nature medicine

Article

https://doi.org/10.1038/s41591-024-03286-y

Prediction of brain metastasis development with DNA methylation signatures

Received: 23 July 2023

Accepted: 5 September 2024

Published online: 8 October 2024

Check for updates

Jeffrey A. Zuccato[1,2], Yasin Mamatjan[1,3,14], Farshad Nassiri[1,2,14], Andrew Ajisebutu[1,14], Jeffrey C. Liu[1,14], Ammara Muazzam[4], Olivia Singh[1], Wen Zhang[4,5], Mathew Voisin[1,2], Shideh Mirhadi[4,5], Suganth Suppiah[1,2], Leanne Wybenga-Groot[4,6], Alireza Tajik[1,7], Craig Simpson[4,6], Olli Saarela[8], Ming S. Tsao[9], Thomas Kislinger[10,11], Kenneth D. Aldape[12], Michael F. Moran[4,5,10], Vikas Patil[1] ✉ & Gelareh Zadeh[1,2,13] ✉

Brain metastases (BMs) are the most common and among the deadliest brain tumors. Currently, there are no reliable predictors of BM development from primary cancer, which limits early intervention. Lung adenocarcinoma (LUAD) is the most common BM source and here we obtained 402 tumor and plasma samples from a large cohort of patients with LUAD with or without BM (n = 346). LUAD DNA methylation signatures were evaluated to build and validate an accurate model predicting BM development from LUAD, which was integrated with clinical factors to provide comprehensive patient-specific BM risk probabilities in a nomogram. Additionally, immune and cell interaction gene sets were differentially methylated at promoters in BM versus paired primary LUAD and had aligning dysregulation in the proteome. Immune cells were differentially abundant in BM versus LUAD. Finally, liquid biomarkers identified from methylated cell-free DNA sequenced in plasma were used to generate and validate accurate classifiers for early BM detection. Overall, LUAD methylomes can be leveraged to predict and noninvasively identify BM, moving toward improved patient outcomes with personalized treatment.

Brain metastases (BMs) are the most common intracranial cancer type and they lead to a very poor median overall survival (OS) of 10–16 months, despite advances in standard-of-care treatment[1,2]. Lung adenocarcinoma (LUAD) primary cancers are the most prevalent source of BMs and more than 30% of patients with LUAD develop BMs[2–4]. Presently, BMs are primarily identified once patients develop headaches or neurological symptoms; unfortunately, treatment at this stage is mainly palliative with surgery or radiotherapy[1,5]. Accordingly, a major limitation in current clinical care of patients with cancer is the inability to reliably identify patients at risk of developing BMs so that early detection and treatment can be initiated, due to the absence of predictive biomarkers. Early detection of BMs while they are smaller in size allows for improved tumor control, decreased morbidity and ultimately longer OS[5]. The Cancer Genome Atlas (TCGA) Research Network[6,7] performed an extensive multi-omic analysis of LUAD that has been informative; however, a limitation is the availability of follow-up clinical data on BM

[1]MacFeeters Hamilton Neuro-Oncology Program, Princess Margaret Cancer Centre, University Health Network and University of Toronto, Toronto, Ontario, Canada. [2]Division of Neurosurgery, Department of Surgery, University of Toronto, Toronto, Ontario, Canada. [3]The Faculty of Science, Thompson Rivers University, Kamloops, BC, Canada. [4]Program in Cell Biology, Hospital for Sick Children, Toronto, Ontario, Canada. [5]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [6]SPARC BioCentre, Hospital for Sick Children, Toronto, Ontario, Canada. [7]School of Medicine, St. George's University, Grenada, Grenada. [8]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada. [9]Department of Pathology, The Princess Margaret Cancer Centre, Toronto, Ontario, Canada. [10]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [11]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [12]Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA. [13]The Arthur and Sonia Labatt Brain Tumour Research Centre, Hospital for Sick Children, Toronto, Ontario, Canada. [14]These authors contributed equally: Yasin Mamatjan, Farshad Nassiri, Andrew Ajisebutu, Jeffrey C. Liu. ✉e-mail: vikas.patil@uhn.ca; gelareh.zadeh@uhn.ca

https://doi.org/10.1038/s41591-024-03286-y

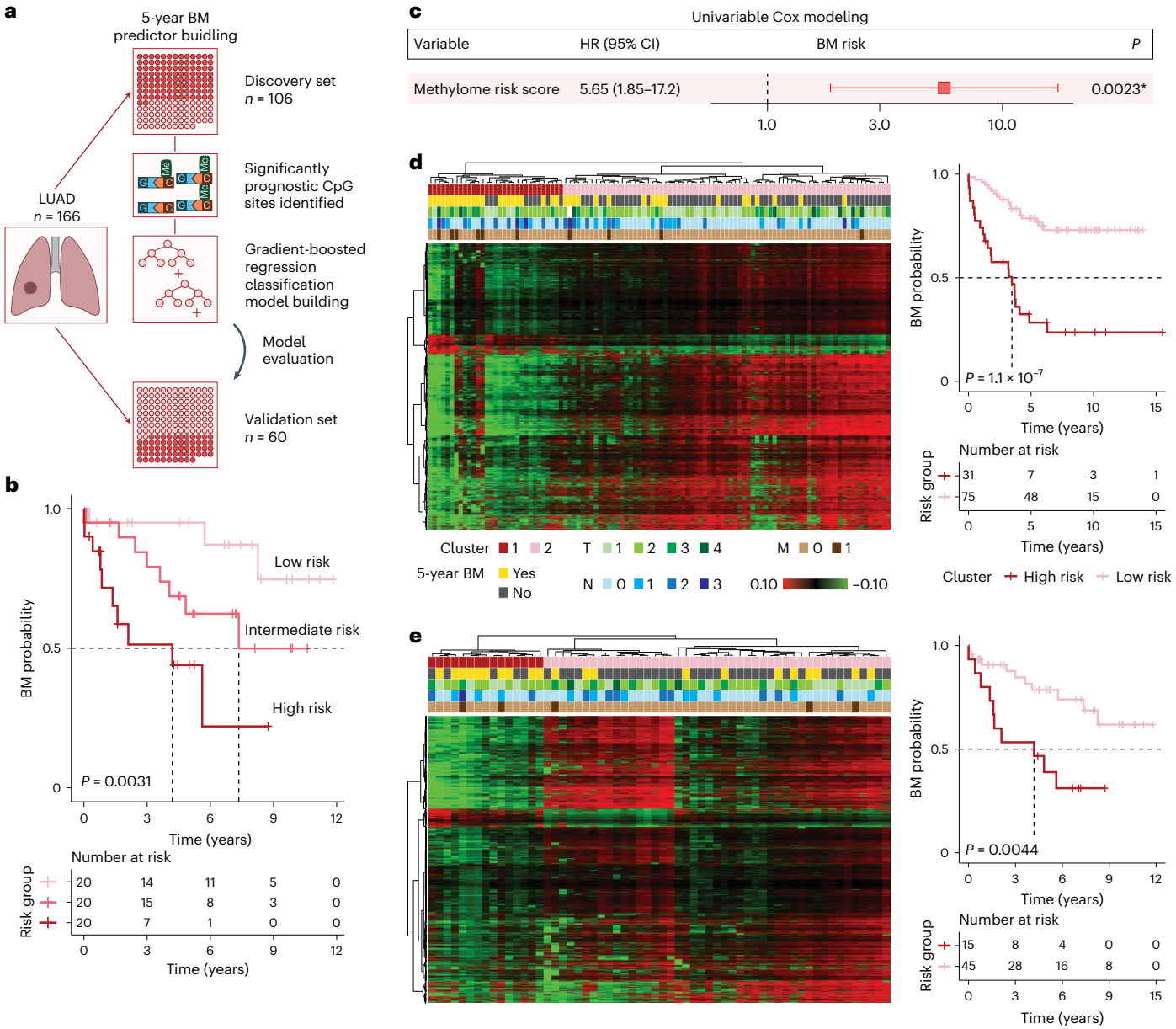


Fig. 1 | DNA methylation-based prediction of BM development from LUAD tissue. a, Flow diagram depicting the machine-learning approach used to build and evaluate the 5-year LUAD-BM predictor. **b**,**c**, Evaluation of the predictor model in the independent validation dataset. **b**, Kaplan–Meier plot and log-rank test using the tertiles of the BM predictor outputs in the independent validation cohort, showing that high risk scores capture early BM events; few low-risk patients developed BM within 5 years. **c**, Results of the univariable Cox proportional hazards model using numeric methylome risk scores from the 5-year predictor in the validation dataset samples, showing significant utility in predicting BM development ($n = 60$). The box with the whiskers displays the HR and 95% CI. **d**,**e**, Identification and validation of the features selected for model building. **d**, Hierarchical clustering of the discovery cohort LUAD tissue samples, using the 5,553 BM-predictive CpG sites identified within this discovery cohort, revealed one cluster (cluster 1) with a higher proportion of BM and shorter time to BM development, as expected. **e**, Hierarchical clustering of the validation cohort LUAD samples using the same features from the discovery dataset also identified more frequent BM in cluster 1, along with a higher event risk shown in a Kaplan–Meier plot with log-rank test for the independent data. * denotes a significant *P* value.

development, leaving a gap in the ability to build predictive models of BM[8].

There is increasing recognition of the clinical utility of DNA methylation signatures to improve the accuracy of brain tumor diagnosis and classification[9–12]. Notably, our team and others showed that predictive models using methylation signatures can also determine survival outcomes for brain tumors, including meningioma and, most recently, chordoma[13–15]. In this study, we investigated leveraging methylation signatures in primary LUAD to generate predictive models of BM development. We generated a large LUAD methylation dataset of 402 tumor tissue and plasma samples from 346 patients, including paired samples of primary LUAD-BM with comprehensive clinical annotation. We built a methylation-based model that predicts BM development in patients with primary LUAD, with utility as a reliable biomarker to advance cancer care toward early BM detection and treatment, reduction of neurological morbidity and prolongation of survival.

Studying the methylation profiles of LUADs that develop or do not develop BMs as well as paired primary LUAD and BM tumors, we established the landscape of methylation and copy number (CN) differences associated with BM development. Proteomics data from LUAD and BM were used to identify gene sets associated with BM that are both differentially methylated and have aligning upregulation or

https://doi.org/10.1038/s41591-024-03286-y

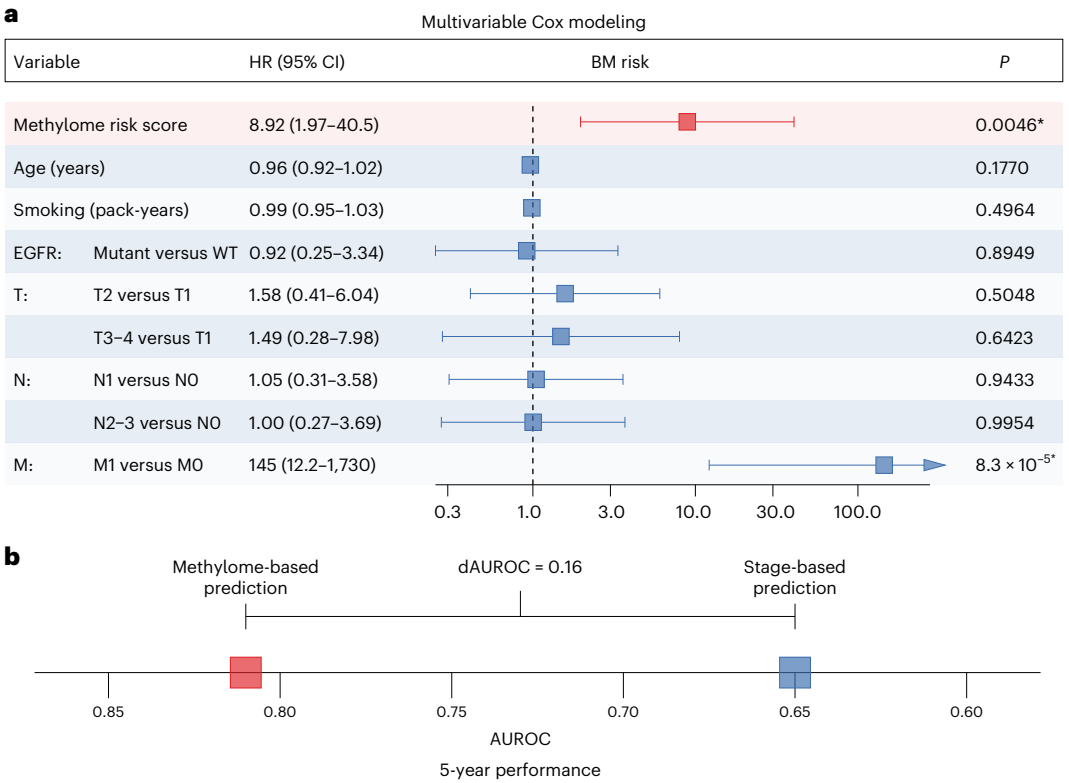


Fig. 2 | Utility of DNA methylation-based prediction of BM development over stage-based prognostication. a, Multivariable Cox proportional hazards modeling of BM development demonstrating that methylome risk scores have prognostic utility that is independent of clinical factors, including TNM component cancer staging scores ($n = 56$). The boxes with whiskers display the HR and 95% CI. **b**, AUROC performance metrics showing higher accuracy in differentiating patients who develop BM within 5 years using methylome-based predictor scores compared to stage-based prediction in the validation dataset ($n = 42$). * denotes a significant P value.

downregulation in the proteome, including immune-related and cell interaction-related pathways.

Additionally, we previously demonstrated that tumor-specific methylome-based biomarkers circulating in the plasma of patients can be used to diagnose and discriminate between gliomas and between systemic cancers[15–19]. In this study, we explore the utility and accuracy of noninvasive plasma biomarkers for early BM identification. We show that plasma cell-free (cf) circulating tumor (ct) DNA methylomes can noninvasively differentiate BM from other main differential diagnoses with high accuracy via peripheral blood sampling, extending beyond previous work to focus on tumors that are difficult to differentiate on neuroimaging[20].

Results

Clinical-methylomic BM prediction

DNA methylation-based BM prediction. We generated and analyzed a clinically annotated cohort of 166 primary LUAD tumors with detailed clinical data, including cancer stage and epidermal growth factor receptor (EGFR) mutational status (Extended Data Table 1) along with time to BM outcome data. We analyzed the methylation profiles of these 166 primary LUAD tumors. The flow diagram in Fig. 1a shows the modeling approach used to build a pipeline that is predictive of BM development in a discovery dataset ($n = 106$) and then evaluated in an independent validation dataset ($n = 60$). A total of 5,553 differentially methylated CpGs significantly predicted BM development in the discovery dataset univariable Cox models and were used for model building. These features were used to develop the methylome-based predictor of BM development within 5 years from primary LUAD diagnosis, using a gradient-boosted regression model within the discovery dataset (Fig. 1a).

This BM predictor model was applied to the independent validation cohort, which was not used for feature selection or model building, to assess its performance. Tertiles of predictor output scores (high, intermediate and low risk) were observed to stratify patients according to BM risk in a Kaplan–Meier plot ($P = 0.0031$; Fig. 1b), demonstrating their predictive utility within independent data. Additionally, output scores predicted BM development in both univariable (hazard ratio (HR) = 5.65, 95% confidence interval (CI) = 1.85–17.2, $P = 0.0023$; Fig. 1c) and multivariable Cox models that controlled for relevant clinical variables (HR = 8.92, 95% CI = 1.97–40.5, $P = 0.0046$; Fig. 2a) in the validation dataset, showing the utility of the numeric output values (not grouped) while also establishing that their utility is independent of clinical factors. The accuracy of the predictor in identifying BM development within 5 years using averaged bootstrapped methylation predictor scores (area under the receiver operating characteristic curve (AUROC) = 0.81) was higher than that of an equivalent stage-based predictor (AUROC = 0.65) in the independent validation dataset (dAUROC = 0.16; Fig. 2b), representing its utility over a representative clinical approach.

The features used for model building were used for hierarchical clustering of both discovery (Fig. 1d) and validation datasets (Fig. 1e) after all model building and evaluation were complete to demonstrate the distribution of the beta values in each cohort. One cluster in each set ('cluster 1') contained higher proportions of primary LUAD tumors that developed BM over 5 years of follow-up (71.0 and 66.7%, respectively) compared to cluster 2 (24.0 and 24.4%, respectively) and had different methylome signals in the heatmaps.

