Circulating cell-free DNA methylation profiles as noninvasive multiple sclerosis biomarkers: A proof-of-concept study.

Hailu Fu[1,2], Kevin Huang[3], Wen Zhu[4], Lili Zhang[4], Ravi Bandaru[1,2], Li Wang[1,2,#], Yaping Liu[1,2,#], Zongqi Xia[4,#]

1. Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611
2. Robert H. Lurie Comprehensive Cancer Center of Northwestern University, Chicago, IL 60611
3. Computational Sciences, gRED, Genentech Inc. South San Francisco, CA 94080
4. University of Pittsburgh, Pittsburgh, PA 15260

Corresponding email:
wangliove2@gmail.com (L.W.),
lyping1986@gmail.com (Y.L.),
zxia1@post.harvard.edu (Z.X.)

Abstract

In multiple sclerosis (MS), there is a critical need for non-invasive biomarkers to concurrently classify disease subtypes, evaluate disability severity, and predict long-term progression. In this proof-of-concept study, we performed low-coverage whole-genome bisulfite sequencing (WGBS) on 75 plasma cell-free DNA (cfDNA) samples and assessed the clinical utility of cfDNA methylation as a single assay for distinguishing MS patients from non-MS controls, identifying MS subtypes, estimating disability severity, and predicting disease trajectories. We identified thousands of differentially methylated CpGs and hundreds of differentially methylated regions (DMRs) that significantly distinguished MS from controls, separated MS subtypes, and stratified disability severity levels. These DMRs were highly enriched in immunologically and neurologically relevant regulatory elements (*e.g.,* active promoters and enhancers) and contained motifs associated with neuronal function and T-cell differentiation. To distinguish MS subtypes and severity groups, we achieved area-under-the-curve (AUC) values ranging from 0.67 to 0.81 using DMRs and 0.70 to 0.82 using inferred tissue-of-origin patterns from cfDNA methylation, significantly outperforming benchmark neurofilament light chain (NfL) and glial fibrillary acidic protein (GFAP) in the same cohort. Finally, a linear mixed-effects model identified “prognostic regions” where baseline cfDNA methylation levels were associated with disease progression and predicted future disability severity (AUC=0.81) within a 4-year evaluation window. As we plan to generate higher-depth WGBS data and validation in independent cohorts, the present findings suggest the potential clinical utility of circulating cfDNA methylation profiles as promising noninvasive biomarkers in MS diagnosis and prognosis.

Introduction

Multiple sclerosis (MS) is a chronic autoimmune disorder causing inflammatory demyelination and progressive neurodegeneration, affecting nearly three million people worldwide.[1,2] Based on disease activity and trajectory patterns, people with MS (pwMS) are classified into three main subtypes: Relapsing-Remitting MS (RRMS), Primary Progressive MS (PPMS), and Secondary Progressive MS (SPMS).[1] As the predominant subtype (~85%) during the initial presentation, people with RRMS experience distinct episodes of new or worsening neurological symptoms (*i.e.,* active relapses) followed by periods of partial improvement or complete recovery (*i.e.,* stable remissions).[3] In contrast, people with progressive MS (PMS), including PPMS and SPMS, experience steadily worsening neurodegeneration and neurological disability from initial symptom onset.[4] In real clinical settings, the distinction of PMS from RRMS could take years to recognize in retrospect and cause delays in effective care. While the clinical courses of pwMS can be highly variable,[5] standard MS monitoring tools, such as magnetic resonance imaging (MRI), have limited prognostic value.[1,6] MRI may also not be readily accessible to all patients due to the high cost, scheduling challenges, and patient discomfort. While non-invasive fluid biomarkers, such as neurofilament light chain (NfL)[7] and glial fibrillary acidic protein (GFAP)[8], can inform MS disease activity, they are insufficient to classify MS subtypes or predict disease progression.[9–11] Thus, robust and non-invasive biomarkers that timely ascertain MS subtypes and predict future progression could inform optimal clinical management, including the selection of effective disease-modifying therapy.

Circulating cell-free DNA (cfDNA) in peripheral blood represents a promising non-invasive avenue for monitoring disease activity.[12–15] Under pathological conditions, cfDNA is released not only by peripheral immune cells but also by damaged tissues undergoing apoptosis or necrosis.[16] Previous studies reported substantial alterations in cfDNA methylation at individual promoter regions when comparing RRMS patients with healthy controls.[17–19] Other studies in MS patients found widespread DNA methylation abnormalities in both the central nervous system (CNS) and peripheral immune cells,[20–26] key sources of cfDNA implicated in MS pathogenesis. By integrating cfDNA methylation patterns with publicly available reference methylomes, it is feasible to estimate the cell-type composition of circulating cfDNA and infer disease-related cell death patterns, thereby simultaneously profiling the peripheral immune system and CNS.[27–31] For instance, an increased contribution of CNS-derived cfDNA *may* indicate a progressive MS subtype, greater disability severity, faster progression, or neurodegeneration.

To demonstrate the clinical utility as MS biomarkers, it is crucial to examine the extent to which genome-wide cfDNA methylation signatures and the inferred tissue-of-origin profiles can accurately diagnose MS, differentiate among MS subtypes, assess clinical severity, and predict future disease progression using affordable methods. Here, as a proof-of-concept, we performed low-coverage (~1×) whole-genome bisulfite sequencing (WGBS) on 75 peripheral blood samples obtained from pwMS across MS subtypes and non-MS controls in a clinic-based cohort. We assessed the genome-wide differences in cfDNA methylation and its inferred tissue-of-origin between MS and controls, across MS subtypes, and between individuals with severe disability versus otherwise. Using the cfDNA methylation profiles, we applied machine learning techniques to differentiate MS patients from controls, across MS subtypes, and between disability severity levels in a held-out test set. Finally, we developed a linear mixed model to identify “prognostic regions” in which baseline cfDNA methylation is associated with subsequent disease progression after adjusting for baseline severity and other covariates.

Results

Differentially methylated CpGs and regions separate MS subtypes and from non-MS controls.

We collected 2 mL of plasma from 57 MS patients and 18 non-MS controls in a clinic-based prospective cohort study (*i.e.,* PROMOTE) (**Table 1**). Among the MS samples, 12 were obtained within 3 months of an active

relapse in RRMS, 27 during stable remission in RRMS, and 18 from PMS (**Fig. 1**, **Supplementary Table 1**). We generated cfDNA WGBS data with a median of approximately 56 million paired-end reads per sample. After removing low-quality reads, we retained a median of around 34 million paired-end reads per sample, covering approximately 22 million CpGs at 1.5X per CpG (**Supplementary Table 2**). Using a beta-binomial regression model (RadMeth[48]), we identified significant DNA methylation changes between MS and non-MS controls as well as across MS subtypes (**Fig. 2A**). Specifically, we detected 662 hypermethylated and 1,445 hypomethylated CpGs in MS when compared with non-MS controls ($q<0.1$, absolute methylation difference >0.1). Applying the same model and thresholds, we found 487 hypermethylated and 704 hypomethylated CpGs in PMS relative to RRMS. Similarly, we observed 692 hypermethylated and 370 hypomethylated CpGs in active-relapse RRMS compared with stable-remission RRMS (**Supplementary Table 3**).

We next merged adjacent differentially methylated CpGs into differentially methylated regions (DMRs). This yielded 672 DMRs when comparing MS versus non-MS controls, 342 DMRs when comparing PMS versus RRMS, and 387 DMRs when comparing active-relapse RRMS versus stable-remission RRMS (**Supplementary Table 4**). Due to the inherent noise in methylation measurements from low-coverage WGBS, which may hinder reliable clinical correlations, we adopted a strategy previously developed for single-cell epigenetic analyses.[32,33] We performed singular value decomposition (SVD)[34] on the DNA methylation matrix for DMRs and selected the top components that explained the largest fraction of variability (**Supplementary Fig. 1A**). We then applied Uniform Manifold Approximation and Projection (UMAP)[35] for dimensional reduction, revealing clear separation between MS and non-MS controls and among different MS subtypes (**Fig. 2B**).

DMRs across MS subtypes are associated with neuronal and immune-related gene regulatory processes.

To assess the regulatory potential of these DMRs, we leveraged chromHMM states characterized in immune cells.[36,37] All identified DMRs showed significant enrichment in active promoter and enhancer states and were markedly depleted in quiescent states (**Fig. 2C,** FDR<0.05, Fisher Exact test). Additionally, DMRs differentiating MS from non-MS controls and those distinguishing active-relapse from stable-remission MS were enriched at transcription start site (TSS) flanking states (FDR<0.01, Fisher Exact test), while DMRs between MS and non-MS controls showed significant depletion in weak Polycomb repressive states (FDR<0.01, Fisher Exact test).

