MONTE: Methylation-based Observation Normalization and Tumor purity Estimation

Mirae Kim[1,†], Wei-Hao Lee[2,†], Vicky Yao[1,3,4,*]

[1] Department of Computer Science, Rice University, Houston, TX 77005
[2] Systems, Synthetic, and Physical Biology, Rice University, Houston, TX 77005
[3] Ken Kennedy Institute, Rice University, Houston, TX 77005
[4] Rice Synthetic Biology Institute, Rice University, Houston, TX 77005
[†]These authors contributed equally to this work.
[*] *Correspondence to:* vy@rice.edu

Abstract

Bulk DNA methylation profiling is widely used to study cancer epigenomics in clinical settings. However, these measurements aggregate signals from malignant and non-malignant cells, introducing composition-dependent confounding that complicates tumor-intrinsic interpretation and cross-cohort analyses. While existing methods can estimate tumor purity and, in some cases, correct methylation measurements, they are often constrained by cancer-specific reference models, predefined probe sets, or limited adaptability to new datasets and purity definitions. We present MONTE (Methylation-based Observation Normalization and Tumor purity Estimation), a cancer label–free method for tumor purity inference and CpG-resolved methylation purification from bulk DNA methylation data. MONTE learns probe-wise relationships between observed methylation and tumor purity using a stabilized linear model with empirical Bayes variance moderation and infers purity in new samples via signal-to-noise weighted aggregation of probe-level effects, without requiring matched normal samples or predefined probe sets. MONTE achieves robust purity estimation across cancer types using a single pan-cancer model and can be efficiently recalibrated to specific contexts through Bayesian transfer learning. MONTE provides a flexible, scalable, and interpretable framework for both methylation-based tumor purity estimation and correction, yielding tumor-intrinsic profiles for downstream analysis.

Introduction

DNA methylation is a major epigenomic modification that plays a central role in regulating gene expression and maintaining cellular identity. These patterns are dynamically shaped by both exogenous environmental exposures, such as smoking, and endogenous biological factors, such as age. In cancer, aberrant methylation landscapes are a hallmark of tumorigenesis, often involving the silencing of tumor suppressor genes and activation of oncogenic pathways, thereby promoting uncontrolled cell proliferation and tumor progression [1]. Interpreting these alterations in bulk tumor samples, however, is complicated by the fact that measured methylation profiles reflect mixtures of malignant and non-malignant cell populations (e.g., immune and stromal cells) [2].

Most cancer DNA methylation studies rely on bulk tumor samples, in which measurements aggregate signals across heterogeneous cellular compositions. DNA methylation microarrays remain widely used in both research and clinical settings due to their robustness and cost-effectiveness. These platforms quantify methylation at hundreds of thousands of predefined genomic sites using probes that target cytosine-guanine (CpG) dinucleotides. However, variation in tumor purity can introduce systematic differences in the observed

Table 1. Comparison of MONTE with existing, executable DNA methylation tumor purity estimation methods.

	Method			
Property	MONTE	PureBeta [6]	InfiniumPurify [5]	PAMES [7]
Core capabilities				
Purity estimation	✓	✓	✓	✓
Methylation purification	✓	✓	✓	✗
Data and annotation requirements				
Does not require normal samples	✓	✓	✗	✗
Does not require cancer labels	✓	✗	✗	✗
Does not require predefined probe sets	✓	✓	✗	✗
Statistical modeling properties				
Modeling framework	OLS + EB	OLS	DMC	DMC
Probe-weighting by statistical reliability	SNR	N/A	N/A	N/A
Transferability and usability				
Transfer learning across datasets/purity definitions	✓	✗	✗	✗
Software availability	Python	R	R	R

Note: MEpurify and RF_purify are not listed because we could not obtain results from them (see Methods). DMC: Differential methylated CpGs; EB: Empirical Bayes; OLS: Ordinary least squares; SNR: Signal-to-noise ratio

methylation profiles. This compositional heterogeneity can bias downstream analyses, leading to spurious or attenuated differential methylation signals and confounded clinical associations [3–5]. Accurate tumor purity estimation followed by appropriate correction of methylation measurements is a critical prerequisite for analyses that aim to recover tumor-intrinsic epigenetic signals.

Several approaches have been proposed to infer tumor purity from DNA methylation data, differing in their core capabilities, data requirements, and modeling assumptions (Table 1). Methods such as InfiniumPurify [5], MEpurity [8], and PAMES [7] rely on differential methylation between tumor and tissue-matched normal references to identify cancer-associated CpG sites. Tumor purity is then estimated using summary statistics [5, 7] or mixture modeling [8] over these predefined, cancer-specific probe sets. While effective in supported settings, these approaches require the availability of appropriate normal samples and are trained separately for each cancer type, limiting their applicability in rare cancers or in cohorts lacking suitable normal references. Other methods model probe-purity relationships directly using regression-based frameworks. RF_purify [9] applies random forest regression to predict purity, but it requires an exact match to the probe set used during model training, limiting robustness to probe filtering and cross-dataset application. PureBeta [6] does not require predefined probe sets, but it combines probe-wise linear regression with a computationally intensive mixture modeling procedure and requires cancer-specific model training. Pretrained PureBeta models are currently available for only a small number of cancer types.

Importantly, tumor purity estimation alone can only identify samples with substantial non-tumor contamination but cannot recover tumor-intrinsic DNA methylation signals. To make effective use of samples with varying purity, purity information must be coupled with explicit correction of methylation measurements to yield a CpG-resolved pure-tumor methylome, required for downstream analyses such as molecular subtyping [2, 10], regulatory interpretation [11], or integration with other omics layers [12, 13]. Only PureBeta [6] and InfiniumPurify [5] currently support both estimation and correction, but both rely on fixed, predefined reference models and purity assumptions that are tightly coupled to specific cancer types and data contexts. These constraints limit adaptability across datasets and purity scoring frameworks, motivating the need for more flexible and scalable approaches.

