generalizability in the future. Additionally, scWGBS-GPT currently does not integrate other omics data, such as scRNA-seq, scHi-C, or scCUT&Tag, which could enhance its applicability across various tasks. Furthermore, incorporating metadata such as individual patient information, clinical imaging, and physiological parameters (e.g., blood pressure) could significantly enhance scWGBS-GPT's ability to predict clinical status at an individual level.

We envision that incorporating more datasets, including scWGBS data from diverse tissues, disease conditions, and cell types, along with other omics data and metadata, will enhance scWGBS-GPT's performance across various tasks. Additionally, we plan to pretrain scWGBS-GPT on time-series and perturbation data, enabling it to learn causal relationships and infer cell behavior in response to DNA methylation alterations. As a foundation model, scWGBS-GPT has the potential to accelerate the discovery of epigenetic dynamics at the single-cell level, benefiting biomedical research.

4 METHODS

4.1 Model

4.1.1 scWGBS Tokenizer

The scWGBS tokenizer is a critical component of scWGBS-GPT, as it transforms raw genomic data into a format suitable for processing by the model. Given that scWGBS data consists of methylation levels at each CpG site across large genomic regions, the tokenizer's primary role is to encode these methylation values into a tokenized sequence that preserves both the spatial and biological characteristics of the input data.

In contrast to traditional NLP tokenization, where words are split into subwords or characters, the scWGBS tokenizer is designed to handle the unique challenges posed by biological data, particularly the representation of DNA sequences and methylation patterns. Each CpG site is treated as a "token," and its methylation state is encoded in a way that retains the crucial epigenetic information.

The core idea behind the scWGBS tokenizer is to treat DNA sequences as strings of "CpG words," where each CpG site is a token. This tokenizer design incorporates both methylated and unmethylated states as part of the token representation, ensuring that methylation patterns are preserved for downstream processing.

To achieve this, each CpG site is encoded as a binary token that represents its methylation state: - A methylated CpG is assigned a token value of 1, - An unmethylated CpG is assigned a token value of 0.

In addition to binary methylation states, we incorporate flanking nucleotide context to account for the surrounding DNA sequence, which can have biological significance in regulating methylation patterns. This allows the tokenizer to generate more informative representations for each CpG site.

Thus, each token is a combination of the methylation state and its local DNA context, forming a sequence of tokens that represent the entire genomic region.

The tokenization process can be broken down into the following steps:

1. **DNA Sequence Splitting**: The genomic data is split into windows of CpG sites, where each window corresponds to a continuous region of the genome (e.g., a 10kb region). Each window contains a sequence of CpG sites, and the surrounding nucleotides are encoded as part of the context.

2. **Encoding Methylation States**: For each CpG site in the window, its methylation state is encoded as a binary value (0 or 1). Additionally, the surrounding nucleotides (e.g., ± 3 bases around each CpG site) are incorporated to form a context-aware token. This results in a k-mer token representation, where each CpG is accompanied by its local sequence context.

3. **Sequence Assembly**: The tokenized sequences are then assembled into a sequence of tokens, where each token represents a CpG site along with its methylation state and surrounding context. The final sequence represents the entire window of genomic data.

$$\text{Token}_i = \left(\text{MethylationState}(\mathbf{x}_i), \text{FlankingContext}(\mathbf{x}_i)\right), \tag{1}$$

where $\mathbf{x}_i$ is the i-th CpG site in the genomic window, and the MethylationState refers to the methylation status (0 or 1),

while the FlankingContext represents the surrounding nucleotides.

4. **[BOS] and [SEP] Token Insertion**: A special [BOS] token is inserted at the beginning of each window to signal the start of a new sequence, as well as to serve as the query in the cross-attention mechanism, which aggregates global context from the entire sequence.

Once the tokenization process is complete, the resulting sequence of tokens can be fed into the Mamba Backbone for further processing. The [SEP] token plays a crucial role in generating the query vector for the cross-attention mechanism, as described in Section A.1.3. This integration of tokenized genomic data with the cross-attention mechanism allows scWGBS-GPT to capture both the methylation state of individual CpG sites and their relationships with distant sites across the genome.

For example, consider a window of CpG sites:

$$\text{Genomic Window} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n].$$

The tokenizer processes this window by encoding each CpG site $\mathbf{x}_i$ as a token that includes its methylation state and flanking nucleotide context. The resulting sequence is then represented as:

$$\text{Tokenized Sequence} = [[\text{BOS}], \mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n, [\text{SEP}]].$$

This tokenized sequence is then passed through the Mamba Backbone, where the [SEP] token generates the query vector for the cross-attention mechanism, which attends to each CpG site (represented by the tokens $\mathbf{t}_i$) based on their methylation states and contextual relationships.

