



RESEARCH ARTICLE OPEN ACCESS

Profiling Epigenetic Aging at Cell-Type Resolution Through Long-Read Sequencing

Alec Eames | Mahdi Moqri | Jesse R. Poganik | Vadim N. Gladyshev

Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA

Correspondence: Vadim N. Gladyshev (vgladyshev@bwh.harvard.edu)

Received: 24 November 2024 | **Revised:** 28 March 2025 | **Accepted:** 8 April 2025

Funding: Supported by grants from the National Institute on Aging and Hevolution to VNG.

Keywords: aging | epigenetics | long-read sequencing

ABSTRACT

DNA methylation can give rise to robust biomarkers of aging, yet most studies profile it at the bulk tissue level, which masks cell type-specific alterations that may follow distinct aging trajectories. Long-read sequencing technology enables methylation profiling of extended DNA fragments, enabling mapping to their cell type of origin. In this study, we introduce a framework for evaluating cell type-specific aging using long-read sequencing data, without the need for cell sorting. Leveraging cell type-specific methylation patterns, we map long-read fragments to individual cell types and generate cell type-specific methylation profiles, which are used as input to a newly developed probabilistic aging model, LongReadAge, capable of predicting epigenetic age at the cell type level. We use LongReadAge to track aging of myeloid cells and lymphocytes from bulk leukocyte data as well as circulating cell-free DNA, demonstrating robust performance in predicting age despite limited shared features across samples. This approach provides a novel method for profiling the dynamics of epigenetic aging at cell type resolution.

1 | Introduction

Aging is characterized by progressive physiological decline and increased incidence of disease, disability, and death (Guo et al. 2022). At the molecular level, epigenetic modifications have emerged as a key biomarker of aging, with widespread shifts observed across the epigenomic landscape as we age (Moqri et al. 2024; Tharakan et al. 2023). The most widely used epigenetic modification in aging research is methylation of cytosines at cytosine-guanine (CpG) dinucleotides, whose patterns enable robust prediction of age, mortality risk, and disease susceptibility (Lu, Fei et al. 2023; Lu et al. 2019; Levine et al. 2018; Ying et al. 2024; Cappozzo et al. 2022).

Emerging evidence suggests that aging involves cell type-specific alterations (Lu, Zhang, et al. 2023; Ren et al. 2023; Chien et al. 2024; Eames and Chandrasekaran 2024; Trapp

et al. 2021), underscoring the importance of tracking aging at cell-type resolution. Existing methylation-based biomarkers of aging, however, largely rely on bulk methylation profiles, rendering them susceptible to confounding from cell composition and obscuring signals from low-abundance cell types (Jaffe and Irizarry 2014). Although algorithms to computationally generate cell-type methylation patterns from bulk profiles have been developed (Rahmani et al. 2019), these approaches are hindered by their dependence on known cell proportions. Cell-type methylation patterns can be determined experimentally through cell sorting and single-cell sequencing; however, such techniques are time-consuming, expensive, and not scalable for routine use.

With the advent of long-read sequencing technology, it is now possible to sequence extended fragments of DNA in a single read (Logsdon et al. 2020). Crucially, this technology enables the simultaneous determination of nucleotide sequence and

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Aging Cell* published by Anatomical Society and John Wiley & Sons Ltd.

CpG methylation, without the need for experimental cytosine-to-uracil conversion. This conversion, commonly performed with bisulfite treatment, often degrades DNA and introduces sequencing biases (Tanaka and Okamoto 2007; Olova et al. 2018). Excitingly, recent studies have demonstrated the feasibility of computationally tracing fragments from long-read sequencing to their cell type of origin using cell type-specific methylation patterns (Choy et al. 2022; Yu et al. 2021; Hannon and Mill 2024), obviating the need for cell sorting. A drawback of these long-read methylation profiles, however, is the small number of intersecting CpG sites across samples, which impedes across-sample age profiling, particularly for cell type-specific methylation profiles derived from a limited number of reads.

In this study, we leverage long-read methylation data paired with a newly developed probabilistic aging model, LongReadAge, to profile epigenetic aging at cell-type resolution. LongReadAge, based on the scAge framework (Trapp et al. 2021), computes the epigenetic age of sparsely intersecting long-read methylation profiles. By mapping fragments from long-read data to their cell type of origin and applying LongReadAge to the resulting methylation profiles, we show that cell type-specific estimates of epigenetic age may be produced from long-read sequencing of bulk samples. We focus our analysis on peripheral blood, a widely studied tissue composed of a heterogeneous mixture of cell types, through which we track aging of myeloid cells and lymphocytes—two fundamental hematopoietic cell types that display distinct features during aging (Aw et al. 2007; Soto-Herederro et al. 2023).

We further apply our model to track aging of circulating cell-free DNA (Figure 1).

2 | Results

2.1 | Development and Application of LongReadAge

Inspired by the scAge framework (Trapp et al. 2021), we developed LongReadAge, a probabilistic model designed to predict epigenetic age based on long-read methylation data (Figure 1). In this model, a reference aging dataset is used to build linear regression models predicting methylation as a function of age across every profiled CpG site. The input to LongReadAge is a long-read methylation profile, consisting of a series of CpG sites along with their methylation status. We identify CpGs common to both the input long-read profile and the reference dataset, then filter these CpGs to include only those with the highest Pearson correlation with chronological age in the reference dataset. Next, for this filtered set of CpGs, we predict methylation patterns over a range of ages, from 0 to 100 years, using the linear age–methylation relationships found in the reference dataset. For each age's predicted methylation profile, we compare it to the input long-read methylation profile using a probability score that quantifies the similarity between the two profiles (see Methods). The age that yields the highest probability score, indicating the closest match between the predicted