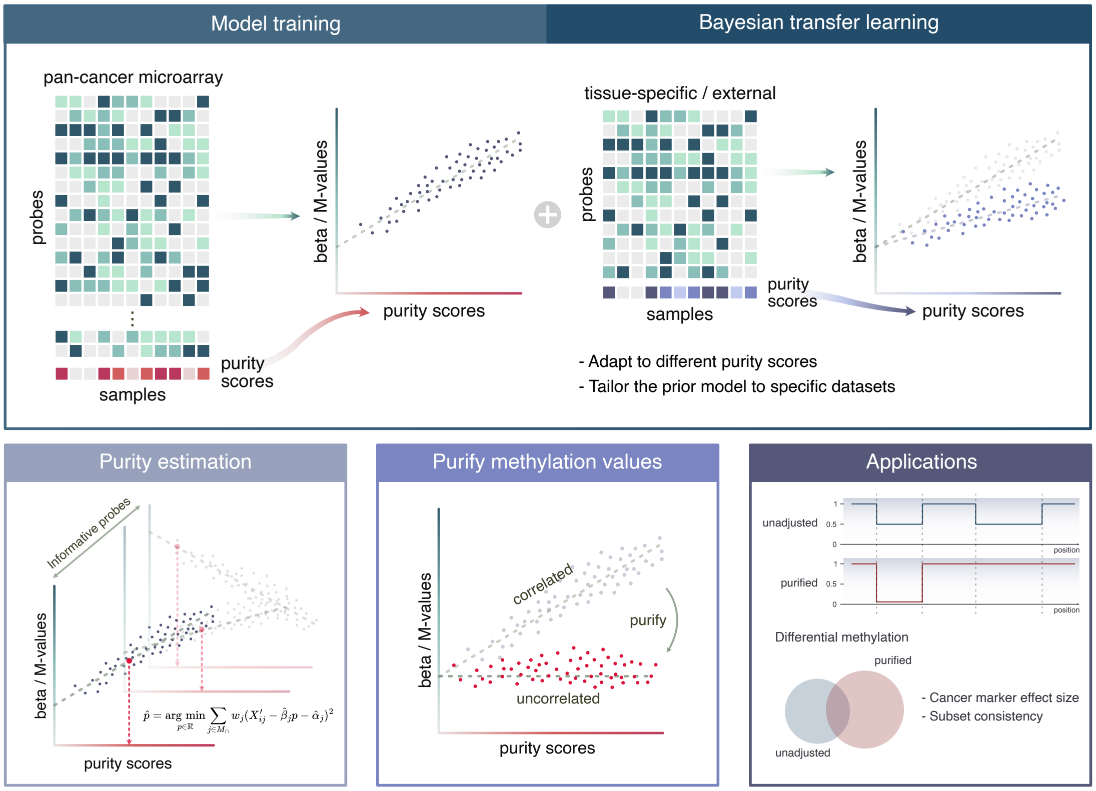

Figure 1. Overview of MONTE. (Top) The pan-cancer model learns probe-purity relationships from data without the use of cancer labels, then optionally refines these coefficients using Bayesian transfer learning on tissue-specific or external datasets. (Bottom) MONTE predicts purity using weighted regression and purifies methylation values by removing purity-associated signal. The resulting profiles better reflect pure-tumor methylation and improve downstream analyses.

Here, we present MONTE (Methylation-based Observation Normalization and Tumor purity Estimation), a flexible and label-free framework for tumor purity estimation and methylation purification (Figure 1). MONTE does not require matched normal samples, cancer type labels, or preemptive feature selection. It models probe-wise relationships between the observed methylation and tumor purity using a regression framework with empirical Bayes variance moderation, resulting in stabilized and uncertainty aware estimates across heterogeneous datasets. Purity inference for new samples is performed via signal-to-noise weighted aggregation of probe-level effects, enabling robust prediction even when only a subset of probes are available. To extend beyond a single cohort or purity metric, MONTE incorporates Bayesian transfer learning, allowing pretrained models to be efficiently recalibrated to new datasets or alternative purity scoring systems using limited labeled data. MONTE's learned probe-purity relationships enable projection to complete tumor purity, producing a CpG-resolved purified tumor methylome suitable for downstream analysis. Across cancer and cross-dataset evaluations, MONTE achieves superior purity estimation accuracy compared to existing methods, reduces confounded probe-purity correlations, and improves recovery of tumor-intrinsic signals.

Results

A single MONTE model provides robust pan-cancer tumor purity estimation

MONTE estimates tumor purity from bulk DNA methylation data by aggregating probe-level associations between methylation and purity learned in a pan-cancer setting (Figure 1). In this framework, probe-wise

