Briefings in Bioinformatics, 2026, 27, bbag261
https://doi.org/10.1093/bib/bbag261
Published: 31 May 2026
Problem Solving Protocol

OXFORD

USADAE: a deep learning approach to disentangle hidden covariates in RNA-seq data

Xu Chen, Luoyuan Guo, Yaosheng Chen, Delin Mo*, Xiaohong Liu*

State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, No. 135, Xingang West Road, Haizhou District, Guangzhou, Guangdong 510275, China

*Corresponding authors. Delin Mo, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, No. 135, Xingang West Road, Haizhou District, Guangzhou, Guangdong 510275, China. E-mail: modelin@mail.sysu.edu.cn; Xiaohong Liu, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, No. 135, Xingang West Road, Haizhou District, Guangzhou, Guangdong 510275, China. E-mail: liuxh8@mail.sysu.edu.cn

Abstract

Integrative analysis of RNA-seq datasets faces critical challenges in disentangling biologically meaningful signals from implicit confounders, such as unmeasured technical variability and nonlinear interactions between biological and technical variables. Traditional methods like Surrogate Variable Analysis (SVA), Probabilistic Estimation of Expression Residuals (PEER), and Remove Unwanted Variation (RUVSeq) rely on linear assumptions, which generally fail under complex nonlinear confounding patterns. Although deep learning approaches show promise in single-cell RNA-seq, they primarily address known batch labels rather than disentangling hidden confounders. Here, we develop a novel autoencoder framework coupled with adversarial learning, UnSupervised Adversarial Deconfounding AutoEncoder (USADAE), specifically designed to separate confounders from biological signals. The model encodes RNA-seq data into distinct biological and confounder latent spaces through adversarial disentanglement, enabling downstream correction of differential expression and eQTL analysis. In comprehensive simulations, USADAE significantly outperforms existing methods in extracting covariates while preserving biological signals. Real-data applications further demonstrate its robustness across diverse scenarios, including cancer genomics and eQTL studies.

Keywords autoencoder, adversarial learning, confounder disentanglement, RNA-seq

Introduction

In recent years, numerous RNA-seq datasets have been generated from different experimental conditions or laboratories [1–4]. However, the utility of integrative analysis depends on separating true biological signals from confounders. For instance, batch effects arising from technical variability in sequencing depth or library preparation can mimic true differential expression, thereby significantly impacting downstream analysis and leading to spurious associations and false positives [5–7]. Similarly, in eQTL (expression quantitative trait loci) studies, data affected by batch effects or unaccounted technical variation may result in spurious associations [8]. In some cases, explicit covariates can be obtained by recording known variables. However, implicit covariates that affect gene expression are often unmeasured or inherently unobservable for various reasons. Therefore, disentangling these implicit covariates from observed signals is important for subsequent differential expression analysis (DEA) or eQTL analysis.

Traditional confounder correction strategies in DEA are divided into supervised and unsupervised methods [9]. Supervised approaches, like Combat and Combat-seq, rely on known covariates [7, 10]. Unsupervised methods include Surrogate Variable Analysis (SVA), which assumes linear confounder effects, and Remove Unwanted Variation (RUV) methods like RUVG and RUVR [11]. Probabilistic Estimation of Expression Residuals (PEER), commonly used in eQTL studies [12–14], models gene expression as a combination of known covariates (biological signals) and latent factors (hidden covariates), estimated via a Bayesian framework [8]. Although these traditional methods have shown significant potential in disentangling unmeasured covariates, they generally assume that observed variables are linear combination of true biological signals and hidden covariates, less effective for nonlinear interactions common in real-word datasets [15, 16].

Recent studies demonstrated the capacity of deep learning to extract meaningful representations from RNA-seq data, especially in single-cell RNA-seq (scRNA-seq). Autoencoder-based frameworks like BERMUDA [17], scVI [18, 19], DESC [20], scANVI [21], and Fugue [22] employ reconstruction losses to learn low-dimensional embeddings that preserve cellular identity while correcting batch effects. deepMNN [23] abandons autoencoders for a residual network, explicitly using cross-batch PCA-space MNN pairs to guide correction.

Received: August 28, 2025. **Revised:** March 22, 2026. **Accepted:** April 26, 2026

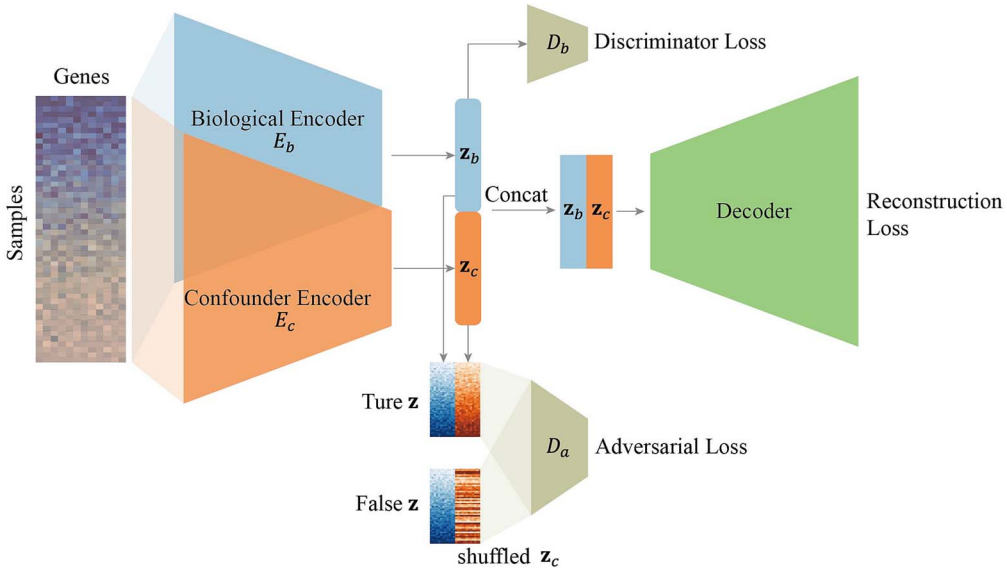

Figure 1 Sketch of USADAE framework. USADAE employs dual encoders to decompose RNA-seq data into biologically relevant ($\mathbf{z}_b$) and confounder-associated ($\mathbf{z}_c$) latent representations, concatenated for decoder-driven reconstruction. The autoencoder operates along the gene-expression feature dimension, reducing the dimensionality of expression profiles while preserving sample-wise relationships. An adversarial discriminator (D_a) enforces statistical independence between $\mathbf{z}_b$ and $\mathbf{z}_c$ by distinguishing true $[\mathbf{z}_b, \mathbf{z}_c]$ from shuffled pairs, while a weak discriminator (D_b) guides the biological encoder to retain label-discriminative features (e.g. case/control). Through adversarial optimization, USADAE explicitly isolates hidden covariates from observed signals.