The scWGBS tokenizer offers several advantages for modeling DNA methylation data:

1. **Preserving Methylation Information**: By encoding both methylation states and surrounding nucleotide contexts, the tokenizer ensures that key epigenetic information is preserved for downstream processing.

2. **Contextualized Tokenization**: The inclusion of flanking nucleotides in the tokenization process allows the model to learn not only the methylation state of each CpG site but also the surrounding sequence context that may influence methylation patterns.

3. **Scalability**: The tokenizer's ability to efficiently process large genomic windows and generate contextually rich tokens enables scWGBS-GPT to scale to millions of CpG sites, making it suitable for analyzing large-scale genomic data.

The scWGBS tokenizer is specifically designed to capture biologically relevant features of the genome. The methylation status of each CpG site is crucial for understanding gene regulation, and the flanking nucleotide context helps reveal potential regulatory motifs that influence methylation patterns. By integrating these features into the tokenization process, we ensure that the model can learn meaningful biological patterns from the data.

4.1.2 Mamba Backbone

Mamba is a state-of-the-art sequence modeling framework designed for efficient processing of ultra-long sequences. Unlike traditional transformer-based architectures, which suffer from quadratic complexity in self-attention, Mamba leverages Selective State Space Models (SSMs) (Gu & Dao, 2023) to achieve linear time complexity while maintaining strong representation capabilities.

The motivation for using Mamba in scWGBS-GPT stems from the inherent challenges of single-cell whole-genome bisulfite sequencing (scWGBS) data. Each cell's methylation profile consists of millions of CpG sites, forming an ultra-long sequence that requires efficient modeling of both local CpG interactions and long-range dependencies. Mamba's ability to capture global dependencies without the prohibitive computational costs of transformers makes it a very natural fit for this problem.

Long genomic sequences present unique challenges in computational biology (Nguyen et al., 2024; Fishman et al., 2023). Methylation patterns at distant CpG sites often exhibit correlations due to chromatin organization, regulatory regions, and epigenetic inheritance. Traditional deep learning models, including recurrent neural networks (RNNs) and CNNs, struggle with these dependencies due to limited receptive fields or memory constraints.

Mamba addresses these limitations through a continuous-time selective state space model (SSM), which allows for efficient and memory-friendly long-sequence modeling. The core state update equation is:

$$\mathbf{h}_t = \mathbf{A}\mathbf{h}_{t-1} + \mathbf{B}\mathbf{x}_t, \tag{2}$$

where: $\mathbf{h}_t$ is the hidden state at time step t, $\mathbf{A}$ is a learned transition matrix controlling sequence memory, $\mathbf{B}$ is an input transformation matrix, $\mathbf{x}_t$ is the input at time step t.

Unlike transformers, which require explicit attention over all previous positions, Mamba's state-space model implicitly propagates long-range dependencies without explicit the token-by-token interactions, significantly reducing computational overhead.

The Mamba Backbone in scWGBS-GPT consists of a deep stack of **Mamba Blocks**, each designed to refine feature representations at increasing levels of abstraction. Each block operates as follows:

1. **Convolutional Input Projection**: The input sequence $\mathbf{X}$ is first processed by a depth-wise convolutional layer to extract local CpG patterns:

$$\mathbf{X}' = \text{Conv1D}(\mathbf{X}). \tag{3}$$

2. **State-Space Model (SSM) Evolution**: Each token is passed through a continuous-time state space layer, updating the internal sequence representation:

$$\mathbf{H}_t = \mathbf{A}\mathbf{H}_{t-1} + \mathbf{B}\mathbf{X}'_t. \tag{4}$$

3. **Selective Gating Mechanism**: A sigmoid gating function dynamically modulates the influence of new information:

$$\mathbf{G}_t = \sigma(\mathbf{W}_g\mathbf{X}_t), \tag{5}$$

where σ is the sigmoid activation and $\mathbf{W}_g$ is a learned weight matrix.

4. **Linear Projection and Normalization**: The output is projected to a new feature space and normalized:

$$\mathbf{Y}_t = \text{LayerNorm}(\mathbf{W}_y\mathbf{H}_t). \tag{6}$$

Each Mamba Block stacks these transformations, progressively refining long-sequence representations.

Network Configuration and Parameters

The Mamba Backbone in scWGBS-GPT consists of 8 stacked Mamba Blocks, leading to a total parameter count of approximately 1M parameters. This configuration balances computational efficiency with modeling capacity.

The architecture can be expressed as a sequential composition:

$$\mathbf{X} \xrightarrow{\text{Mamba Block 1}} \mathbf{Y}_1 \xrightarrow{\text{Mamba Block 2}} \mathbf{Y}_2 \ldots \xrightarrow{\text{Mamba Block 8}} \mathbf{Y}_8. \tag{7}$$

This deep stacking mechanism enables the model to capture both short-range CpG patterns and long-range regulatory dependencies in scWGBS data. The Mamba Block computations are optimized using CUDA kernels for efficient parallelization, allowing scWGBS-GPT to process ultra-long sequences efficiently.