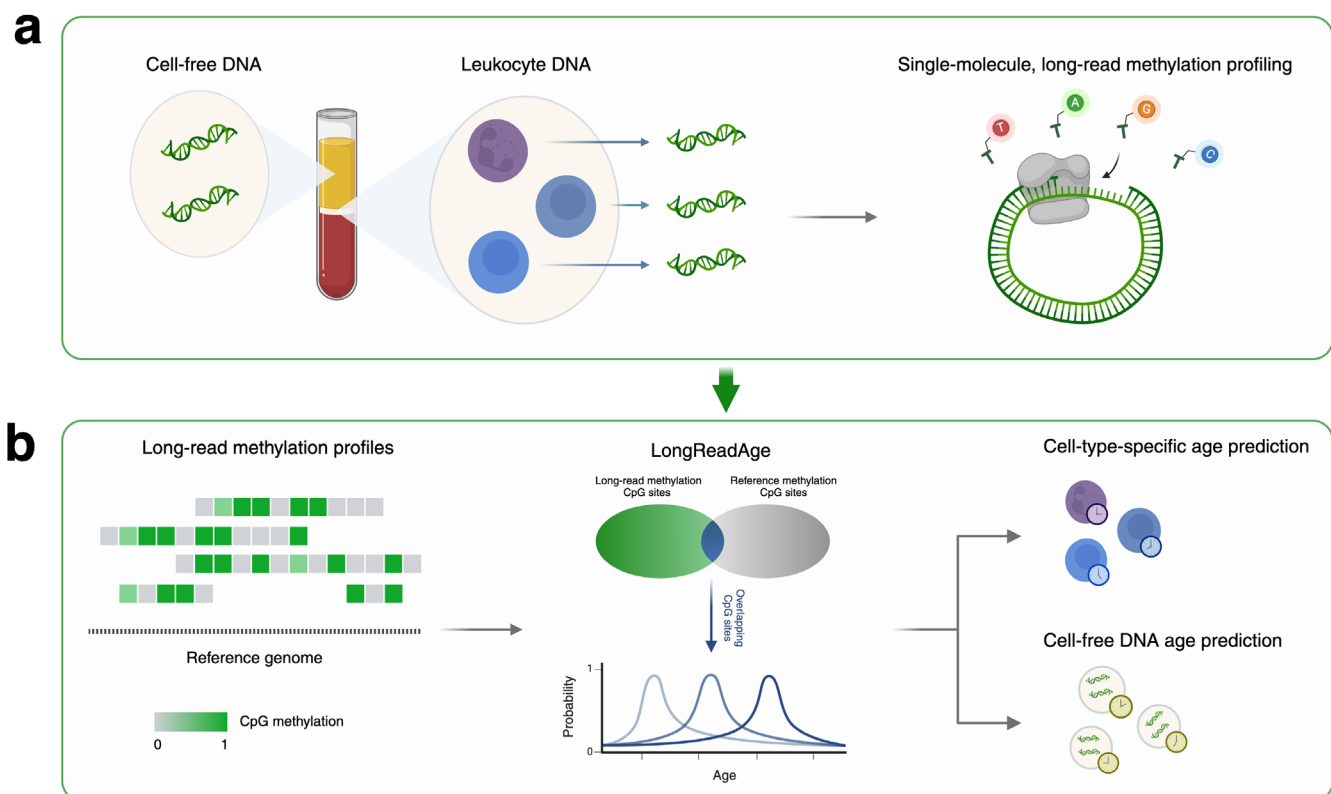


FIGURE 1 | Schematic overview of the study. (a) CpG methylation of peripheral leukocytes and circulating cell-free DNA was analyzed through long-read sequencing. (b) A probabilistic model, LongReadAge, was developed to predict epigenetic age using sparsely intersecting long-read methylation profiles. We applied this model to profile immune-cell-specific and cell-free DNA aging.

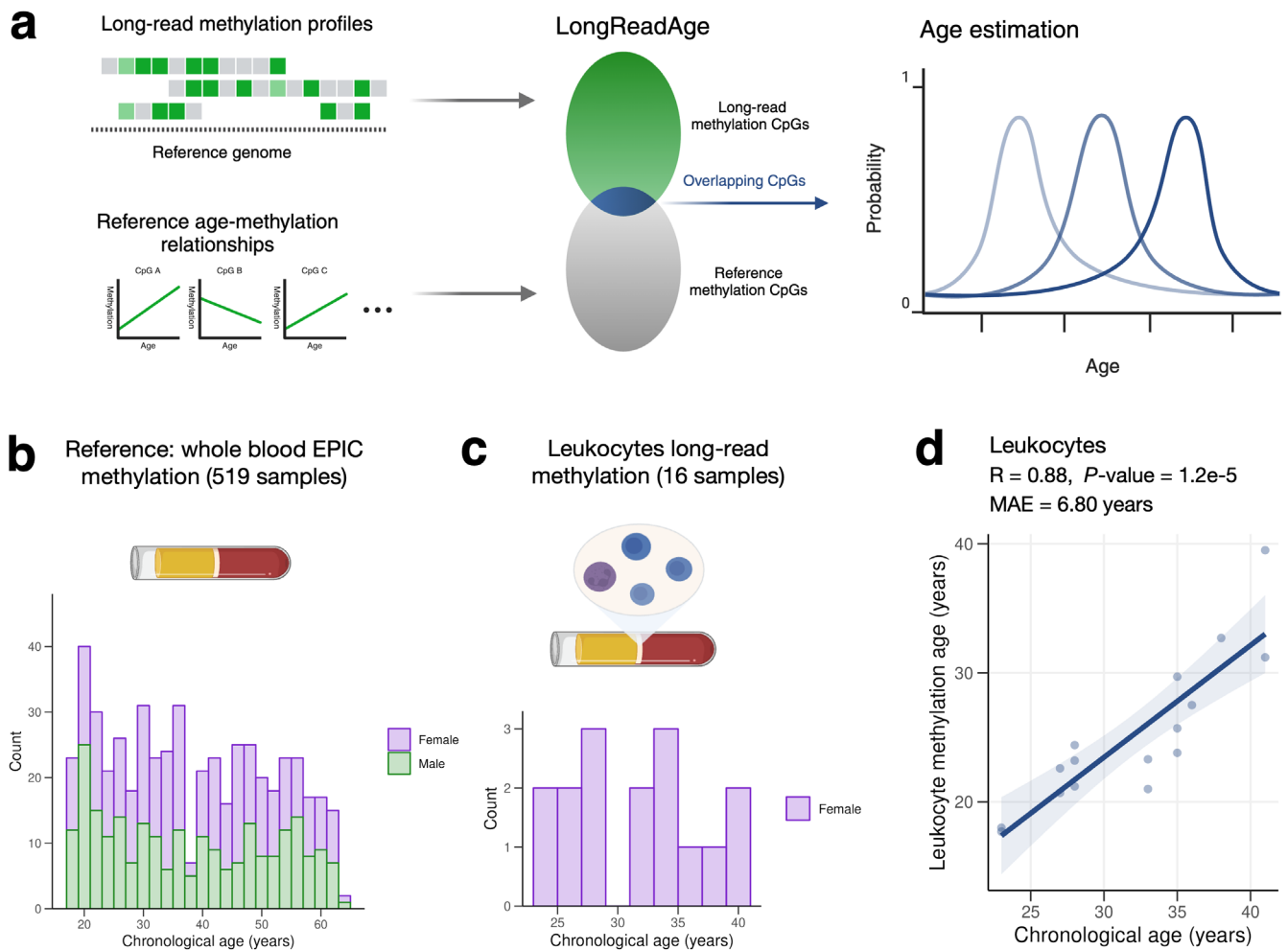


FIGURE 2 | Development of LongReadAge. (a) Schematic depicting LongReadAge framework. In this model, CpG sites from long-read methylation profiles are intersected with CpGs from a reference dataset, where the relationship between chronological age and CpG methylation is well-characterized. For these intersecting CpG sites, methylation status is predicted using chronological age. The age that best predicts the observed long-read methylation pattern is considered the epigenetic age of the sample. (b) Our reference dataset used in this model consisted of CpG methylation from 519 healthy individuals analyzed using the EPIC microarray platform. (c) Long-read single-molecule methylation sequencing data was obtained from 16 samples. DNA was extracted from peripheral leukocytes (buffy coat). (d) Application of LongReadAge to long-read leukocyte data.

and observed long-read profile, is considered the epigenetic age of the sample (Figure 2A).