AD-AE [24] and cr-XVAE [16], designed for bulk RNA-seq, leverages known covariates information to remove confounders. While these methods are powerful, they are primarily applied to scRNA-seq data for tasks such as cell clustering in latent space or batch effect correction when batch labels are known. However, they often fail to explicitly disentangle hidden covariates, which could improve downstream analysis when covariates are unknown. Therefore, there is a need for a novel method capable of identifying complex implicit covariates for subsequent analysis. To address this problem, we introduced a deep-learning-based model named UnSupervised Adversarial Deconfounding AutoEncoder (USADAE), designed to disentangle biological signal and hidden covariates from RNA-seq data. USADAE employs two autoencoders to encode observed signals into distinct biological and confounder latent representations. Subsequently, an adversarial network is utilized to disentangle these biological and confounder latent spaces. Then, the extracted hidden covariates can be utilized in subsequent DEA or eQTL analysis to correct spurious associations. USADAE demonstrated better performance compared to competing methods in simulation studies and real data analysis across diverse downstream applications.

Results

Framework of USADAE algorithm

USADAE is a model leveraging a deep learning strategy for the explicit disentanglement of hidden covariates from RNA-seq data. The model employs two encoders to transform the gene expression profile into distinct low-dimensional biological ($\mathbf{z}_b$) and hidden confounder ($\mathbf{z}_c$) latent representations. These two latent representations are subsequently concatenated as $[\mathbf{z}_b, \mathbf{z}_c]$ and fed as input to a decoder (Fig. 1).

To achieve the disentanglement of the biological and confounder latent spaces, an adversarial learning framework is utilized. Specifically, a discriminator, denoted as D_a, is introduced. This discriminator is trained to distinguish between "true" data, formed by concatenating the original biological and confounder latent ($[\mathbf{z}_b, \mathbf{z}_c]$) and "false" data generated by concatenating the biological latent with a shuffled version of the confounder latent ($[\mathbf{z}_b, \text{shuffled}(\mathbf{z}_c)]$). The encoders are simultaneously optimized to deceive D_a, thereby ensuring the statistical independence of the biological and confounder latent spaces.

Furthermore, to encourage the biological encoder to exclusively capture biological signals rather than confounder information, a weak discriminator, D_b, is additionally incorporated. This discriminator is designed to distinguish biological labels (e.g. control versus case groups) as effectively as possible.

Through the synergistic introduction of D_a and D_b discriminators and the application of adversarial learning, USADAE can explicitly disentangle biological and confounding signals within the latent space.

Model evaluation on simulated differential expression data

To evaluate our model's ability to disentangle covariates, we simulated an RNA-seq dataset, with strong separation patterns driven by batch, sex, and age (Fig. 2a).

The USADAE's covariates (Fig. 2b) maintained alignment with true covariates without exhibiting group-level separation. In the biological latent space, the separation related to true covariates was no longer apparent, while the separation between case and control groups became more distinct (Fig. 2c). The data corrected by USADAE show the random separation by covariates, while retaining the separation pattern similar to the pure biological signal before adding confounders (Supplementary Fig. S1).

Specific confounder dimensions (e.g. Conf1, Conf2, Conf3) strongly correlated with known covariates (Fig. 2d). To further assess the effectiveness of USADAE's covariates, classification or regression models were developed using them. Predictive models achieved near-perfect for categorical covariates (Fig. 2e) and accurately reconstructed the continuous covariate age ($R^2 = 0.87 \pm 0.01$), significantly