Gene Ontology analysis further indicated that these DMRs are highly enriched in both neuronal and immune-related processes (**Fig. 2D, Supplementary Fig. 2, Supplementary Table 5)**. For example, DMRs separating MS from non-MS controls were enriched in “anatomical structure development” and “T cell differentiation,” whereas DMRs separating active-relapse RRMS and stable-remission RRMS were enriched in “synaptic signaling.” Motif analysis revealed multiple transcription factor binding sites (TFBSs) to be significantly overrepresented in all DMR sets (**Supplementary Fig. 3, Supplementary Table 5**). Notably, ZBTB14, a regulator of monocyte and macrophage development,[38] was the most enriched TFBS in DMRs, distinguishing between MS and controls. ZNF609, which modulates myoblast proliferation,[39] was the most enriched TFBS in DMRs, distinguishing between PMS and RRMS. Finally, ZNF37A, previously implicated in myotonic dystrophy,[40] was the top-enriched TFBS in DMRs, distinguishing active-relapse RRMS from stable-remission RRMS.

Tissue-of-origin inferred from cfDNA methylation varied significantly across MS subtypes.

Recent studies have identified the tissue-of-origins of cfDNA molecules through their DNA methylation profiles.[27,29,31] By deconvoluting cfDNA methylation haplotypes, we observed significantly increased neuronal cell death in the PMS patient group compared to other groups and elevated T cell death in both PMS and active-relapse RRMS groups relative to the other MS subtypes (**Fig. 2E, Supplementary Fig. 4, Supplementary Table 6,** one-side Mann-Whitney U test, $p<0.05$). These findings align with previous reports of heightened neuronal

cell death in PMS and T cell-mediated pathogenesis in MS.[41] Interestingly, we also detected markedly lower monocyte cell death in PMS and active-relapse RRMS groups, higher megakaryocyte cell death in those same groups, and reduced erythroid progenitor cell death specifically in the active-relapse RRMS group (**Supplementary Fig. 4,** one-side Mann-Whitney U test, $p<0.05$). While MS pathogenesis research has traditionally centered on T cells, emerging evidence suggests that additional cell types, including peripheral monocytes[42] and megakaryocytes[43], may also contribute to disease development. We interpret these results with caution as they derive from low-coverage WGBS data, which may limit the power to detect tissue-of-origin differences due to the high variability in cfDNA source estimation. Based on simulation analysis, higher-coverage (≥8×) WGBS data will be necessary to validate these observations (**Supplementary Fig. 5**).

cfDNA methylation level and its inferred tissue-of-origin distinguish among MS subtypes and from controls.

Next, we examined whether DMR-based cfDNA methylation levels, together with tissue-of-origin patterns, could distinguish MS from controls and among MS subtypes. Using 10-fold cross-validation, we first identified DMRs and significantly differentiated tissue-of-origin patterns (in CNS and immune cells) within a training set. We then trained machine-learning models on these features and evaluated their performance in a held-out test set. For distinguishing MS from non-MS controls, DMRs alone achieved an area under the curve (AUC) of 0.81 (standard deviation [SD] ± 0.17), while tissue-of-origin features alone yielded AUC 0.70 (SD ± 0.19). DMRs and tissue-of-origin features outperformed neurofilament light chain (NfL, AUC: 0.66) or glial fibrillary acidic protein (GFAP, AUC: 0.58) as reported in a previous, larger study from the same clinic-based cohort as historical comparison.[44] For distinguishing PMS from RRMS, DMRs achieved AUC 0.76 (SD ± 0.20) while tissue-of-origin features achieved AUC 0.78 (SD ± 0.21), outperforming NfL (AUC: 0.67) and GFAP (AUC: 0.61) in historical comparisons. For distinguishing active-relapse RRMS from stable-remission RRMS, DMRs achieved AUCs 0.67 (SD ± 0.29) while tissue-of-origin features achieved AUC 0.77 (SD ± 0.26), outperforming NfL (AUC: 0.53) and GFAP (AUC: 0.59) in historical comparisons. As we plan for validation in an independent dataset, these findings suggest that cfDNA methylation and its inferred tissue-of-origin are promising biomarkers for MS diagnosis and subtyping.

cfDNA methylation level and its inferred tissue-of-origin distinguish disability severity.

We further evaluated whether cfDNA methylation profiles and tissue-of-origin patterns could distinguish MS-related disability severity level, analogous to NfL and GFAP. Using a well-validated and clinically implemented patient-reported outcome (PRO), Patient Determined Disease Steps (PDDS), we categorized MS patients in this study cohort into severe (PDDS≥4) versus normal-mild-moderate (PDDS<4) disability groups based on the requirement for ambulatory assistance around blood collection. We identified 521 hypermethylated and 923 hypomethylated CpGs in the higher disability group compared to the lower disability group (q value <0.1 and absolute methylation difference >0.1) (**Fig. 3A**, **Supplementary Table 7**). Merging adjacent differentially methylated CpGs into DMRs clearly distinguished these two disability severity groups (**Fig. 3B, Supplementary Table 8, Supplementary Fig. 1B**). Notably, the DMRs were highly enriched in active promoter and TSS-flanking regions and depleted in the quiescent regions (**Fig. 3C**), indicating strong gene-regulatory potential. Gene Ontology analysis identified significant enrichment of neuronal processes, *e.g.,* nervous system development and anatomical structure development (**Fig. 3D, Supplementary Fig. 6, Supplementary Table 9)**. In addition, motif analysis revealed substantial overrepresentation of multiple transcription factor binding sites, including members of the AP-2 family previously implicated in cranial neural crest cell development[45] (**Supplementary Fig. 7, Supplementary Table 9**). Comparing tissue-of-origin patterns between the severe and less severe disability groups demonstrated notably lower monocyte cell death and significant variability in epithelial cell death (**Fig. 3E, Supplementary Fig. 8, Supplementary Table 6,** one-side Mann-Whitney U test, $p<0.05$). These

observations corroborate prior studies linking peripheral blood monocyte counts to MS severity[46] and suggesting a potential role for epithelial cells in disease progression.[47]

To evaluate the predictive performance of distinguishing severe from less severe disability status, we used 10-fold cross-validation. First, we identified DMRs and differentially inferred tissue-of-origin patterns (in immune and epithelial cells) in each training fold. Then, we applied machine learning models for evaluation using the corresponding held-out test fold. cfDNA methylation DMRs alone achieved AUC 0.74 (SD ± 0.34), while tissue-of-origin patterns alone yielded AUC 0.82 (SD ± 0.23) (**Fig. 3F)**. Both approaches performed comparably to historical comparisons with NfL and GFAP (AUC: 0.77 each) from the same cohort, suggesting that cfDNA methylation and its inferred tissue-of-origin can accurately assess the severity of MS disease at the time of blood draw.

cfDNA methylation level at prognostic regions can predict future disability progression.

Finally, we evaluated whether baseline cfDNA methylation profile could predict future disability progression in MS. Using longitudinal PDDS data (**Supplementary Table 10**) and baseline cfDNA methylation levels, we constructed a linear mixed-effects model to estimate the interaction between time and baseline cfDNA methylation. After adjusting for covariates (age, gender, and race) and correcting for multiple hypotheses, we identified 7,096 potentially "prognostic regions" (FDR<0.01) whose baseline methylation levels were associated with the trajectory of subsequent disability severity over time (**Fig. 4A, Supplementary Fig. 9, Supplementary Table 11**). To assess predictive utility, we used a cross-validation approach with a gradient boosting model trained on baseline cfDNA methylation in these potentially prognostic regions as features and binary PDDS values (PDDS>5 vs. <2) as outcomes either at baseline (at sample collection) or ~1,500 days post–blood collection. Repeating this procedure 100 times, the performance was inadequate for predicting baseline disability severity (AUC: 0.54, SD ± 0.28) but strong for predicting future disability severity at ~1,500 days (AUC: 0.81, SD ± 0.19) (**Fig. 4B**). Taken together, these findings suggest that baseline cfDNA methylation at prognostic regions holds promise as a biomarker for forecasting long-term disability progression.

Discussions

In this proof-of-concept study to assess the potential clinical utility of cfDNA methylation profiles as MS biomarkers, we generated low-coverage (~1×) WGBS data from plasma cfDNA in 75 samples, including common MS subtypes and non-MS controls. Using a beta-binomial regression model, we identified thousands of differentially methylated CpGs and hundreds of DMRs despite the challenges due to low-coverage sequencing. By integrating cfDNA methylation profiles at these DMRs and their inferred tissue-of-origin patterns, we effectively distinguished MS patients from non-MS controls, differentiated among common MS subtypes, and assessed association with disability severity. We further identified putative gene-regulatory functions for these DMRs and distinct tissue-of-origin patterns among patient groups. Notably, baseline cfDNA methylation profiles showed promising performance in predicting future disability trajectories within a 4-year evaluation window.