We evaluated our model by applying it to data from a cohort of 11 pregnant women for which 16 long-read methylation profiles of peripheral leukocytes were generated in a previous study (Figure 2C) (Yu et al. 2021). As our reference aging dataset, we used a whole blood methylation dataset comprising 519 healthy individuals balanced between males and females measured with the EPIC microarray (Hannon et al. 2021) (Figure 2B). We selected this dataset as our reference due to its large number of healthy subjects, broad and even age distribution, and compositional balance between men and women. LongReadAge yielded epigenetic age estimates showing a strong correlation with chronological age and a mean absolute error below 7 years, demonstrating the potential of our approach in estimating age from sparsely intersecting long-read methylation data (Figure 2D). We validated the robustness of our approach to an alternative reference methylation dataset

in Figure S1 and found comparable performance. Notably, since LongReadAge is trained on the reference dataset, it can be applied to long-read datasets with a small number of subjects, as is the case here.

2.2 | LongReadAge Captures Immune-Cell-Specific Aging

We next applied LongReadAge to profile cell type-specific aging. Beginning with the same 16 long-read bulk samples measuring leukocyte methylation, we generated cell-type methylation profiles by tracing individual fragments to their cell type of origin, following a previously developed and validated approach (Choy et al. 2022; Yu et al. 2021). For this, we leveraged a comprehensive methylation atlas measuring genome-wide methylation across a range of cell types (Loyfer et al. 2023). From this atlas, we identified differentially methylated CpG sites distinguishing two distinct cell types found in blood: lymphocytes and myeloid

cells. We identified these sites using the *wgbstools* suite (Loyer et al. 2024), applying several criteria to identify robustly differentially methylated CpG sites.

Next, we overlapped each long-read fragment's CpG sites with the differentially methylated CpG sites in the reference atlas.

Among fragments with at least five overlapping reference CpG sites, we computed a score describing the likelihood of a fragment originating from each cell type based on how well the fragment's methylation patterns matched each respective cell type. Fragments with a 1.05× greater likelihood of originating from one cell type versus another were considered to derive from that