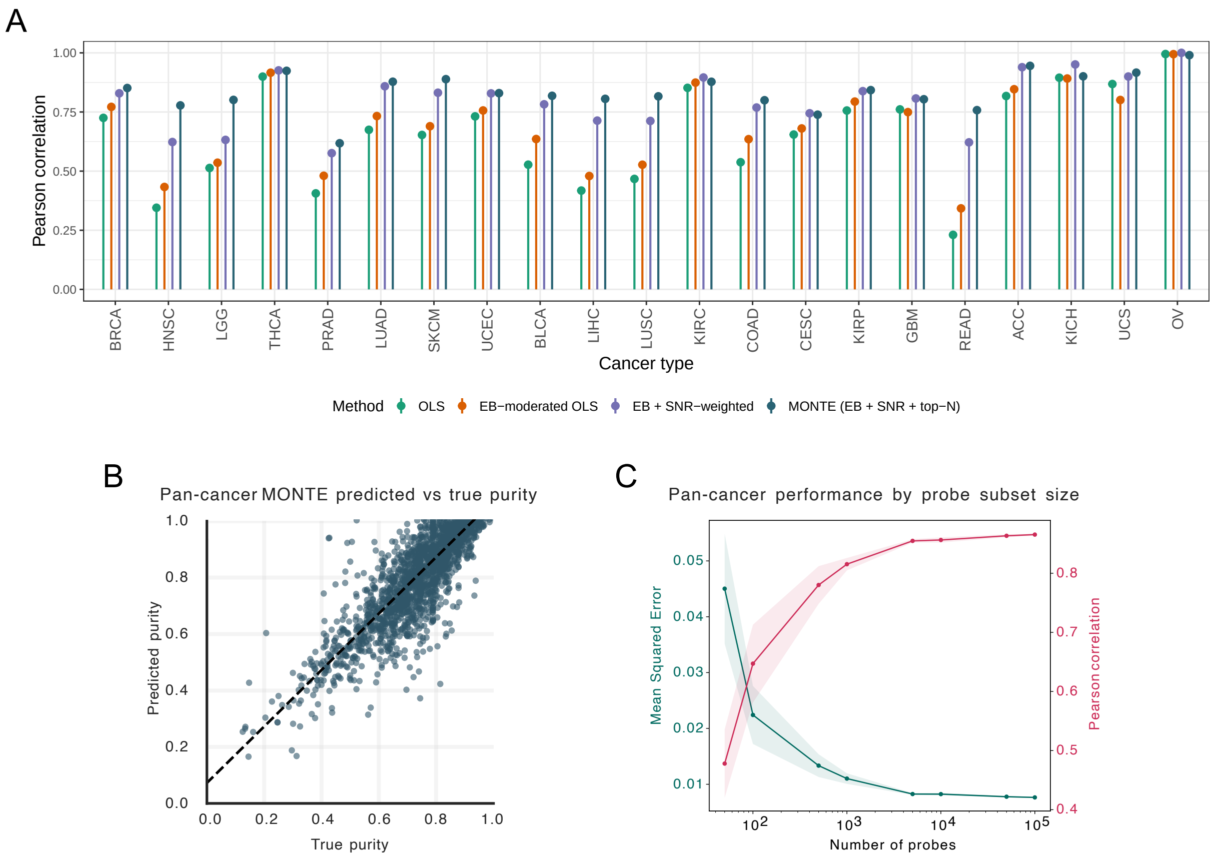

Figure 2. Purity estimation performance. (A) Pearson correlation between estimated and true tumor purity across cancer types with MONTE ablations. (B) Pan-cancer MONTE predictions compared against true TCGA CPE purity values on the TCGA test set (Pearson r: 0.865, MSE: 0.008). The top_n hyperparameter was selected through 5-fold cross-validation on the training set, resulting in optimal top_n of 250. (C) Pearson correlation and mean squared error between predicted and true purity as a function of number of probes in input samples showing performance robustness with any probe set with larger than 1000 probes.

regression coefficients are estimated using ordinary least squares with empirical Bayes variance moderation, and probes are weighted by their signal-to-noise ratio when inferring purity for new samples. An optional top-N probe selection step is used to regularize estimation when large or heterogeneous probe sets are available (see Methods).

We evaluated MONTE using DNA methylation data from The Cancer Genome Atlas (TCGA) [14], which provides one of the largest publicly available DNA methylation resources spanning diverse cancer types. We used Illumina HumanMethylation 450K data from TCGA samples with consensus purity estimates (CPE), which integrate multiple purity measures derived from copy number, gene expression, DNA methylation, and histopathology [2]. After quality control, this resulted in a dataset of 7,040 samples across 21 cancer types. While CPE is an imperfect and potentially noisy proxy for true tumor purity and incorporates methylation-based criteria, it provides a standardized reference that is commonly used for comparative evaluation across methods. To mitigate concerns about circularity and data leakage, samples were split into disjoint training and test sets using a 70%/30% split stratified by cancer type, and all evaluations were performed exclusively on held-out test samples, notably without providing any cancer labels.

To understand how individual components of the MONTE framework contribute to robust purity estimation, we performed an ablation analysis comparing alternative formulations of the same modeling framework, including baseline ordinary least squares (OLS) regression, empirical Bayes-moderated OLS (Var), empirical

Bayes with signal-to-noise (SNR)-based probe weighting, and the full MONTE model incorporating top-N probe selection. Across cancer types, each successive component yielded significant improvements in purity estimation accuracy as measured by Pearson correlation (Figure 2A, one-sided Wilcoxon signed-rank test; MONTE vs OLS p=9.5e-7, MONTE vs Var p=1.4e-6, MONTE vs SNR p=4.5e-3). Although the magnitude of improvement varies across cancer types, MONTE consistently matches or outperforms simpler variants with no clear dependence on cohort size. Larger gains are observed primarily in cancers with lower baseline OLS performance, suggesting that MONTE preferentially stabilizes purity estimation in more challenging settings.

The full MONTE model achieves strong agreement with reference purity values on the held-out test set (Figure 2B; Pearson r=0.865, MSE=0.008), demonstrating that a single pan-cancer model can accurately estimate tumor purity without using cancer labels. To assess robustness to probe availability, we further evaluated MONTE using randomly subsampled input probe sets across several orders of magnitude, with five random seeds per setting. Performance improves rapidly as the number of probes increased and stabilized within a few thousand probes (Figure 2C), corresponding to <1% of the ∼450K probes captured on the array, showing that MONTE is robust to substantial probe filtering and does not rely on predefined probe panels.

MONTE improves the accuracy and scalability of pan-cancer tumor purity estimation

We next compared MONTE with existing DNA methylation–based tumor purity estimation methods to assess accuracy, coverage, and computational scalability in a pan-cancer setting (Figure 3). For methods that rely on cancer-specific reference models or probe sets, we evaluated performance under two settings: pretrained and dataset-fitted. In the pretrained setting, methods were applied using provided reference models or probe sets, reflecting their intended out-of-the-box usage, though we note that both methods use TCGA data for training, so inevitably may overlap with the test samples. In the dataset-fitted setting, methods were trained using the same TCGA training split used for MONTE, providing a best-case comparison under matched data availability. All methods were evaluated on the same TCGA test samples (held out from MONTE and the dataset-fitted settings).

Across cancer types, MONTE achieves consistently stronger performance than competing methods in nearly every cancer type where direct comparison is possible (Figure 3A). Specifically, MONTE significantly outperformed PAMES (difference in medians = 0.288, Wilcoxon $p = 2.44 * 10^{-4}$), and both variants of InfiniumPurify (difference in medians = 0.145 (dataset-fitted) and 0.114 (pretrained), Wilcoxon $p = 4.88 * 10^{-4}$ and $p = 9 * 10^{-6}$). The performance gain relative to PureBeta was also positive (differences in medians = 0.269 (dataset-fitted) and 0.057 (pretrained)) though statistical testing was not performed due to the limited number of cancer types for which PureBeta models were available (n=3). Notably, MONTE achieves this performance using a single pan-cancer model applied uniformly across all cancer types, whereas existing approaches rely on cancer-specific reference models or probe sets that limit their generalizability.

We further evaluated the computational scalability of MONTE relative to existing methods for purity estimation using pretrained models as well as fitting and estimating purity with a given input dataset (Figure 3B). In both scenarios, MONTE and InfiniumPurify had comparable, efficient runtimes, whereas PureBeta required significantly more computation, reflecting the higher cost of its mixture modeling-based fitting procedure (median runtimes, pretrained-dataset-fitted: MONTE, 0.02s–3.79m; InfiniumPurify, 0.09s–1.79m; PAMES, 0.029s-n/a; PureBeta, 8.6m–25.9h). In the dataset-fitted setting, MONTE requires only a single pan-cancer model fit followed by rapid purity estimation across all cancers, resulting in substantially lower total runtime for the full dataset (3.113m). In contrast, PureBeta and InfiniumPurify require cancer-specific fitting procedures and correspondingly higher cumulative computational cost when applied at scale (79.965h and 22.513m, respectively).