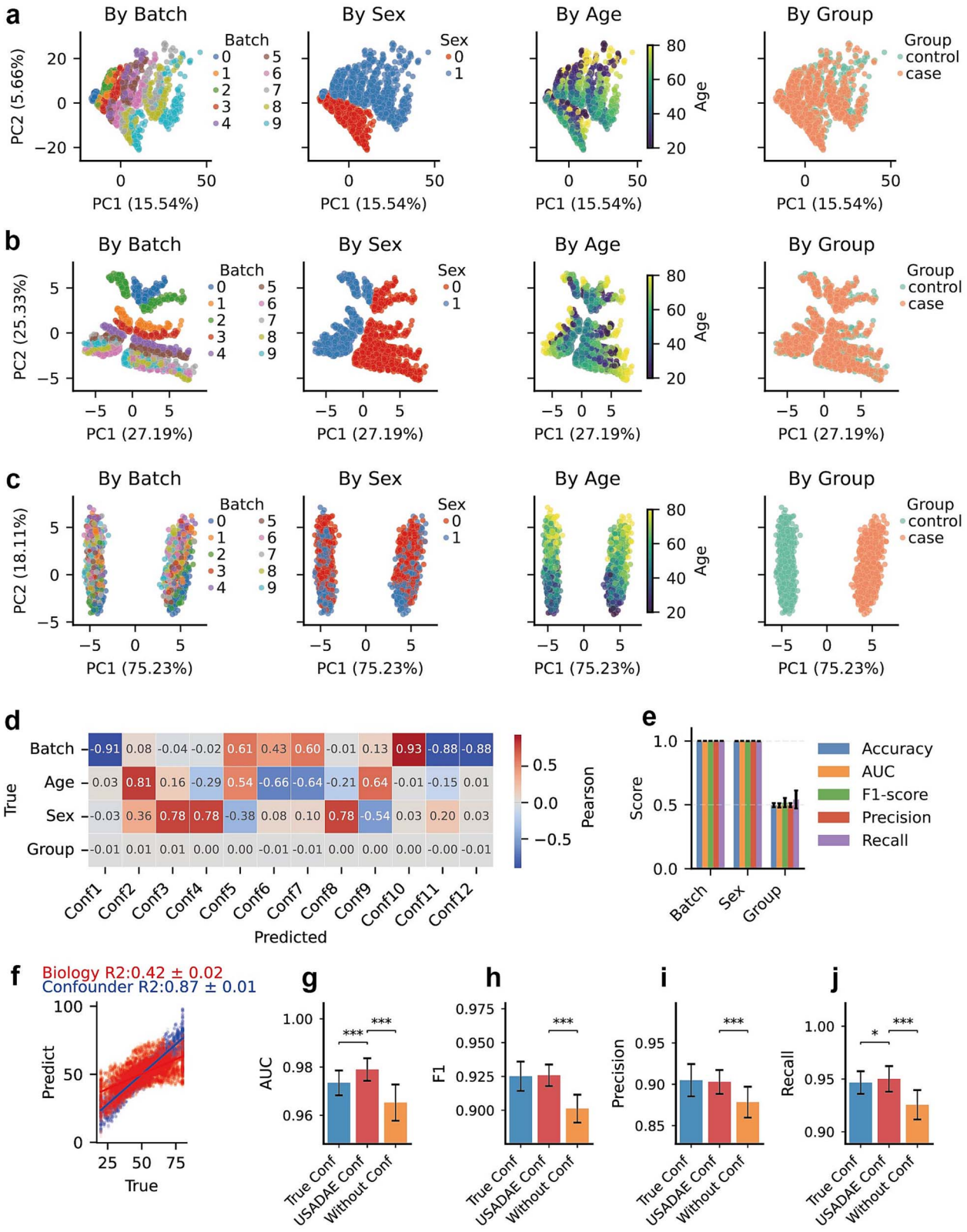

Figure 2 Performance evaluation of USADAE on simulated dataset. (a) PCA plot colored by batch, age, sex, and group, with PC1 and PC2 variance explained. (b, c) PCA plots of confounder and biological latent spaces respectively. (d) Correlation heatmap between USADAE's covariates and true covariates and group. (e) Predict performance for batch, sex, and group using USADAE's covariates across 5-fold cross-validation. (f) Scatter plot compares true and predicted age from confounder (blue) and biological latent spaces (red). (g–j) Evaluation of DEGs detection by different metrics. Error bars represent the standard deviation across 10 independent analyses; significance markers (paired t-tests and Benjamini–Hochberg adjusted, *adjusted $P < .05$, **adjusted $P < .01$, ***adjusted $P < .001$) indicate statistical differences between USADAE conf and other groups. AUC is based on full ROC curves computed across all *P*-value thresholds, while others use the threshold in Methods.

outperforming the biological latent space ($R^2 = 0.42 \pm 0.02$) (Fig. 2f). Moreover, the confounder space showed random-level prediction for the biological group, indicating minimal information leakage.

We performed 10 independent simulations using distinct random seeds. DEA incorporating USADAE's covariates (USADAE Conf) yields significantly higher area under the curve (AUC), F1-score, precision, and recall than model without covariates (Fig. 2g–j). Notably, USADAE improved AUC over true covariates by capturing nonlinear effect that may distort *P*-value rankings, while others remained unchanged at fixed thresholds where true confounders already perform adequately. These findings indicate USADAE's enhanced DEA accuracy through confounder separation.

Comparative performance of USADAE in confounder separation and differential expression analyses

To evaluate how the USADAE algorithm performs relative to RUVR and SVA, a comparative analysis using simulated data was conducted. PCA plot revealed that the covariates estimated by RUVR (k = 12) and SVA exhibited weaker stratification patterns compared to those extracted by USADAE, in which latent embedding dimension was set to 12 (Figs 2b, 3a, and b). Then, the effectiveness of confounder separation was evaluated by prediction score obtained from model constructed by estimated covariates.

USADAE and USADAE-linear show slightly higher performance than RUVR and SVA in batch and sex prediction (Fig. 3c–e). Here, USADAE-linear refers to a variant of USADAE in which activation functions are removed from the decoder, allowing us to assess whether USADAE effectively learns nonlinear confounding structure. For biological group, all performed at chance level, with USADAE showing lower variance. For age, USADAE's covariates yielded $R^2 = 0.871 \pm 0.01$, surpassing RUVR (0.713 ± 0.02) and SVA (0.692 ± 0.02) (Fig. 3f, Supplementary Table S1), indicating more stable confounder capture.

To evaluate the significance, 10 independent analyses were conducted using the simulated data described above. USADAE achieved significantly higher score than others (Fig. 3g–j, Supplementary Fig. S1F). These results show the robustness of USADAE in isolating covariates while preserving biological signals.

To evaluate the computational efficiency, we compared memory usage and runtime on simulated datasets with varying sample sizes. USADAE demonstrated superior computational efficiency, with substantially shorter runtime and lower memory usage than RUVR and SVA, particularly at larger sample sizes (Supplementary Fig. S2A and B).

Validation of USADAE performance on real breast cancer datasets

To validate the performance of the USADAE on real-world data, we applied it to breast cancer datasets (GSE2034 [25], GSE7390 [26], GSE20194 [27, 28], GSE25066 [29], GSE62725 [30], and GSE20271 [31]), which exhibit diverse technical and biological variability.

T-SNE projections of gene expression data showed distinct dataset-specific clusters, with no clear separation by estrogen receptor (ER) status (Fig. 4a and b, left). For USADAE's covariates, dataset-specific clustering persisted, capturing technical confounders without encoding ER status (Fig. 4a and b, middle). Conversely, the biological latent space reduced dataset-specific clustering and revealed clear ER+ versus ER- separation (Fig. 4a and b, right), demonstrating USADAE's ability to suppress confounders while amplifying biological signals.

To evaluate the accuracy of different algorithms in separating covariates, logistic regression and random forest classification models were constructed using covariates estimated by various methods for dataset, grade, and race, while linear and random forest regression models were developed for age and tumor size (Supplementary Table S2). USADAE outperformed RUVR and SVA in predicting grade and race, showed comparable performance for dataset, age, and tumor size (Fig. 4c–g), but had lower performance for ER status, indicating minimal biological information in its confounder estimates (Fig. 4h).