This study has several strengths and novelties. First, to our knowledge, this represents the first whole-genome investigation of plasma cfDNA methylation and its tissue-of-origin in MS and MS common subtypes. The tissue-of-origin patterns in MS subtypes and pwMS with different severity levels are not only largely concordant with previous knowledge but also crucially shed light on the potential unknown roles of other peripheral immune cells and epithelial cell types in informing MS subtypes and disease progression. Second, it demonstrates the potential of cfDNA methylation and its tissue-of-origin as biomarkers to *concurrently* diagnose MS, identify MS subtypes, evaluate disability severity, and predict long-term progression. Our preliminary evidence suggests circulating cfDNA methylation profile as a single-assay for both diagnosis and prognosis in MS outperforming other emerging non-invasive blood biomarkers such as NfL and GFAP from the same cohort. Third, we used samples

collected from a real-world clinic cohort and disability outcome measure (*i.e.,* PDDS) already deployed in routine clinical care. The novel use of this longitudinal clinical data to identify prognostic regions underscores the potential of this approach to inform individualized clinical guidance to ultimately improve patient outcomes.

Our pilot study also has several limitations that inform future study design. First, we chose low-coverage WGBS as an affordable method for testing the feasibility of the eventual clinical implementation. While low-coverage WGBS is cost-effective for large cohorts, it also introduces substantial noise, reducing the power to detect DMRs and impeding robust estimation of cfDNA tissue-of-origin. Given that CNS-derived cfDNA is a relatively small fraction of total circulating cfDNA, this limitation is especially relevant for MS. Simulation analyses suggest that effective coverage of at least 8× (~10× raw coverage) would be necessary for reliable tissue-of-origin estimates in this cohort (**Supplementary Fig. 5**). We plan for higher coverage WGBS data generation to identify more definitive biomarkers and facilitate targeted methylation assays for clinical use.

Second, the variable MS disability progression trajectories among pwMS would require larger sample sizes and independent cohort validation to clinically establish biomarkers. Although we compared the cfDNA methylation profiling with other well-recognized MS fluid-based biomarkers (*i.e.,* NfL, GFAP) reported in the same clinic-based cohort, these comparisons did not always involve the exact same subsets of participants and should be viewed as approximate benchmarks. Future studies Incorporating multi-modal data (*e.g.*, cfDNA methylation, tissue-of-origin, other blood protein biomarkers, and standard clinical and demographic features) to train multi-class (rather than binary) classifiers could enhance predictive performance and clinical applicability.

Third, the MS samples included a greater proportion of RRMS than PMS, consistent with the broader clinical population and the clinic-based cohort. For analyses, we grouped PPMS and SPMS. There was insufficient sample size to perform analyses to differentiate PPMS from SPMS, which may have distinct cfDNA methylation patterns.

Finally, the observed methylation differences might reflect either cell-type-specific methylation changes or shifts in the proportion of cells contributing to cfDNA. Future investigation using purified cells or disease-relevant tissues (e.g., CNS and peripheral immune cells) could elucidate the underlying molecular mechanisms.

In conclusion, this proof-of-concept study underscores the potential for circulating cfDNA methylation profiles to serve as a robust and cost-effective fluid-based biomarker for MS diagnosis and prognosis. Future studies are required to validate these findings and establish the clinical utility.

Materials and Methods

Ethics approval

This research study was approved by the University of Pittsburgh Institutional Review Board (STUDY19080007) in accordance with the Declaration of Helsinki. All participants provided written informed consent to participate in the research.

Clinic cohort and sample collection

Participants enrolled in a clinic-based long-term natural history cohort study (Prospective Investigation of Multiple Sclerosis in the Three Rivers Region [PROMOTE], Pittsburgh) during 2017-2021. The current study (Fig. 1) included 57 samples from adults with a neurologist-confirmed diagnosis of MS according to the 2017 McDonald criteria and 18 samples from adult non-MS controls. Clinical and demographic data were collected through a

review of electronic health records. MS subtype and relapse status are determined by the treating neurologist. An acute relapse could be a clinical and/or radiologic event within 3 months of sample collection. We define clinical relapses as having new or recurrent neurological symptoms (deemed MS-relevant by clinicians) lasting persistently ≥24 hours without fever or infection (≥30 days from the onset of a preceding event). We define a radiological relapse as having either a new T1-enhancing lesion and/or a new or enlarging T2-fluid-attenuated inversion recovery (FLAIR) hyperintense lesion based on clinical radiology reports of routine brain, orbit, or spinal cord magnetic resonance imaging studies. Participants completed patient-reported outcomes (PROs), including Patient Determined Disease Steps (PDDS), via electronic or paper questionnaires. Research venous blood samples were collected during clinic visits. Plasma samples were isolated within four hours of phlebotomy by centrifugation at 2,000g for 10 minutes, followed by 15,000g for 10 minutes at room temperature, and stored at -80°C until cfDNA extraction and sequencing.

Whole-genome bisulfite sequencing of cfDNA

cfDNA was extracted from 2mL plasma using the MagMAX™ Cell-Free DNA Isolation Kit (Applied Biosystems A29319) following the manufacturer's protocol. The concentration and size distribution of cfDNA was measured by Qubit 1X dsDNA High Sensitivity kit (Invitrogen) and BioAnalyzer (Agilent). 5ng cfDNA was subjected to bisulfite conversion using the EZ-96 DNA Methylation-Gold Kit (Zymo D5008). Subsequently, cfDNA methylation libraries were prepared using the Accel-NGS® Methyl-Seq DNA Library Kit (Swift 30096) and Methyl-Seq Unique Dual Indexing Kit (SWIFT 390384). To minimize batch effects, all cfDNA samples were processed in a single batch for both bisulfite conversion and library preparation. Qualified libraries were pooled and sequenced on the Illumina NovaSeq S6000 PE150 platform with a 20% spike of PhiX.

Preprocess of whole-genome bisulfite sequencing data

Paired-end WGBS from cfDNA was processed using an internal pipeline. Based on FastQC results on the distribution of four nucleotides along the sequencing cycle, the adapter was trimmed by Trim Galore! (v0.6.0)[49] with cutadapt (v2.1.0)[50] and with parameters "--clip_R1 8 --clip_R2 15 --three_prime_clip_R1 5 --three_prime_clip_R2 5". After the adapter trimming, reads were aligned to the human genome (GRCh37, human_g1k_v37.fa) by Biscuit (v0.3.16.20200420) with default parameters.[51] PCR-duplicate reads were marked by samblaster[52]. Only high-quality reads were used for all the downstream analyses (uniquely mapped, properly paired, mapping quality score of 30 or greater, and not a PCR duplicate). The methylation level at each CpG was called by Bis-SNP (v0.90)[53] with default parameters in bissnp_easy_usage.pl. Only CpGs at autosomes were used for downstream analyses.

Identification of differentially methylated CpGs and DMRs from WGBS data

The beta-binomial model implemented in RADMeth[48] was used to calculate DMRs between different subtypes of MS, between MS and non-MS controls, and between MS patients with high and low PDDS scores. Sex was utilized as a covariate in the modeling. "-bins 1:100:1" was used as parameters for command *adjust*. *dmrs* was used with -p 0.1 to merge differentially methylated CpGs nearby. A threshold of q value <0.1 and absolute methylation difference >0.1 was used to filter DMRs.

Visualization of DMRs from WGBS data

DMRs from different MS subtypes were merged into a single location file. DNA methylation levels at these DMRs from 75 samples were calculated. Only DMRs with <80% missing data were kept for the analysis. The *impute.knn* function from *impute* package (v1.70.0) in R (v4.2.0) was used to impute the missing data with parameters of

“maxp="p", rowmax = 0.8”. After *scale* transformation, *svd* function in R was applied to the methylation matrix. We plotted the first 50 singular values against the proportion of variance explained and chose the elbow point. For DMRs when comparing different MS subtypes, we chose the top six singular value decomposition components (SVDs). For DMRs when comparing between MS patients with high vs low PDDS scores, we chose the top four SVDs. Then, we used *umap* (v0.2.10.0) in R for dimension reduction and visualization.