4.1.3 Cross Attention

Cross-attention is a powerful mechanism that allows a model to focus on relevant parts of the input sequence while processing a different sequence. In traditional attention mechanisms, such as in the Transformer model, a query vector interacts

with all keys and values within the same sequence. However, in cross-attention, the query vector is derived from one sequence, and it attends to another sequence's keys and values (Gheini et al., 2021). This approach is especially useful when the relationships between two different data representations are essential, such as in multi-modal tasks or when integrating sparse information like DNA methylation patterns.

In our scWGBS-GPT framework, we leverage cross-attention to capture long-range dependencies and contextualize CpG site interactions across the entire genome. Specifically, we use a [SEP] token to generate query, key, and value representations (denoted as $\mathbf{Q}, \mathbf{K}, \mathbf{V}$) for the cross-attention mechanism. The [SEP] token's final hidden state serves as the query, while each CpG site's hidden state is used to compute the keys and values.

The mechanism of the employed cross-attention is clarified as follows: Given an input sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, where $\mathbf{x}_i$ represents the hidden state corresponding to the i-th token in the sequence, cross-attention computes the interactions between the query $\mathbf{Q}$ and the keys $\mathbf{K}$, along with their corresponding values $\mathbf{V}$.

The cross-attention operation can be described as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{8}$$

where: $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$ is the query matrix, $\mathbf{K} \in \mathbb{R}^{n \times d_k}$ is the key matrix, $\mathbf{V} \in \mathbb{R}^{n \times d_v}$ is the value matrix, d_k is the dimension of the key vectors.

The softmax function ensures that the attention weights sum to 1, providing a probabilistic distribution over the values $\mathbf{V}$ based on the relevance of the queries to the keys.

In scWGBS-GPT, we utilize cross-attention to capture the intricate dependencies between CpG sites across the entire genome, focusing on biologically meaningful interactions. To achieve this, we employ the [SEP] token in the following way:

1. The [SEP] token's final hidden state is used as the query $\mathbf{Q}$ for cross-attention. This token is specially designed to aggregate global contextual information from the entire sequence.

2. Each CpG site's hidden state, generated from the Mamba Backbone, serves as both the key $\mathbf{K}$ and the value $\mathbf{V}$.

Thus, for each position in the sequence (representing a CpG site), the attention weights are computed between the [SEP] token's query and the CpG site's key. This allows the [SEP] token to attend to the relevant CpG sites and capture their relationships in the final representation.

The cross-attention operation for a single CpG site can be described as:

$$\mathbf{H}_{\text{att}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{9}$$

where: $\mathbf{H}_{\text{att}}$ represents the attention-weighted hidden state for each CpG site after the cross-attention computation, $\mathbf{Q}$ is the hidden state of the [SEP] token, $\mathbf{K}$ and $\mathbf{V}$ are the hidden states of each CpG site.

This attention-weighted hidden state $\mathbf{H}_{\text{att}}$ is then used for further processing in downstream layers, where it contributes to the final genomic sequence representation.

The use of cross-attention in scWGBS-GPT offers several benefits for modeling genomic data:

1. **Capturing Long-Range Dependencies**: Cross-attention allows the [SEP] token to attend to distant CpG sites, which is essential for modeling the long-range interactions that occur in DNA methylation patterns. These long-range dependencies often play crucial roles in gene regulation and epigenetic modifications.

2. **Contextualizing CpG Interactions**: By attending to the entire genomic sequence, the cross-attention mechanism helps the model contextualize the interaction between CpG sites, ensuring that biologically relevant correlations are learned.

3. **Efficient Representation Learning**: The ability of the [SEP] token to serve as a global query ensures that even sparse methylation data can be processed effectively, enabling the model to capture meaningful patterns from incomplete or sparse data.

The cross-attention mechanism is particularly effective in processing large-scale genomic datasets, such as those in scWGBS-GPT, where each input sequence can span millions of CpG sites. By using the [SEP] token to generate the query vector, we avoid the need for expensive pairwise attention calculations across all tokens, enabling the model to efficiently capture global dependencies while maintaining computational feasibility.

In scWGBS-GPT, the cross-attention mechanism is integrated within the architecture to handle ultra-long sequences efficiently. By applying attention selectively through the [SEP] token query, the model can scale to handle large genomic sequences with millions of CpG sites while still extracting biologically meaningful relationships between distant CpG sites.

FUNDING

This work was supported by the National Natural Science Foundation of China (82370148), the R&D Program of Guangzhou Laboratory (GZNL2023A02003; GZNL2023A02011), the Guangzhou Key Research and Development Program of Guangzhou Municipal Science and Technology Bureau (2024B03J0046), the grant of State Key Laboratory of Respiratory Disease (SKLRD-Z-202307), and the Plan on Enhancing Scientific Research in GMU (02-410-2302244XM).