Since true DEGs are unknown in real data, the performance of confounder separation algorithms was evaluated by examining the relevance of enriched pathways to known biological processes. DEA incorporating covariates from USADAE, RUVR, and SVA was conducted, with adjusted thresholds to ensure comparable DEG counts ($FDR < 0.05$, $|FC| > 1.2$ for USADAE; $FC > 1.5$ or < -1.2 for others) (Supplementary Fig. S3A–D). Notably, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis revealed that RUVR, SVA, and "no covariate" methods enriched irrelevant pathways (e.g. "Coronavirus disease - COVID-19"), while USADAE avoided such enrichments and significantly enriched breast cancer-relevant pathways like IL-17 signaling [32] (Supplementary Fig. S3E and F). To further evaluate the biological relevance and specificity of USADAE, enrichment analysis was conducted using the top up-regulated or down-regulated 50–500 DEGs derived from ER+/ER- breast cancer-related pathways (estrogen signaling [33, 34], IL-17 signaling, cytokine–cytokine receptor interaction [35, 36]) and unrelated pathways (e.g. Parkinson's disease, COVID-19). USADAE consistently yielded lower *P*-values for breast cancer pathways, particularly for upregulated genes in estrogen signaling (ER+ specific) and downregulated genes in IL-17 signaling and cytokine interactions (ER- specific) (Fig. 4i–k, Supplementary Fig. S4). For unrelated pathways, USADAE showed higher *P*-values, indicating greater specificity by avoiding spurious enrichments (Fig. 4l–n). To evaluate whether USADAE captures nonlinear confounding structure beyond linear model, we compared the biological embeddings of USADAE and USADAE-linear. Under PCA, a linear dimensionality reduction method, both models separated samples by ER status but not by dataset (Supplementary Fig. S5A and B, left and middle). Under t-SNE, a nonlinear embedding method, USADAE-linear exhibited dataset-specific clustering similar to SVA and RUVR (Supplementary Fig. S5A–D), a pattern absent in USADAE (Fig. 4a and b, right). Accordingly, differential expression results with USADAE-linear covariates resembled those from linear methods (Supplementary Fig. S5E and F).

Application of USADAE in eQTL analysis with simulated gene expression data

To evaluate the utility of the USADAE algorithm in eQTL analysis, we simulated gene expression data across different tissues using real genotype data. USADAE was applied to extract covariates for eQTL analysis, compared to PEER, PCA [37], and Control (covariates = 1).

PCA plot showed that samples clustered by batch, sex, and age, but not tissue (Fig. 5a). In USADAE's confounder latent space, clustering by true covariates remained evident (Fig. 5b). In the biological latent space and corrected data, confounder clustering disappeared, revealing distinct tissue-type clustering aligned with

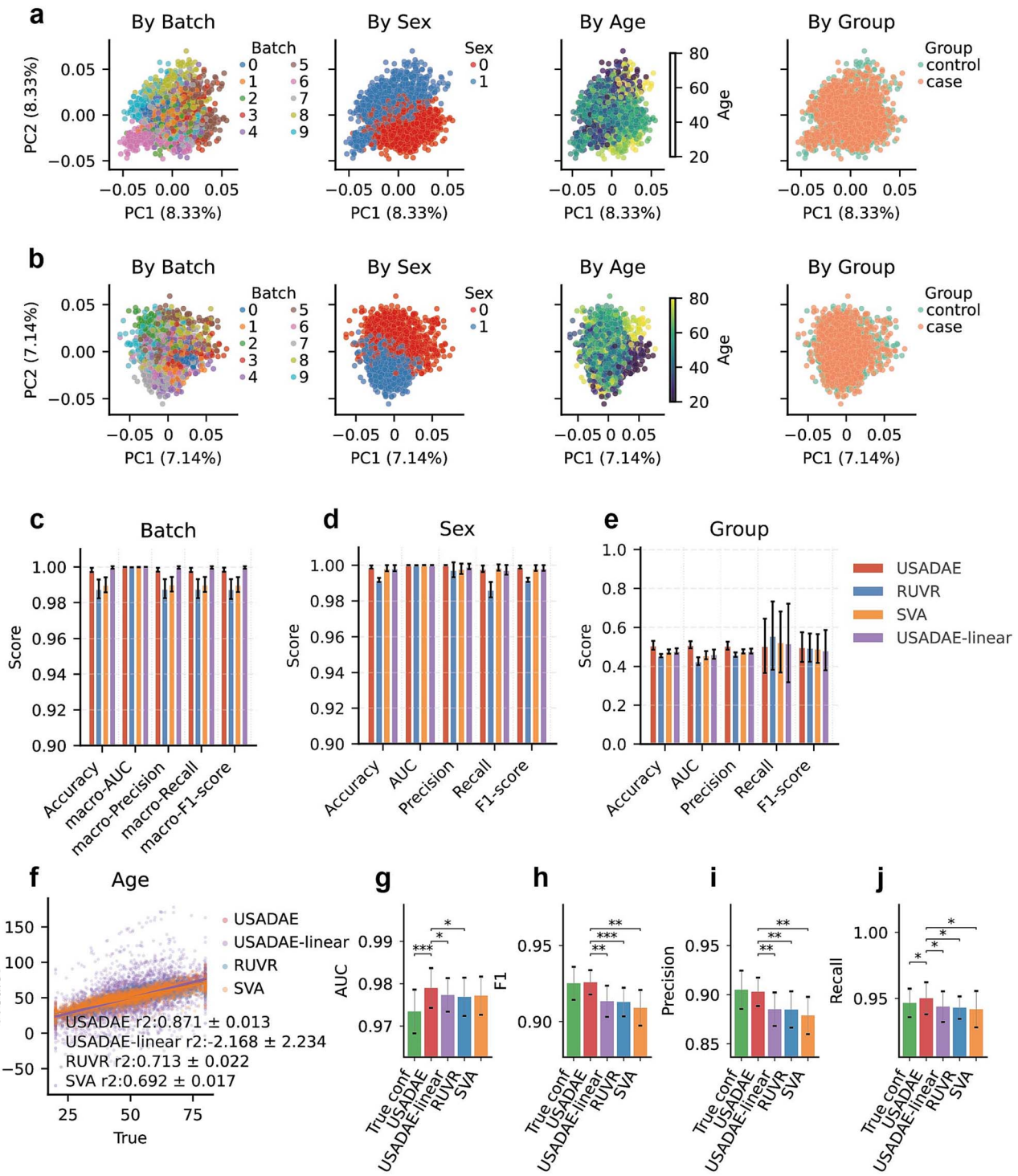

Figure 3 Performance comparison of USADAE, USADAE-linear, RUVR, and SVA for DEA on simulated dataset. (a, b) PCA plot of covariates estimated by RUVR and SVA, respectively. (c–e) Classification performance of logistic regression models using covariates from USADAE, RUVR, and SVA, averaged over 5-fold cross-validation with standard errors. (f) Mean R^2 comparison of linear regression models using covariates from USADAE, USADAE-linear, RUVR, and SVA, averaged over 5-folds with standard errors. (g–j) Metrics comparison for DEA using covariates from USADAE, USADAE-linear, RUVR, SVA, and True Conf.