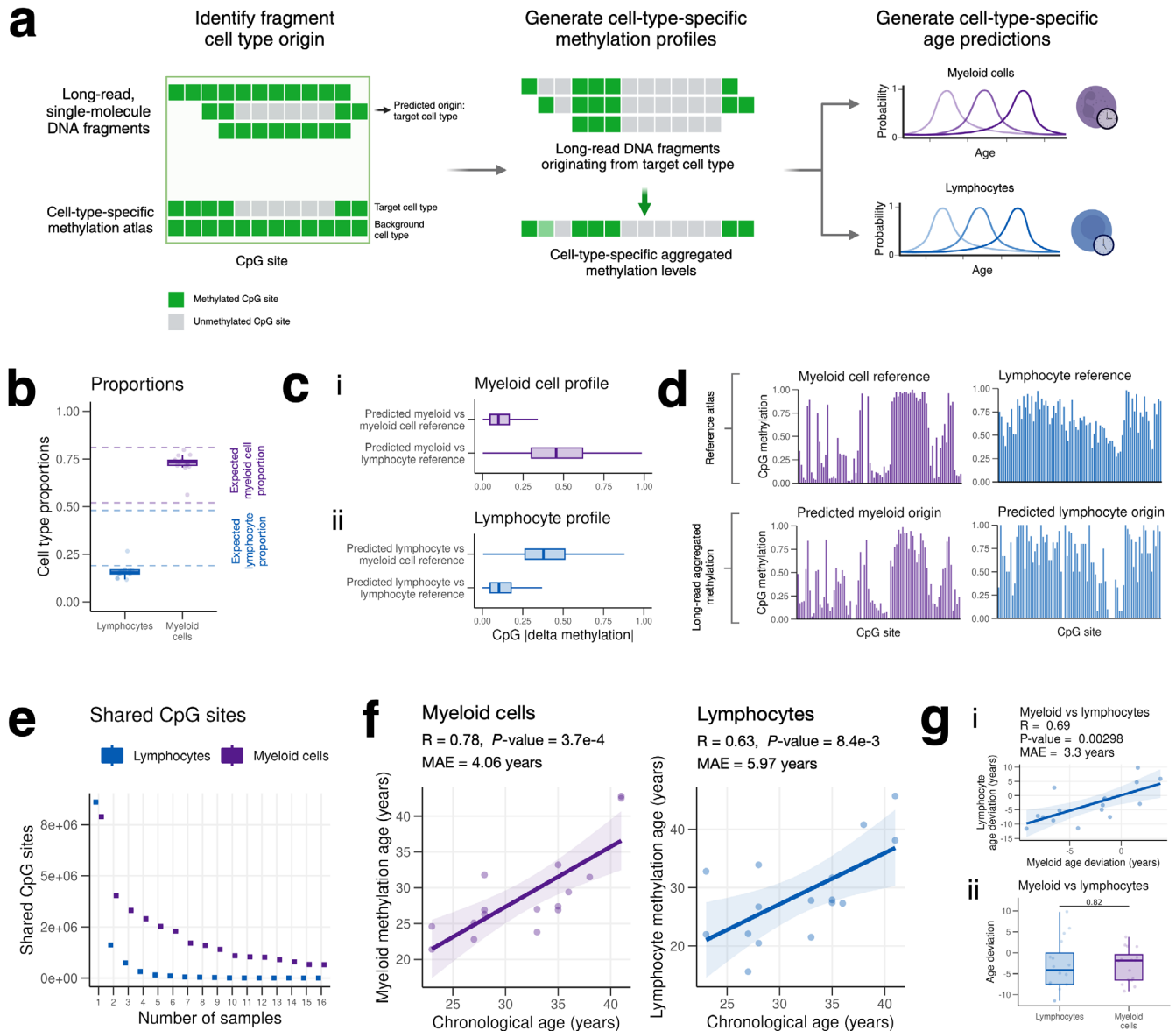


FIGURE 3 | LongReadAge captures cell type-specific aging. (a) Schematic depicting cell type-specific methylation profile generation and age prediction. Using single-molecule long-read DNA methylation data from leukocytes, we identified the cellular origin of individual DNA fragments and obtained cell type-specific methylation profiles. These profiles were applied as input to LongReadAge to generate cell type-specific epigenetic age predictions. (b) Proportions of reads predicted to originate from lymphocytes and myeloid cells. Dotted lines depict ranges of typical lymphocyte and myeloid cell proportions in peripheral blood. (c) Average magnitude of CpG methylation difference between (i) predicted myeloid cell methylation profile vs. myeloid cell reference and vs. lymphocyte reference, and (ii) predicted lymphocyte methylation profile vs. myeloid cell reference and vs. lymphocyte reference. (d) Example methylation profiles from cell-specific methylation atlas (myeloid cells and lymphocytes) and averaged long-read methylation data for reads predicted to originate from myeloid and lymphocyte cells. Displayed here is a random selection of CpGs present in both lymphocyte and myeloid differentially methylated CpG sites and in at least one sample of myeloid- and lymphocyte-predicted long-read methylation profiles. (e) Number of shared CpGs between progressively increasing numbers of samples for both (1) lymphocyte long-read methylation profiles and (2) myeloid cell long-read methylation profiles. (f) Results of applying LongReadAge to cell type-specific methylation profiles. Two-step percentile, fixed number cutoff is used here to generate predictions. (g) Age deviation comparison between lymphocytes and myeloid cells for the same individuals shown as scatterplot (i) and boxplot (ii).

cell type. The distribution of cell type likelihoods across reads and samples is plotted in Figure S3.

We aggregated all fragments corresponding to a given cell type to generate cell type-specific methylation profiles (Figure 3A). The accuracy of generating cell type-specific methylation profiles was verified with two approaches. First, we confirmed the proportion of reads predicted to originate from myeloid cells and lymphocytes matched typical proportions of these cell types in peripheral blood (Figure 3B) (Seo and Lee 2022). Among prediction-eligible fragments, approximately 75% were predicted to originate from myeloid cells, and around 20% were predicted to originate from lymphocytes, with a slight underrepresentation due to a conservative 1.05× probability cutoff. Second, we compared long-read cell type-specific methylation to reference cell-type methylation profiles, which revealed a strong concordance between the two. With methylation represented as a value ranging from 0 to 1 (universally unmethylated to universally methylated), the long-read lymphocyte profile displayed a median deviation of approximately 0.1 from the reference lymphocyte profile versus about 0.45 from the reference myeloid profile. The long-read myeloid profile displayed similar results, matching the myeloid reference much more closely than the lymphocyte reference (Figure 3C). Visual examination of long-read

aggregated methylation profiles versus cell type-specific reference profiles underscored their similarity (Figure 3D).

Given that only reads with at least 5 CpG site overlaps with the reference cell-type atlas were eligible to be mapped to a particular cell type, the resulting cell type-specific long-read methylation profiles contained a small number of intersecting CpG sites (Figure 3E), necessitating a reference-based aging model that uses different sets of CpGs for each sample. Applying LongReadAge to these cell type-specific methylation profiles demonstrated a robust ability to track aging across both myeloid cells and lymphocytes, with mean errors below 6 years for both cell types (Figure 3F). Comparing the epigenetic age estimates between myeloid cells and lymphocytes across the same individuals, we observed a moderately strong correlation between the two and no significant average difference (Figure 3G). Thus, long-read sequencing data has the potential to resolve epigenetic ages within a bulk sample at cell-type resolution.

2.3 | LongReadAge Captures Aging of Circulating Cell-Free DNA

Circulating cell-free DNA (cfDNA) consists of DNA fragments outside of cells that circulate in plasma and other biofluids,