To define risk groups using methylome risk scores, an optimal methylome risk score cutoff of 0.456 was identified in an analysis to maximize the Youden index in the discovery cohort data and applied

to the validation dataset to derive high-risk and low-risk groups. The corresponding high-risk LUAD group had statistically significantly poorer BM outcomes than the low-risk LUAD group in a log-rank test (Kaplan–Meier plot, $P = 0.0034$; Extended Data Fig. 1a), univariable Cox model (HR = 3.53, 95% CI = 1.44–8.63, $P = 0.0058$; Extended Data Fig. 1b) and multivariable Cox model (HR = 3.56, 95% CI = 1.27–9.95, $P = 0.0156$; Extended Data Fig. 1c) using the validation dataset.

We evaluated the potential utility of methylome risk scores for predicting extracranial metastases. For patients without BM in the validation cohort, we assessed whether BM risk scores were predictive of the development of extracranial metastases. Methylome risk scores did not statistically significantly predict extracranial metastases in a univariable Cox model in the validation dataset (HR = 2.1, 95% CI = 0.2–20.5, $P = 0.517$), as expected given that they were trained to predict BM. Additionally, we did not find an association between methylome risk scores and patient age (Pearson $R^2 = 0.0024$, $P = 0.71$).

Representative clinical cases that illustrate the opportunity for potential escalation of treatment based on patient methylome-based BM risk scores are shown in Extended Data Fig. 1d. The first case is a stage I patient considered lower risk for BM development based on clinical factors; however, they had a high methylation-based score and in fact developed a BM after 1.8 years. The second case is a stage III patient considered as a high-risk individual based on clinical factors; however, they did not develop a BM over 8.1 years; the BM predictor score was indeed low.

Comprehensive clinical-methylomic BM-predictive nomogram. Next, we aimed to build a comprehensive nomogram model that incorporates methylation-based predictor scores with prognostic clinical factors independently in univariable Cox models (Extended Data Fig. 2a), complementary in a minimum redundancy maximum relevance ensemble analysis, and contributory to nomogram outputs together with predictor scores. Using this approach, TNM staging scores were identified for inclusion together with numeric methylome-based predictor values, which better discriminated risk compared to risk score groups. Figure 3a shows the final nomogram incorporating these variables to estimate composite 5-year BM development probabilities.

The performance of the nomogram was then evaluated in the independent validation dataset using the nomogram output scores. Nomogram scores stratified according to the median into higher-risk and lower-risk groups had significantly different BM event rates (Kaplan–Meier plot, $P < 0.0001$; Fig. 3b). In a univariable Cox model, numeric nomogram scores significantly predicted BM development (HR = 17.2, 95% CI = 4.13–71.3, $P < 0.0001$; Fig. 3c), with a higher HR than that of methylome scores alone (Fig. 1e) or individual clinical factors alone (Extended Data Fig. 2a), including T and N scores, which would currently be used clinically to predict outcomes in patients with non-metastatic cancer (M0) who have a more favorable prognosis and may benefit most from BM prevention or early management, compared to M1 patients who have already developed metastases. Additionally, nomogram evaluation with a mean-time-dependent 5-year AUROC in the validation dataset showed reliable detection of tumors that developed BM within 5 years (AUROC = 0.82, 95% CI = 0.77–0.86; Fig. 3d).

Predictor and nomogram utility across the spectrum of disease. We next assessed whether the methylome-based BM predictor was built to discriminate BM risk in different cancer stages representing the degree of disease burden. We also compared these scores stratified according to EGFR mutant and wild-type (WT) status as major molecular subtypes of LUAD. Predictor score tertiles stratified patients into high, intermediate and low BM risk subsets across all these patient groups, showing that the predictor was developed for the spectrum of LUAD disease (Fig. 4a). High BM risk determined by the optimal Youden cutoff was also well correlated with BM development within 5 years across the spectrum of LUAD disease (Extended Data Table 2). Nomogram risk scores were also built to discriminate BM risk in both EGFR mutant and WT tumors (Fig. 4b).

Methyloproteomic/CN alterations associated with BM

Methyloproteomic alterations in primary LUAD that develop BM. To characterize alterations in LUAD that develop BM, we analyzed the methylation signatures and CN alterations in primary LUAD that subsequently developed BM over 5 years versus those that did not develop BM. We identified differentially methylated CpGs in primary tumors that developed BM and characterized gene sets differentially methylated at promoters using a gene set enrichment analysis (GSEA). The key pathways identified in the GSEA related to hypomethylated immune and hypermethylated differentiation and development pathways in LUAD that developed BM (Extended Data Fig. 3a).

We then generated proteomics data from LUAD tumors and identified gene sets upregulated and downregulated in the proteome in LUAD that developed BM over 5 years compared to those that did not. Upregulated gene sets were mainly immune-related (Extended Data Fig. 4a); downregulated gene sets were mainly cell interaction and differentiation and development-related (Extended Data Fig. 4b). Aligning gene sets upregulated in the proteome and hypomethylated in LUAD with BM or downregulated in the proteome and hypermethylated in LUAD with BM are shown in Extended Data Fig. 3b, with immune pathway upregulation and hypomethylation in LUAD with BM being the main finding.

Additional LUAD proteomics data were acquired from the publicly available National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) dataset for assessment in an independent cohort with matched DNA methylation data. The methylome-based predictor from Figs. 1 and 2 was used to obtain methylome-based BM risk scores for this dataset; patients were split into high-risk and low-risk groups based on the optimal risk score cutoff (Extended Data Fig. 1a–c). We identified gene sets upregulated and downregulated in the proteome in LUAD with high versus low BM risk scores. Upregulated gene sets were related to mainly immune and cell cycle pathways (Extended Data Fig. 4c) while downregulated gene sets were related to mainly immune and cell interaction pathways (Extended Data Fig. 4d). The alignment of gene sets both upregulated in the proteome and hypomethylated in LUAD with BM or downregulated in the proteome and hypermethylated in CPTAC LUAD with high BM risk scores is shown in Extended Data Fig. 3c. Immune pathway upregulation and hypomethylation in LUAD with high BM risk scores is again the main finding that aligns closely with the result derived from the study proteomic data (Extended Data Fig. 3b).

CN alterations in primary LUAD that develop BM. Next, we evaluated all significant chromosome-level and gene-level CN alterations in primary tumors that develop BM ($q < 0.25$ in a, $P < 0.05$ in b; Extended Data Fig. 5a,b) within univariable Cox models to assess their prognostic utility for BM. Those that were predictive of BM development were 8q23.2, 21q21.3, *ACTG2*, *BLM*, *FGFR3*, *MCL1*, *CCND1*, *EGFR*, *MDM2* and *MYC* (Extended Data Fig. 5c); only the last four have been well described in BM from LUAD previously[21]. These CN alterations were then evaluated in separate multivariable Cox models using the validation dataset, together with the nomogram risk scores, to determine whether any of the CN alterations predicted BM development independently of our nomogram scores and found that none were independently prognostic (Extended Data Fig. 5c). Therefore, these CN alterations were considered genomic correlates of nomogram risk scores with amplifications of *EGFR*, 8q23.2, *MCL1*, *MYC*, *MDM2*, *CCND1* and *FGFR3* plus deletions of *ACTG2* correlating with higher BM risk scores while deletions of *BLM* and 21q21.2 were correlates of lower scores (Extended Data Fig. 5c).

Cell type abundance in primary LUAD that develop BM. We also deconvoluted methylation profiles into fractions of microenvironmental cells for both LUAD that develops BM and LUAD that did not

https://doi.org/10.1038/s41591-024-03286-y

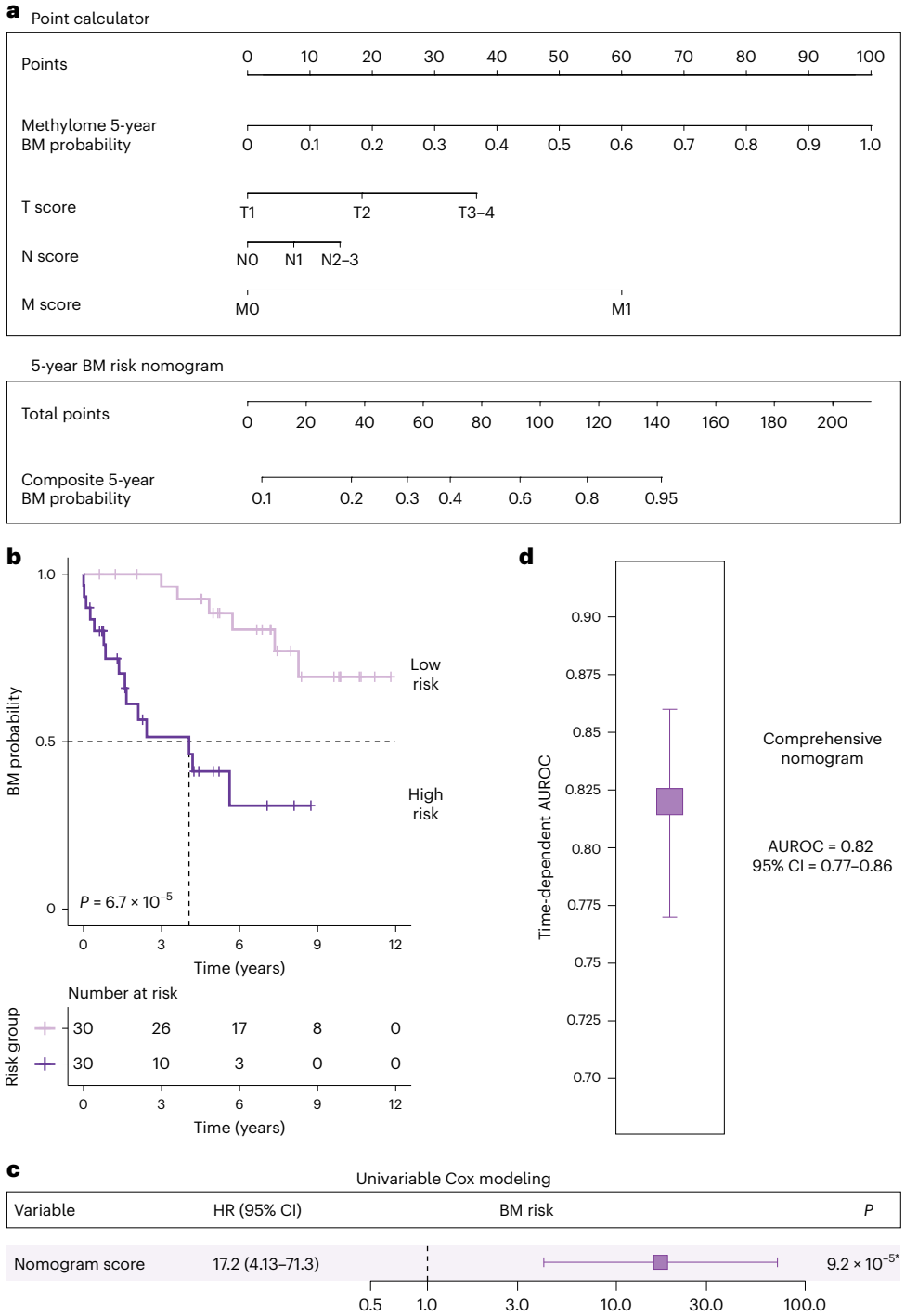


Fig. 3 | Comprehensive clinical-methylomic nomogram that predicted BM development from LUAD tissues. a, Point calculator diagram assigning point values to methylome and TNM scores, followed by the nomogram that uses total summed point values to derive composite probabilities of 5-year BM development. **b**, Kaplan–Meier plot and log-rank test of high-risk and low-risk nomogram scores, based on the median, demonstrating in the independent validation dataset that high-risk scores capture most of the early BM events plus BM that occurs within 5 years, while most low-risk scores occur in patients who do not develop BM early or within 5 years. **c**, Results of the univariable Cox proportional hazards model using nomogram scores demonstrating the significantly increased BM risk with increasing nomogram score ($n = 60$). The box with the whiskers displays the HR and 95% CI. **d**, Time-dependent mean AUROC with associated 95% CI in the validation cohort, using a bootstrap resampling approach to derive nomogram scores, showing accurate differentiation of patients who developed BM within 5 years ($n = 42$). The box with the whiskers displays the AUROC and 95% CI. * denotes a significant P value.

develop BM. Many immune cell types plus fibroblasts and endothelial cells were differentially abundant in the microenvironment of LUAD that developed BM (Extended Data Fig. 6), particularly cytotoxic T lymphocytes and monocytes. Cell fractions of cytotoxic T lymphocytes and monocytes were not significantly predictive of BM development in univariable Cox models ($P = 0.1510$ and $P = 0.1570$, respectively) or in multivariable Cox models adjusting for DNA methylation-based risk scores ($P = 0.2713$ and $P = 0.4711$, respectively). A multivariable Cox model incorporating methylome risk scores with the ratio of cytotoxic T lymphocytes to monocytes incorporating both cell types

https://doi.org/10.1038/s41591-024-03286-y

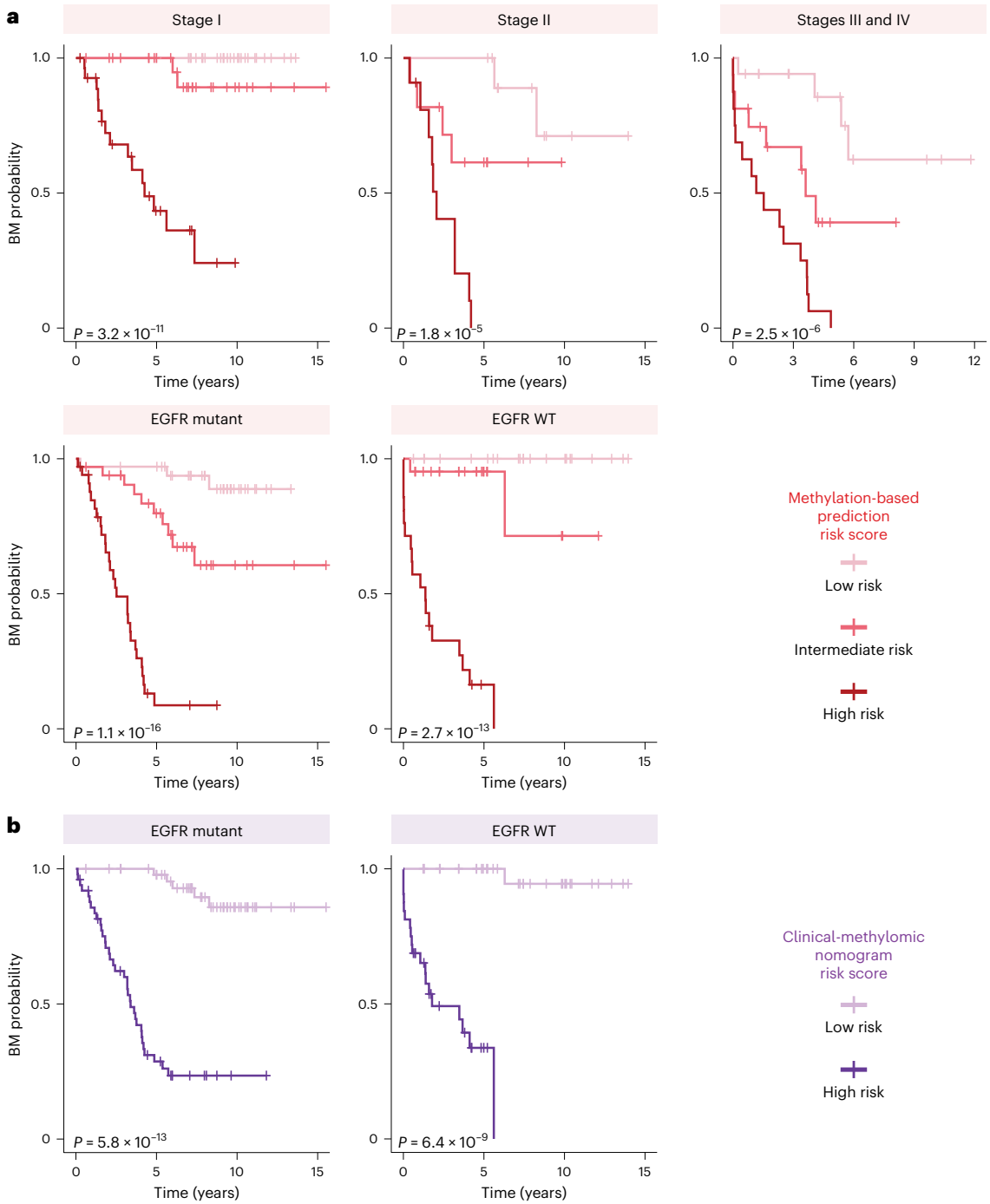


Fig. 4 | Confirmation that the methylome-based predictor and clinical-methylomic nomogram were developed to discriminate BM outcomes across the spectrum of LUAD disease. a, Kaplan–Meier plots with log-rank tests showing that BM predictor output tertiles distinguish patients with high, intermediate and low risk for BM development across cancer stages (top) and both EGFR mutation statuses (bottom). **b**, Kaplan–Meier plots with log-rank tests showing that BM nomogram scores stratified according to the median distinguish high-risk and low-risk patients across both EGFR mutation statuses.

demonstrated the prognostic utility of methylome scores (HR = 7.5, 95% CI = 2.4–23.5, $P = 0.0006$) independently of the ratio of immune cell fractions that was not prognostic ($P = 0.7479$). Of the 5,553 BM-predictive CpGs (Fig. 1a), 106 were included within the 1,184 CpG signatures used for methylCIBERSORT modeling, representing the overlap between these two signature sets.