intrinsic biological signals (Fig. 5c, Supplementary Fig. S6A and B). In PEER's covariates, sex and age clustering remained, but batch effects were less evident (Supplementary Fig. S6C and D). For confounder prediction, USADAE's covariates outperformed PEER in predicting batch and sex, with comparable age prediction and tissue prediction (Fig. 5d–g). These indicate that the confounder latent space successfully concentrates non-biological variation, which can be used as covariates in downstream eQTL analysis. Across 10 independent

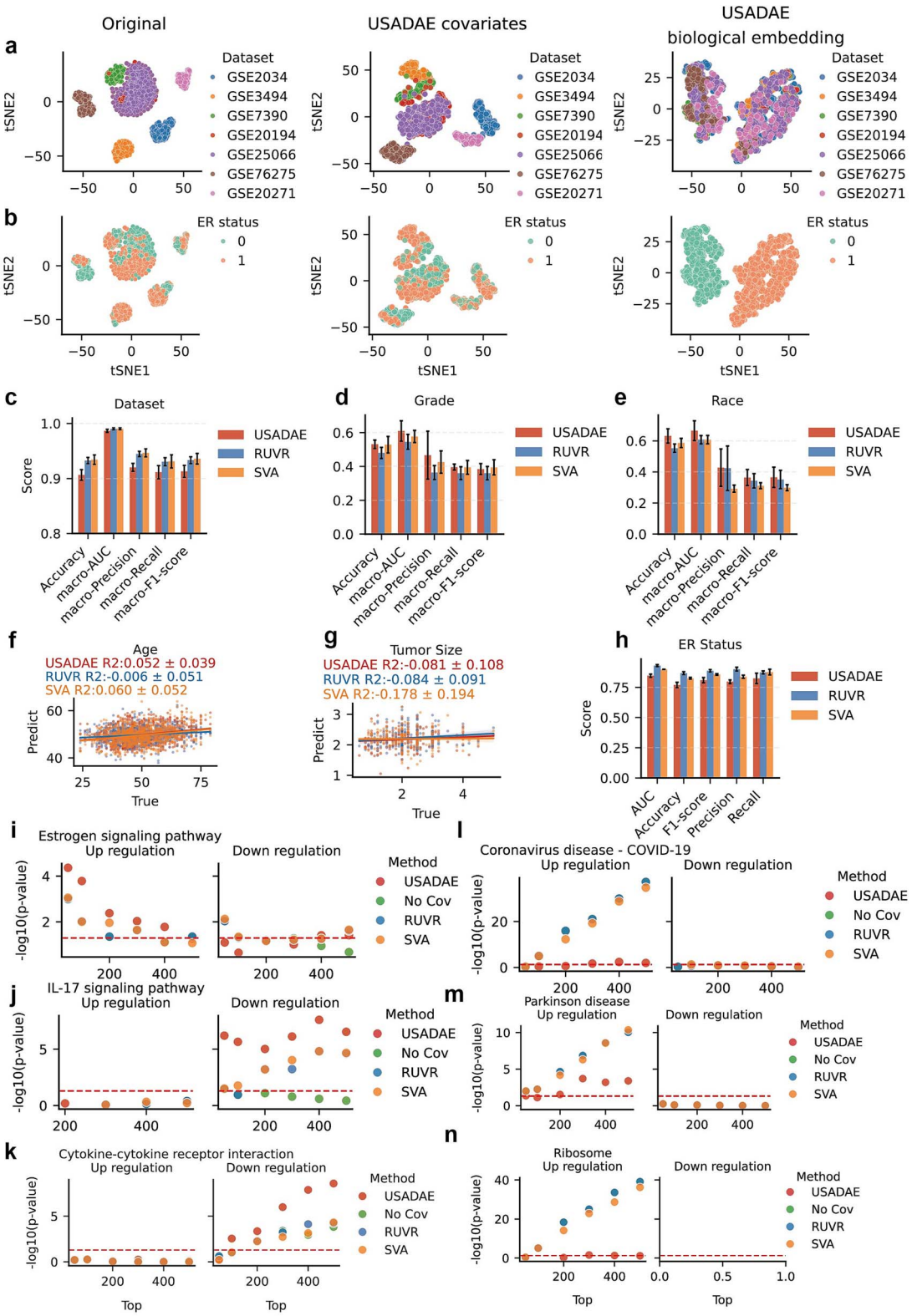

Figure 4 Performance comparison of USADAE, RUVR, and SVA on breast cancer datasets. (a) t-SNE plot colored by dataset. From left to right, the panels display: (1) the original data, (2) USADAE's covariates, and (3) USADAE's biological latent. (b) t-SNE plot, similar to (a), colored by ER status. (c–e) Mean classification performance (dataset, grade, race) of random forest models using covariates from USADAE, RUVR, SVA, averaged over 5-fold cross-validation, with standard error bars. (f, g) Scatter plots of predicted and true covariates, showing mean R^2 with standard errors. (h) Bar charts comparing random forest performance for ER status. (i–n) Scatter plots of -log10 (*P*-value) for pathway enrichment of top differentially expressed genes, with (i–k) ER-related and (l–n) non-ER-related pathways.

eQTL analyses, USADAE demonstrated similar performance compared to PEER, while showing significantly higher recall and F1-score than the PCA and Control at low h^2 (Fig. 5h–j). On the simulated dataset, USADAE showed significantly shorter cumulative runtime than PEER across three tissues, while PEER required less memory for single-tissue analysis (Supplementary Fig. S2C–E).

Application of USADAE in eQTL analysis with PigGTEx datasets

To assess the performance of the USADAE algorithm in real-world eQTL analysis, it was applied to PigGTEx datasets, focusing on evaluating confounder removal, stability across data splits, and *cis*-heritability (cis-h^2) of eGenes, compared to PEER (optimal factor number = 10), PCA (optimal PC number = 18) and Control methods.