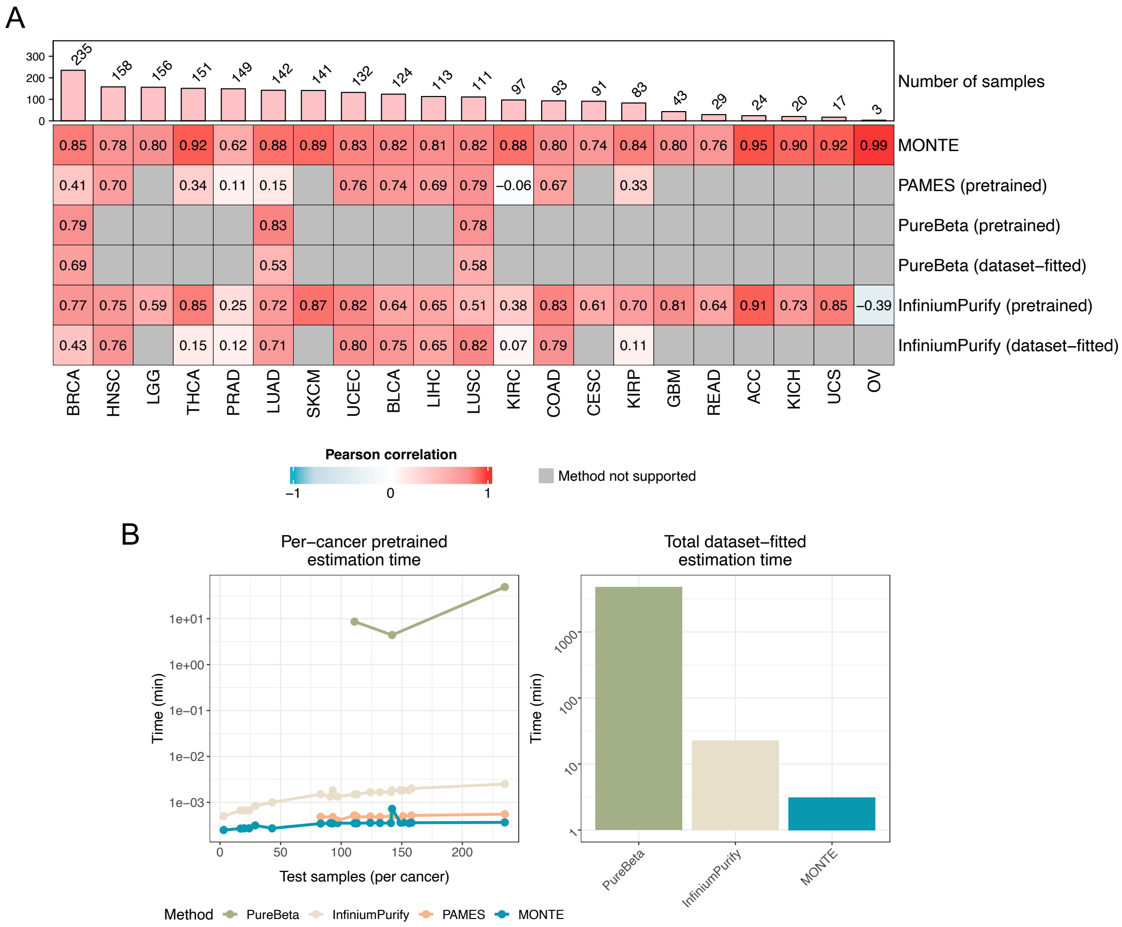

Figure 3. Benchmarking the performance of purity estimation against existing methods. (A) Test-set purity correlation by TCGA cancer type comparing MONTE (pan-cancer), PAMES, PureBeta, and InfiniumPurify. Pretrained models use method-provided models and DMCs to estimate purity, and dataset-fitted models estimate using model and DMC fitted on our data split. MONTE pan-cancer outperform others in most cancers, with noticeable gains in cancers with limited training samples such as OV. (B) Pretrained estimation time (left) shown as function of number of test samples per cancer for each method. Dataset-fitted runtime (right) shows the total time required to estimate purity for the entire test set across all cancers for each method with fitting functionality.

Bayesian transfer learning enables methylation purification across cancer types and cohorts

While the pan-cancer MONTE model provides accurate tumor purity estimates, Bayesian transfer learning allows the model to be further adapted for methylation purification and for application to datasets with differing purity definitions. This framework updates probe-level coefficients from a pan-cancer prior using a limited number of samples with known purity, enabling context-specific recalibration without retraining the model from scratch. To evaluate methylation purification in a cancer-specific setting while avoiding data leakage, we examined TCGA-BRCA, LUAD, and LUSC using different pan-cancer priors constructed from training data excluding the target cancer (see Methods). These cancer types were selected because they were the only ones available for both PureBeta and InfiniumPurify.

In BRCA, purification with MONTE sharpens the distribution of probe–purity correlations by pulling correlations toward zero (Figure 4A). This behavior is consistent with the expectation that once tumor and non-tumor component are disentangled and adjusted for pure tumor, probe-level methylation should no longer track sample-level purity. For a more systematic evaluation, probes were stratified by their unadjusted