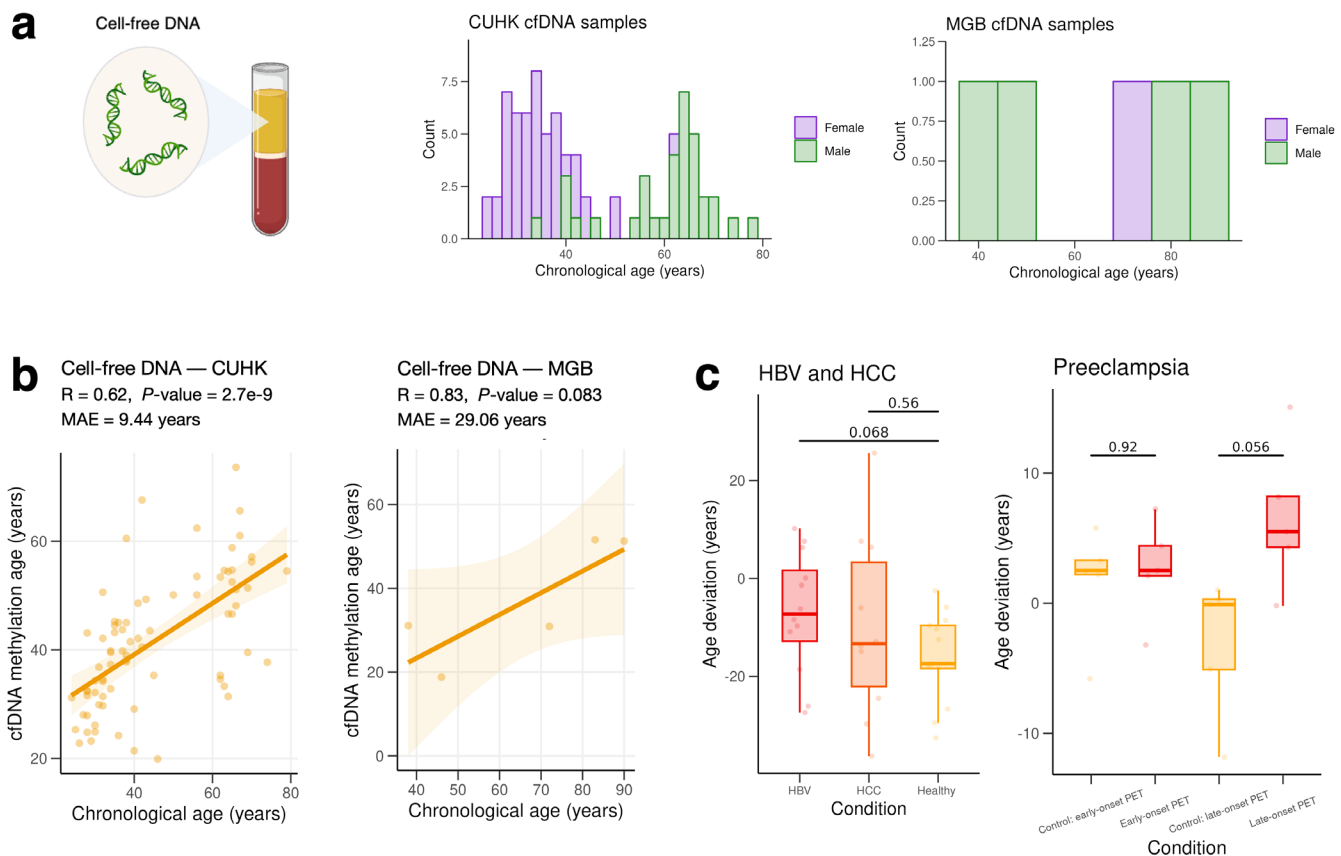


FIGURE 4 | LongReadAge captures aging of circulating cell-free DNA. (a) Cell-free DNA derived from plasma was profiled in two cohorts: One from the Chinese University of Hong Kong (CUHK) consisting of 84 samples (healthy + pregnancy + disease) and one from the Mass General Brigham (MGB) Biobank (5 samples). (b) Application of LongReadAge to CUHK and MGB cohort. Samples with fewer than 300 CpG sites intersecting the reference CpGs were omitted. (c) Comparison of age deviation between (1) hepatitis B virus (HBV) carriers, hepatocellular carcinoma (HCC) patients, and healthy controls and (2) early-onset and late-onset preeclampsia (PET) vs. respective controls.

often as a result of cell death. Molecular alterations in cfDNA are observed with aging and may serve as valuable biomarkers of the aging process (Shtumpf et al. 2024; Teo et al. 2019; Li et al. 2024). We applied LongReadAge to track epigenetic aging in long-read plasma cfDNA data derived from two cohorts: one from the Chinese University of Hong Kong (CUHK) (Choy et al. 2022; Yu et al. 2021) and another from the Mass General Brigham (MGB) Biobank, which was newly generated in this study (Figure 4A). To generate this data, we obtained plasma samples from the MGB Biobank, extracted cfDNA, and profiled long-read methylation using PacBio Single Molecule Real-Time (SMRT) sequencing. Across both cohorts, LongReadAge demonstrated a modest ability to track aging (Figure 4B). In the CUHK cohort, our method tracked aging for both women and men when analyzed separately (Figure S4).

The CUHK cfDNA cohort included patients with a broad spectrum of diseases and health conditions, including hepatitis B virus (HBV) carriers, hepatocellular carcinoma (HCC) patients, healthy controls, and cases of early- and late-onset preeclampsia (PET) with respective controls. When comparing the cfDNA age deviation (predicted age minus chronological age) between HCC subjects and controls, as well as between HBV subjects and controls, we observed a marginally significant increased age deviation in the HBV group compared to the healthy group and no difference between HCC patients and healthy controls (Figure 4C). In the comparison between PET subjects and controls, no differences were found in early-onset PET cases, whereas a marginally significant increase in cfDNA age deviation was observed in late-onset PET cases (Figure 4C).

3 | Discussion

Understanding cell-type biology has been indispensable in characterizing various age-related diseases, from dopamine-neuron-specific deterioration in Parkinson's disease (Bloem et al. 2021) to beta-cell-specific dysfunction in type 2 diabetes (Galicía-García et al. 2020). A comprehensive understanding of aging will require similarly granular elucidation of how individual cell types deteriorate over time, and how cell-specific changes relate to aging at tissue, organ, and organismal levels. Presently, however, options for profiling cell type-specific features of aging are limited, and our understanding of aging trajectories between different cell types remains incomplete.

In this study, we present a novel framework, LongReadAge, that predicts the epigenetic age of sparsely intersecting long-read methylation profiles. LongReadAge shows robust performance in capturing the aging of peripheral leukocytes, and its reference-based architecture enables accurate age profiling despite limited numbers of long-read samples and limited shared CpGs across samples.

From bulk leukocyte long-read data, we generated immune-cell-specific methylation profiles that demonstrated a strong concordance with reference cell types. Extending LongReadAge to these methylation profiles revealed robust performance in profiling cell type-specific aging despite vanishingly low numbers of shared CpG sites between samples. We further demonstrated

the ability of LongReadAge to capture epigenetic aging of plasma cfDNA. Although the performance is more modest here, the observed marginally significant increase in cfDNA age deviation in disease states, including hepatitis and preeclampsia, suggests potential applications of our model to evaluating aging in the context of disease. Notably, cfDNA is composed predominantly of white blood cells and erythrocyte progenitors (Moss et al. 2018), which explains why our blood methylation dataset—composed of similar cell types—works as a reference dataset for plasma cfDNA.

Although our study focuses on peripheral blood, the versatility of our approach enables its applications to additional tissues and disease states. It could be used, for instance, to study neuron-specific aging in long-read brain data, or endothelial cell-specific aging in long-read heart data. Another promising application of this approach is tracking organ-specific aging through plasma cfDNA. As cells die throughout the body, their DNA is released into the bloodstream as circulating cfDNA. Thus, although cfDNA is largely composed of white blood cell DNA, it also comprises DNA from various organs throughout the body, a feature leveraged by recently developed liquid biopsies, which have shown promise in diagnosing early-stage solid-tissue cancers based on epigenetic signals from plasma cfDNA (Li et al. 2024; Bloem et al. 2021; Galicía-García et al. 2020). Extending this concept, our approach could be applied to infer the epigenetic age of different solid-tissue organs, tissues, cell types, and perhaps individual cells, from a single blood sample. As long-read sequencing throughput continues to improve and comprehensive reference datasets become more widely available, the prospect of accurately determining the epigenetic age of multiple cell types, including those from solid organs, from a single blood draw is becoming increasingly feasible. This approach could revolutionize our ability to monitor aging at organ-, tissue-, and cell-specific resolution.

Looking ahead, several key steps could enhance this approach. First, developing supervised machine learning models to predict the origin of DNA fragments could improve the robustness of cell-type assignment and may enable more fragments to be mapped to their origin cell type (Hannon and Mill 2024). Second, employing whole-genome reference aging datasets would greatly expand the capabilities of our approach. Currently, our EPIC array-based aging reference covers only 3% of the genome, leaving the vast majority of CpG sites measured in long-read sequencing ineligible for age prediction. Access to whole-genome reference aging datasets would allow us to profile a broader range of cell types and may even enable age prediction at the level of individual DNA fragments. Although whole-genome cell type-specific methylation atlases, such as those developed recently (Loyfer et al. 2023), do exist, their current sample sizes and age ranges are inadequate for use in our model as an aging reference dataset. However, with the rapidly declining costs of DNA sequencing, generating comprehensive, whole-genome, cell type-specific references will likely become increasingly feasible in the future.