Methyloproteomic alterations in matched LUAD-BM pairs. We next built a unique cohort of BM tumors paired to the matched primary LUAD tumor and evaluated differential methylation, CN and cell deconvolution between BM and primary lung tumors. Differentially methylated CpGs between paired tumors were mapped to gene promoters and evaluated in a GSEA that showed immune, differentiation and development, and cell interaction-related pathways differentially methylated and mainly hypermethylated at promoters in BM tumors (Extended Data Fig. 7a).

Publicly available paired BM and primary LUAD proteomic data were acquired[22]. Gene sets upregulated and downregulated in BM versus paired LUAD were identified; aligning gene sets upregulated in the proteome and hypomethylated in BM or downregulated in

the proteome and hypermethylated in BM are shown in Extended Data Fig. 7b. Alignment was mainly in immune-related and cell interaction-related pathways downregulated in the proteome and hypermethylated in BM compared to paired primary LUAD.

We then generated proteomics data from BM tumors and performed nonnegative matrix factorization for clustering to further assess and characterize proteomic alterations within the BM dataset. Five main metaprograms were identified, which were related to immune function, cell interaction, cell cycle and central nervous system (CNS) development (Extended Data Fig. 7c).

CN and cell type abundance in matched LUAD-BM pairs. Additionally, we examined chromosome-level and gene-level CN alterations in BM relative to paired LUAD tumors to further characterize alterations in LUAD-BM pairs. BM samples showed deletions across 7q, 8p, 9, 10, 12–16 and 22q ($q < 0.25$) and greater deletion to *GLI2* and *FGFR1*, greater amplification in *NF1* and *KDM2A*, less deletion to *ALK* and less amplification in *STK11* relative to paired LUAD tumors ($P < 0.05$) (Extended Data Fig. 8a). We also deconvoluted methylation profiles into fractions of cancer cells and microenvironmental cell types, and showed that many immune cell types plus fibroblasts made up significantly different fractions of the BM tumor microenvironment than that of paired primary LUAD tumors (Extended Data Fig. 8b). The overall lower abundance of immune and extracellular matrix-related cell types in the BMs aligns with the hypermethylation of promoters and downregulation in the proteome of related pathways (Extended Data Fig. 7b).

Plasma methylome biomarkers for BM liquid biopsy

We next aimed to determine whether plasma methylation signatures can be used to noninvasively distinguish BM from other brain lesions, which may be used to diagnose BM once they develop in patients predicted to be at high risk using our nomogram. For this analysis, we developed and evaluated 50 iterations of plasma methylome-based glmnet classifiers of BM versus main tumors within a typical differential diagnosis (gliomas and CNS lymphoma), plus non-brain-tumor controls, as shown in the flow diagram in Fig. 5a and as we have used previously[15–17]. For each model iteration, plasma samples were split into a random 80% discovery dataset used for feature selection for model building. Features used were the combination of the top 300 differentially methylated regions (DMRs) in each pairwise class comparison within the discovery dataset; these features were observed to clearly cluster BM from other classes in a multidimensional scaling (MDS) plot of all samples, demonstrating their discriminative value (Fig. 5b).

The 20% independent validation datasets were used to assess model performance where the BM-versus-others classifiers accurately distinguished BM samples from those of main clinical differential diagnoses and non-CNS cancer controls (median AUROC = 0.80, 95% CI = 0.68–0.93; Fig. 5c). Therefore, the plasma methylome-based models were able to noninvasively identify and confirm BM diagnoses in patients with LUAD who developed brain lesions with high accuracy. Additionally, BM plasma methylome signals were well correlated with BM tissue methylation values in patients with matched paired plasma and tissue samples (median Spearman rank correlation coefficient = 0.62, 95% CI = 0.55–0.69, $P < 2.2 \times 10^{-16}$; Fig. 5d). This correlation is similar to what we showed previously for other tumor types and provides further support for the recovery of ct methylomes using our plasma-based approach[15–17], which we now leveraged for ctDNA-based BM classification from clinical differential diagnoses and non-CNS tumor controls.

Representative cases of the BM-versus-others classifiers accurately identifying a BM in Fig. 5e and a non-BM (that is, 'other') in Fig. 5f are shown, with the box plots depicting probability values of a BM classification for each of the cf methylated DNA immunoprecipitation and high-throughput sequencing (cfMeDIP–seq)-based models where the case was within the testing dataset and not used in model building.

Additional independent assessment and validation

All three methylome-based models underwent further validation in additional independent datasets. The methylome-based BM predictor and clinical-methylomic nomogram were both evaluated in the external publicly available TCGA validation cohort for their utility in predicting distant metastases from primary LUAD, which is the most relevant available clinical variable in this cohort as BMs are the most common site of distant lung metastases[23,24]. When applied to the TCGA cohort for further validation, metastasis development was predicted by both the methylome-based predictor (multivariable Cox HR = 2.8, 95% CI = 1.1–7.4, $P = 0.0391$; Fig. 6a) and the composite nomogram (mean time-dependent AUROC = 0.78 95% CI = 0.76–0.79; Fig. 6b).

The methylome-based BM predictor was also evaluated in TCGA methylation data from lung squamous cell carcinoma (LUSC) with distant metastasis clinical data to determine whether it is generalizable to another lung cancer subtype. Methylome risk scores did not statistically significantly predict distant metastases (univariable Cox HR = 1.20, 95% CI = 0.48–2.95, $P = 0.698$; Extended Data Fig. 2b). Additionally, TCGA LUAD clustered distinctly from TCGA LUSC using the 5,553 BM-predictive CpGs sites (Fig. 1a), showing significant differences in the methylomes of these two tumor types (Extended Data Fig. 2c).

Furthermore, the plasma methylome-based BM-versus-others (glioma, CNSL, non-tumor control) classifiers were applied to an additional independent BM plasma cohort, where they reliably classified BM samples as BM (median accuracy = 74.2%, 95% CI = 63.0–85.3%; Fig. 6c) for further validation. The set of DMRs identified between all four entities in plasma cfMeDIP–seq data were used for MDS plotting of an independent external publicly available tissue methylation dataset for additional validation and demonstration of tumor specificity of plasma-based features. The resulting MDS plot in Fig. 6d shows distinct clustering of BM tissue samples from gliomas, CNSLs and non-brain tumor controls using independently identified, plasma-based features.

Discussion

A continued limitation in the management of patients with cancer is the inability to reliably predict who is at risk of developing BM and thereby experiencing worse outcomes. In this study, we demonstrated the first use of primary lung cancer tissue methylation signatures to accurately predict and estimate the risk of BM development, using a validated predictive model that has significantly greater utility than current clinical approaches using cancer staging to predict BM. Building on the strength of this methylation-based predictor is our validated comprehensive clinical-methylomic nomogram that also takes into consideration and integrates TNM scores as statistically significant complementary clinical variables. The composite nomogram robustly determined patient-specific composite BM risk probabilities; we illustrate the nomogram as a visual calculator to demonstrate feasibility for future clinical use. We believe that accurate prediction of BM development will move us toward a transformation in care, with individualized, biomarker-driven management decision. Specifically, high-risk patients may receive more aggressive initial primary cancer management to reduce BM risk, potentially offering radiotherapy and chemotherapy in early-stage patients with high risk scores, as well as frequent early surveillance brain imaging to detect BMs that develop at an early stage rather than once they enlarge and cause neurological symptoms, where treatment has less morbidity and is more likely to achieve tumor control[25]. More patient-specific risk profile information may also be used for patient counseling regarding the prognosis and outcomes of their disease.

The methylation landscape of metastatic LUAD to the brain has not been comprehensively explored previously. Pathways related to immunity, differentiation and development, and cell interaction were differentially methylated at promoters in BM compared to paired primary LUAD tumors. Immune-related and differentiation and development-related gene sets were also differentially methylated

https://doi.org/10.1038/s41591-024-03286-y

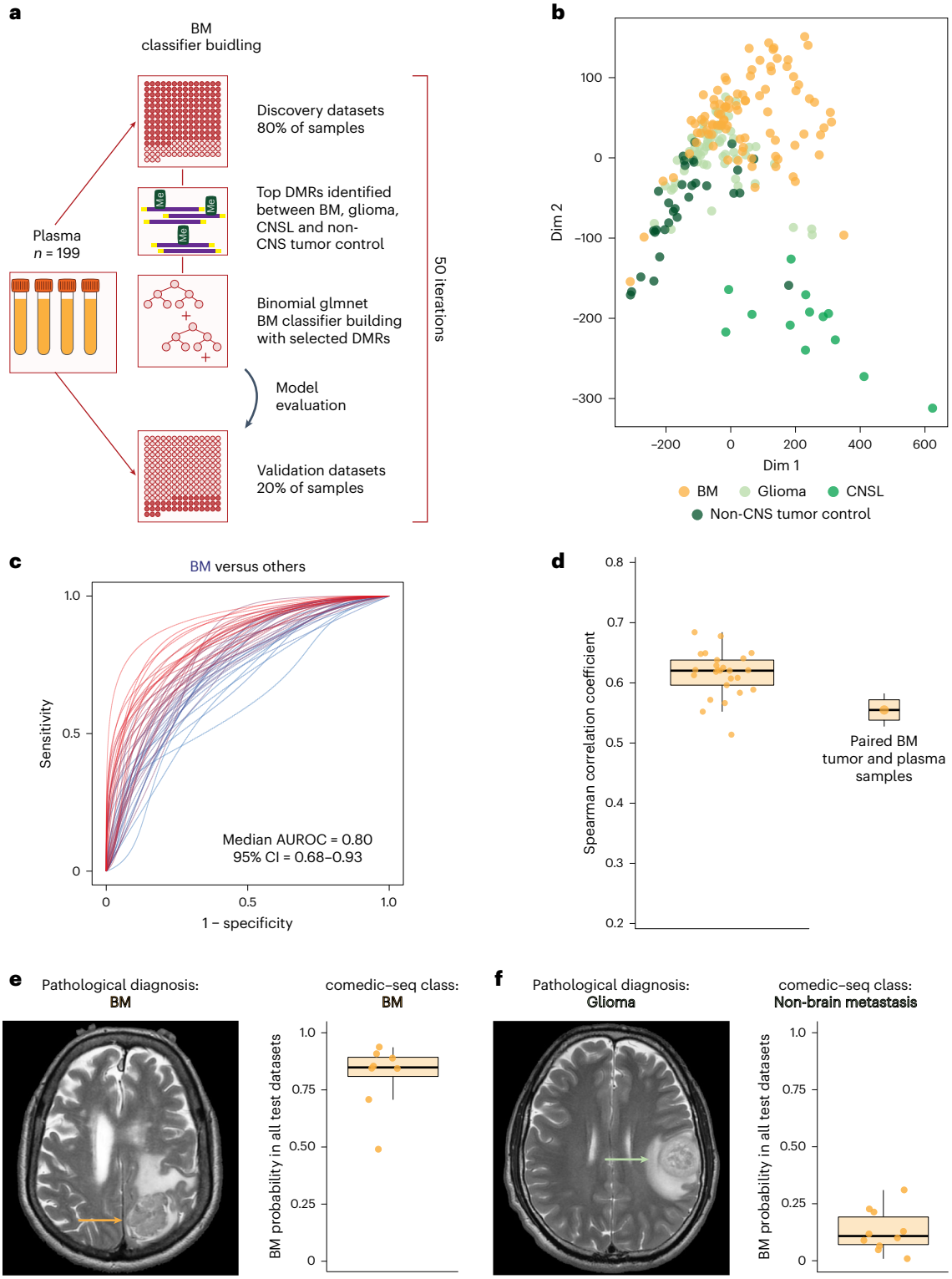


Fig. 5 | Plasma methylome-based detection of BM development. a, Flow diagram of the approach used to build and evaluate classifiers of BM versus other main brain tumors in a clinical differential diagnosis and non-CNS tumor controls. **b**, MDS plot using DMRs selected between each class (BM, glioma, CNSL and control) during model building, with BM plasma samples clustering separately from other entities. **c**, Ensemble of ROC curves evaluating 50 iterations of BM-versus-others classifiers in the validation dataset plasma, demonstrating reliable discrimination of BM from others, including glioma, CNSL and non-CNS tumor controls. **d**, Box plot of Spearman correlation coefficient values, showing alignment between methylation signatures in paired BM tumor and plasma samples ($n = 25$). **e**,**f**, Representative individual clinical cases of a BM (**e**) and a glioma (**f**) that were accurately identified as a BM and non-BM (that is, 'other'), respectively, based on testing dataset class probabilities from the BM-versus-others cfMeDIP–seq models. In the box plots, the central bars correspond to the median; the upper and lower distribution quartiles are displayed using boxes; and the 1.5× interquartile range (IQR) is represented by the whiskers.

https://doi.org/10.1038/s41591-024-03286-y

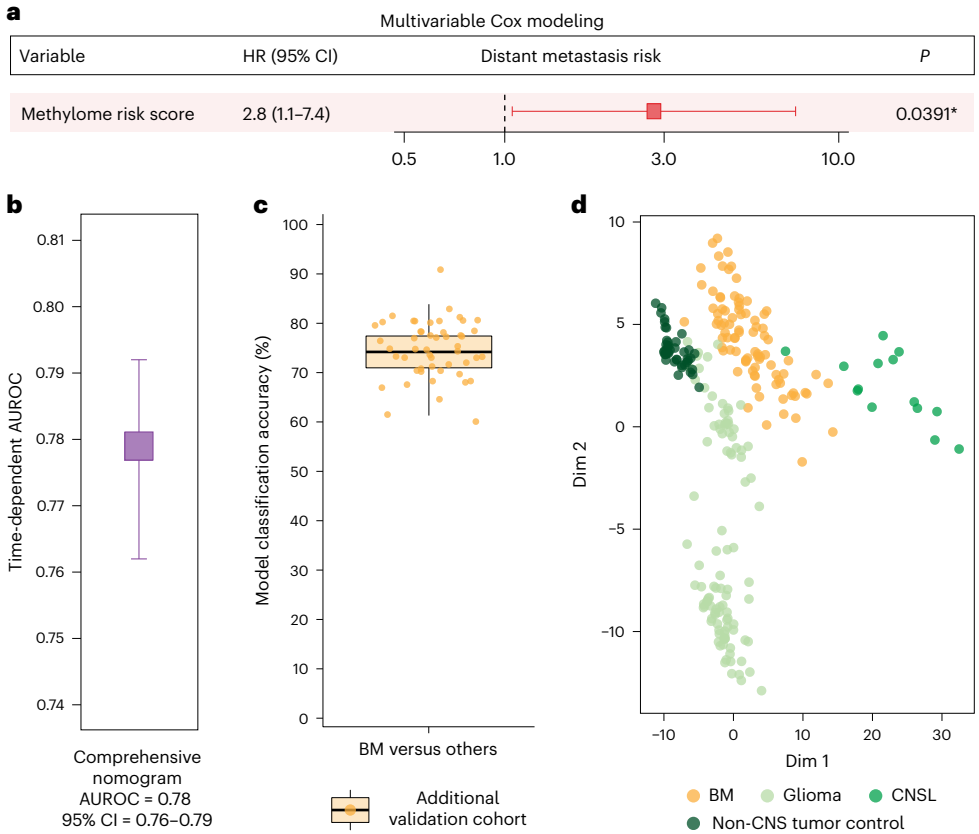


Fig. 6 | Additional independent validation of methylome-based models. **a**, Application of the 5-year BM predictor to the external TCGA LUAD tissue dataset yielded methylome risk scores that predicted distant metastasis development after controlling for clinical variables with Cox proportional hazards modeling ($n = 244$). The box with the whiskers displays the HR and 95% CI. **b**, The clinical-methylomic BM nomogram was applied to the TCGA LUAD tissue samples for distant metastasis prediction; the time-dependent AUROC analysis using a bootstrap resampling approach showed nomogram utility in this external dataset ($n = 244$). The box with whiskers displays the AUROC and 95% CI. **c**, Application of the ensemble of plasma methylome-based BM-versus-others classifiers to the additional independent validation cohort of BM samples, showing their accurate identification as BM and not others ($n = 31$). In the box plots, the central bars correspond to the median; the upper and lower distribution quartiles are displayed using boxes; and the 1.5× IQR is represented by the whiskers. **d**, MDS plot of the external brain tumor and non-brain tumor control tissue cohorts using independent plasma-based features (DMRs) identified in pairwise comparisons between BM, glioma, CNSL and control samples, showing distinct clustering of BM samples from others. * denotes a significant *P* value.

early in the tumor evolution process (at the time of the initial LUAD diagnosis) in patients who developed BMs. These findings add a new epigenetic layer to our understanding of the metastatic process, a layer that to date has not been well described. Immune and cell interaction differentially methylated pathways associated with BM also showed aligning dysregulation in the proteome. Additionally, many immune cell types were differentially abundant in BM versus paired LUAD, and in LUAD that developed BM compared to those that did not, in cell deconvolution analyses. Overall, immune pathways were consistently dysregulated across multiple datasets and analyses in this study; thus, they may be preferential targets of future studies on potential preventive treatments. Accordingly, we await with great interest the outcomes of ongoing clinical trials of immunotherapy in LUAD, which are increasingly including more patients with BMs[26–28].