Enrichment analysis of DMRs

The 15-state chromHMM result at GM12878 from Epigenome Roadmap was used to calculate the enrichment of DMRs. The p-value was calculated against random intervals with the same chromosome and sizes using the Fisher exact test after multiple hypothesis corrections (via Benjamini-Hochberg). DMRs were lifted up to the hg38 genome by liftOver[54] and Gene Ontology (GO) and motif enrichment were computed by g:profile.[55]

Tissue-of-origin analysis by using cfDNA WGBS

Only high-quality reads mapped to the cell-type-specific marker regions were used for the tissue-of-origin analysis. The cell-type-specific DNA methylation markers (Atlas.U25.l4.hg19.tsv) were downloaded from https://raw.githubusercontent.com/nloyfer/UXM_deconv/main/supplemental/. The *wgbstools (v0.2.2) bam2pat* was used with default parameters to convert the bam files to the pat format.[29] Then, *uxm deconv* was used estimate the tissue-of-origin fraction in each sample with default parameters.[29]

WGBS coverage simulation for assessing tissue-of-origin accuracy

We merged processed bam files from non-MS control samples, filtered low-quality reads, and achieved ~20X effective coverage representing the population average. We randomly sampled from the merged bam files at different genomic coverages ranging from 0.01x to 12x and estimated tissue-of-origin fraction as mentioned above at each step. We repeated this process 100 times to build a distribution and estimate variance around the sample mean at different coverages.

The linear mixed-effects model reveals the interaction effect between baseline cfDNA methylation level and time on the trajectory of disability severity

To evaluate the association between baseline biomarker (cfDNA methylation) levels and disability progression over time, we conducted a linear mixed-effects analysis using the *lmer* function from the *lme4* package[56] in R (v4.2.0). The outcome variable was repeatedly assessed patient-reported disability severity (*e.g.*, PDDS), collected at multiple time points for each patient after baseline sample collection (for cfDNA isolation). Fixed effects included: (1) time (treated as a continuous variable), (2) baseline biomarker group (categorized as “High” vs. “Low” based on median cfDNA methylation levels in specific genomic regions), and (3) an interaction term (time × biomarker group) to assess potential differences in disability severity trajectories between groups. We controlled for age (continuous), gender, race, and ethnicity as additional covariates. (This study population included only one ethnicity group, *i.e.,* non-Hispanic). A random intercept for each patient accounted for repeated within-patient observations.

The model specification is as follows:

$$\text{Severity}_{ij} = \beta_0 + \beta_1 \times \text{Time}_{ij} + \beta_2 \times \text{Group}_i + \beta_3 \times (\text{Time}_{ij} \times \text{Group}_i) + \beta_4 \times \text{Age}_i + \beta_5 \times \text{Gender}_i + \beta_6 \times \text{Race}_i + \beta_7 \times \text{Ethnicity}_i + u_i + \epsilon_{ij}$$

Where:

- $Severity_{ij}$: Disability severity for patient i at time j.
- β_0: Fixed intercept.
- β_1: Fixed effect of time.
- β_2: Fixed effect of biomarker group (cfDNA methylation level).
- β_3: Interaction effect between time and biomarker group.
- β_4, β_5, β_6, and β_7 represent the effects of age, gender, race, and ethnicity, respectively.
- u_i: Random intercept for patient i.
- ϵ_{ij}: Residual error term.

We used the *lmerTest* package[57] to obtain p-values for the fixed effects using Satterthwaite's method for approximating degrees of freedom. Model assumptions were checked by examining residual plots for homoscedasticity and normality. Statistical significance was set at a two-sided alpha level of 0.01. Adjustments for multiple comparisons were made using the Bonferroni procedure when testing multiple biomarkers to control the false discovery rate.

PDDS interpolation

Given the pragmatic study design of the clinic-based cohort, PDDS scores were not always collected at identical time points for all participants. To address this, we performed linear interpolation to facilitate comparisons of baseline cfDNA methylation profiles with longitudinal trajectories of PDDS scores. Using the *interp1d* function from *SciPy* (v1.12.0)[58], we interpolated each participant's scores at 50-day intervals, covering the period from day 0 to day 2,000 after blood sample collection. For patients with observed values beyond 2,000 days, we also include the participant's maximum observed PDDS score.

Candidate genomic regions for the linear mixed model

We first selected genomic regions without completely missing data between MS patients with high PDDS (PDDS≥4) and low PDDS score (PDDS< 4) at baseline. We then merged regions located within 300 bp of one another using bedtools,[59] resulting in 2,316,371 total regions. Samples exhibiting >90% missing values across these regions were excluded from subsequent analyses.

Machine learning models: general comments

For all machine learning studies described below, we evaluated seven commonly used machine learning models: LogisticRegression with "l2" or "elasticnet," support vector machine (SVM) with "rbf" or "linear" kernels, XGBoost, GradientBoostingClassifier, and LightGBM's LGBMClassifier. Parameter tuning was conducted within each model. For DMR features, we tested various transformations (*e.g.*, raw input, top 2–20 SVD components, recursive feature elimination). A unified model or an ensemble approach is needed in future studies.

Machine learning models for comparing MS vs non-MS controls

We employed a 10-fold cross-validation scheme. In each fold, 90% of the MS cases and 90% of the controls constituted the training set, while the remaining 10% served as the held-out test set. This approach ensured balanced case/control ratios in both training and test sets. Using a beta-binomial model (RadMeth), we identified differentially methylated CpGs and merged adjacent sites to form DMRs (≥3 CpGs, length <100 bp). Methylation densities were computed for each DMR. Missing data were marked as NA and processed using NumPy's

nan_to_num function. We then performed singular value decomposition (SVD) and selected the top two SVD components as input features. We trained a logistic regression model (LogisticRegression from scikit-learn, v1.1.1)[60] with the following parameters:

"penalty='l2', dual=False, tol=1e-5, C=120, fit_intercept=True, intercept_scaling=1, class_weight='balanced', solver='lbfgs', max_iter=1000".

To calculate all possible differential tissue-of-origin patterns in CNS and immune cells, we conducted a one-sided Mann-Whitney U test to identify cell types (within CNS and immune cells) that differed significantly ($p<0.05$) between group comparisons (*e.g.*, PMS vs other subtypes, PMS + A-RRMS vs. S-RRMS + controls). Cell types meeting this criterion in any comparison were included as features. We trained a gradient boosting classifier (GradientBoostingClassifier from scikit-learn) with the following parameters:

"n_estimators=200, learning_rate=0.2, max_depth=10, min_samples_split=2,min_samples_leaf=1,subsample=0.8,max_features='log2'".

Because each participant had only one NfL or GFAP value, we applied a simple logistic regression model (C=1.0, using *StratifiedKFold* for the 10-fold cross-validation).

Machine learning models for comparing PMS vs RRMS

We used the same 10-fold cross-validation scheme and preprocessing steps (as for the MS vs non-MS comparison). For DMRs features, instead of taking the top SVD components, we employed recursive feature elimination with cross-validation (RFECV from scikit-learn) in each fold to refine the final features ("step=1,scoring='accuracy',min_features_to_select=1"). We trained a LightGBM classifier (LGBMClassifier in *lightgbm*) with the following parameters:

"n_estimators=100,learning_rate=0.1,num_leaves=31,max_depth=-1,reg_alpha=0.1,reg_lambda=0.1,subsample=0.8, colsample_bytree=0.8,min_child_samples=10".

For differential tissue-of-origin features, given the smaller sample size of PMS relative to RRMS, we relaxed the tissue-of-origin p-value cutoff to 0.1. We again performed a one-sided Mann-Whitney U test ($p<0.1$) to select significantly altered cell types. We trained a SVM with the following parameters: "C=8.0, kernel='rbf', probability=True".

Machine learning models for comparing active relapse RRMS vs stable remission RRMS

For DMR features, we retained the top three SVD components of DMR methylation levels as input. We trained an SVM model with the following parameters: "kernel='rbf', probability=True,gamma='scale',C=0.1".

For differential tissue-of-origin features, given the smaller sample size of active relapse RRMS relative to stable remission RRMS, we relaxed the p-value cutoff to 0.2. We then used a logistic regression model with the following parameters: "penalty='l2', dual=False, tol=0.0001, C=100, fit_intercept=True, intercept_scaling=1, class_weight='balanced', solver='lbfgs', max_iter=100".

Machine learning models to distinguish different disability severity levels

We excluded samples lacking PDDS scores, retaining 67. To ensure clear phenotypic distinction, we only used samples with PDDS <2 or >5, further reducing the dataset to 46. We applied the same 10-fold cross-validation and preprocessing strategies (as in MS vs non-MS comparisons). For DMRs features, we utilized xgboost[61] with the following parameters: "max_depth=3, learning_rate=0.4, n_estimators=1000, class_weight='balanced', objective='binary:logistic',subsample=0.667".

For differential tissue-of-origin features, we tested all immune and epithelial cell types using a p-value cutoff of 0.1, then trained a GradientBoostingClassifier with following parameters:

“n_estimators=200, learning_rate=0.15, max_depth=2, min_samples_split=2,min_samples_leaf=2,subsample=0.8,max_features='log2'”.