The t-SNE plot of the original data revealed partial clustering by tissues, with testis, ovary, embryo, and morula proving challenging to distinguish (Fig. 6a, left), while identical tissues (e.g. muscle) from different BioProject (PRJNA486202, PRJNA488311, PRJNA511588) clustered together according to their BioProject origins (Fig. 6b, left). In USADAE's confounder latent space, BioProject clustering improved, while tissue clustering became more random (Fig. 6a and b, middle). In the biological latent space, BioProject clustering diminished (e.g. muscle), and tissue separation increased (Fig. 6a and b, right).

DEA was conducted between muscle and adipose, as well as between muscle and heart (cardiac muscle), with covariates estimated by different algorithms. KEGG pathway analysis from muscle versus adipose showed consistent enrichment across methods (Supplementary Fig. S7A–C), likely due to significant biological differences outweighing technical confounders (Fig. 6a). For the more similar muscle versus heart comparison, USADAE reduced false-positive findings. Notably, pathways unrelated to muscle or heart biology—such as pancreatic secretion and bile secretion—were substantially suppressed (Supplementary Fig. S7D–F).

To evaluate the effectiveness of the extracted covariates, classification models were constructed for categorical covariates including sex, library layout, bio-project, breed group, breed, and sequencing model, while regression models were developed for continuous covariates such as age, clean reads, mapped reads, and mapping rate, consistent with the methodology described in the simulated data results. Classification models using USADAE's covariates outperformed PEER for categorical variables, and regression models showed greater stability for continuous variables (Supplementary Table S3).

To evaluate the consistency of eGene detection, a comparative analysis was performed in muscle tissue between eGenes identified using USADAE's covariates and PEER's covariates as covariates. USADAE detected 9779 eGenes in muscle, of which 8555 overlapped with PEER (9838 eGenes). USADAE-specific eGenes enriched more muscle-related pathways (e.g. calcium signaling), unlike PEER's less relevant pathways (Supplementary Fig. S8). Additionally, cis-h^2 of eGenes across chromosomes were calculated to assess the proportion of expression variance explained by local genetic variants after confounder correction. USADAE showed similar cis-h^2 values compared to PEER, but significantly lower cis-h^2 values than Control (Fig. 6d and e). To assess stability, the dataset was separated into five equal parts using k-fold strategy and eQTL analysis was performed on each subset. USADAE showed slightly lower slope and slope SE correlations than PEER but outperformed Control, with comparable *P*-value correlations (Fig. 6f–h). The π replication rate was comparable to PEER but higher than Control (Fig. 6i), demonstrating USADAE's robustness in reproducible eQTL detection across splits.

Discussion

In summary, we propose a deep learning-based algorithm that explicitly extracts covariates from observed RNA-seq signals by encoding them into distinct biologically relevant and confounding latent representations via adversarial learning. It outperforms traditional methods like RUVR, SVA, and PEER by capturing nonlinear confounding structures, offering precise correction and stable performance. In breast cancer, USADAE enhances ER-related pathway enrichment, suggesting potential for biomarker discovery. In PigGTEx, it matches PEER's performance while enabling multi-tissue confounder extraction, facilitating integrative analysis.

USADAE's disentanglement capability, which relies on enforcing statistical independence in the latent space, requires large sample sizes. This constraint, shared by many deep learning models, confines its optimal use to large RNA-seq or integrated multi-source datasets. A recent study showed that PCA outperforms several methods under the additive model assumption: expression = genotype effect + covariate effect + noise [37]. However, its performance is limited in the presence of interactions between covariates and genotype effects (Supplementary Fig. S9). Furthermore, extending USADAE to jointly model genomic and transcriptomic variation represents a promising direction that may further improve the resolution and accuracy of latent covariate extraction.

Finally, dimensionality reduction via k-means++ is currently used to mitigate overfitting caused by high input dimensionality, but this inevitably introduces some information loss (Supplementary Fig. S10). As larger RNA-seq and multi-omics datasets become increasingly available, such preprocessing may no longer be necessary.

Method

Model architecture

We proposed a USADAE model to disentangle covariates from biological signals in observed data with unknown confounders. We assumed that the observed data are generated by a combination of biological signals and latent covariates. Based on this, we introduced two parallel encoders: a biological encoder E_b consisting of fully connected layers ($D \rightarrow 512 \rightarrow 256 \rightarrow L$) with batch normalization and ReLU activations, producing a latent representation $\mathbf{z}_b$, and a confounder encoder E_c ($128 \rightarrow 64 \rightarrow L_c$) that captures hidden covariates $\mathbf{z}_c$. Here, D denotes the input dimension, L represents the dimensionality of the biological latent space, and L_c indicates the dimensionality of the hidden covariate space. These representations are jointly decoded through *Decoder* ($[\mathbf{z}_b; \mathbf{z}_c]$), composed of symmetric fully connected layers ($L + L_c \rightarrow 256 \rightarrow 512 \rightarrow D$), to reconstruct the input. The reconstruction loss is defined as:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\mathbf{x}}\left[\| \mathbf{x} - \text{Decoder}\left([E_b(\mathbf{x}); E_c(\mathbf{x})]\right) \|_2^2\right]$$

To enforce statistically independent representations of biological factors and hidden covariates, we implemented an adversarial disentanglement framework [38]. Specifically, we introduced a discriminator network D_a ($(L + L_c) \rightarrow 32 \rightarrow 32 \rightarrow 1$) that learns to distinguish between two embedding types: joint embeddings $\mathbf{z}_{\text{joint}} =$