Several limitations of this study should be noted. First, our reference aging methylation dataset was derived from whole blood samples. To better capture cell type-specific aging, the reference aging dataset would ideally be cell-specific as well

(e.g., a myeloid cell aging reference for myeloid cell long-read methylation profiles). Although such references do not exist presently, future incorporation could improve the ability of our approach to capture cell type-specific features of aging. Second, our reference-based method, in which different CpGs are used in different samples, introduces inter-sample variability, which may impair accuracy and test-retest reliability. Third, our model currently measures only linear changes corresponding to the passage of chronological time. Although these changes serve as a reasonable first approximation of aging, future models could be constructed to better capture functional fitness of individual cell types or cell-specific deleterious changes. Fourth, the cohort in which our cell type-specific model is validated is limited in demographic diversity, consisting entirely of pregnant women around the age of 20–40 years. Validating this framework in more diverse cohorts will be essential in the future.

In sum, our study introduces a novel approach for profiling epigenetic aging at cell-type resolution using long-read methylation sequencing. This method opens new avenues for investigating cell type-specific aging across various cell types and tissues. As long-read sequencing technologies continue to advance and reference datasets improve, we anticipate that this approach will become an increasingly valuable tool in aging research and biomarker development. Future applications may provide valuable insights into cell type-specific aging in both health and disease and facilitate the development of increasingly biologically precise aging biomarkers.

4 | Methods

4.1 | Ethics

This study was approved by the Mass General Brigham Institutional Review Board under Human Protocol #2021P003059. Details of ethics approval for previously generated datasets used in this study are found in each dataset's respective publication: (Choy et al. 2022; Yu et al. 2021, 2023; Hannon et al. 2021; Loyfer et al. 2023).

4.2 | Long-Read Datasets

Our study utilized long-read datasets from the Chinese University of Hong Kong, which are detailed in publications by Choy et al. (2022), Yu et al. (2023), and Yu et al. (2021). Across these datasets, we used 16 long-read buffy coat samples and 84 cfDNA samples. For long-read data, DNA was extracted from buffy coat and plasma using Qiagen's QIAamp DNA Blood Mini Kit and QIAamp Circulating Nucleic Acid Kit, respectively. Long-read sequencing was then conducted on PacBio's Sequel II platform using the SMRT Cell 8M and the SMRTbell Express Template Prep Kit 2.0.

4.3 | Microarray Datasets

For our reference aging dataset, we used a blood methylation dataset available on GEO under accession GSE152026, as

described in Hannon et al. (2021). Briefly, DNA was extracted from whole blood, bisulfite-converted using the EZ-96 DNA Methylation-Gold Kit, and methylation was profiled on the Illumina EPIC microarray platform. We filtered this dataset to include only control subjects, resulting in 519 samples. To verify the robustness of our method to the reference dataset, we used an additional blood methylation dataset described in Robinson et al. (2020) and available on GEO under accession GSE147740.

4.4 | Whole-Genome Bisulfite Sequencing Datasets

As our cell type-specific methylation atlas, we employed whole-genome methylation sequencing data available on GEO under accession GSE186458 and described in more depth in Loyfer et al. (2023). Specifically, we utilized 36 methylation profiles measured in different blood cell types.

4.5 | MGB Long-Read cfDNA Cohort

Our study generated a new dataset profiling cfDNA methylation using PacBio long-read technology. Plasma samples and health information were obtained from the Mass General Brigham (MGB) Biobank, a biorepository of consented patient samples at MGB, from healthy subjects spanning a broad age range. DNA was extracted from 7 to 8 mL of plasma using Qiagen's Circulating Nucleic Acid Kit, and cfDNA concentration and genomic DNA contamination were assessed using the Agilent TapeStation 4150 automated electrophoresis platform and Agilent Cell-free DNA ScreenTape. DNA was filtered to enrich for longer fragments using Zymo's Select-a-Size MagBead Kit.

For long-read sequencing, five plasma DNA samples (ranging from 34 to 75 ng of DNA) were sequenced on PacBio's Sequel II platform using Single Molecule Real-Time Sequencing. Samples were treated with RNase to eliminate RNA contamination and concentrated using 3× AMPure Beads. Then, libraries were prepared with PacBio's SMRTbell Template Prep Kit 3.0, samples were barcoded and pooled with two or three samples per pool, and polymerase was bound to templates using the Sequel II Binding Kit 3.2. Finally, samples were multiplexed onto two SMRT Cells (8M) and sequenced on the Sequel II with 30-h movie times.

4.6 | Long-Read Processing

All long-read data were processed using PacBio's SMRT Tools software (v12.0). Raw reads (subreads) were converted to circular consensus sequence (CCS) reads using the *ccs* tool, with the *hifi-kinetics* parameter set to output fluorescence kinetics signals, which are used to infer CpG methylation. For the MGB data, samples were demultiplexed using *lima* along with their respective barcodes. CpG methylation information was extracted from the kinetic signals using the *primrose* tool, and the CCS reads were aligned to the hg19 reference genome using *pbmm2 align* under the CCS preset. The CCS reads were then converted to BED files describing CpG methylation across the

genome using *pb-CpG-tools* (v2.3.1) with a pileup mode of count and a minimum coverage of 1.

4.7 | LongReadAge Model Development

LongReadAge is based on the scAge framework. Like scAge, we employed a reference methylation dataset, referred to here as the reference aging dataset, and built linear regression models to predict CpG methylation as a function of age across every CpG site in the reference aging dataset. Our reference was constructed using blood EPIC microarray data. The input to our model is a long-read methylation profile, consisting of a series of CpGs and their methylation statuses, which range from 0 to 1 and represent the proportion of CpGs with a methyl group. These CpGs were intersected with CpGs in the reference EPIC dataset and filtered to include only the top age-correlating CpG sites.

To select the top age-correlating CpG sites, we identified the top 1% of CpGs with the highest absolute Pearson R correlation coefficient with age, unless otherwise noted, consistent with the scAge framework. We validated the robustness of our approach to varied percentage cutoffs in Figure S2. For cell-specific methylation profiles, we observed better performance when using a fixed number of top CpG sites rather than a percentile cutoff. The fixed number of CpG sites was determined by finding the mean number of CpGs used at a 1% percentile cutoff. For the cfDNA samples, several had fewer than 300 intersecting, top-correlating CpG sites at a 1% percentile cutoff. These samples were removed from further analysis.

Using the top-correlating CpG sites, we predicted methylation as a function of age across ages ranging from 0 to 100 years in increments of 0.1 years. For each predicted methylation pattern, we determined how well it matched the observed methylation in the long-read methylation profile by computing a probability score using the following formula, where n is the number of top age-correlating CpG sites in the intersected methylation profile, k is a particular CpG site index, p is the predicted methylation vector, and m is the observed methylation vector.

$$P = \sum_{k=1}^n \log(p_k^{m_k} * (1-p_k)^{(1-m_k)})$$

The age that best predicted the observed methylation pattern, yielding the highest probability score, was designated as the sample's epigenetic age. Further details of this algorithm are described in Trapp et al. (2021).