Machine learning approach to predict the future disease severity

We included only samples with PDDS values at ≥3 distinct time points. We focused on 7,096 potentially “prognostic regions” identified by the linear mixed-effects model. At day 0 or ~1,500 days, samples with PDDS < 2 or > 5 were retained for analysis. We used a RepeatedStratifiedKFold (five folds, repeated 100 times). The GradientBoostingClassifier was trained with the following parameters:

“n_estimators=100, learning_rate=0.1, max_depth=5, min_samples_split=2,min_samples_leaf=1,subsample=0.8,max_features='log2'”.

Statistics and Reproducibility

No statistical method was used to predetermine the sample size. The experiments were randomized to generate cfDNA sequencing libraries. The investigators were not blinded to allocation during experiments and outcome assessment.

Data Availability

The raw WGBS sequencing data generated in this study have been deposited in the Sequence Read Archive with controlled access from dbGaP under accession code (pending). The data are available under restricted access, and access can be obtained by contacting the Data Access Committee in dbGaP. The raw sequencing data are protected by data privacy laws. Processed DNA methylation level is available at Gene Expression Omnibus (GEO) with access ID (pending). The processed and de-identified data are available at zenodo.org (doi: https://doi.org/10.5281/zenodo.14803482). The remaining data are available within the Article, Supplementary Information, and Source Data file. The sample id in the supplementary information and source data cannot reveal the identity of the study subjects.

Code Availability

Analysis scripts are publicly available on GitHub under the MIT license for academic researchers: https://github.com/epifluidlab/ms_cfnda_wgbs_manuscript. The zipped code and meta-data are also available on zenodo.org (doi: https://doi.org/10.5281/zenodo.14803482).

Acknowledgments

This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University, which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. This research was also supported in part through the computational resources and staff contributions provided by the Genomics Compute Cluster, which is jointly supported by the Feinberg School of Medicine, the Center for Genetic Medicine, and Feinberg's Department of Biochemistry and Molecular Genetics, the Office of the Provost, the Office for Research, and Northwestern Information Technology. The Genomics Compute Cluster is part of Quest, Northwestern University's high performance computing facility, with the purpose to advance research in genomics. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant number ACI-1548562. This work used the XSEDE at the Pittsburgh Supercomputing Center (PSC) through allocation MCB190124P and MCB190006P. L.W. and Y.L. are supported by the startup grant to Y.L. from Northwestern University, Robert H. Lurie Comprehensive Cancer Center of Northwestern University, and the iDEA-TECH award from Sanofi Inc. (to Y.L. and Z.X.). Z.X. is supported by NINDS (R01NS098023 and R01NS124882).

Author Contributions

L.W., Y.L., and Z.X. conceived the study. H.F. and L.W. performed the cfDNA extraction and library constructions. Y.L., R.B., and K.H. performed the data analysis with input from Z.X.. Z.X., W.Z., and L.Z. provided the clinical samples and guidance related to the clinic applications. H.F., R.B., W.Z., L.W., Y.L., and Z.X. wrote the manuscript together. All authors read and approved the final manuscript.

Competing Interests Statement

Y.L. and Z.X. filed a provisional patent. Y.L. owns stocks from Freenome Inc. The remaining authors declare no competing interests.

References

1. Reich, D. S., Lucchinetti, C. F. & Calabresi, P. A. Multiple Sclerosis. *N. Engl. J. Med.* **378**, 169–180 (2018).
2. Walton, C. *et al.* Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. *Mult. Scler.* **26**, 1816–1821 (2020).
3. Dutta, R. & Trapp, B. D. Relapsing and progressive forms of multiple sclerosis: insights from pathology. *Curr. Opin. Neurol.* **27**, 271–278 (2014).
4. Koch, M., Kingwell, E., Rieckmann, P. & Tremlett, H. The natural history of primary progressive multiple sclerosis. *Neurology* **73**, 1996–2002 (2009).
5. Amin, M. & Hersh, C. M. Updates and advances in multiple sclerosis neurotherapeutics. *Neurodegener. Dis. Manag.* **13**, 47–70 (2023).
6. Yang, J. *et al.* Current and Future Biomarkers in Multiple Sclerosis. *Int. J. Mol. Sci.* **23**, (2022).
7. Kuhle, J. *et al.* Serum neurofilament light chain in early relapsing remitting MS is increased and correlates with CSF levels and with MRI measures of disease severity. *Mult. Scler.* **22**, 1550–1559 (2016).
8. Högel, H. *et al.* Serum glial fibrillary acidic protein correlates with multiple sclerosis disease severity. *Mult. Scler.* **26**, 210–219 (2020).
9. Ayrignac, X. *et al.* Serum NfL and GFAP are weak predictors of long-term multiple sclerosis prognosis: A 6-year follow-up. *Mult. Scler. Relat. Disord.* **89**, 105747 (2024).
10. Ferreira-Atuesta, C., Reyes, S., Giovanonni, G. & Gnanapavan, S. The Evolution of Neurofilament Light Chain in Multiple Sclerosis. *Front. Neurosci.* **15**, 642384 (2021).
11. Lei, J. *et al.* Glial fibrillary acidic protein as a biomarker in severe traumatic brain injury patients: a prospective cohort study. *Crit. Care* **19**, 362 (2015).
12. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).
13. Foda, Z. H. *et al.* Detecting Liver Cancer Using Cell-Free DNA Fragmentomes. *Cancer Discov.* **13**, 616–631 (2023).
14. Mathios, D. *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat. Commun.* **12**, 5060 (2021).

15. Liu, Y. At the dawn: cell-free DNA fragmentomics and gene regulation. *Br. J. Cancer* (2022).

16. Lo, Y. M. D., Han, D. S. C., Jiang, P. & Chiu, R. W. K. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372**, (2021).

17. Liggett, T. *et al.* Methylation patterns of cell-free plasma DNA in relapsing-remitting multiple sclerosis. *J. Neurol. Sci.* **290**, 16–21 (2010).

18. Lehmann-Werman, R. *et al.* Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1826–34 (2016).

19. Dunaeva, M., Derksen, M. & Pruijn, G. J. M. LINE-1 Hypermethylation in Serum Cell-Free DNA of Relapsing Remitting Multiple Sclerosis Patients. *Mol. Neurobiol.* **55**, 4681–4688 (2018).

20. Celarain, N. & Tomas-Roig, J. Aberrant DNA methylation profile exacerbates inflammation and neurodegeneration in multiple sclerosis patients. *J. Neuroinflammation* **17**, 21 (2020).

21. Chomyk, A. M. *et al.* DNA methylation in demyelinated multiple sclerosis hippocampus. *Sci. Rep.* **7**, 8696 (2017).

22. Maltby, V. E. *et al.* Genome-wide DNA methylation profiling of CD8+ T cells shows a distinct epigenetic signature to CD4+ T cells in multiple sclerosis patients. *Clin. Epigenetics* **7**, 118 (2015).

23. Graves, M. C. *et al.* Methylation differences at the HLA-DRB1 locus in CD4+ T-Cells are associated with multiple sclerosis. *Mult. Scler.* **20**, 1033–1041 (2014).

24. Kulakova, O. G. *et al.* Whole-genome DNA methylation analysis of peripheral blood mononuclear cells in multiple sclerosis patients with different disease courses. *Acta Naturae* **8**, 103–110 (2016).

25. Huynh, J. L. *et al.* Epigenome-wide differences in pathology-free regions of multiple sclerosis–affected brains. *Nat. Neurosci.* **17**, 121–130 (2014).

26. Bos, S. D. *et al.* Genome-wide DNA methylation profiles indicate CD8+ T cell hypermethylation in multiple sclerosis. *PLoS One* **10**, e0117403 (2015).

27. Moss, J. *et al.* Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 5068 (2018).

28. Liu, Y. *et al.* FinaleMe: Predicting DNA methylation by the fragmentation patterns of plasma cell-free DNA. *Nat. Commun.* **15**, 2790 (2024).

29. Loyfer, N. *et al.* A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).

30. Li, S. *et al.* Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2305236120 (2023).

31. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5503–12 (2015).

32. Nichols, R. V. *et al.* High-throughput robust single-cell DNA methylation profiling with sciMETv2. *Nat. Commun.* **13**, 7627 (2022).

33. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).

34. Klema, V. & Laub, A. The singular value decomposition: Its computation and some applications. *IEEE Trans. Automat. Contr.* **25**, 164–176 (1980).

35. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).

36. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

37. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

38. Deng, Y. *et al.* Zbtb14 regulates monocyte and macrophage development through inhibiting pu.1 expression in zebrafish. *Elife* **11**, (2022).