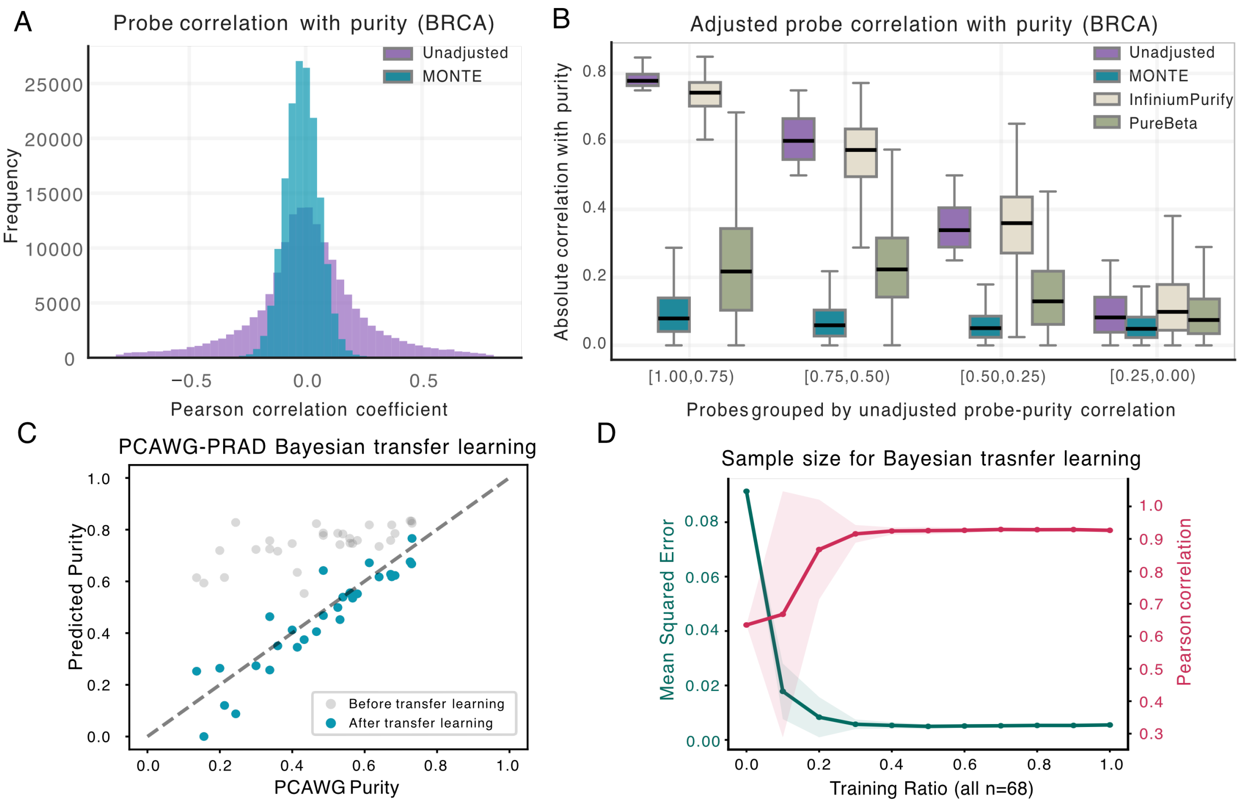

Figure 4. Methylation purification. (A) Distribution of probe-purity correlations in BRCA before and after purification. Purification collapses purity-dependent variation while preserving biological signal, producing a tighter distribution centered near zero. (B) Probes grouped by their unadjusted purity correlation, comparing the correlation after purification using MONTE, PureBeta, and InfiniumPurify. MONTE consistently produces the largest reduction toward zero across all correlation bins. (C) Purity estimation performance of MONTE before and after Bayesian transfer learning on external PCAWG-PRAD dataset. Bayesian transfer learning calibrates to different purity measurements and provide performance gain in purity estimation. (D) Purity estimation performance as a function of sample size in Bayesian transfer learning on PCAWG-PRAD. Bayesian transfer learning stabilizes in performance after sample size of 14.

purity correlations, and we quantified the extent to which each method reduced these correlations (Figure 4B). Across all correlation bins, MONTE consistently produces the strongest reduction toward zero, with the largest adjustments occurring for probes that were most strongly confounded by purity in the unadjusted data. These trends were also observed in LUAD and LUSC (Supplementary Figure 1).

To assess whether MONTE's Bayesian transfer learning generalizes to external cohorts, we evaluated it on data from the Pan-Cancer Analysis of Whole Genomes (PCAWG) [15, 16]. PCAWG provides a well-controlled setting for evaluation because a subset of samples overlaps with TCGA but is annotated with purity estimates based on copy number variation, which differ systematically from TCGA's CPE measure. These overlapping samples enable direct assessment of model behavior before and after recalibration to an alternative purity definition. To avoid leakage, all overlapping samples between TCGA and PCAWG were excluded from the pan-cancer prior construction as well as samples used for transfer learning and used exclusively for evaluation. We find that Bayesian transfer learning recalibrates probe-level coefficients to better align with the external purity definition while retaining information from the original model, correcting systematic offsets and substantially improving agreement with PCAWG purity estimates (Figure 4C).

We further examined how recalibration performance changes as a function of the number of samples used during Bayesian transfer learning. Both correlation and MSE compared with PCAWG purity values improve rapidly as the number of calibration samples increases, with substantial gains achieved using as few as

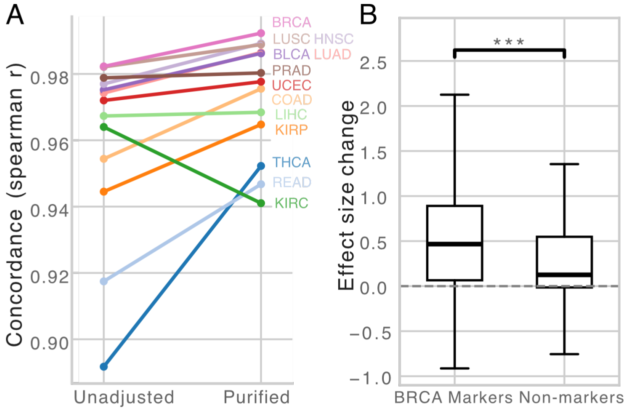

Figure 5. Purification increases cancer marker effect sizes and biological signal consistency. (A) Median spearman correlation of differential methylation effect sizes across cancer types between 5 iterations of randomly divided halves. Calculated for cancer types with larger than 5 samples in both test and normal sample sets (n cancers=13) and shows improved consistency in purified profiles. (B) Change in BRCA-normal breast differential methylation effect sizes in known BRCA marker gene promoters compared to non-markers (n CpGs=865). Statistical significance was calculated by comparing to 5,000 permutations of random non-marker promoter CpG sets of same size (n=865), and one representative iteration is shown (***: permutation p-value<1e-3.)

20 samples and stabilizing soon after (Figure 4D). The sample-efficient behavior reflects the fact that the pan-cancer prior already captures genome-wide relationships between DNA methylation and tumor purity, allowing effective adaptation to alternative purity definitions with limited cohort-specific data.

Methylation purification enhances tumor differential methylation signal and consistency

We first examined the impact of methylation purification on the consistency of differential methylation analyses across cancer types. For each cancer, tumor samples were randomly divided into two equal halves (five random seeds), and differential methylation against normal samples was computed independently for each split using both purified and unadjusted methylation profiles. Concordance between splits was quantified using Spearman correlation of probe-level effect sizes, and median correlations were compared before and after purification. Across almost all cancer types, purification increased concordance, with the exception of KIRC (with a modest decrease from 0.964 to 0.941), indicating improved reproducibility of differential methylation effects (Figure 5A). Notably, THCA saw the largest increase in concordance (from 0.892 to 0.952). The relatively large improvement post-purification on concordance in THCA may be due to its high immune infiltration rate [17].