4.8 | Generating Cell Type-Specific Methylation Profiles

To generate cell type-specific methylation profiles from bulk long-read methylation data, we developed a three-step process. First, differentially methylated regions were identified between cell types of interest. Second, long-read fragments were traced to their cell type of origin. Third, methylation profiles were generated from all reads predicted to originate from a given cell

type. This process was based on methods developed by Choy et al. (2022) and Yu et al. (2021), with several refinements.

4.8.1 | Identifying Differentially Methylated Regions

We first identified differentially methylated regions between cell types of interest by leveraging a cell type-specific reference methylation atlas published in Loyfer et al. (2023) and available at GSE186458. Specifically, we used the 36 samples measured in blood encompassing all major leukocyte cell types.

We categorized each sample into its respective target cell type (myeloid or lymphocyte) (as described in Table S1), then found genomic regions whose methylation patterns differentiated these cell types by applying the software suite *wgbstools* (Loyfer et al. 2024). This process involved two steps: first, identifying genomic segments (or blocks) of similarly methylated regions based on input *beta* files with *segment*, then finding differentially methylated blocks using *find_markers*, which applied several criteria to identify robust differentially methylated regions, including cutoffs for mean methylation, quantiles, *t*-test *p*-value, and coverage. The exact parameters used for these steps are listed in Table S2.

Once these markers were determined, they were mapped from genomic blocks to individual CpGs to enable CpG-specific comparison with long-read fragments, yielding a final table with mean methylation of each cell type across each differentially methylated CpG site, referred to here as the reference cell-type atlas. Of note, a reference cell-type atlas is generated for each cell type of interest.

4.8.2 | Tracing Fragments to Their Cell Type of Origin

To identify long-read fragments' origin cell type, we sought to determine how well a given fragment matched the methylation of a particular cell type in the reference cell-type atlas. To achieve this, we first filtered the fragments to include only autosomal fragments with a minimum mapping quality of 20.

We then identified the long-read fragments with at least 5 CpGs that overlapped with the reference atlas for each cell type. We chose the value 5 as the minimum overlap based on a computational analysis conducted by Yu et al. (2021) that found 5 sites as the minimum for area under the receiver operating characteristic (AUROC) values exceeding 82% in performing tissue-of-origin determination. From these overlapping CpGs, we computed a likelihood score describing how well the long-read fragment matched each cell type in the reference cell-type atlas. The likelihood score was calculated using the following formula where n is the number of overlapping CpGs, k is a particular CpG site index, r is the long-read methylation profile, and c is the cell-type methylation profile in the reference cell-type atlas.

$$L = \sum_{k=1}^n (1 - |r_k - c_k|)$$

Fragments with a likelihood score for one cell type at least (1/0.95) or 1.053× greater than all other cell types were

considered to originate from that cell type. We raised this cutoff from 1.00× (as used in previous studies (Choy et al. 2022; Yu et al. 2021)) to increase accuracy in identifying target reads.

4.8.3 | Generating Cell Type-Specific Methylation Profiles

Finally, we filtered the original CCS long-read files to include only reads predicted to originate from a given cell type using Picard's *FilterSamReads* and generated methylation profiles from the resulting reads using *pb-CpG-tools* (v2.3.1) with a pileup mode of *count* and a minimum coverage of 1.

4.9 | Quantification and Statistical Analysis

All statistical analyses were performed in R (version 4.0.2) and Python (version 3.12). Correlations were computed using the Pearson method, and a Wilcoxon signed-rank test was used to determine the statistical significance of differences between groups. Sample sizes were based on the number of replicates in each dataset, and no power analyses were conducted. A *p*-value threshold of 0.05 was used for determining significance.

Author Contributions

A.E. extracted and sequenced cfDNA in the MGB cohort, obtained access to and processed CUHK datasets, developed the LongReadAge model, performed all analyses, and drafted and revised the manuscript. M.M. guided computational analyses and contributed to manuscript preparation. J.R.P. contributed to sample procurement and assisted with manuscript writing. V.N.G. conceived the study, oversaw research, and wrote and revised the manuscript.

Acknowledgments

We thank the Yale Center for Genome Analysis (YCGA) and Keck Microarray Shared Resource (KMSR) at Yale University for providing the necessary PacBio sequencing services, which are funded in part by the National Institutes of Health instrument grant 1S10OD028669-01. This study makes use of data generated by The Chinese University of Hong Kong (CUHK) Circulating Nucleic Acids Research Group, as reported by Choy et al. in Clin Chem (DOI: [10.1093/clinchem/hvac086](https://doi.org/10.1093/clinchem/hvac086)), Yu et al. in Clin Chem (DOI: [10.1093/clinchem/hvac180](https://doi.org/10.1093/clinchem/hvac180)), and Yu et al. in Proc Natl Acad Sci USA (DOI: [10.1073/pnas.2114937118](https://doi.org/10.1073/pnas.2114937118)).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The CUHK datasets are not publicly available; however, they can be requested by contacting the CUHK Circulating Nucleic Acids Research Group. Data from the newly generated MGB cfDNA dataset is available from GEO under accession [GSE296282](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE296282). Code is available at the following GitHub repository: [LongReadAge](https://github.com/LongReadAge).

References

Aw, D., A. B. Silva, and D. B. Palmer. 2007. "Immunosenescence: Emerging Challenges for an Ageing Population." *Immunology* 120, no. 4: 435–446. <https://doi.org/10.1111/j.1365-2567.2007.02555.x>.

Bloem, B. R., M. S. Okun, and C. Klein. 2021. "Parkinson's Disease." *Lancet* 397, no. 10291: 2284–2303. [https://doi.org/10.1016/S0140-6736\(21\)00218-X](https://doi.org/10.1016/S0140-6736(21)00218-X).

Cappozzo, A., C. McCrory, O. Robinson, et al. 2022. "A Blood DNA Methylation Biomarker for Predicting Short-Term Risk of Cardiovascular Events." *Clinical Epigenetics* 14: 121.

Chien, J. F., H. Liu, B. A. Wang, et al. 2024. "Cell-Type-Specific Effects of Age and Sex on Human Cortical Neurons." *Neuron* 112: 2524–2539. e5. <https://doi.org/10.1016/j.neuron.2024.05.013>.

Choy, L. Y. L., W. Peng, P. Jiang, et al. 2022. "Single-Molecule Sequencing Enables Long Cell-Free DNA Detection and Direct Methylation Analysis for Cancer Patients." *Clinical Chemistry* 68: 1151–1163.

Eames, A., and S. Chandrasekaran. 2024. "Leveraging Metabolic Modeling and Machine Learning to Uncover Modulators of Quiescence Depth." *PNAS Nexus* 3, no. 1: 13. <https://doi.org/10.1093/pnasnexus/pgae013>.