REFERENCES

Bian, S., Hou, Y., Zhou, X., Li, X., Yong, J., Wang, Y., Wang, W., Yan, J., Hu, B., Guo, H., et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. Science, 362(6418):1060–1063, 2018.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature Methods, pp. 1–11, 2024.

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. Single-cell dna methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell reports, 10(8):1386–1397, 2015.

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):808–822, 2016.

Fishman, V., Kuratov, Y., Petrov, M., Shmelev, A., Shepelin, D., Chekanov, N., Kardymon, O., and Burtsev, M. Gena-lm: A family of open-source foundational models for long dna sequences. bioRxiv, pp. 2023–06, 2023.

Garmire, L. X., Li, Y., Huang, Q., Xu, C., Teichmann, S. A., Kaminski, N., Pellegrini, M., Nguyen, Q., and Teschendorff, A. E. Challenges and perspectives in computational deconvolution of genomics data. Nature Methods, 21(3):391–400, 2024.

Gheini, M., Ren, X., and May, J. Cross-attention is all you need: Adapting pretrained transformers for machine translation. arXiv preprint arXiv:2104.08771, 2021.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023.

Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., and Song, L. Large-scale foundation model on single-cell transcriptomics. Nature Methods, pp. 1–11, 2024.

Houseman, E. A., Kile, M. L., Christiani, D. C., Ince, T. A., Kelsey, K. T., and Marsit, C. J. Reference-free deconvolution of dna methylation data and mediation by cell composition effects. BMC bioinformatics, 17:1–15, 2016.

Kremer, L. P., Braun, M. M., Ovchinnikova, S., Küchenhoff, L., Cerrizuela, S., Martin-Villalba, A., and Anders, S. Analyzing single-cell bisulfite sequencing data with methscan. Nature Methods, 21(9):1616–1623, 2024.

Liu, H., Zhou, J., Tian, W., Luo, C., Bartlett, A., Aldridge, A., Lucero, J., Osteen, J. K., Nery, J. R., Chen, H., et al. Dna methylation atlas of the mouse brain at single-cell resolution. Nature, 598(7879):120–128, 2021.

Liu, H., Zeng, Q., Zhou, J., Bartlett, A., Wang, B., Berube, P., Tian, W., Kenworthy, M., Altshul, J., Nery, J., et al. Single-cell dna methylome and 3d multi-omic atlas of the adult mouse brain. biorχiv. 2022.

Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., Fox-Fisher, I., Shabi-Porat, S., Hecht, M., Pelet, T., et al. A dna methylation atlas of normal human cell types. Nature, 613(7943):355–364, 2023.

Lutsik, P., Slawski, M., Gasparoni, G., Vedeneev, N., Hein, M., and Walter, J. Medecom: discovery and quantification of latent components of heterogeneous methylomes. Genome biology, 18:1–20, 2017.

Mulqueen, R. M., Pokholok, D., Norberg, S. J., Torkenczy, K. A., Fields, A. J., Sun, D., Sinnamon, J. R., Shendure, J., Trapnell, C., O'Roak, B. J., et al. Highly scalable generation of dna methylation profiles in single cells. Nature biotechnology, 36(5):428–431, 2018.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. Advances in neural information processing systems, 36, 2024.

Onuchic, V., Hartmaier, R. J., Boone, D. N., Samuels, M. L., Patel, R. Y., White, W. M., Garovic, V. D., Oesterreich, S., Roth, M. E., Lee, A. V., et al. Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types. Cell reports, 17(8):2075–2086, 2016.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. 2018.

Scherer, M., Nazarov, P. V., Toth, R., Sahay, S., Kaoma, T., Maurer, V., Vedeneev, N., Plass, C., Lengauer, T., Walter, J., et al. Reference-free deconvolution, visualization and interpretation of complex dna methylation data using decomp-pipeline, medecom and factorviz. Nature Protocols, 15(10):3240–3263, 2020.

Tian, W., Zhou, J., Bartlett, A., Zeng, Q., Liu, H., Castanon, R. G., Kenworthy, M., Altshul, J., Valadon, C., Aldridge, A., et al. Single-cell dna methylation and 3d genome architecture in the human brain. Science, 382(6667):eadf5357, 2023.

Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., and Yao, J. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. Nature Machine Intelligence, 4(10):852–866, 2022.

Zhou, W., Dinh, H. Q., Ramjan, Z., Weisenberger, D. J., Nicolet, C. M., Shen, H., Laird, P. W., and Berman, B. P. Dna methylation loss in late-replicating domains is linked to mitotic cell division. Nature genetics, 50(4):591–602, 2018.