Having established improved consistency across cancers, we next asked whether methylation purification also enhances recovery of biologically meaningful tumor-specific signal. We focused on BRCA, where well-characterized breast cancer marker genes are available, and performed differential methylation analysis between the full test set of tumor versus normal samples pre- and post-purification. For each CpG site, effect size differences were calculated as the change in absolute differential methylation coefficients between the two settings. CpGs located in the promoter regions of literature-curated breast cancer marker genes [18] exhibited significantly larger increases in differential methylation effect sizes compared to size-matched CpGs from promoter regions of non-marker genes (p=1.99e-4; Figure 5B), demonstrating the amplification of tumor-relevant regulatory signal following purification.

Discussion

In this study, we present MONTE, a unified framework that addresses several long-standing challenges in the analysis of bulk DNA methylation data by jointly enabling accurate tumor purity estimation, CpG-resolved methylation purification, and flexible adaptation across cancer types and datasets. MONTE extends classical linear modeling approaches by incorporating probe-wise empirical-Bayes variance moderation and signal-to-noise weighting, stabilizing effect estimates across heterogeneous CpGs. Compared to existing approaches, MONTE provides improved accuracy and broader applicability while remaining computationally efficient and interpretable.

A key feature of MONTE is the ability to operate as a single pan-cancer model without requiring tumor labels or normal samples, achieving strong out-of-the-box purity estimation performance across diverse cancer types. Bayesian transfer learning further extends this framework by allowing probe-purity relationships to be recalibrated using limited samples in new contexts without retraining the model from scratch. This design is particularly advantageous for underrepresented and external cohorts, where large reference datasets may not be available. Beyond purity estimation, MONTE enables CpG-resolved purification of methylation profiles, facilitating downstream analyses aimed at recovering tumor-intrinsic epigenetic signal.

As with any modeling framework, MONTE operates under a set of assumptions that define its current scope. The method models bulk methylation measurements as a linear combination of tumor and non-tumor signal, corresponding to a two-component mixture. This abstraction is intentionally simple, capturing the dominant source of confounding in bulk tumor data while remaining interpretable and data-efficient. Nevertheless, more expressive extensions, such as explicit modeling of multiple non-malignant compartments or incorporation of uncertainty in purity estimates, could further refine this framework and enable deeper understanding of tumor-microenvironment interactions.

More broadly, MONTE provides a statistical scaffold for integrating DNA methylation data. Although our analyses have focused on array-based methylation data here due to its broader availability, MONTE's design is not platform-specific and could be extended to sequencing-based assays through the same transfer learning framework. Thus, MONTE is particularly well-suited for integrative cancer epigenomics studies across heterogeneous data to more effectively understand tumor-intrinsic epigenetic signal at scale.

Methods

Data preprocessing

Sample downloading and curation. All publicly available DNA methylation data from The Cancer Genome Atlas (TCGA) [14] were collected using the TCGAbiolinks package [19] in R. For each TCGA project, Illumina HumanMethylation450 (450K) beta values processed and normalized by the TCGA consortium were used directly. Samples without CPE purity annotations were excluded from analysis. CpG probes with missing beta values or located on sex chromosomes were excluded, following best practices for quality control [20]. For pan-cancer modeling, the intersection of probes across all projects with purity annotations was used.

Sample partitioning. For the pan-cancer model, we used a 70:30 train-test split stratified by cancer type. For cancer-specific modeling with Bayesian transfer learning, samples were partitioned further into train:val:test sets, where the cancer-specific train set had train samples from non-target cancers and the cancer-specific validation and cancer-specific test sets respectively contained train and test samples of the target cancer (Supplementary Table 1). To remove the possibility of data leakage from physiological or tissue-of-origin resemblance, related cancer types were grouped during prior construction, including cancers of the digestive tract (COAD and READ), kidney (KIRC, KIRP, and KICH), uterine (UCEC and UCS), brain (GBM and LGG), and lung (LUAD and LUSC).

External datasets. We used data from the PRAD-CA project from the Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort [15, 16] as an external validation, which provides 450K data with purity estimates derived from whole-genome sequencing (WGS). Only donors with a single tumor WGS sample was used to ensure one-to-one correspondence between methylation and purity measurements.

Model design of MONTE. MONTE models probe-wise associations between bulk DNA methylation and tumor purity using a linear regression framework. Let $X \in \mathbb{R}^{n\times m}$ denote centered methylation measurements (can be beta or M-values) for n samples and m CpG probes, and let $p \in \mathbb{R}^n$ denote the vector of centered tumor purity values. For each probe j, MONTE fits a linear model:

$$X_{ij} = \alpha_j + \beta_j p_i + \epsilon_{ij}, \tag{1}$$

where α_j, β_j are probe-specific intercept and coefficient parameters, and $\varepsilon_j \sim N(0, \sigma_j)$ is the probe residual error and the σ_j is population variance for the probe j. Probe-wise coefficients are estimated using ordinary least squares (OLS), treating tumor purity as the independent variable and methylation values as the response.

To stabilize estimates across heterogeneous probes and variable sample sizes, MONTE applies empirical Bayes variance moderation to the OLS residual variances, shrinking probe-specific variance estimates toward a shared prior. As in limma [21], we assume that the population variance of each probe, σ_j, is drawn from a prior distribution, an inverse chi-squared distribution:

$$\sigma_j^2 \sim \frac{s_0^2 d_0}{\chi_{d_0}^2}, \tag{2}$$

where s_0 is the pooled variance estimate and d_0 is the prior degrees of freedom.

The expectation of the probe's variance σ_j given the observed variance s_j is

$$s_{j,post}^2 = \mathbb{E}[\sigma_j^2 | s_j^2] = \frac{d_0 s_0^2 + d_j s_j^2}{d_0 + d_j}, \tag{3}$$

where d_j is the degrees of freedom for the probe. Using the posterior variance $s_{j,post}^2$, we can obtain the moderated t-statistics for each probe

$$t_j = \frac{\hat{\beta}_j}{s_{j,post}\sqrt{\frac{1}{p^\top p}}}. \tag{4}$$

Purity estimation with signal-to-noise (SNR) weighting. Given new samples with methylation measurements $X' \in \mathbb{R}^{n' \times m'}$, where n' is the number of samples and m' is the number of features, MONTE estimates tumor purity using the intersection of probes between the trained model and the input data, allowing pre-trained models to be appleid across datasets with partially overlapping probe sets. After centering X' using training set means, sample-wise purity estimates can be solved using a weighted least squares objective function

$$\hat{p} = \arg\min_{p\in\mathbb{R}} \sum_{j\in M_\cap} w_j (X'_{ij} - \hat{\beta}_j p - \hat{\alpha}_j)^2, \tag{5}$$

where $M_\cap$ denotes the intersecting probe set.