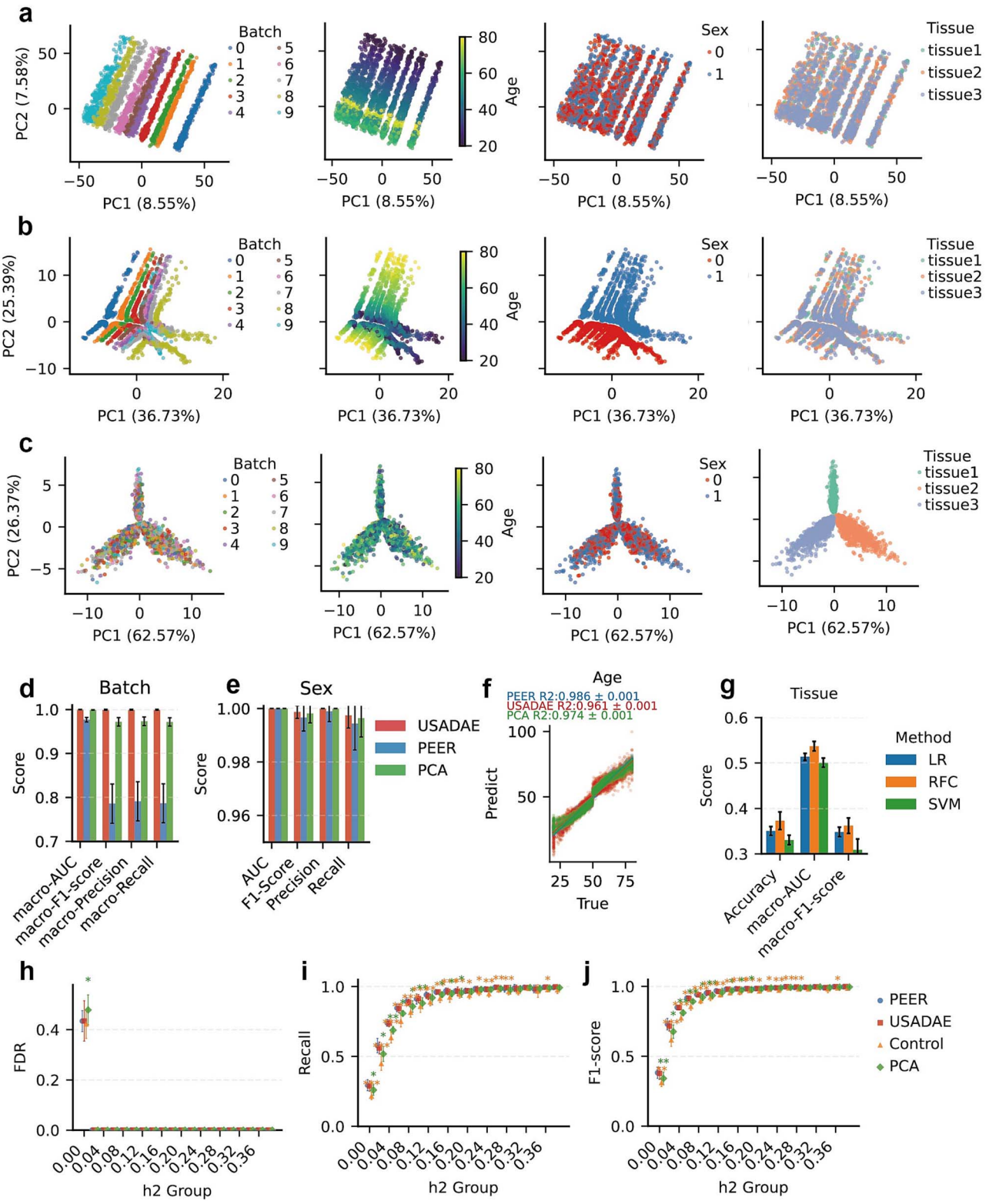

Figure 5 Performance comparison of USADAE and PEER in eQTL analysis with simulated gene expression data. (a) PCA plot colored by batch, age, sex, and tissue. (b, c) PCA plots for USADAE's covariates and biological latent respectively. (d, e) Bar charts comparing random forest classification performance for batch and sex using USADAE and PEER covariates, averaged over 5-fold cross-validation with error bars. (f) Scatter plots of predict and true age values, showing mean R^2 with standard errors. (g) Predict performance for tissue of different model constructed using covariates estimated by USADAE. LR, logistic regression; RF, random forest classifier; SVM, support vector machine classifier. (h–j) Scatter plots evaluate eGene detection performance across h^2 groups using USADAE, PEER, PCA covariates, and control (covariate = 1). Significance markers (paired t-tests with Benjamini–Hochberg adjustment, *adjusted $P < .05$, **adjusted $P < .01$) represent differences between USADAE Conf and other groups: yellow asterisks for USADAE versus control, blue for USADAE versus PEER (no significance observed).