In this study, we also took advantage of plasma methylation signatures in patients with BMs to confirm the diagnosis of BM without the need for invasive diagnostic neurosurgical tissue biopsies[29]. This would allow for a BM diagnosis to be made based on ctDNA before treatment to ensure that other clinical differential diagnoses are ruled out before the initiation of treatment and that optimal treatment is promptly initiated. A noninvasive BM diagnosis would be particularly valuable for patients without active disease after initial LUAD treatment who develop brain lesions much later in the follow-up and for patients who have undiagnosed cancer at the time of BM. Importantly, we also provide data to support additional work to establish methylation-based plasma biomarkers for the early detection of BM diagnosis, as well as serving as a biomarker of response to treatment or recurrence.

This work is limited in sample size, with samples obtained from a single institution; the dataset was collected retrospectively. Application of the predictive nomogram prospectively in additional centers and in future prospective trials will allow for its incorporation into management decisions and to allow for an assessment of its impact on patient outcomes. These models have utility in LUAD only and they are expected to lead to the development of additional methylome-based BM predictors for other cancer types. This work explores dysregulated pathway networks in LUADs that develop BMs at the methylome and proteome levels, but it is limited in the evaluation of these pathways. It will be important for future mechanistic studies to assess their potential role in the metastatic process or as potential new treatment targets. CN assessment from methylome data is limited; thus, further work evaluating the CN alterations shown in this study in genomic data is required. The liquid biopsy models shown distinguish BM from other representative clinical differential diagnoses but do not differentiate between patients with LUAD with and without BM. Future work optimizing model features to those specific to BM may allow for utility in screening for BM.

Overall, we showed an invaluable nomogram that can predict BM development and has the potential to transform care for patients with

LUAD. Implementing the DNA methylation-based nomogram into clinical practice in future trials will allow for its power and clinical utility to be established. Moreover, the DNA methylation and proteomics datasets generated in this study serve as an important research resource for the field[30,31], providing an improved understanding of the biology of BM development that can enable future treatment advance.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03286-y.

References

1. Sperduto, P. W. et al. Survival in patients with brain metastases: summary report on the updated diagnosis-specific Graded Prognostic Assessment and definition of the eligibility quotient. *J. Clin. Oncol.* **38**, 3773–3784 (2020).
2. Ostrom, Q. T., Wright, C. H. & Barnholtz-Sloan, J. S. Brain metastases: epidemiology. *Handb. Clin. Neurol.* **149**, 27–42 (2018).
3. Travis, W. D., Brambilla, E., Burke, A. P., Marx, A. & Nicholson, A. G. Introduction to the 2015 World Health Organization Classification of Tumors of the Lung, Pleura, Thymus, and Heart. *J. Thorac. Oncol.* **10**, 1240–1242 (2015).
4. Wang, G. et al. Distribution of brain metastasis from lung cancer. *Cancer Manag. Res.* **11**, 9331–9338 (2019).
5. Stelzer, K. J. Epidemiology and prognosis of brain metastases. *Surg. Neurol. Int.* **4**, S192–S202 (2013).
6. Collisson, E. A. et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
7. Chen, F. et al. Multiplatform-based molecular subtypes of non-small-cell lung cancer. *Oncogene* **36**, 1384–1393 (2017).
8. Cai, Q. et al. Identification of a novel prognostic DNA methylation signature for lung adenocarcinoma based on consensus clustering method. *Cancer Med.* **9**, 7488–7502 (2020).
9. Capper, D. et al. DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
10. Capper, D. et al. Practical implementation of DNA methylation and copy-number-based CNS tumor diagnostics: the Heidelberg experience. *Acta Neuropathol.* **136**, 181–210 (2018).
11. Karimi, S. et al. The central nervous system tumor methylation classifier changes neuro-oncology practice for challenging brain tumor diagnoses and directly impacts patient care. *Clin. Epigenetics* **11**, 185 (2019).
12. Orozco, J. I. J. et al. Epigenetic profiling for the molecular classification of metastatic brain tumors. *Nat. Commun.* **9**, 4627 (2018).
13. Olar, A. et al. Global epigenetic profiling identifies methylation subgroups associated with recurrence-free survival in meningioma. *Acta Neuropathol.* **133**, 431–444 (2017).
14. Nassiri, F. et al. DNA methylation profiling to predict recurrence risk in meningioma: development and validation of a nomogram to optimize clinical management. *Neuro. Oncol.* **21**, 901–910 (2019).
15. Zuccato, J. A. et al. DNA methylation-based prognostic subtypes of chordoma tumors in tissue and plasma. *Neuro. Oncol.* **24**, 442–454 (2022).
16. Shen, S. Y. et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).
17. Nassiri, F. et al. Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nat. Med.* **26**, 1044–1047 (2020).
18. Nuzzo, P. V. et al. Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. *Nat. Med.* **26**, 1041–1043 (2020).
19. Shen, S. Y., Burgener, J. M., Bratman, S. V. & De Carvalho, D. D. Preparation of cfMeDIP-seq libraries for methylome profiling of plasma cell-free DNA. *Nat. Protoc.* **14**, 2749–2780 (2019).
20. Nussbaum, E. S., Djalilian, H. R., Cho, K. H. & Hall, W. A. Brain metastases. Histology, multiplicity, surgery, and survival. *Cancer* **78**, 1781–1788 (1996).
21. Brastianos, P. et al. Genomic characterization of brain metastases and paired primary tumors reveals branched evolution and potential therapeutic targets. *Eur. J. Cancer* **3**, S586 (2015).
22. Woldmar, N. et al. Proteomic analysis of brain metastatic lung adenocarcinoma reveals intertumoral heterogeneity and specific alterations associated with the timing of brain metastases. *ESMO Open* **8**, 100741 (2023).
23. Riihimäki, M. et al. Metastatic sites and survival in lung cancer. *Lung Cancer* **86**, 78–84 (2014).
24. Ceresoli, G. L. et al. Brain metastases in locally advanced nonsmall cell lung carcinoma after multimodality treatment: risk factors analysis. *Cancer* **95**, 605–612 (2002).
25. Stankiewicz, M., Tomasik, B. & Blamek, S. A new prognostic score for predicting survival in patients treated with robotic stereotactic radiotherapy for brain metastases. *Sci. Rep.* **11**, 20347 (2021).
26. Hu, H. et al. Brain metastases status and immunotherapy efficacy in advanced lung cancer: a systematic review and meta-analysis. *Front. Immunol.* **12**, 669398 (2021).
27. Teixeira Loiola de Alencar, V., Guedes Camandaroba, M. P., Pirolli, R., Fogassa, C. A. Z. & Cordeiro de Lima, V. C. Immunotherapy as single treatment for patients with NSCLC with brain metastases: a systematic review and meta-analysis—the META-L-BRAIN Study. *J. Thorac. Oncol.* **16**, 1379–1391 (2021).
28. Forde, P. M. et al. Neoadjuvant nivolumab plus chemotherapy in resectable lung cancer. *N. Engl. J. Med.* **386**, 1973–1985 (2022).
29. National Comprehensive Cancer Network. *NCCN Guidelines: Central Nervous System Cancers* (v.2) https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1425 (2021).
30. Tan, A. C. et al. Clinical trial eligibility criteria and recently approved cancer therapies for patients with brain metastases. *Front. Oncol.* **11**, 780379 (2022).
31. Nieblas-Bedolla, E., Zuccato, J., Kluger, H., Zadeh, G. & Brastianos, P. K. Central nervous system metastases. *Hematol. Oncol. Clin. North Am.* **36**, 161–188 (2022).

Methods

Patient cohort

Institutional research ethics board (REB) (University Health Network) approval was obtained for all samples used in this study and written informed consent was obtained as required by the REB. A unique set of 402 tumor and plasma samples was compiled from 346 patients. A total of $n = 166$ primary LUAD tumors, $n = 68$ BM tumors (30 paired to primary LUAD tumors) and $n = 123$ BM plasma (25 paired to patients with paired BM and LUAD tumors) were included. We obtained plasma on 13 CNSL together with 32 non-brain tumor controls and used published plasma cfMeDIP–seq data on 62 gliomas[17] to build a reference cohort to compare to BM. All patients were managed at the University Health Network in Toronto and were selected based on availability of tissue samples and comprehensive clinical data. An Excel (v.16) spreadsheet was created with patient demographics; cancer staging with traditional tumor (T), lymph node (N) and metastasis (M) involvement and composite stage I–IV; tumor EGFR mutation status (mutant or WT) as clinically relevant LUAD molecular subgroups; and details of BM development and survival. LUAD samples were acquired procedurally between 2004 and 2017. Time to BM development was defined as the duration between LUAD and BM first identification on imaging.

Publicly available LUAD cell line ($n = 63$) methylation data were used for the cell deconvolution analyses[32]. External tumor tissue methylation datasets were assembled for model validation including TCGA LUAD[6], BM[12] and German Cancer Research Centre gliomas, CNSL and normal brain tissue[9]. The external National Cancer Institute's CPTAC LUAD cohort with matched DNA methylation and proteomics data available was used for the differential gene expression analyses[33] along with a cohort of 20 paired BM and LUAD samples with proteomics data[22]. External tumor tissue methylation data were also obtained from 330 TCGA LUSC samples to evaluate another lung cancer subtype[34].

Tissue methylation profiling

Tumor tissue from $n = 166$ LUAD and $n = 30$ paired BM samples was profiled on the Infinium MethylationEPIC BeadChip (850k) array (Illumina) according to the manufacturer's instructions after DNA extraction, bisulfite conversion, and restoration. For methylation profiling, all LUAD samples were from formalin-fixed and paraffin-embedded (FFPE) tissue, 29 of 30 BM samples were obtained from FFPE and 1 of 30 BM samples were from fresh-frozen tissue. Both FFPE and frozen methylation profiles obtained from the same six BM samples clustered together according to patient with hierarchical clustering using the top 1,000 most variable CpG sites, showing close alignment between methylomes obtained from FFPE and frozen tissue for each patient pair (Extended Data Fig. 2d). The external dataset tumor tissue methylation data outlined above were previously profiled on the Infinium HumanMethylation450 BeadChip (450k) array (Illumina). Methylation array data were processed using the minfi package in R v.4.0.3 (R Foundation for Statistical Computing) and with ssNoob normalization. Low-quality data (detection $P > 0.01$), single-nucleotide polymorphisms, and cross-reactive CpGs were removed[35]. Overlapping CpGs between the 850k and 450k data were selected for analyses that used samples from both arrays[36].

Methylome-based BM predictor development

Primary LUAD tumors were randomly divided into discovery (106 of 166) and validation (60 of 166) datasets, which were used for model development and evaluation only, respectively. A 'predictor' model was developed to predict the risk of 5-year BM development from LUAD using the gbm package as follows with utilization of Partek Genomics Suite software in feature selection. CpG sites predicting time to BM development in the discovery dataset (univariable Cox false discovery rate (FDR) $P_{adj} < 0.025$) were used to build a generalized boosted regression classification model in the discovery dataset that provided methylome-based scores reflecting the risk of developing BM within 5 years. Predictor BM risk scores were evaluated in the validation dataset to assess performance independently in Kaplan–Meier plots with score tertiles (high, intermediate and low risk) and univariable and multivariable Cox models (numeric predictor scores and relevant clinical variables).

A bootstrap resampling approach was then performed using the boot package to obtain mean methylation risk scores across 50 bootstraps of different discovery and validation dataset splits, which were evaluated using 5-year AUROC metrics in each dataset. The same approach described was used to develop a stage-based BM predictor representing standard-of-care practice; the AUROC differential (dAUROC) comparing the methylation-based and stage-based predictors was calculated.

An optimal predictor risk score cutoff was identified by maximizing the Youden index in the discovery cohort data only using the cutpointr package. High-risk and low-risk score groups stratified according to the optimal cutoff in the validation dataset were assessed using a Kaplan–Meier plot, univariable Cox model, multivariable Cox model with relevant clinical factors and Yule correlation coefficients between risk group and the development of BM within 5 years.

After model generation and evaluation, heatmaps of the probes included in model generation were plotted in both discovery and validation datasets to demonstrate the distribution of beta values in each cohort. Methylation predictor score tertiles across cancer stages and EGFR mutation status were assessed in Kaplan–Meier plots as relevant clinical and molecular subgroups of LUAD.

Comprehensive nomogram development

A multistep, iterative nomogram variable selection approach was used to identify clinical variables that significantly predicted BM in a multivariable Cox model, had an absolute value of importance greater than 0.1 in a minimum redundancy, maximum relevance ensemble analysis with methylome scores using the mRMRe package, and ultimately contributed more than 5% of total nomogram points to ensure that features were relevant, reliable, and complementary. The final nomogram was built on Cox modeling with the hdnom package using methylome scores plus each TNM staging category, with tenfold cross-validation used for tuning and a one standard error selection rule.

The nomogram diagram functions to scale variable values to a linear point calculator that provides a 5-year BM risk value. Nomogram score performance was evaluated in a Kaplan–Meier plot (split from the median) and univariable Cox analysis (numeric scores) within the validation dataset only. Time-dependent 5-year AUROCs were calculated in both sets separately using a 10,000 bootstrap resampling approach. Kaplan–Meier plots display nomogram utility for both EGFR mutation status to establish their value for both the EGFR mutant and WT molecular subgroups of LUAD.

Differential methylation, copy number and cell type abundance analyses

Differentially methylated CpGs (delta beta > 0.1, FDR-corrected $P < 0.05$) identified using the limma package, between (1) primary LUAD tumors with and without BM development across 5 years and (2) BM tissue and paired primary LUAD, were mapped to gene promoters, averaged per promoter, and evaluated in a GSEA using rank scores incorporating P and delta values with the Human_GOBP_AllPathways_no_GO_iea_November_01_2020_symbol.gmt set and 2,000 permutations. Enrichment maps were built with a Jaccard coefficient greater than 0.25 and $P < 0.001$ (refs. 37,38).

CN data were inferred from raw methylation data and normalized using the conumee package. Significant CN amplifications and deletions ($q < 0.25$) between groups 1 and 2 were identified using genomic identification of significant targets in cancer (GISTIC) 2.0 software with the default parameters of focal length cutoff = 0.5 and confidence level = 0.9 (ref. 39). Relative CN differences between groups within

specific genes of interest related to LUAD[6], LUAD metastasis[40] and CNS malignancy[9] were identified by comparing the $\log_2$ CN ratios. CN alterations between LUAD with and without BM through 5 years of follow-up were explored for potential inclusion in the comprehensive clinical nomogram, including those with $q < 0.25$ in one group (GISTIC) or $P < 0.05$ (at the gene level). Prognostic CN alterations in univariable Cox models, also with aligning HR directionality to the GISTIC or gene-level results, were assessed in separate multivariable Cox models together with nomogram scores using the validation dataset to assess whether they were independently prognostic.