39. Legnini, I. *et al.* Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis. *Mol. Cell* **66**, 22–37.e9 (2017).

40. Gauthier, M. *et al.* A defective Krab-domain zinc-finger transcription factor contributes to altered myogenesis in myotonic dystrophy type 1. *Hum. Mol. Genet.* **22**, 5188–5198 (2013).

41. Kaskow, B. J. & Baecher-Allan, C. Effector T cells in multiple sclerosis. *Cold Spring Harb. Perspect. Med.* **8**, (2018).

42. Ajami, B., Bennett, J. L., Krieger, C., McNagny, K. M. & Rossi, F. M. V. Infiltrating monocytes trigger EAE progression, but do not contribute to the resident microglia pool. *Nat. Neurosci.* **14**, 1142–1149 (2011).

43. Dziedzic, A., Miller, E., Bijak, M., Przyslo, L. & Saluk-Bijak, J. Increased pro-thrombotic platelet activity associated with thrombin/PAR1-dependent pathway disorder in patients with secondary progressive multiple sclerosis. *Int. J. Mol. Sci.* **21**, 7722 (2020).
44. Zhu, W. *et al.* Association between serum multi-protein biomarker profile and real-world disability in multiple sclerosis. *Brain Commun.* **6**, fcad300 (2024).
45. Eckert, D., Buhl, S., Weber, S., Jäger, R. & Schorle, H. The AP-2 family of transcription factors. *Genome Biol.* **6**, 246 (2005).
46. Akaishi, T., Takahashi, T. & Nakashima, I. Peripheral blood monocyte count at onset may affect the prognosis in multiple sclerosis. *J. Neuroimmunol.* **319**, 37–40 (2018).
47. Peng, H.-R. *et al.* Intestinal epithelial dopamine receptor signaling drives sex-specific disease exacerbation in a mouse model of multiple sclerosis. *Immunity* **56**, 2773–2789.e8 (2023).
48. Dolzhenko, E. & Smith, A. D. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics* **15**, 215 (2014).
49. Krueger, F. Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* **516**, 517 (2015).
50. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
51. Zhou, W. *et al.* BISCUIT: an efficient, standards-compliant tool suite for simultaneous genetic and epigenetic inference in bulk and single-cell studies. *Nucleic Acids Res.* **52**, e32 (2024).
52. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
53. Liu, Y., Siegmund, K. D., Laird, P. W. & Berman, B. P. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.* **13**, R61 (2012).
54. Perez, G. *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* **53**, D1243–D1249 (2025).
55. Kolberg, L. *et al.* g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res.* **51**, W207–W212 (2023).

56. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models Usinglme4. *J. Stat. Softw.* **67**, (2015).

57. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. LmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* **82**, (2017).

58. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

59. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

60. Buitinck, L. *et al.* API design for machine learning software: experiences from the scikit-learn project. *arXiv [cs.LG]* (2013).

61. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv [cs.LG]* (2016) doi:10.48550/ARXIV.1603.02754.

Table and Figures

	Control (N=18)	Actively Relapsing (N=12)	Stable Remitting (N=27)	Progressive (N=18)	Total (N=75)	p value
Gender						0.017
Female	9 (50.0%)	7 (58.3%)	23 (85.2%)	15 (83.3%)	54 (72.0%)	
Male	9 (50.0%)	5 (41.7%)	4 (14.8%)	3 (16.7%)	21 (28.0%)	
Race						0.522
African American	1 (5.6%)	3 (25.0%)	2 (7.4%)	3 (16.7%)	9 (12.0%)	
Asian	0 (0.0%)	0 (0.0%)	1 (3.7%)	0 (0.0%)	1 (1.3%)	
White	17 (94.4%)	9 (75.0%)	24 (88.9%)	15 (83.3%)	65 (86.7%)	
Age at sample collection (yrs)						0.053
Mean (SD)	57.333 (10.954)	41.500 (9.511)	48.333 (11.582)	61.056 (7.689)	52.453 (12.295)	
Range	27 - 70	27 - 57	28 - 71	46 - 71	27 - 71	
PDDS at sample collection						<0.001
Mean (SD)	0.607 (1.883)	2.375 (2.175)	2.861 (2.215)	4.991 (1.755)	2.812 (2.484)	
Range	0 - 7	0 - 5.800	0 - 7.200	1.875 - 7	0 - 7.200	
Missing	4	0	1	2	7	

Table 1. Meta-data of the pilot study cohort

Note: *PDDS,* patient determined disease steps

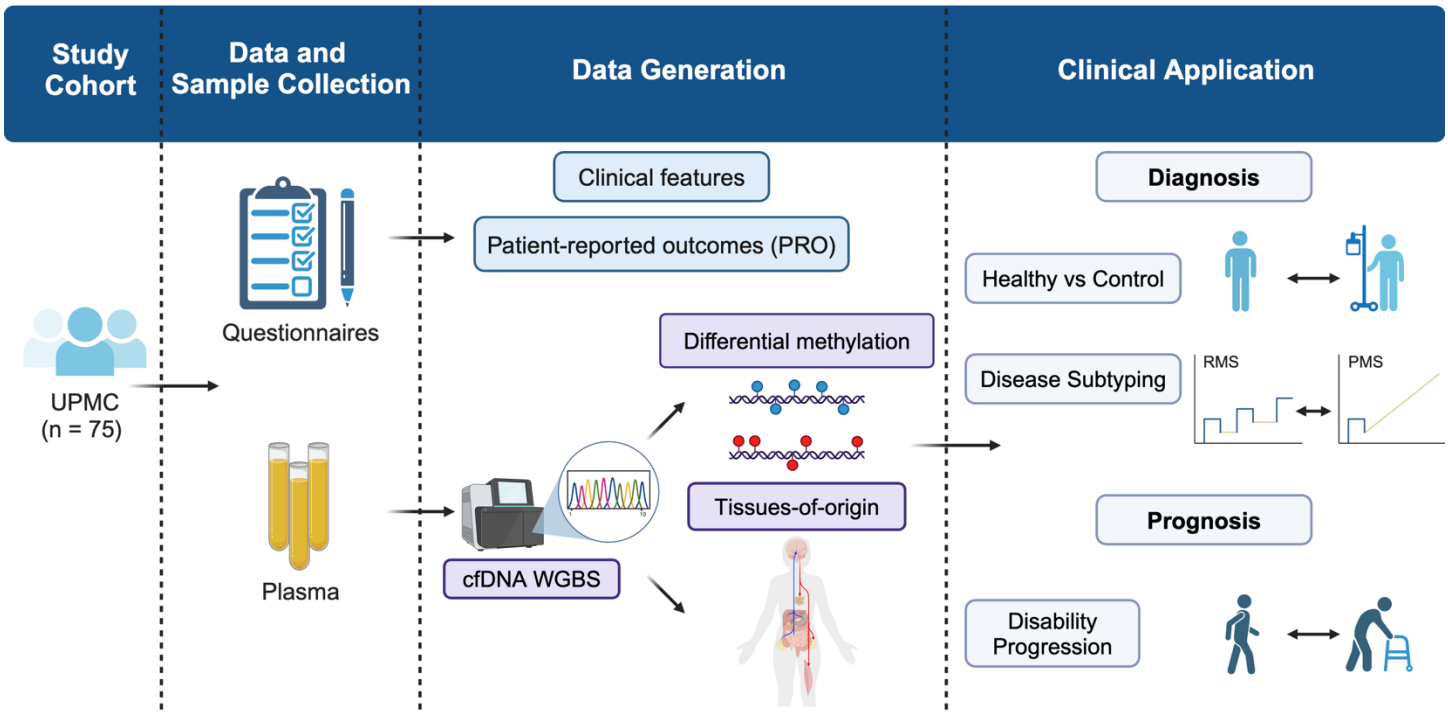


Figure 1. Overview of the study design.