Galicia-Garcia, U., A. Benito-Vicente, S. Jebari, et al. 2020. "Pathophysiology of Type 2 Diabetes Mellitus." *International Journal of Molecular Sciences* 21, no. 17: 6275. <https://doi.org/10.3390/ijms21176275>.

Guo, J., X. Huang, L. Dou, et al. 2022. "Aging and Aging-Related Diseases: From Molecular Mechanisms to Interventions and Treatments." *Signal Transduction and Targeted Therapy* 7, no. 1: 391. <https://doi.org/10.1038/s41392-022-01251-0>.

Hannon, E., E. L. Dempster, G. Mansell, et al. 2021. "Dna Methylation Meta-Analysis Reveals Cellular Alterations in Psychosis and Markers of Treatment-Resistant Schizophrenia." *eLife* 10: e58430.

Hannon, E., and J. Mill. 2024. "Leveraging Epigenetic Signatures to Determine the Cell-Type of Origin From Long Read Sequencing Data." *bioRxiv*.

Jaffe, A. E., and R. A. Irizarry. 2014. "Accounting for Cellular Heterogeneity Is Critical in Epigenome-Wide Association Studies." *Genome Biology* 15: R31.

Levine, M. E., A. T. Lu, A. Quach, et al. 2018. "An Epigenetic Biomarker of Aging for Lifespan and Healthspan." *Aging* 10: 573–591.

Li, S.-J., X. Gao, Z. H. Wang, et al. 2024. "Cell-Free DNA Methylation Patterns in Aging and Their Association With Inflamm-Aging." *Epigenomics* 16: 715–731.

Logsdon, G. A., M. R. Vollger, and E. E. Eichler. 2020. "Long-Read Human Genome Sequencing and Its Applications." *Nature Reviews Genetics* 21, no. 10: 597–614. <https://doi.org/10.1038/s41576-020-0236-x>.

Loyfer, N., J. Magenheimer, A. Peretz, et al. 2023. "A DNA Methylation Atlas of Normal Human Cell Types." *Nature* 613: 355–364.

Loyfer, N., J. Rosenski, and T. Kaplan. 2024. "Wgbstools: A Computational Suite for DNA Methylation Sequencing Data Representation, Visualization, and Analysis." *bioRxiv*.

Lu, A. T., A. Quach, J. G. Wilson, et al. 2019. "DNA Methylation GrimAge Strongly Predicts Lifespan and Healthspan." *Aging* 11: 303–327.

Lu, A. T., Z. Fei, A. Haghani, et al. 2023. "Universal DNA Methylation Age Across Mammalian Tissues." *Nature Aging* 3, no. 9: 1144–1166. <https://doi.org/10.1038/s43587-023-00462-6>.

Lu, Z., M. Zhang, J. Lee, et al. 2023. "Tracking Cell-Type-Specific Temporal Dynamics in Human and Mouse Brains." *Cell* 186: 4345–4364.e24.

Moqri, M., A. Cipriano, D. J. Simpson, et al. 2024. "PRC2-AgeIndex as a Universal Biomarker of Aging and Rejuvenation." *Nature Communications* 15: 5956.

Moss, J., J. Magenheimer, D. Neiman, et al. 2018. "Comprehensive Human Cell-Type Methylation Atlas Reveals Origins of Circulating Cell-Free DNA in Health and Disease." *Nature Communications* 9: 5068.

- Olova, N., F. Krueger, S. Andrews, et al. 2018. "Comparison of Whole-Genome Bisulfite Sequencing Library Preparation Strategies Identifies Sources of Biases Affecting DNA Methylation Data." *Genome Biology* 19: 33.
- Rahmani, E., R. Schweiger, B. Rhead, et al. 2019. "Cell-Type-Specific Resolution Epigenetics Without the Need for Cell Sorting or Single-Cell Biology." *Nature Communications* 10: 3417.
- Ren, Y., X. Wang, S. Zhang, et al. 2023. "Deconvolution Reveals Cell-Type-Specific Transcriptomic Changes in the Aging Mouse Brain." *Scientific Reports* 13: 16855.
- Robinson, O., M. Chadeau Hyam, I. Karaman, et al. 2020. "Determinants of Accelerated Metabolomic and Epigenetic Aging in a UK Cohort." *Aging Cell* 19: e13149.
- Seo, I. H., and Y. J. Lee. 2022. "Usefulness of Complete Blood Count (CBC) to Assess Cardiovascular and Metabolic Diseases in Clinical Settings: A Comprehensive Literature Review." *Biomedicine* 10, no. 11: 2697. <https://doi.org/10.3390/biomedicines10112697>.
- Shtumpf, M., S. Jeong, M. Bikova, H. Mamayusupova, L. Ruje, and V. B. Teif. 2024. "Aging Clock Based on Nucleosome Reorganisation Derived From Cell-Free DNA." *Aging Cell* 23: e14100.
- Soto-Herederó, G., M. M. De Gómez Las Heras, J. I. Escrig-Larena, and M. Mittelbrunn. 2023. "Extremely Differentiated T Cell Subsets Contribute to Tissue Deterioration During Aging." *Annual Review of Immunology* 41: 181–205. <https://doi.org/10.1146/annurev-immunol-101721-064501>.
- Tanaka, K., and A. Okamoto. 2007. "Degradation of DNA by Bisulfite Treatment." *Bioorganic & Medicinal Chemistry Letters* 17: 1912–1915.
- Teo, Y. V., M. Capri, C. Morsiani, et al. 2019. "Cell-Free DNA as a Biomarker of Aging." *Aging Cell* 18: e12890.
- Tharakan, R., C. Ubaida-Mohien, C. Dunn, et al. 2023. "Whole-Genome Methylation Analysis of Aging Human Tissues Identifies Age-Related Changes in Developmental and Neurological Pathways." *Aging Cell* 22: e13847.
- Trapp, A., C. Kerepesi, and V. N. Gladyshev. 2021. "Profiling Epigenetic Age in Single Cells." *Nature Aging* 1: 1189–1201.
- Ying, K., H. Liu, A. E. Tarkhov, et al. 2024. "Causality-Enriched Epigenetic Age Uncouples Damage and Adaptation." *Nature Aging* 4: 231–246.
- Yu, S. C. Y., J. Deng, R. Qiao, et al. 2023. "Comparison of Single Molecule, Real-Time Sequencing and Nanopore Sequencing for Analysis of the Size, End-Motif, and Tissue-Of-Origin of Long Cell-Free DNA in Plasma." *Clinical Chemistry* 69: 168–179.
- Yu, S. C. Y., P. Jiang, W. Peng, et al. 2021. "Single-Molecule Sequencing Reveals a Large Population of Long Cell-Free DNA Molecules in Maternal Plasma." *Proceedings of the National Academy of Sciences of the United States of America* 118: e2114937118.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.