MONTE uses signal-to-noise (SNR) [22] based weights

$$w_j = \frac{\hat{\beta}_j^2}{s_{j,post}}, \tag{6}$$

which upweight probes with strong purity association and low residual variance while downweighting noisy or weakly associated probes. This yields the closed-form purity estimator

$$\hat{p}_i = \frac{\sum_{j\in M_\cap} w_j \hat{\beta}_j (X'_{ij} - \hat{\alpha_j})}{\sum_{j\in M_\cap} w_j \hat{\beta}_j}. \tag{7}$$

To reduce the influence of probes weakly associated with purity, MONTE optionally restricts purity estimation to the top n probes ranked by their moderated t-statistics. The number of probes can be specified by the user or selected via a built-in cross validation function. This built-in function automatically runs k-fold cross validation (default: k=5) on the training set across a set of n values (default: 5, 10, 25, 50, 100, 250, 500, 1000, 2500, 5000, 10000, 25000, 50000, 100000, 150000) and returns the n that corresponds to the lowest average mean squared error across folds.

Bayesian transfer learning. MONTE supports adaptation to new datasets through Bayesian transfer learning, which updates probe-level purity coefficients using limited samples with known purity. We assume a Gaussian distribution on $\beta_j^{(t)}$ for probe j in the new dataset, $\beta_j^{(t)} \sim N(\hat{\beta}_j^{prior}, \tau^2)$, where τ^2 captures the uncertainty of the pan-cancer coefficient estimate for probe j.

Given a target dataset with labeled samples, probe-level coefficients $\hat{\beta}_j^{(t)}$ and associated variance $\sigma_j^{(t)2}$ are estimated using the same linear model. The posterior distribution of $\beta_j^{(t)}$ thus has the closed-form mean

$$\mu_j^{post} = \frac{\hat{\beta}_j^{prior}/\tau^2 + \sum_i X_{ij}^{(t)} p_i^{(t)}/\sigma^{(t)2}}{1/\tau^2 + \sum_i p_i^{(t)2}/\sigma^{(t)2}} \tag{8}$$

and variance,

$$v_j^{post} = \left(\frac{1}{\tau^2} + \frac{1}{\sigma^{(t)2}} \sum_i p_i^2 \right)^{-1}. \tag{9}$$

This update corresponds to a precision-weighted combination of the pan-cancer prior and the target-specific estimate, allowing probe-purity relationships to be recalibrated efficiently without retraining the model from scratch.

Purification of methylation values. MONTE enables CpG-resolved purification of bulk DNA methylation profiles by adjusting observed methylation values to a target tumor purity. Given estimated sample purity $\hat{p}$ and probe-level coefficients $\hat{\beta}_j$, methylation values for sample i and probe j are adjusted as

$$X'_{ij\,adjusted} = X'_{ij} + \hat{\beta}_j(p'_i - \hat{p}_i), \tag{10}$$

where p' denotes the target purity, typically set to 1 to obtain a pure tumor methylation profile. This adjustment corresponds to projecting each probe's methylation value along its estimated purity-associated direction to the desired purity level. The operation is applied efficiently in matrix form across all probes and samples. Because not all probes are necessarily strongly associated with tumor purity, MONTE provides optional controls to restrict purification to informative probes based on probe-level significance thresholds. MONTE also provides easy access to detailed probe statistics that include confidence intervals, allowing users to assess the reliability of individual adjustments. These optional adjustments were not used for the results shown in the paper but allow purification to be tailored to the desired balance between sensitivity and robustness.

Differential methylation, effect size, and subset concordance.

Differential methylation (DM) was performed using the *methylize* [23] package in Python (v3.13.3). Significant probes were defined using FDR-adjusted p-values (< 0.05). CpGs were mapped to promoter regions (±1.5 kb from transcription start sites) using the hg38 manifest [20].

For each CpG probe, tumor–normal effect sizes were obtained from the linear DM coefficient. Absolute effect size changes were calculated as the difference between purified and unadjusted coefficients. Cancer marker genes were obtained from MethMarkerDB [18]. Marker-associated CpGs were defined as probes mapped to promoters of curated marker genes. To assess whether marker CpGs exhibited larger effect size increases than compared to those of random non-marker sets, a permutation test was performed by randomly sampling non-marker probes 5,000 iterations.

To assess the consistency of DM results between purified and unadjusted methylation, test tumor samples were randomly split into two halves five times for each cancer type. For each split, DM was performed against normal samples using *methylize* [23]. Only cancer types with at least five tumor and five normal samples were included. Concordance between splits was quantified using Spearman correlation of effect sizes. Median correlation across iterations was compared between unadjusted and purified data.

Existing purity estimation and purification methods

We compared MONTE against three existing DNA methylation array–based tumor purity estimation tools and their corresponding beta value purification procedures: PAMES [7], InfiniumPurify [5], and PureBeta [6]. All methods were run using pretrained and dataset-fitting modes following recommended settings whenever possible. We note that all methods include TCGA data in their training, so the pretrained models inevitably have the advantage of being trained on some of the testing samples.

PAMES was evaluated using the provided informative probe collections, which supports 12 cancer types (BLCA, BRCA, COAD, HNSC, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, THCA, UCEC). PureBeta under both pretrained and dataset-fitted settings. Pretrained models used the precomputed regression parameters provided for BRCA, LUAD, and LUSC. Dataset-fitting was also performed for these same cancer types, which the authors have described as satisfying the method's modeling assumptions. For InfiniumPurify, pretrained analyses were restricted to tumor types supported by the package so READ was missing, but as suggested in the method documentation, we used the COAD model for READ predictions. In the dataset-fitting setting, model fitting required at least 20 tumor and 20 normal samples, which limited applicability in rarer cancer types with small sample sizes. Beta value purification also required at least 20 normal samples, restricting the number of cancer types that we could compare against.

We additionally attempted to evaluate RF_purify [9] and MEpurity [8]; however, RF_purify could not handle missing values introduced during standard quality control, and MEpurity produced empty outputs when run according to the provided documentation. These methods were therefore excluded from further analysis.

Data and code availability

Our implementation of MONTE package is available on GitHub at `https://github.com/ylaboratory/MONTE`, with additional analysis code and tutorials available at `https://github.com/ylaboratory/MONTE-analysis`, both released under the BSD 3-clause license for open source use.

Declaration of interests

The authors declare no competing interests.