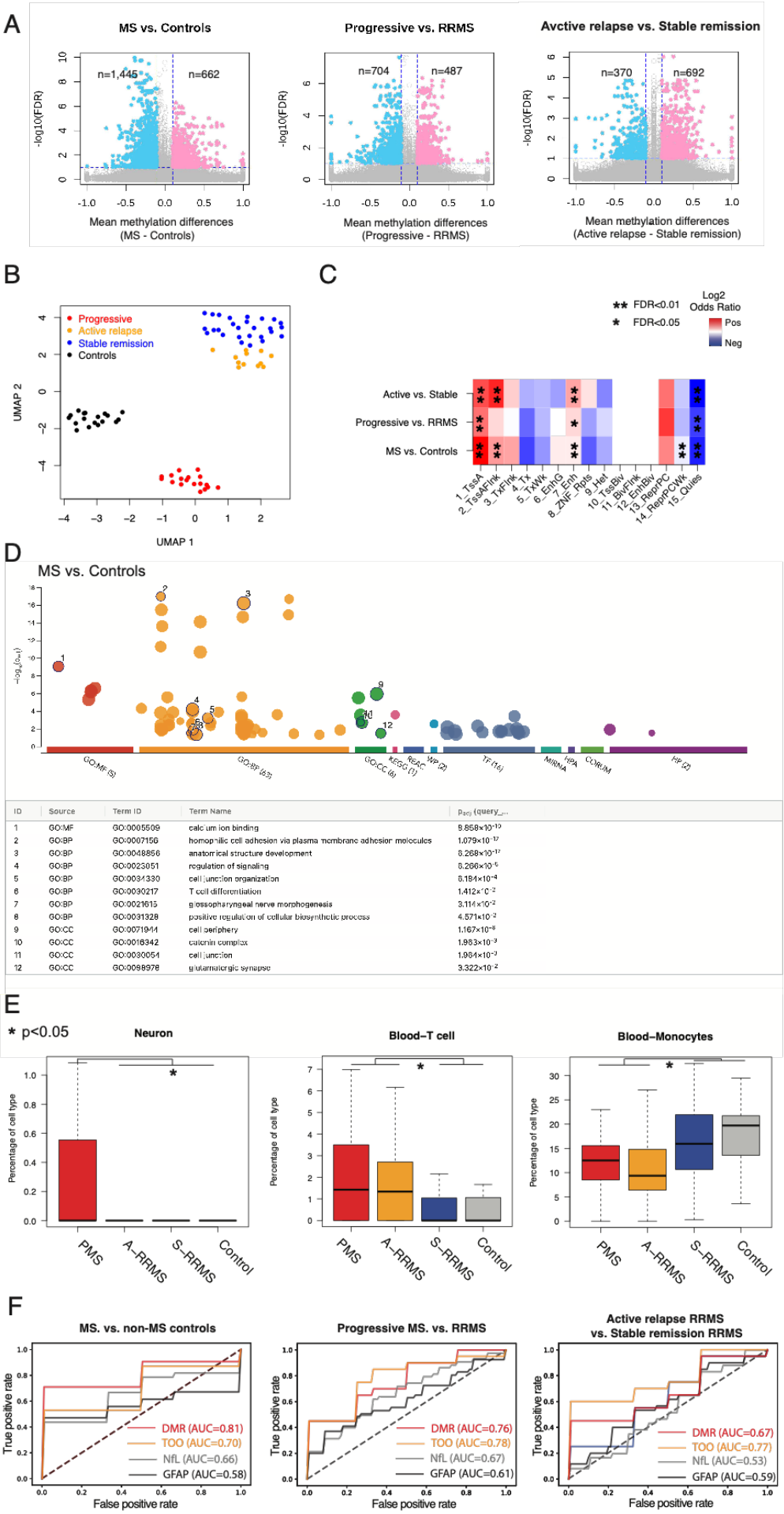
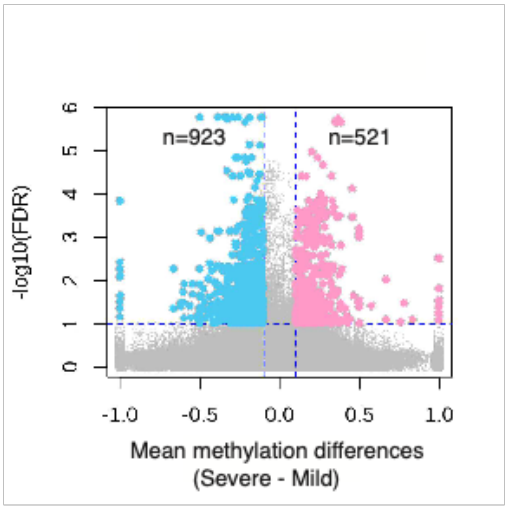

A
MS vs. Controls
n=1,445
n=662
-log10(FDR)
Mean methylation differences
(MS - Controls)
Progressive vs. RRMS
n=704
n=487
Mean methylation differences
(Progressive - RRMS)
Avctive relapse vs. Stable remission
n=370
n=692
Mean methylation differences
(Active relapse - Stable remission)
B
Progressive
Active relapse
Stable remission
Controls
UMAP 2
UMAP 1
C
** FDR<0.01
* FDR<0.05
Log2
Odds Ratio
Pos
Neg
Active vs. Stable
Progressive vs. RRMS
MS vs. Controls
1_TssA
2_TssAFlnk
3_TxFlnk
4_Tx
5_TxWk
6_EnhG
7_Enh
8_ZNF_Rpts
9_Het
10_TssBiv
11_BivFlnk
12_EnhBiv
13_ReprPC
14_ReprPCWk
15_Quies
D
MS vs. Controls
GO:MF (3)
GO:BP (8,3)
GO:CC (6)
KEGG (1)
REAC
WP (2)
TF (15)
MIRNA
HPA
CORUM
HP (2)
ID | Source | Term ID | Term Name | p_{adj} (query_...
1 | GO:MF | GO:0005509 | calcium ion binding | 9.858×10^{-19}
2 | GO:BP | GO:0007156 | homophilic cell adhesion via plasma membrane adhesion molecules | 1.079×10^{-17}
3 | GO:BP | GO:0048856 | anatomical structure development | 8.268×10^{-17}
4 | GO:BP | GO:0023051 | regulation of signaling | 8.266×10^{-5}
5 | GO:BP | GO:0034330 | cell junction organization | 8.184×10^{-4}
6 | GO:BP | GO:0030217 | T cell differentiation | 1.412×10^{-2}
7 | GO:BP | GO:0021615 | glossopharyngeal nerve morphogenesis | 3.114×10^{-2}
8 | GO:BP | GO:0031328 | positive regulation of cellular biosynthetic process | 4.571×10^{-2}
9 | GO:CC | GO:0071944 | cell periphery | 1.167×10^{-8}
10 | GO:CC | GO:0016342 | catenin complex | 1.983×10^{-3}
11 | GO:CC | GO:0030054 | cell junction | 1.984×10^{-3}
12 | GO:CC | GO:0098978 | glutamatergic synapse | 3.322×10^{-2}
E
* p<0.05
Neuron
Blood–T cell
Blood–Monocytes
Percentage of cell type
PMS
A–RRMS
S–RRMS
Control
F
MS. vs. non-MS controls
DMR (AUC=0.81)
TOO (AUC=0.70)
NfL (AUC=0.66)
GFAP (AUC=0.58)
Progressive MS. vs. RRMS
DMR (AUC=0.76)
TOO (AUC=0.78)
NfL (AUC=0.67)
GFAP (AUC=0.61)
Active relapse RRMS
vs. Stable remission RRMS
DMR (AUC=0.67)
TOO (AUC=0.77)
NfL (AUC=0.53)
GFAP (AUC=0.59)
True positive rate
False positive rate

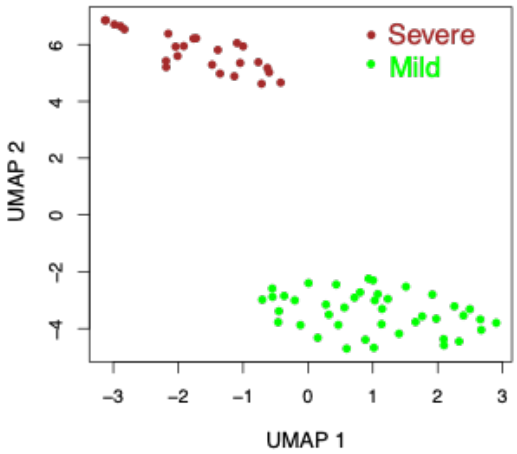
Figure 2. Circulating cfDNA methylation and its inferred tissue-of-origin patterns can distinguish MS from non-MS controls and among different MS subtypes.

A. Scatter plots of -log10(FDR) vs. mean methylation differences at differentially methylated CpGs between MS and non-MS controls (left panel), between progressive MS (PMS: PPMS and SPMS) and RRMS (middle panel), and between active relapse RRMS and stable remission RRMS (right panel). Colors indicate CpGs with FDR<0.01 and absolute mean methylation difference >0.1. **B**. Uniform Manifold Approximation and Projection (UMAP) representation of cfDNA methylation level in DMRs distinguish MS patients across MS subtypes. **C**. The enrichment of DMRs in different chromHMM states (that were previously computed from GM12878 cells in the Epigenome Roadmap project). **D**. The gene ontology enrichment analysis of DMRs between MS and non-MS controls. **E**. Differences in cell death contribution from brain Neuron (left panel), T cells (middle panel), and Monocytes (right panel) across different MS subtypes (PMS, PPMS, and SPMS; A-RRMS, active relapsing RRMS; S-RRMS: stable remission RRMS) and non-MS controls. **F**. Receiver operating characteristic (ROC) curve by DMR only (red), Tissue-of-origin only (TOO, orange), NfL only (grey), and GFAP only (black) to distinguish MS and non-MS controls (left panel), progressive MS and RRMS (middle panel), and active relapse RRMS and stable remission RRMS (right panel). NfL and GFAP were measured from a subset of the patients from the same cohort as historical controls.