The microenvironmental composition in primary LUAD versus paired BM tissue, and in LUADs that develop BMs within 5 years versus those that did not, was estimated using a methylCIBERSORT-based cell deconvolution of methylation data incorporating the LUAD cell line data (delta beta > 0.2, FDR < 0.01, features and comparison ≤ 100)[41]. The abundances of cancer cells, neutrophils, B lymphocytes, natural killer cells, cytotoxic T lymphocytes, monocytes and macrophages, regulatory T lymphocytes, effector T lymphocytes, endothelial cells, fibroblasts and eosinophils were compared between groups. The MethylCIBERSORT package has undergone pan-cancer validation, including in LUAD[42–44].

Tissue proteomic profiling

For 49 LUAD tumors, proteins were extracted from tissue samples, mixed with equal amounts of super-SILAC standard and otherwise prepared and analyzed as done previously[45,46]. Samples were reduced with dithiothreitol (4.5 mmol l^{-1} at 60 °C for 20 min), alkylated with iodoacetamide (10 mmol l^{-1} in the dark for 30 min), digested overnight with trypsin and desalted and purified with Pierce C18 Spin Tips (cat. no. 84850, Thermo Fisher Scientific). Purified peptides were separated using an ultrahigh performance liquid chromatography (LC) device (Easy-nLC 1000, Thermo Fisher Scientific) coupled to a mass spectrometer (Orbitrap Elite, Thermo Fisher Scientific) through a nanoelectrospray ion source (EASY-Spray, Thermo Fisher Scientific). Peptides were resuspended in 0.1% formic acid, loaded into a 50-cm column (75 µmol l^{-1} inner diameter, packed with PepMap RSLC C18 resin) and heated to 60 °C. The flow rate was 250 nl min^{-1} from 5% to 30% acetonitrile in 0.1% formic acid over 224 min, then 30–90% acetonitrile in 0.1% formic acid over 2 min and finally 90% acetonitrile in 0.1% formic acid over 12 min for column cleanup. One full mass spectrometry (MS) scan (400–1,200 *m/z*) in the mass analyzer plus ten data-dependent MS2 scans were used to obtain the mass spectra. The MS scan approach included an automatic gain control target of 1,000,000, maximum ion injection time of 200 ms and a resolution of 240,000 (full-width, half-maximum) at 400 *m/z*. The MS2 scan approach included up to ten of the most intensive ions selected and fragmented, normalized collision energies of 30 in the ion trap, an automatic gain control target of 10,000 and a maximum ion injection time of 50 ms. A 50-s dynamic exclusion duration and a 500 maximum exclusion list were used.

For 50 BM samples from LUAD, samples were processed as described previously[47] with the following modifications: lysis buffer for protein isolation for each sample was 100 µl of 0.5 M Tris, pH 8.0, 50 mM NaCl, 2% SDS, 1% NP-40, 1% Triton X-100, 5 mM EDTA, 10 mM Tris(2-carboxyethyl)phosphine hydrochloride, 40 mM chloroacetamide, 1 mM β-glycerophosphate, 1 mM NaF and 0.5 mM Na_3VO_4. Samples were sonicated for 10–15 s and incubated at 95 °C (20 min with mixing at 900 rpm) before centrifugation (20,000*g*, 5 min, 18 °C). A total of 140 ml of supernatant was transferred to microcentrifuge tubes, pelleted with chloroform and methanol precipitation, and incubated overnight in 200 µl of 100 mM HEPES, pH 8.5, and 6 mg of trypsin/Lys-C (cat. no. V5072, Promega Corporation). Trifluoroacetic acid was added to 1% final and samples were centrifuged (10,000*g*, 5 min, at room temperature).

For BM spectral library generation, 2 µl of each sample was pooled, dried using a SpeedVac vacuum and suspended in 4.5 mM ammonium formate in 2% acetonitrile. Peptides were fractionated using high-pH reverse phase high-pressure LC, dissolved in 2% acetonitrile and 0.1% formic acid, and analyzed using LC tandem MS with an EASY-nLC 1000 system (Thermo Fisher Scientific) with a 1-h analysis and a timsTOF Pro mass spectrometer (Bruker). The LC analysis used an IonOpticks Aurora Ultimate column, a 250 nl min^{-1} flow rate, buffer A (0.1% formic acid in water) and buffer B (0.1% formic acid in 80% acetonitrile). The buffer B gradient increased from 3% at 0 min to 22% at 40 min, then to 30% at 55 min, then to 100% at 56 min; it was held at 100% until 60 min. The timsTOF Pro mass spectrometer was set to parallel accumulation–serial fragmentation (PASEF) mode; precursor ions with a 2–5 charge state were selected, ten PASEF tandem MS (MS/MS) scans were acquired for each MS/MS cycle, target intensity was 14,500, ramp time was 100 ms and dynamic exclusion time was 0.4 min. MS and MS/MS spectra were recorded from 100 to 1,700 *m/z*, with an ion mobility window between 0.70 and 1.5 Vs cm^{-2}. Advanced collision energy ranged from 20.0 to 77.32 eV.

For individual BM analysis, dried peptides in the remaining sample were dissolved in 2% acetonitrile plus 0.1% formic acid and quantified using absorbance at 205 nm with Scopes correction on a NanoDrop One (Thermo Fisher Scientific). Peptides were diluted to 325 ng 5 $µl^{-1}$ with 100 fmol of Pierce Peptide Retention Time Calibration Mixture (cat. no. 88321, Thermo Fisher Scientific) and analyzed using LC–MS/MS, separated as above but over 90 min with a buffer B gradient increased from 3% at 0 min to 20% at 65 min, then to 35% at 85 min and finally to 100% at 86 min; it was held at 100% until 90 min. The timsTOF mass spectrometer was used in data-independent acquisition PASEF mode. Defined isolation windows in the *m/z* versus ion mobility plane were optimized based on the spectral library data using pydiAID[48]. Spectra were recorded from 100 to 1,700 *m/z*; the ion mobility window was 0.70–1.5 Vs cm^{-2}, the advanced collision energy ranged from 20.0 to 77.32 eV and ramp and accumulation times were 100 ms.

Analysis of BM-related gene set enrichment in proteome

Raw LUAD sample MS files were analyzed using MaxQuant software with the UniProt human reference proteome from September 2023 (https://www.uniprot.org/proteomes/) with the parameters we outlined previously[45]. The BM sample spectral library was generated with FragPipe software using raw MS files as done previously but with mass tolerances of 50 ppm for precursor and 0.1 Da for fragment; up to three missed trypsin cleavages were allowed, and phosphorylation of serine, threonine and tyrosine were additional variable modifications[49]. The independent BM sample analysis of the data-independent acquisition was searched with DIA-NN software using the generated spectral library. Default settings were used except for the included charge states being +1 to +5; the fragment ion range was 100–1,700 *m/z*, precursor and fragment ion mass tolerances were 20 ppm, match between runs was enabled, all identifications were filtered for FDR ≤ 0.1%, the cleavage reagent was trypsin and up to three missed cleavages were allowed. Carbamidomethylation of cysteine was a fixed modification while oxidation of methionine, acetylation of protein N termini, and phosphorylation of serine, threonine and tyrosine were variable modifications[50].

In addition to proteomics data generated in this study, the external CPTAC[33] and paired LUAD-BM[22] proteomics datasets described in the 'Patient cohort' section of the Methods were acquired. Missing values were imputed from the Gaussian distribution centered around a minimal value of the protein using DEP. Differential protein expression analyses were conducted using the limma package between primary LUAD tumors with and without BM development over 5 years in the study LUAD data, between LUAD tumors with high and low methylome risk scores in the CPTAC data and between BM and paired LUAD in the data in ref. 22. Differentially expressed proteins were evaluated in GSEAs. Rank scores incorporating *P* and $\log_2$ fold change values were assessed with the Human_GOBP_AllPathways_no_GO_iea_November_01_2020_symbol.

gmt set and with 2,000 permutations. Normalized enrichment scores were generated and enrichment maps were built with a Jaccard coefficient greater than 0.25 and $P < 0.001$ (refs. 37,38). A nonnegative matrix factorization (NMF) analysis was carried out using the study BM data for proteins detected in more than 75% of BM samples. Expression values were centered by subtracting the average expression across all samples. The NMF package was applied to relative expression levels after transforming negative values to zeros, with factor k ranging from two to six. Proteins were ranked using NMF scores for each identified expression program. Seventeen expression programs were identified in the tumor samples, with an emphasis on robustness through a 70% overlap criterion. Hierarchical clustering of programs, based on shared proteins, was conducted using the top 50 genes in each program to identify meta-signatures.

Methylated cfDNA sequencing

All plasma for BM, CNSL and non-brain tumor controls was obtained at the time of surgery at the University Health Network from peripheral blood, which had undergone centrifugation at 2,500*g* and 20 °C for 15 min using tubes containing acid citrate dextrose. Samples were subsequently cryopreserved. cfDNA extracted from 0.5–3 ml of plasma (QIAamp Circulating Nucleic Acid Kit, cat. no. 55114, QIAGEN) underwent the cfMeDIP–seq protocol; data were processed as described elsewhere[15–17,19]. In summary, 1–10 ng per sample of cfDNA quantified using a Qubit fluorometer (Qubit 4, Thermo Fisher Scientific) underwent library preparation (KAPA HyperPrep Kit, cat. no. KK8504, Roche), methylated cfDNA immunoprecipitation with 5-mC monoclonal antibody 33D3 (MagMeDIP Kit, 0.16 µg used per reaction, cat. no. C02010021, Diagenode), purification (IPure Kit v2, cat. no. C03010015, Diagenode), library amplification, and cleanup. Optimal fragment size selection was confirmed on Bioanalyzer traces (2100 Bioanalyzer, Agilent). Samples were sequenced with the Illumina NovaSeq 6000 (100-bp paired end reads, median 62 million reads per sample). Sequencing reads were aligned to the human genome with the Bowtie2 software, deduplicated, indexed with SAMtools software and converted to counts per million values using the MEDIPS package at 300-bp genomic windows covering regulatory CpG islands, shores, shelves and FANTOM5 enhancers.

Noninvasive BM detection

A total of 92 of 123 (75%) BM plasma samples along with all glioma, CNSL and non-brain cancer control plasma were split into fifty 80% discovery and 20% validation datasets using a random class balanced split approach. The remaining 31 (25%) BM plasma samples were used as an additional independent validation dataset after being randomly split from the above cohort. Binomial glmnet models classifying BMs from others were built with the glmnet package in each discovery dataset iteration using the top 300 DMRs in each pairwise comparison between entities using moderated t-statistic with limma-trend as we have described previously[15–17]. All BM-versus-others models underwent three iterations of tenfold cross-validation for optimization. Model performance was assessed with AUROC metrics in the 20% validation dataset. All DMRs from model building were used to assess sample clustering in an MDS plot. For each of 25 patients with BMs with matched plasma cfMeDIP–seq data and tumor tissue EPIC array data, the correlation between plasma counts per million values for each window and mean CpG beta value within the window was evaluated with Spearman correlation coefficients. For representative cases, BM classification probabilities from all BM-versus-others models with the case in the validation dataset were evaluated.

Assessment of model features in additional datasets

The methylation predictor of BM development and clinical-methylomic nomogram of BM development were applied to the external TCGA LUAD set using distant metastases as the most relevant available outcome for further validation, most of which have been shown to be BM[23,24]. A multivariable Cox model assessed methylome risk score utility after controlling for age, sex, smoking status, EGFR status, and TNM scores. Time-dependent 5-year AUROC values using 10,000 bootstraps assessed nomogram utility.

BM-versus-others cfMeDIP–seq classifiers were applied to the additional $n = 31$ BM plasma validation dataset for further validation, and classifier accuracy was calculated. The top 300 DMRs in each pairwise comparison between all four plasma entities were used for MDS plotting of independent publicly available tissue data from the same entities using CpGs within these DMRs.

Statistical analysis and data presentation

Descriptive statistics are expressed as n (percentage) or mean (range). Groups were compared using Wilcoxon's rank sum (interval variables) and log-rank (Kaplan–Meier plots) tests. Pearson or Spearman (numeric data) and Yule's Q (nominal data) correlation coefficients were used as indicated. Two-tailed $P < 0.05$ was considered statistically significant unless stated otherwise, after correction for multiple comparisons with FDRs where required. Cohort sizes for model building and evaluation were determined based on our previous experience[14–18]. In the box plots the central bars represent the median, the boxes represent the upper and lower distribution quartiles, and the whiskers represent 1.5× the IQR. The GSEA gene sets are represented by nodes, whose size correlates to the number of genes; the internodal connections represent gene set overlap. The following colors have been used to aid with the interpretation of the results: blue denotes clinical factors; red shows the methylation predictor; purple is used for the nomogram; and yellow or orange are used for BM.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The glioma plasma cfMeDIP–seq data were previously published and can be accessed at *Zenodo* https://doi.org/10.5281/zenodo.3715312 (ref. 51). The LUAD cell line methylation data for cell deconvolution were accessed from the Gene Expression Omnibus (GEO) under accession no. GSE68379. The TCGA LUAD and LUSC tissue methylation datasets were acquired from the Genomic Data Commons Data Portal at https://portal.gdc.cancer.gov/analysis_page?app=Projects. External BM methylation data were accessed from the GEO under accession no. GSE108576. The German Cancer Research Centre methylation data are accessible from the GEO under accession no. GSE109381. The CPTAC LUAD tissue proteomics data were acquired from the Proteomic Data Commons Data Portal at https://pdc.cancer.gov/pdc/cptac-pancancer (Proteome_BCM_GENCODE_v34_harmonized_v1.zip). External LUAD and LUAD-BM tissue proteomics data were acquired from ref. 22 (Supplementary Table S1). The DNA methylation data for all tissue samples in this study have been deposited with the GEO under accession no. GSE220838. The cfMeDIP–seq data for the plasma samples can be accessed via *Zenodo* (https://doi.org/10.5281/zenodo.7338401)[52]. The proteomics data in the study are accessible from PRIDE in the ProteomeXchange Consortium under accession no. PXD048525.

Code availability

Publicly available open source R (v.4.0.3) packages used for the analysis include minfi (v.1.46.0), caret (v.6.0-94), gbm (v.2.1.8), survival (v.3.5-8), hdnom (v.6.0.0), boot (v.1.3-25), cutpointr (v.1.1.2), mRMRe (v.2.1.2), limma (v.3.56.2), conumee (v.1.38.0), MethylCIBERSORT (v.0.2.1), DEP (v.1.22.0), NMF (v.0.26), MEDIPS (v.1.22.0), glmnet (v.4.1-8) and Rtsne (v.0.16). Publicly available open source software used for the analysis includes Partek Genomics Suite (v.6.6), GSEA (v.4.0.3), cytoscape (v.3.6), GISTIC (v.2.0), MaxQuant (v.2.4.7.0), FragPipe (v.22.0), DI-ANN

(v.1.8.1), FASTQC (v.0.11.5), bowtie2 (v.2.5.1) and SAMtools (v.1.3.1). Custom code has been deposited at https://github.com/patilvikas/BrainMetsPrediction/tree/main.