Acknowledgments

The authors would like to thank members of the ylaboratory for helpful discussions. This work was supported by the Cancer Prevention & Research Institute of Texas [CPRIT RR190065 to VY] and the National Science Foundation [NSF DBI-2144534 to VY]. VY is a CPRIT Scholar in Cancer Research.

References

1. Geissler, F. *et al.* The role of aberrant DNA methylation in cancer initiation and clinical impacts. en. *Therapeutic Advances in Medical Oncology* **16,** 17588359231220511. ISSN: 1758-8359, 1758-8359. https://journals.sagepub.com/doi/10.1177/17588359231220511 (2026) (Jan. 2024).
2. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. en. *Nature Communications* **6,** 8971. ISSN: 2041-1723. https://www.nature.com/articles/ncomms9971 (2025) (Dec. 2015).
3. Arneson, D., Yang, X. & Wang, K. MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. en. *Communications Biology* **3,** 422. ISSN: 2399-3642. https://www.nature.com/articles/s42003-020-01146-2 (2025) (Aug. 2020).
4. Wu, Z. *et al.* Impact of the methylation classifier and ancillary methods on CNS tumor diagnostics. en. *Neuro-Oncology* **24,** 571–581. ISSN: 1522-8517, 1523-5866. https://academic.oup.com/neuro-oncology/article/24/4/571/6374546 (2025) (Apr. 2022).
5. Qin, Y., Feng, H., Chen, M., Wu, H. & Zheng, X. InfiniumPurify: An R package for estimating and accounting for tumor purity in cancer methylation research. *Genes & Diseases* **5,** 43–45. ISSN: 2352-4820. https://pmc.ncbi.nlm.nih.gov/articles/PMC6147081/ (2025) (Feb. 2018).
6. Sasiain, I., Nacer, D. F., Aine, M., Veerla, S. & Staaf, J. Tumor purity estimated from bulk DNA methylation can be used for adjusting beta values of individual samples to better reflect tumor biology. en. *NAR Genomics and Bioinformatics* **6,** lqae146. ISSN: 2631-9268. https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqae146/7874836 (2025) (Sept. 2024).
7. Benelli, M., Romagnoli, D. & Demichelis, F. Tumor purity quantification by clonal DNA methylation signatures. *Bioinformatics* **34,** 1642–1649. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/bty011 (2025) (May 2018).
8. Liu, B. *et al.* MEpurity: estimating tumor purity using DNA methylation data. en. *Bioinformatics* **35** (ed Schwartz, R.) 5298–5300. ISSN: 1367-4803, 1367-4811. https://academic.oup.com/bioinformatics/article/35/24/5298/5531131 (2025) (Dec. 2019).
9. Johann, P. D., Jäger, N., Pfister, S. M. & Sill, M. RF_Purify: a novel tool for comprehensive analysis of tumor-purity in methylation array data based on random forest regression. en. *BMC Bioinformatics* **20,** 428. ISSN: 1471-2105. https://doi.org/10.1186/s12859-019-3014-z (2025) (Aug. 2019).
10. Zhang, W., Feng, H., Wu, H. & Zheng, X. Accounting for tumor purity improves cancer subtype classification from DNA methylation data. en. *Bioinformatics* **33,** 2651–2657. ISSN: 1367-4803, 1367-4811. https://academic.oup.com/bioinformatics/article/33/17/2651/3796398 (2025) (Sept. 2017).
11. Haider, S. *et al.* Systematic Assessment of Tumor Purity and Its Clinical Implications. en. *JCO Precision Oncology*, 995–1005. ISSN: 2473-4284. https://ascopubs.org/doi/10.1200/PO.20.00016 (2025) (Nov. 2020).
12. Liang, W.-W. *et al.* Integrative multi-omic cancer profiling reveals DNA methylation patterns associated with therapeutic vulnerability and cell-of-origin. en. *Cancer Cell* **41,** 1567–1585.e7. ISSN: 15356108. https://linkinghub.elsevier.com/retrieve/pii/S1535610823002532 (2025) (Sept. 2023).
13. Petralia, F. *et al.* A new method for constructing tumor specific gene co-expression networks based on samples with tumor purity heterogeneity. en. *Bioinformatics* **34,** i528–i536. ISSN: 1367-4803, 1367-4811. https://academic.oup.com/bioinformatics/article/34/13/i528/5045768 (2025) (July 2018).
14. The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. en. *Nature Genetics* **45,** 1113–1120. ISSN: 1061-4036, 1546-1718. https://www.nature.com/articles/ng.2764 (2025) (Oct. 2013).
15. The International Cancer Genome Consortium. International network of cancer genome projects. en. *Nature* **464,** 993–998. ISSN: 0028-0836, 1476-4687. https://www.nature.com/articles/nature08987 (2025) (Apr. 2010).

16. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium *et al.* Pan-cancer analysis of whole genomes. en. *Nature* **578,** 82–93. ISSN: 0028-0836, 1476-4687. https://www.nature.com/articles/s41586-020-1969-6 (2025) (Feb. 2020).
17. Thorsson, V. *et al.* The Immune Landscape of Cancer. en. *Immunity* **48,** 812–830.e14. ISSN: 10747613. https://linkinghub.elsevier.com/retrieve/pii/S1074761318301213 (2026) (Apr. 2018).
18. Zhu, Z. *et al.* MethMarkerDB: a comprehensive cancer DNA methylation biomarker database. en. *Nucleic Acids Research* **52,** D1380–D1392. ISSN: 0305-1048, 1362-4962. https://academic.oup.com/nar/article/52/D1/D1380/7331013 (2026) (Jan. 2024).
19. Silva, T. C. *et al.* TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. en. *F1000Research* **5,** 1542. ISSN: 2046-1402. https://f1000research.com/articles/5-1542/v2 (2025) (Dec. 2016).
20. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. en. *Nucleic Acids Research,* gkw967. ISSN: 0305-1048, 1362-4962. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw967 (2025) (Oct. 2016).
21. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. en. *Nucleic Acids Research* **43,** e47–e47. ISSN: 1362-4962, 0305-1048. http://academic.oup.com/nar/article/43/7/e47/2414268/limma-powers-differential-expression-analyses-for (2025) (Apr. 2015).
22. Gu, Q., Li, Z. & Han, J. *Generalized Fisher Score for Feature Selection* arXiv:1202.3725 [cs]. Feb. 2012. http://arxiv.org/abs/1202.3725 (2025).
23. FOXO Bioscience. *methylize* https://life-epigenetics-methylprep.readthedocs-hosted.com/en/latest/index.html.