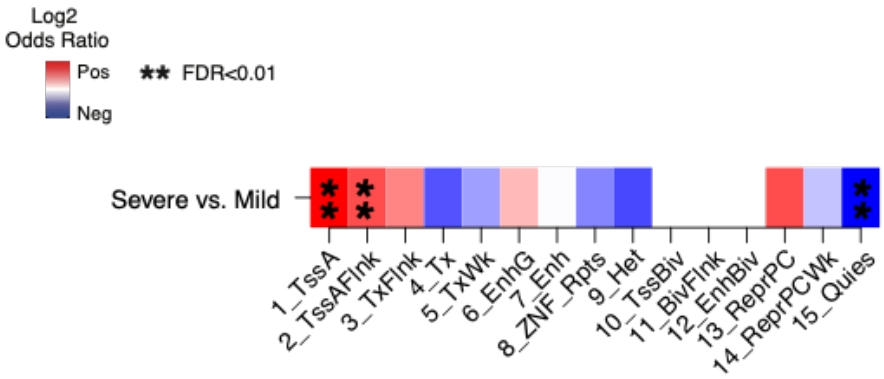
A



B



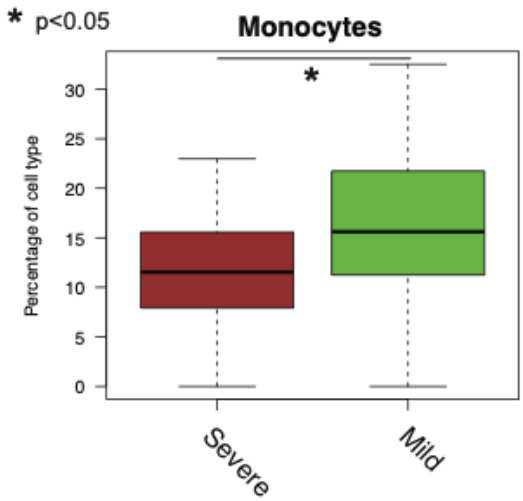
C



D

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
homophilic cell adhesion via plasma membrane adhesion molecules	GO:0007156	2.773×10^{-16}	
cell-cell adhesion via plasma-membrane adhesion molecules	GO:0098742	7.688×10^{-13}	
nervous system development	GO:0007399	9.470×10^{-10}	
anatomical structure development	GO:0048856	2.332×10^{-9}	
system development	GO:0048731	2.719×10^{-9}	
developmental process	GO:0032502	5.845×10^{-9}	
multicellular organism development	GO:0007275	3.305×10^{-8}	
cell-cell adhesion	GO:0098609	2.306×10^{-7}	
cell adhesion	GO:0007155	9.905×10^{-7}	
multicellular organismal process	GO:0032501	3.586×10^{-6}	

E



F

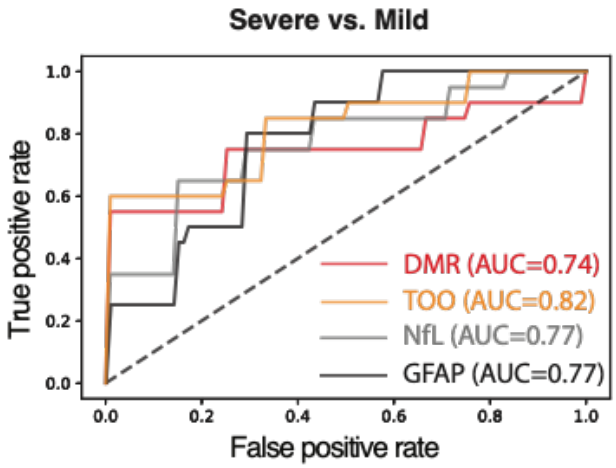


Figure 3. Circulating cfDNA methylation and its inferred tissue-of-origin patterns can distinguish patient-reported disability levels.

A. Scatter plots of -log10(FDR) vs. mean methylation differences at differentially methylated CpGs between two disability severity levels based on the requirement for ambulatory assistance around sample collection: high PDDS scores (severe, PDDS≥4) and low PDDS scores (moderate/mild/normal, PDDS<4). Colors indicate CpGs with FDR<0.01 and absolute mean methylation difference >0.1. PDDS score (*i.e.,* patient determined disease steps) indicates ambulatory disability status. **B**. UMAP representation of cfDNA methylation level in DMRs can distinguish MS patients based on the binary threshold of patient-reported disability. **C**. The enrichment of DMRs in different chromHMM states (that were previously computed from GM12878 cells in the Epigenome Roadmap project). **D**. The gene ontology enrichment analysis of DMRs between MS patients with severe/moderate PDDS scores and mild/normal PDDS scores by g:profile. **E**. Differences in cell death contribution from Monocytes (left panel) and Epithelial cells (right panel) between MS patients with severe and moderate/mild/normal patient-reported disability status. **F**. Receiver operating characteristic (ROC) curve by DMR only (red), Tissue-of-origin only (TOO, orange), NfL only (grey), and GFAP only (black) to distinguish MS patients based on patient-reported disability (PDDS>5 vs. <2). NfL and GFAP were measured from a subset of the patients from the same cohort as historical controls.

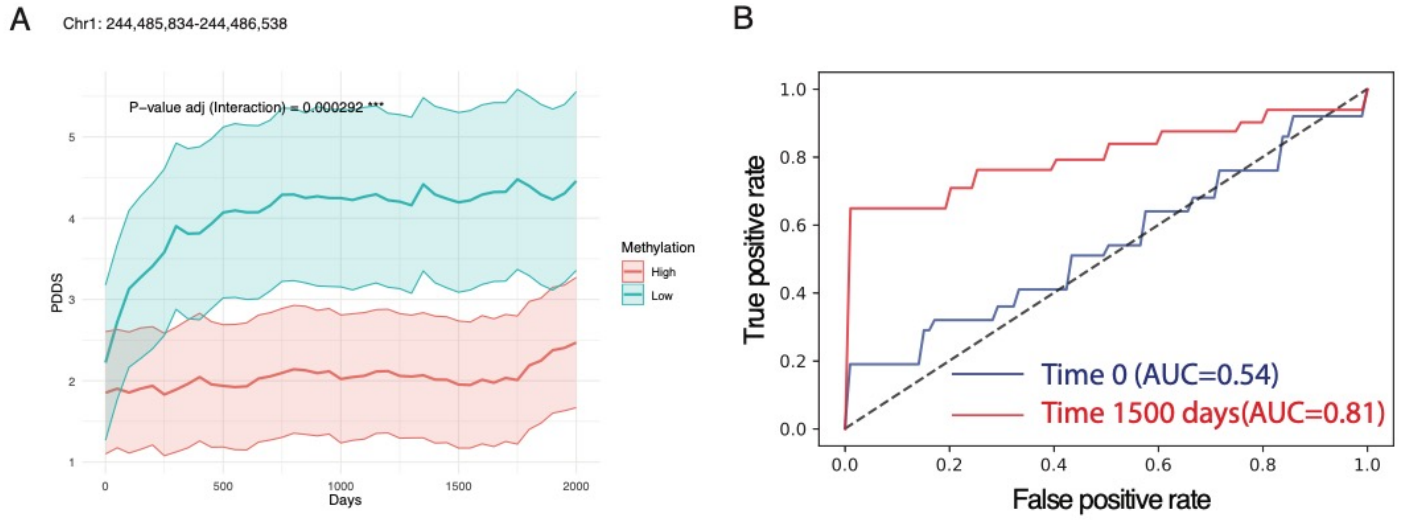


Figure 4. Baseline circulating DNA methylation profiles are informative of future disability severity trajectory in MS patients.

A. Representative plots for the patient-reported disability severity (*i.e.,* PDDS) over time stratified according to baseline cfDNA methylation level (high, ≥median, red; vs. low, <median, green) at one of the prognostic genomic regions. The subgroup with higher methylation levels showed increases in patient-reported disability severity over time. The p-value has been corrected for multiple testing. The shaded area represents the 95% confidence interval. **B**. ROC curve by using the baseline cfDNA methylation level at prognostic regions to predict the binary status (PDDS>5 vs. <2) of patient-reported disability severity at 0 days (blue line) or 1500 days (red line) after the blood sample collection. *PDDS*, patient determined disease steps.