References

32. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
33. Gillette, M. A. et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* **182**, 200–225 (2020).
34. Hammerman, P. S. et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
35. Morris, T. J. & Beck, S. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods* **72**, 3–8 (2015).
36. Kling, T., Wenger, A., Beck, S. & Carén, H. Validation of the MethylationEPIC BeadChip for fresh-frozen and formalin-fixed paraffin-embedded tumours. *Clin. Epigenetics* **9**, 33 (2017).
37. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
38. Mootha, V. K. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
39. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
40. Tang, W.-F. et al. Timing and origins of local and distant metastases in lung cancer. *J. Thorac. Oncol.* **16**, 1136–1148 (2021).
41. Baghban, R. et al. Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Commun. Signal.* **18**, 59 (2020).
42. Chakravarthy, A. et al. Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat. Commun.* **9**, 3220 (2018).
43. Grabovska, Y. et al. Pediatric pan-central nervous system tumor analysis of immune-cell infiltration identifies correlates of antitumor immunity. *Nat. Commun.* **11**, 4324 (2020).
44. Guidry, K. et al. DNA methylation profiling identifies subgroups of lung adenocarcinoma with distinct immune cell composition, DNA methylation age, and clinical outcome. *Clin. Cancer Res.* **28**, 3824–3835 (2022).
45. Mirhadi, S. et al. Mitochondrial aconitase ACO2 links iron homeostasis with tumorigenicity in non-small cell lung cancer. *Mol. Cancer Res.* **21**, 36–50 (2023).
46. Zhang, W. et al. Proteomic profiles of human lung adeno and squamous cell carcinoma using super-SILAC and label-free quantification approaches. *Proteomics* **14**, 795–803 (2014).
47. Krieger, J. R. et al. Evosep One enables robust deep proteome coverage using tandem mass tags while significantly reducing instrument time. *J. Proteome Res.* **18**, 2346–2353 (2019).
48. Skowronek, P. et al. Rapid and in-depth coverage of the (phospho-)proteome with deep libraries and optimal window design for dia-PASEF. *Mol. Cell. Proteomics* **21**, 100279 (2022).
49. Demichev, V. et al. dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nat. Commun.* **13**, 3944 (2022).
50. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).
51. Chakravarthy, A. Reproducibility archive for MeDIP analyses of plasma DNA from brain tumour patients. *Zenodo* https://zenodo.org/records/3715312 (2020).
52. Zuccato, J., Patil, V. & Zadeh, G. Predictive modelling of brain metastasis risk and non-invasive biomarker detection using DNA methylation signatures. *Zenodo* https://zenodo.org/records/7338401 (2022).

Acknowledgements

We would like to thank the University Health Network and the Princess Margaret Cancer Centre for their support of our work and research program. J.A.Z. is supported by the Canadian Institute of Health Research (CIHR) Canada Graduate Scholarship Doctoral Award, Strategic Training in Transdisciplinary Radiation Science for the 21st Century Program Scholarship, and the Hold'em for Life Oncology Fellowship. G.Z. is funded by a National Institutes of Health R01 grant (no. 1R01CA263196-01), a CIHR project grant (no. 175335), a Canadian Cancer Society (CCS)/CIHR/Brain Canada Spark grant (no. 707139), a CCS/CIHR/Brain Canada Breakthrough grant (no. 707701), the MacFeeters Hamilton Centre for Neuro-Oncology Research Fund and the Princess Margaret Foundation. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript

Author contributions

G.Z. supervised the study. J.A.Z., Y.M., K.D.A., V.P. and G.Z. designed the study. J.A.Z., A.A., A.M., O. Singh, W.Z., M.V., S.M., L.W.-G., A.T. and C.S. performed the laboratory processing of the biospecimens. J.A.Z. performed the statistical analyses. Y.M., J.C.L. and V.P. performed the bioinformatic analyses with input and planning from J.A.Z. and G.Z. J.A.Z., Y.M., F.N., K.D.A., V.P. and G.Z. interpreted the data. O. Saarela provided biostatistical support. J.A.Z. and G.Z. performed the initial manuscript preparation, figure organization, manuscript revision and wrote the response to the reviewers. All authors were involved in the final interpretation of the data and manuscript revisions.

Competing interests

The authors declare no competing interests.

Additional information

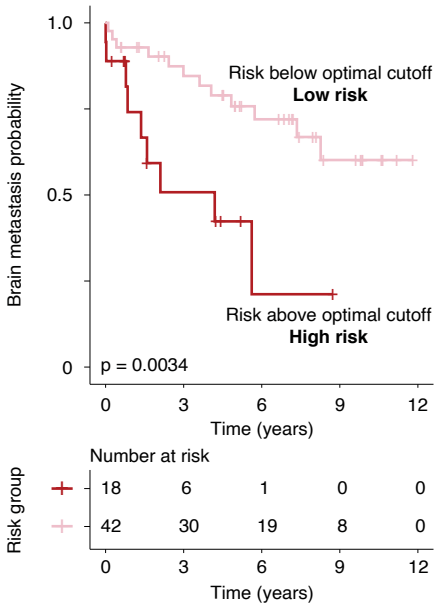
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03286-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03286-y.

Correspondence and requests for materials should be addressed to Vikas Patil or Gelareh Zadeh.

Peer review information *Nature Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

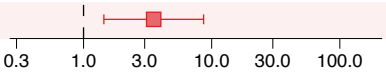
https://doi.org/10.1038/s41591-024-03286-y



b

Univariable Cox Modeling

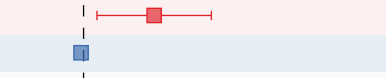
Variable		HR (95% CI)	Brain metastasis risk	p-value
Methylome risk	> vs. < cutoff	3.53 (1.44–8.63)		0.0058*



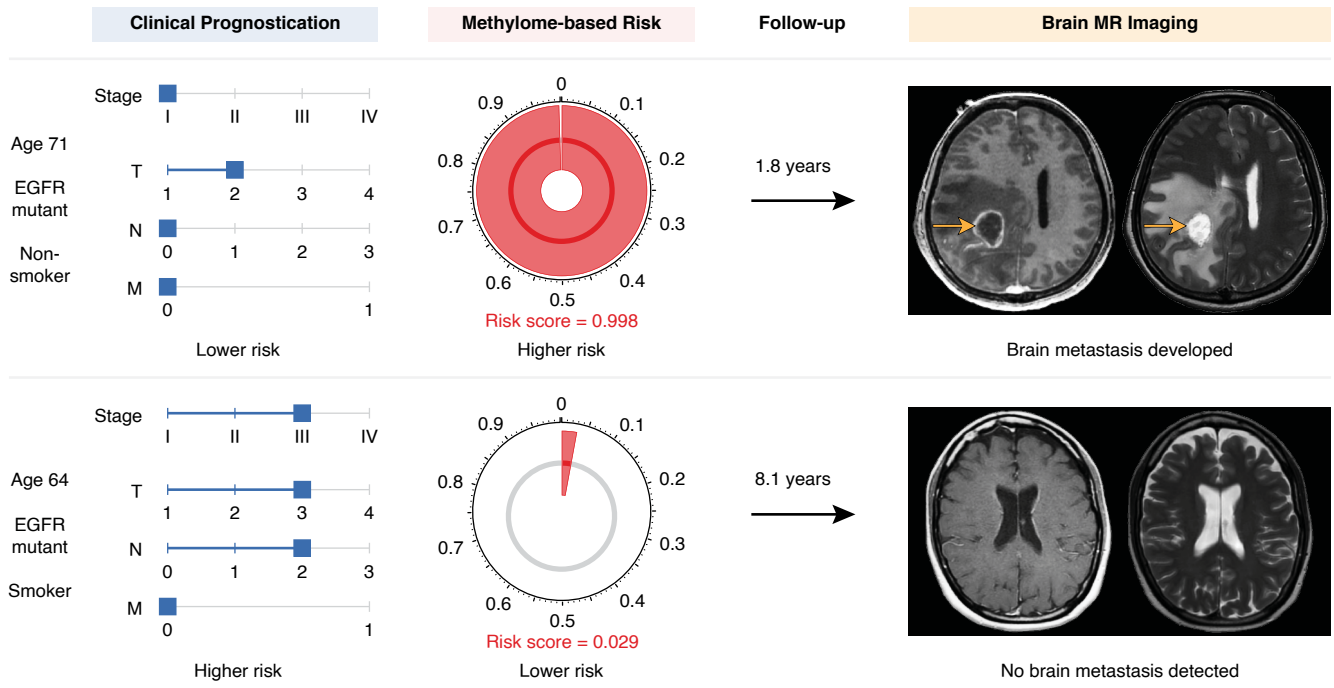
c

Multivariable Cox Modeling

Variable		HR (95% CI)	Brain metastasis risk	p-value
Methylome risk	> vs. < cutoff	3.56 (1.27–9.95)		0.0156*
Age	Years	0.96 (0.91–1.00)		0.0554
Smoking	Pack-years	0.99 (0.97–1.02)		0.7167
EGFR:	Mut vs. Wt	0.85 (0.31–2.34)		0.7468
T:	T2 vs. T1	1.85 (0.73–4.69)		0.1958
	T3–4 vs. T1	1.71 (0.57–5.12)		0.3389
N:	N1 vs. N0	1.28 (0.48–3.44)		0.6249
	N2–3 vs. N0	1.02 (0.33–3.14)		0.9677
M:	M1 vs. M0	104.2 (11.6–940)		3.5x10^{-5}*

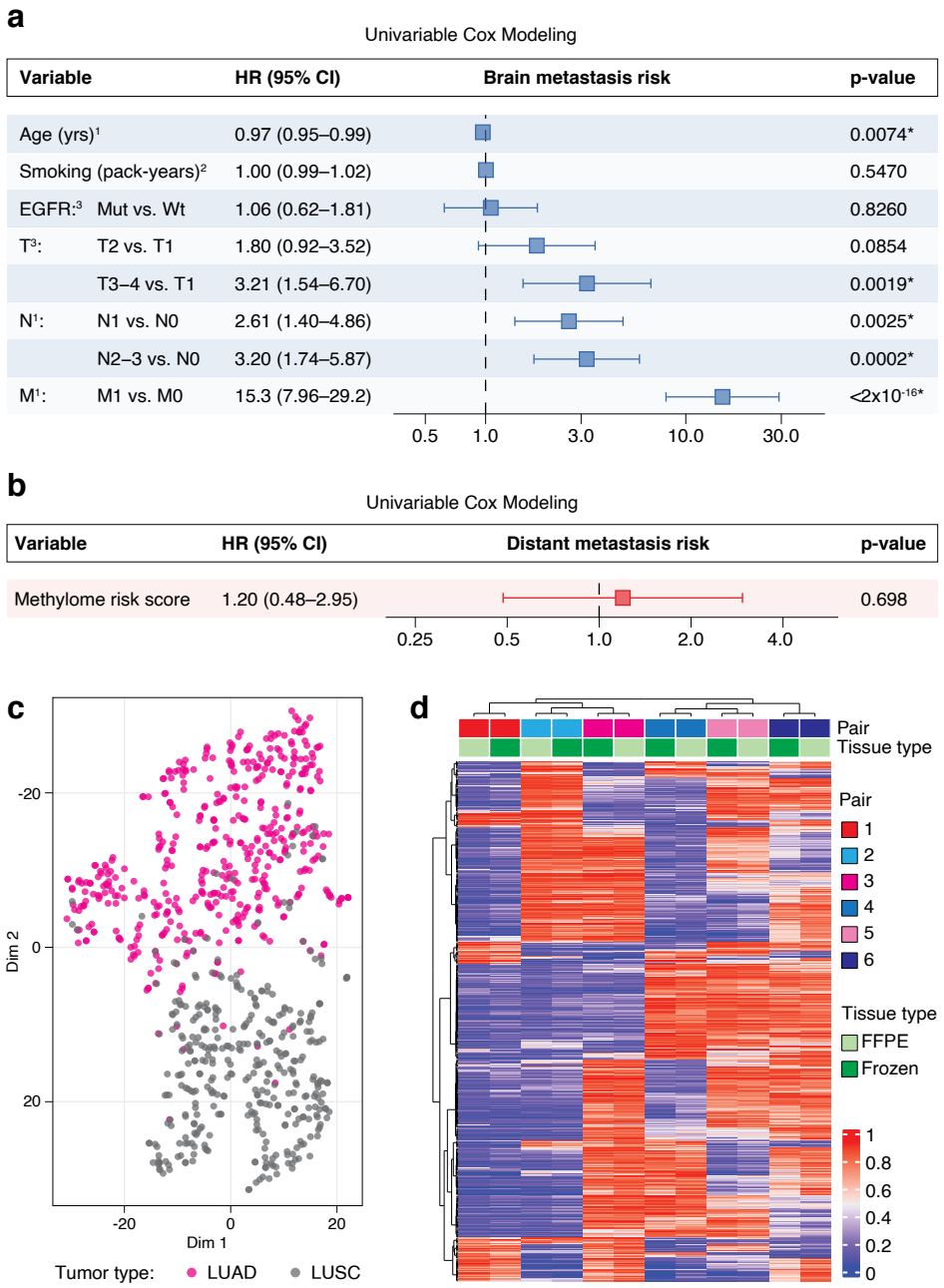


d



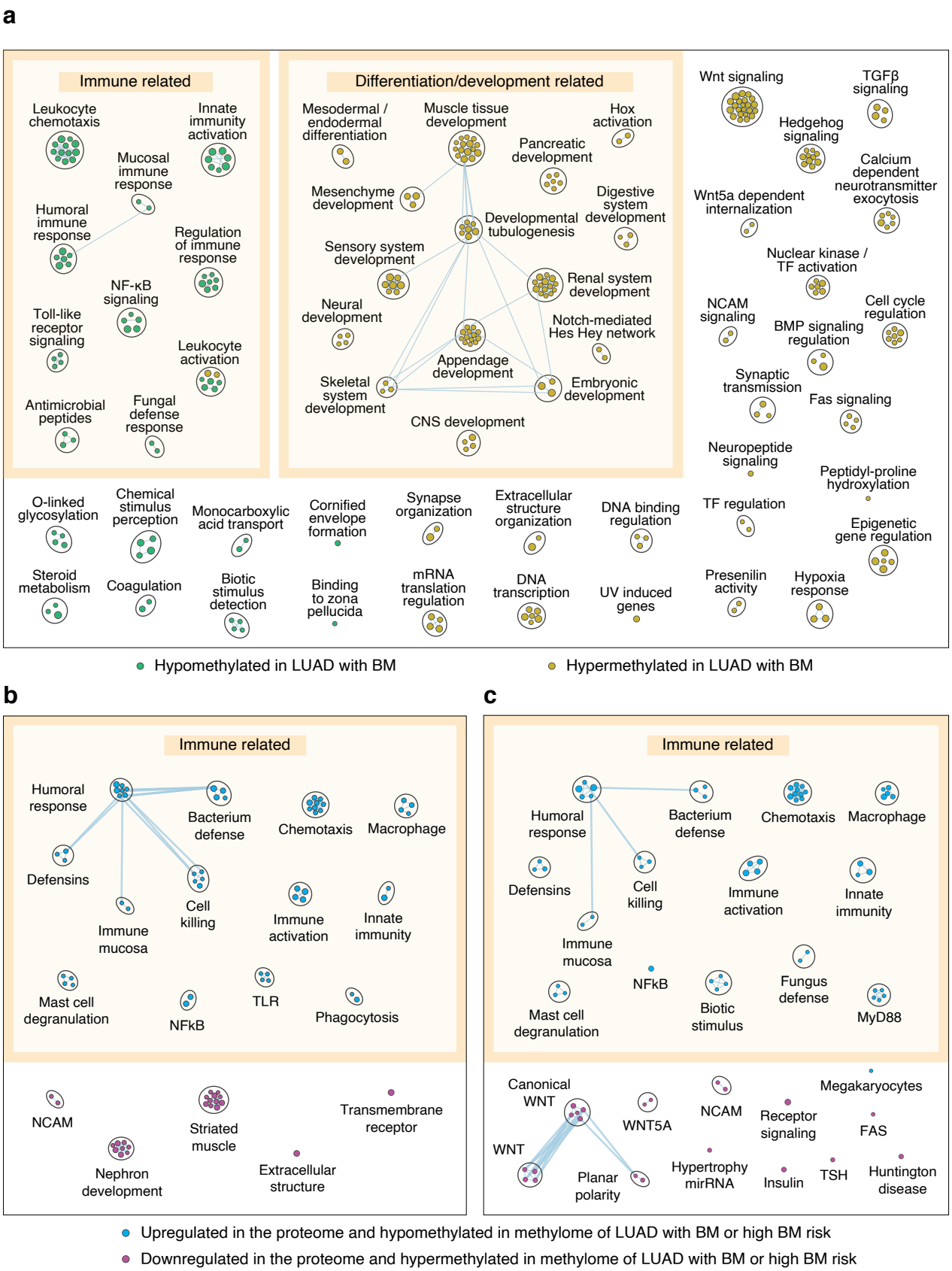
Extended Data Fig. 1 | Evaluation of optimal 5-year methylome-based BM predictor risk score cutoff in validation set and representative clinical cases illustrating use of the predictor. a, Kaplan Meier plot and log-rank test in validation cohort using risk groups stratified by the optimal cutoff, showing worse BM outcomes in the high risk group. **b**, Univariable Cox proportional hazards model of risk groups stratified by the optimal cutoff in validation set samples showing BM prediction (N = 60). **c**, Multivariable Cox proportional hazards modeling of BM development in validation set samples, demonstrating that methylome risk groups stratified by the optimal cutoff have prognostic utility independent of clinical factors (N = 56). **d**, Representative patient cases demonstrating the opportunity for potential escalation of treatment based on patient methylome-based predictor scores. In the upper panel case, a BM developed 1.8 years after diagnosis, which was not predicted by clinical factors as they were overall lower risk, but was consistent with the high methylome risk score. In the lower panel case, the patient did not develop a BM across 8.1 years of follow-up which was accurately predicted with a low methylome score but may not have been expected based on higher risk clinical factors. Boxes with whiskers display HR and 95% CI. * denotes a significant *P* value.

https://doi.org/10.1038/s41591-024-03286-y



Extended Data Fig. 2 | Utility of clinical factors for BM prediction in LUAD, evaluation of LUAD BM predictive modeling in lung squamous cell cancer (LUSC), and comparison between fresh frozen and FFPE paired methylation data. a, Univariable Cox proportional hazards modeling of BM development using clinical factors demonstrates that TNM component cancer staging scores have prognostic utility. **b**, Univariable Cox proportional hazards model results using numeric methylome risk scores from the 5-year LUAD-BM predictor in TCGA LUSC samples (N = 330). **c**, tSNE plot using the 5553 BM predictive CpG sites in both TCGA LUAD and LUSC cohorts for clustering. **d**, Hierarchical clustering using the top 1,000 most variably methylated CpGs across the 12 samples from 6 tumors. 1, N = 166; 2, N = 157; 3, N = 165. Boxes with whiskers display HR and 95% CI. * denotes a significant *P* value.

https://doi.org/10.1038/s41591-024-03286-y

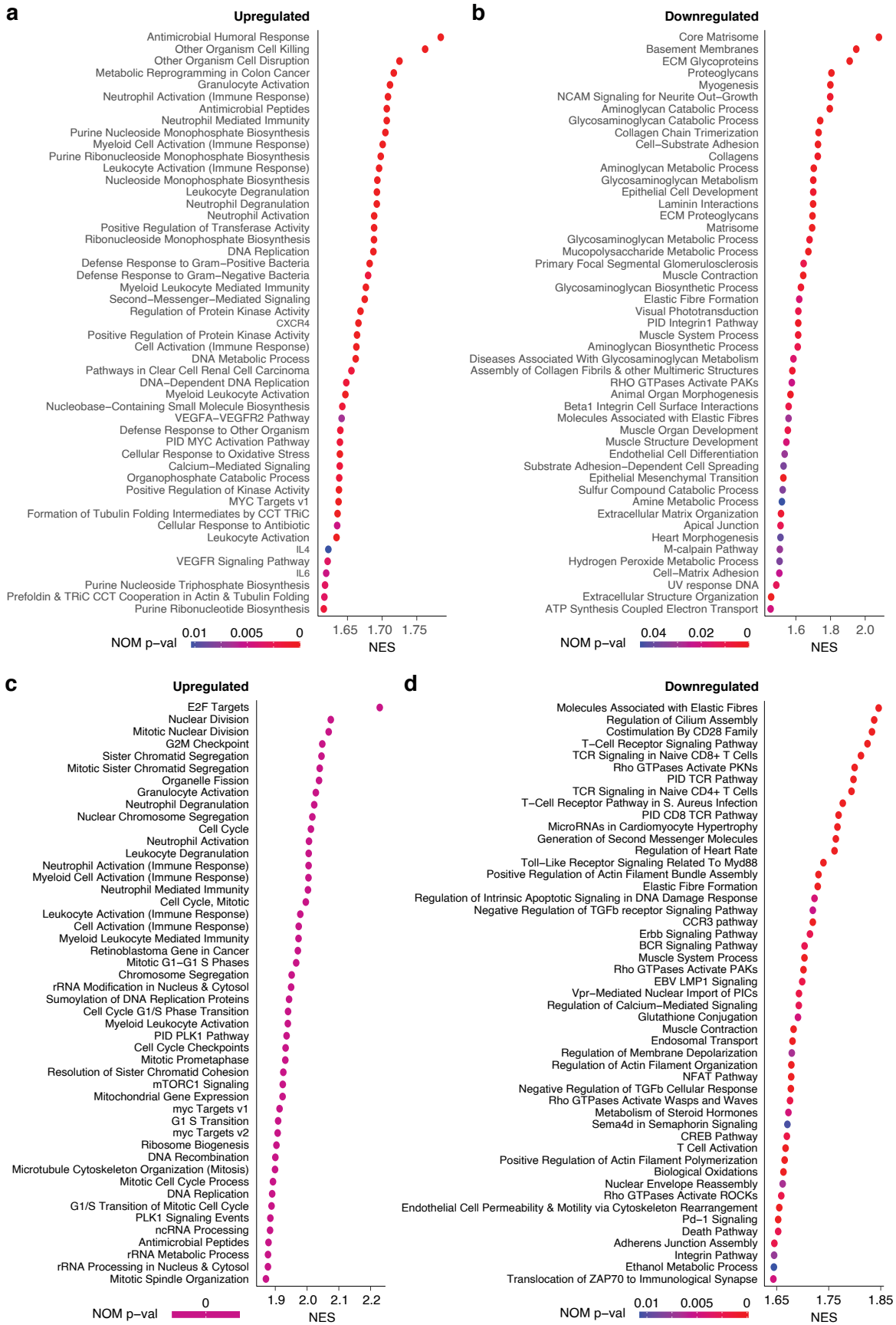


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41591-024-03286-y

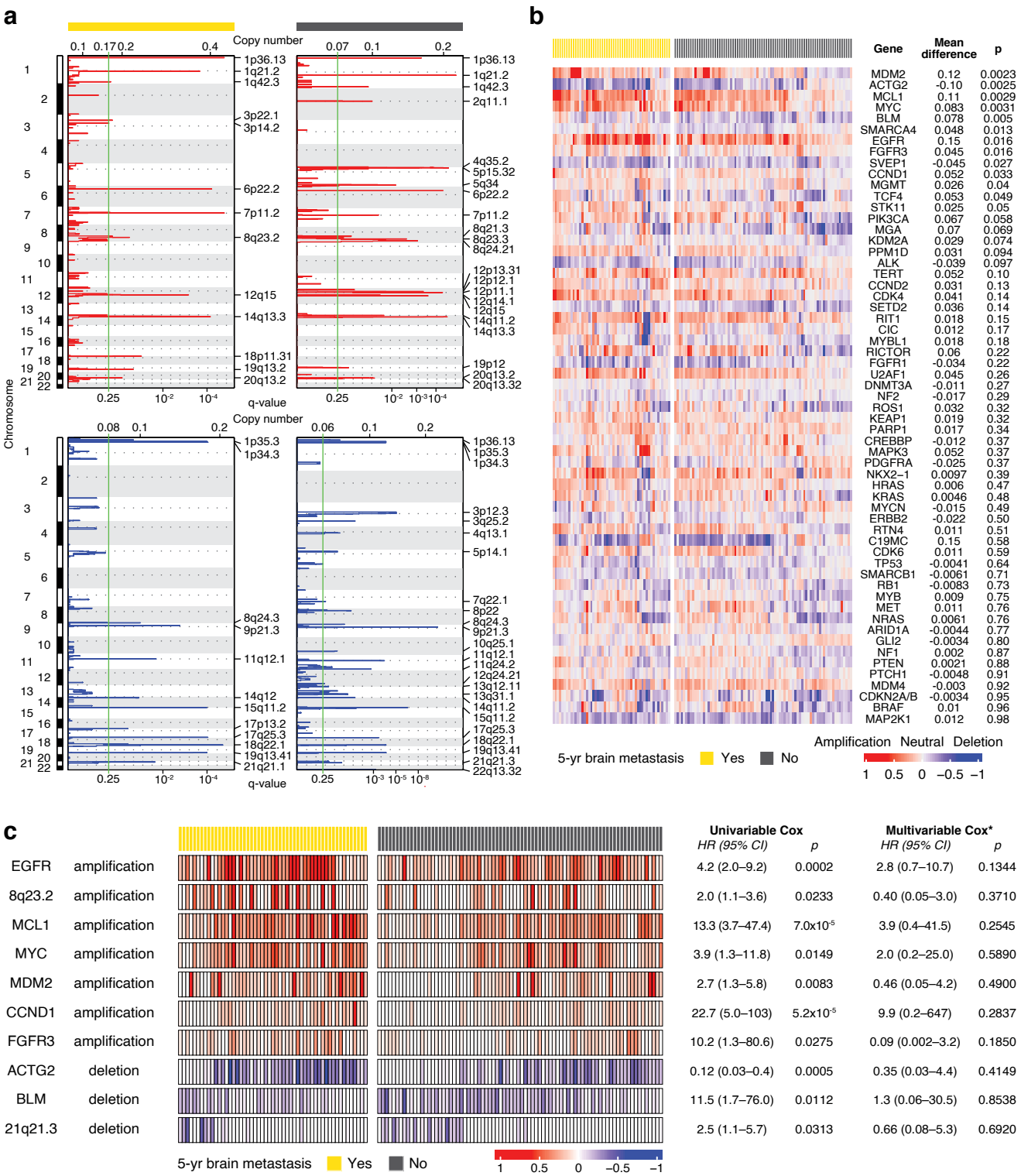
Extended Data Fig. 3 | Characterization of pathways differentially methylated at gene promoters and also dysregulated in the proteome within LUAD tumors that developed BM within 5 years or had high methylome-based risk scores. a, GSEA of pathways differentially methylated at gene promoters in LUAD that developed BM compared to LUAD without BM throughout 5-years of follow-up. Pathways shown are mainly immune and differentiation and development related as shown with yellow boxes. **b**, GSEA showing shared pathways of differential proteomics and methylation analyses within the study cohort comparing LUAD that developed BM versus those without BM over 5 years. **c**, GSEA of shared pathways of differential proteomics and methylation analyses using CPTAC dataset comparing proteomes from LUAD with high versus low methylome-based BM predictor risk scores.

https://doi.org/10.1038/s41591-024-03286-y



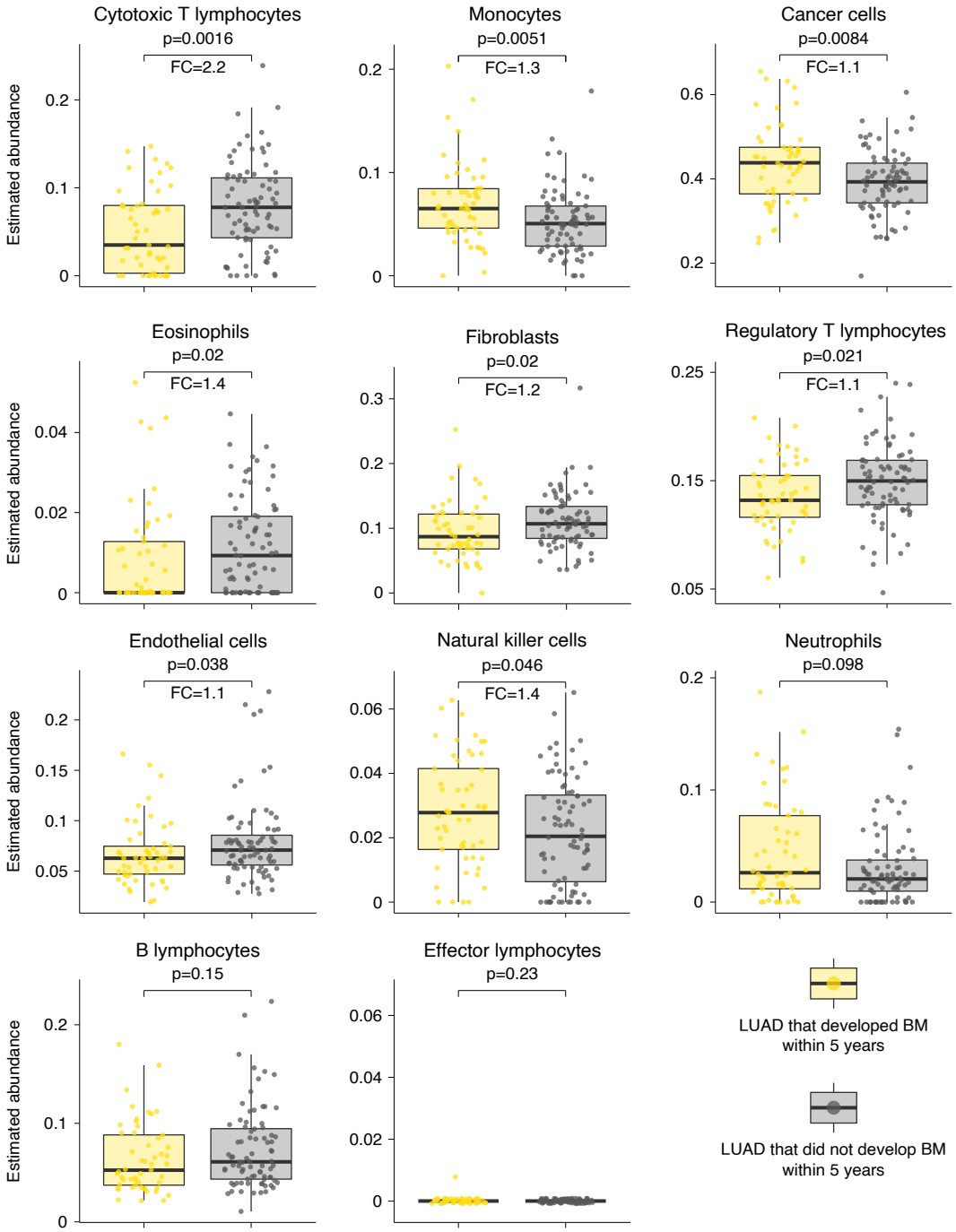
Extended Data Fig. 4 | Differential expression analysis between LUAD with and without BM risk. a-b, GSEA using study LUAD cohort comparing LUAD that developed BM versus those without BM over 5 years. **c-d**, GSEA using CPTAC dataset comparing proteomes from LUAD with high versus low methylome-based BM predictor risk scores. NES, normalized enrichment score.

https://doi.org/10.1038/s41591-024-03286-y



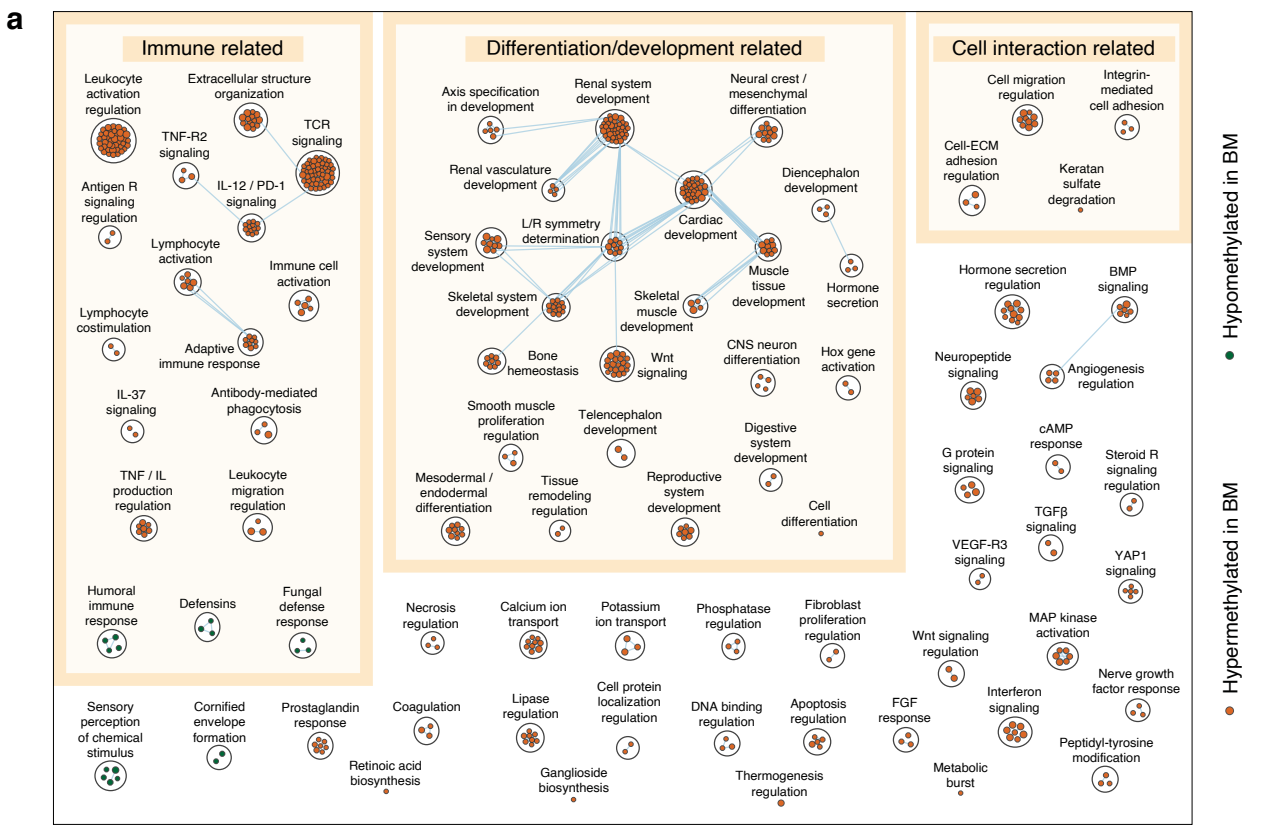
Extended Data Fig. 5 | Characterization of chromosome and gene-level copy number alterations in LUAD tumors that developed BM within 5 years compared to LUAD that did not develop BM within 5 years of follow-up. a, GISTIC analysis results with significantly amplified (red) and deleted (blue) chromosome regions depicted as peaks extending beyond the green significance cutoff (q < 0.25). **b**, Heatmap comparing gene-level copy number data between LUAD with and without 5-year BMs along with associated mean difference (yes vs. no BM) and Wilcoxon's rank sum test p-values. **c**, Chromosome-level and gene-level CN alterations as genomic correlates of nomogram risk scores (prognostic in univariable Cox proportional hazards models but not independent of nomogram scores in multivariable analyses). Correlates of increased BM risk include amplifications of *EGFR*, 8q23.2, *MCL1*, *MYC*, *MDM2*, *CCND1*, and *FGFR3* as well as deletions of *ACTG2*. Correlates of a lower BM risk were *BLM* and 21q21.2 deletions. *Multivariable models include each amplification or deletion together with nomogram risk scores.

https://doi.org/10.1038/s41591-024-03286-y



Extended Data Fig. 6 | Cell deconvolution analysis comparing cell type fractions between LUAD tumors that did and did not develop BM within 5 years. Boxplots of deconvoluted cell compositions from bulk tissue methylation data in BM and paired LUAD showing greater fractions of monocytes and natural killer cells as well as lower fractions of cytotoxic T lymphocytes, eosinophils, fibroblasts, regulatory T lymphocytes, and endothelial cells in LUAD that develop BM (N = 133). Wilcoxon's rank sum test p-values shown. FC, fold change. Boxplots display medians with central bars, upper and lower distribution quartiles using boxes, and 1.5x interquartile range with whiskers.

https://doi.org/10.1038/s41591-024-03286-y

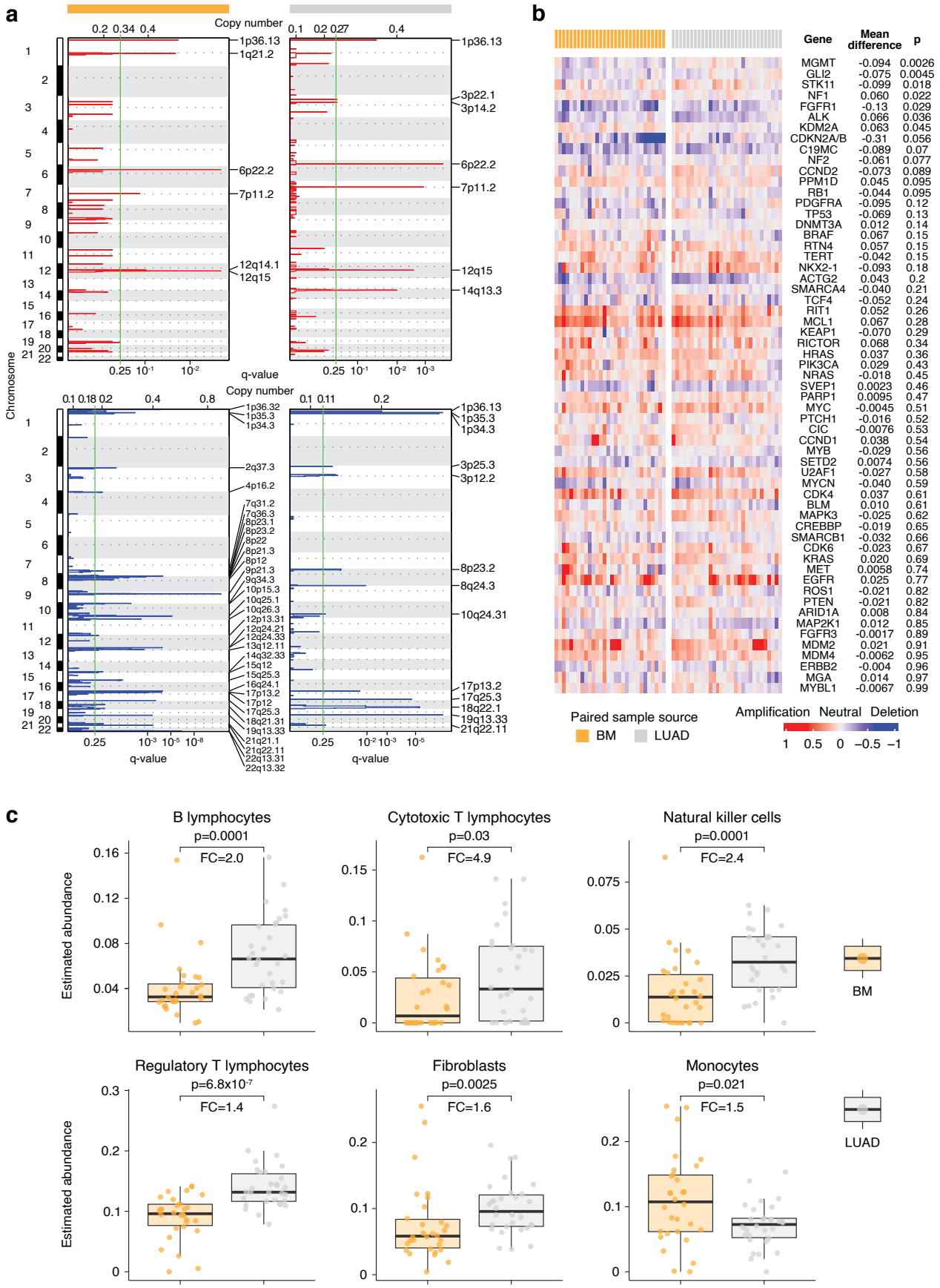


Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s41591-024-03286-y

Extended Data Fig. 7 | Characterization of pathways differentially methylated at gene promoters and also dysregulated in the proteome within paired patient BM tissue and primary LUAD tumors. a, GSEA of pathways differentially methylated at gene promoters in BM samples compared to LUAD including those related to immune function, tissue differentiation and development within the nervous system and other systems, and cell interaction. **b**, Proteomics analysis in paired primary LUAD and matched BM dataset showing shared pathways between differential proteomics analysis (data in ref. 22) and differential methylation analysis (study data). **c**, Metaprograms characterized by non-negative matrix factorization for clustering of proteomics data.

https://doi.org/10.1038/s41591-024-03286-y



Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41591-024-03286-y

Extended Data Fig. 8 | Chromosome-level and gene-level copy number alterations as well as cell deconvolution analysis comparing cell type fractions in BM tumor samples compared to paired LUAD primary tumors. **a**, GISTIC analysis results showing significantly amplified (red) and deleted (blue) chromosome regions as peaks with extension beyond the green significance cutoff ($q < 0.25$). **b**, Heatmap comparing gene-level copy number data between BM and paired LUAD with mean difference (BM vs. LUAD) and p-values. **c**, Boxplots of deconvoluted cell compositions from bulk tissue methylation data in BM and paired LUAD showing an overall greater immune cell and fibroblast abundance in LUAD tumors compared to paired BM samples ($N = 60$). FC, fold change. Boxplots display medians with central bars, upper and lower distribution quartiles using boxes, and 1.5x interquartile range with whiskers. Wilcoxon's rank sum test p-values shown.

https://doi.org/10.1038/s41591-024-03286-y

Extended Data Table 1 | Characteristics of the methylome-based predictor patient cohort

Characteristic		*Full cohort*	*BM within 5 years*	*No BM over 5 years*	*Discovery Set*	*Validation Set*
		Mean (Range)	*Mean (Range)*	*Mean (Range)*	*Mean (Range)*	*Mean (Range)*
Age	Years	64.1 (41–84)	60.7 (42–80)	65.6 (41–84)	64.5 (41-84)	63.3 (42-84)
Smoking extent	Pack-years	16.8 (0–100)	17.8 (0–100)	14.7 (0–60)	17.6 (0-100)	15.3 (0-80)
		Number (%)	*Number (%)*	*Number (%)*	*Number (%)*	*Number (%)*
EGFR mutational status	Mutant	101 (61.2)	35 (66.0)	55 (69.6)	63 (60.0)	38 (63.3)
	Wildtype	64 (38.8)	18 (34.0)	24 (30.4)	42 (40.0)	22 (36.7)
T component score	T1	47 (28.5)	8 (15.4)	32 (40.0)	30 (28.6)	17 (28.3)
	T2	81 (49.1)	27 (51.9)	37 (46.2)	52 (49.5)	29 (48.3)
	T3–4	37 (22.4)	17 (32.7)	11 (13.8)	23 (21.9)	14 (23.3)
N component score	N0	104 (62.7)	23 (43.4)	62 (77.5)	66 (62.3)	38 (63.3)
	N1	28 (16.9)	14 (26.4)	11 (13.8)	17 (16.0)	11 (18.3)
	N2–3	34 (20.5)	16 (30.2)	7 (8.8)	23 (21.7)	11 (18.3)
M component score	M0	151 (91.0)	39 (73.6)	80 (100)	96 (90.6)	55 (91.7)
	M1	15 (9.0)	14 (26.4)	0 (0)	10 (9.4)	5 (8.3)
Cancer stage	Stage I	84 (50.6)	13 (24.5)	55 (68.8)	53 (50.0)	31 (51.7)
	Stage II	33 (19.9)	14 (26.4)	15 (18.8)	21 (19.8)	12 (20.0)
	Stage III	33 (19.9)	10 (18.9)	10 (12.5)	21 (19.8)	12 (20.0)
	Stage IV	16 (9.6)	16 (30.2)	0 (0)	11 (10.4)	5 (8.3)
Initial treatment	Resection	161 (97.0)	48 (90.6)	80 (100)	102 (96.2)	59 (98.3)
	Radio-therapy	23 (13.9)	9 (17.0)	9 (11.2)	17 (16.0)	6 (10.0)
	Systemic therapy	79 (47.6)	36 (67.9)	27 (33.8)	51 (48.1)	28 (46.7)

! "

https://doi.org/10.1038/s41591-024-03286-y

Extended Data Table 2 | Proportion of patients who developed BM within 5 years in high-risk and low-risk groups based on maximizing the Youden index according to EGFR mutational status and staging variables

Characteristic		*High risk (above cutoff)*	*Low risk (below cutoff)*	*Coefficient of correlation*
		BM proportion (%)	*BM proportion (%)*	*Yule's Q*
EGFR mutational status	Mutant	4/5 (80.0)	8/26 (30.8)	0.80
	Wildtype	4/6 (66.7)	1/5 (20.0)	0.78
T component score	T1	2/4 (50.0)	1/10 (10.0)	0.80
	T2	3/3 (100)	6/17 (35.3)	1.00
	T3–4	3/4 (75.0)	2/4 (50.0)	0.50
N component score	N0	5/8 (62.5)	3/18 (16.7)	0.79
	N1	2/2 (100)	4/7 (57.1)	1.00
	N2–3	1/1 (100)	2/6 (33.3)	1.00
M component score	M0	5/8 (62.5)	7/29 (24.1)	0.68
	M1	3/3 (100)	2/2 (100)	–
Cancer stage	Stage I	3/5 (60.0)	1/15 (6.7)	0.91
	Stage II–III	2/3 (66.7)	6/14 (42.9)	0.45
	Stage IV	3/3 (100)	2/2 (100)	–

nature portfolio

Corresponding author(s): Gelareh Zadeh

Last updated by author(s): September 4, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Microsoft Excel v16.
Data analysis	Publicly available open source R (v. 4.0.3) packages used for analysis include minfi (v. 1.46.0), caret (v. 6.0-94), gbm (v. 2.1.8), survival (v. 3.5-8), hdnom (v. 6.0.0), boot (v. 1.3-25), cutpointr (v. 1.1.2), mRMRe (v. 2.1.2), limma (v. 3.56.2), conumee (v. 1.38.0), MethylCIBERSORT (v. 0.2.1), DEP (v. 1.22.0), NMF (v. 0.26), MEDIPS (v. 1.22.0), glmnet (v. 4.1-8), and Rtsne (v. 0.16). Publicly available open source software used for analysis include GSEA (v. 4.0.3), cytoscape (v.3.6), GISTIC (v. 2.0), MaxQuant (v. 2.4.7.0), FragPipe (v. 22.0), DI-ANN (v. 1.8.1), FASTQC (v. 0.11.5), bowtie2 (v. 2.5.1), and SAMtools (v. 1.3.1). Custom code has been deposited in the following repository: https://github.com/patilvikas/BrainMetsPrediction/tree/main.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The glioma plasma cfMeDIP-seq data was previously published and accessible from the Zenodo archive at https://doi.org/10.5281/zenodo.3715312. LUAD cell line methylation data for cell deconvolution was accessed from the Gene Expression Omnibus (GEO) under accession code GSE68379. TCGA LUAD and LUSC tissue methylation datasets were acquired from the Genomic Data Commons Data Portal at the following link https://portal.gdc.cancer.gov/analysis_page?app=Projects. External BM methylation data was accessed from GEO under accession code GSE108576. DKFZ methylation data is accessible from GEO accession code GSE109381. CPTAC LUAD tissue proteomics data was acquired from the Proteomic Data Commons Data Portal at the following link https://pdc.cancer.gov/pdc/cptac-pancancer (Proteome_BCM_GENCODE_v34_harmonized_v1.zip). External LUAD and LUAD BM tissue proteomics data was acquired from https://doi.org/10.1016/j.esmoop.2022.100741 (Supplementary Table S1).

DNA methylation data for all tissue samples in this study have been deposited to GEO under accession code GSE220838. cfMeDIP-seq data for plasma samples are available from the Zenodo archive at http://doi.org/10.5281/zenodo.7338401. Proteomics data in the study are accessible from PRIDE in the ProteomeXchange Consortium under accession code PXD048525.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	All patients with available samples were included in the study without any limitations placed on recruitment related to sex and gender.
Reporting on race, ethnicity, or other socially relevant groupings	All patients with available samples were included in the study without any limitations placed on recruitment related to race, ethnicity, or other social groupings.
Population characteristics	Baseline characteristics of this cohort of patients are shown in Extended Data Table 1.
Recruitment	All patients with tumor or plasma samples available to us were included in the study.
Ethics oversight	All samples were obtained upon approval of the institutional ethics committee (University Health Network)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined based on stored tumor and plasma availability. Sample selection was performed retrospectively and sample numbers were sufficient for analyses performed.
Data exclusions	No patient samples that underwent methylation profiling or cfMeDIP-seq were excluded from analyses.
Replication	No patient samples that underwent methylation profiling or cfMeDIP-seq were done in repeat given limited tissue availability per patient.
Randomization	Samples were allocated to disease status based on pathology reports regarding tumor type. Covariates were controlled for in multivariate analyses. Discovery and validation sets were separated using a random class balanced split approach.
Blinding	Tumor and plasma samples were processed and sequenced/profiled in a blinded manner. Unsupervised analyses were performed blinded to clinical data. Supervised analyses were unblinded due to the need to compare or partition the cohort by disease status.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	5-mC monoclonal antibody 33D3 (Cat#C15200081, lot# RD0004) from the MagMeDIP kit from Diagenode. 0.16 ug used per reaction.
Validation	Antibody validation for against 5-mC DNA shown by Diagenode: https://www.diagenode.com/files/products/antibodies/Datasheet_5-mC33D3_C15200081.pdf, using MeDIP-seq, Dot blot, immunofluorescence and Surface plasmon resonance (SPR) analysis.