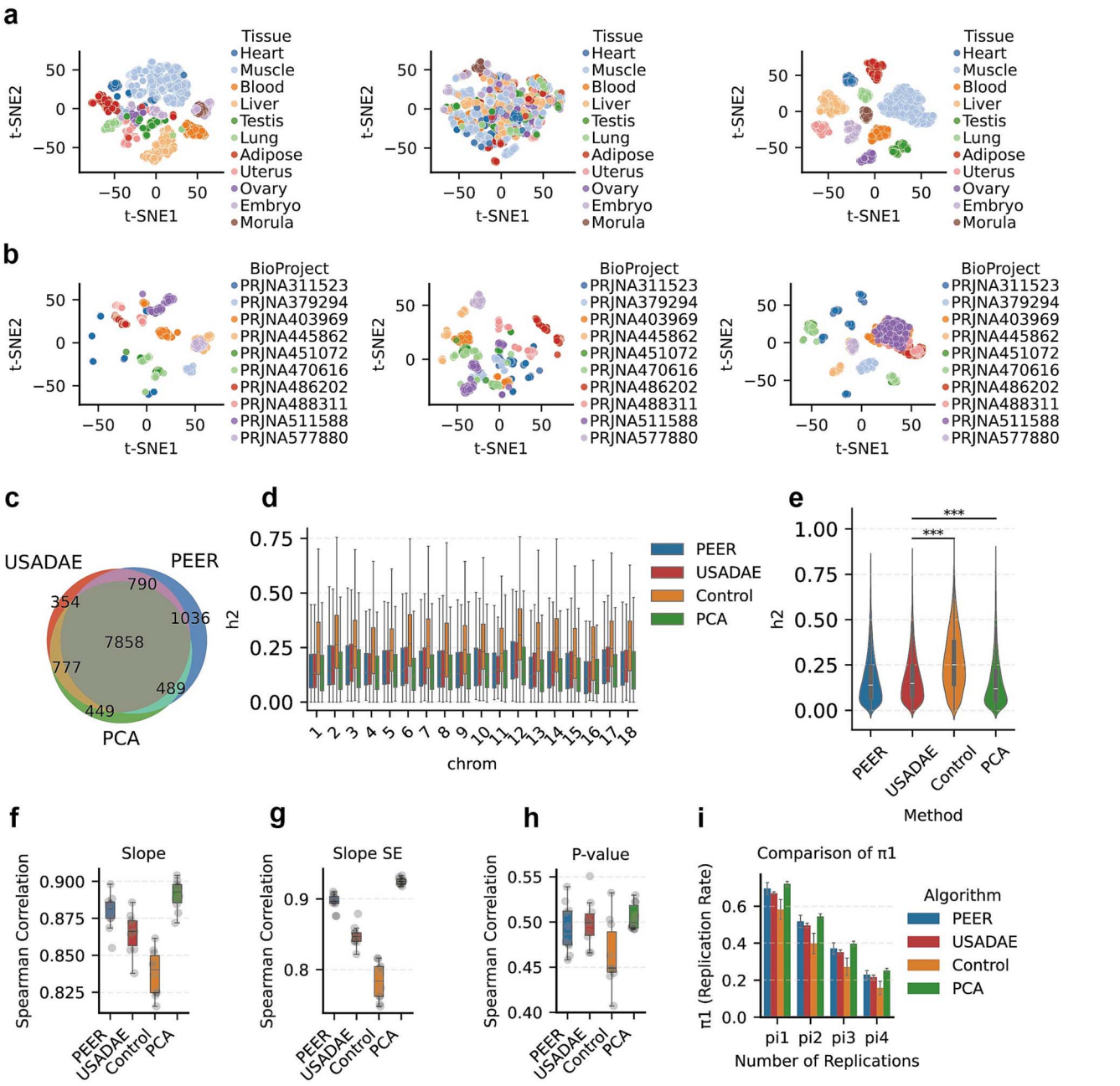

Figure 6 Performance comparison of USADAE, PCA, and PEER in eQTL analysis with PigGTEx datasets. (a) t-SNE plot colored by tissue. From left to right, the panels display: (1) the original data, (2) USADAE's covariates, and (3) USADAE's biological latent. (b) t-SNE plot, similar to (a), colored by BioProject. (c) Venn plot of eGenes. (d) Boxplot compares cis-h^2 values of eGenes detected using covariates estimated by different methods. (e) Comparison of cis-h^2 distribution of all eGenes. USADAE and PEER have similar distributions ($P = .176$, Wilcoxon test). (f–h) Boxplots display Spearman correlations for (f) slope estimates, (g) slope standard errors, and (h) *P*-values of eGene-eQTL pairs across folds. (i) Bar plots show eGene-eQTL pair replication rates across folds, with error bars indicating standard errors from-fold cross-validation.

$[E_b(\mathbf{x}); E_c(\mathbf{x})]$ generated from the same sample, representing the natural co-occurrence of biological signals and hidden covariates [39, 40]. Shuffled embeddings $\mathbf{z}_{\text{shuffled}} = \left[E_b\left(\mathbf{x}^{(i)}\right); E_c\left(\mathbf{x}^{(j)}\right)\right]$ for $i \neq j$ constructed by pairing biological embeddings with randomly permuted confounder embeddings, where $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ represent input sample i and j, respectively.

The adversarial training follows a minimax objective [38]:

$$\min_{E_b, E_c}\max_{D_a} \mathcal{V}(D_a, E_b, E_c) = \mathbb{E}_{\mathbf{x}}\left[\log D_a\left([E_b(\mathbf{x}); E_c(\mathbf{x})]\right)\right]$$
$$+\mathbb{E}_{\mathbf{x}}\left[\log\left(1 - D_a\left(\left[E_b\left(\mathbf{x}^{(i)}\right); E_c\left(\mathbf{x}^{(j)}\right)\right]\right)\right)\right]$$

Through this adversarial formulation, the model enforces the statistical independence condition:

$$P\left(E_b\left(\mathbf{x}\right), E_c\left(\mathbf{x}\right)\right) \rightarrow P\left(E_b\left(\mathbf{x}\right)\right) P\left(E_c\left(\mathbf{x}\right)\right)$$

thus, minimizing the Jensen–Shannon divergence between the joint distribution and the product of marginal distributions [41].

To ensure that E_b specifically captures biological signals but not others, we added a weak discriminator D_b that guides the biological encoder to retain label-discriminative features. The tissue discriminator loss is defined as:

$$\mathcal{L}_{\text{tissue}} = \mathbb{E}_{(\mathbf{x}, y_t)}\left[\text{CrossEntropy}\left(y_t, D_b\left(E_b\left(\mathbf{x}\right)\right)\right)\right]$$

where y_t represents the true tissue-type or condition label (e.g. ER status in breast cancer datasets or tissue type in PigGTEx). Finally, by combining these three synergistic objectives, the model optimization can achieve biologically meaningful disentanglement. The model training consists of three stages: (i) Autoencoder training, (ii) Autoencoder training + adversarial training, and (iii) Autoencoder training + adversarial training + tissue discrimination training. The epochs for stages 1 and 2 are set to 100, while the epochs for stage 3 are set to 500. Dropout layers are applied at the bottleneck layer of the model, and dropout rate is set to 0.4 by default [42]. The activation function used throughout the model is ReLU. Adam is employed across all training phases, with learning rates set to $1e-3$ for autoencoder training, and $1e-4$ for both adversarial and tissue discrimination training. The main tunable parameters in USADAE include the latent embedding dimension (L), the dropout rate, the λ values for adversarial training, and the number of centroids used in the preliminary k-means++ clustering step. In practice, we recommend setting L between 12 and 32, dropout between 0.2 and 0.6, λ_{adv} for adversarial between 0.5 and 2, λ_{tissue} for tissue equal to λ_{adv} plus 0.5, and the number of k-means++ centroids between 500 and 2500, depending on dataset size and signal complexity. All neural network models were implemented in PyTorch and trained on an NVIDIA GeForce RTX 3090 GPU with 24 GB memory.

Simulated data for differential expression analyses

To assess the performance of USADAE in separating covariates, we simulated RNA-seq data comprising 3000 genes across 1000 samples (500 cases, 500 controls). Gene expression was influenced by three confounders-batch, age, and sex. Genes with $|\log_2 FC| > 0.263$ before adding confounders were treated as true significant differential expression genes (Supplementary Information).

GEO datasets

We utilized seven breast cancer datasets from the GEO database (GSE2034, GSE3494, GSE7390, GSE20194, GSE20271, GSE25066, and GSE76275), comprising a total of 1544 samples with 12 548 genes. These datasets all contained ER status labels, which served as biological groups for downstream DEA.

To validate confounder accuracy, subsets with clinical annotations were extracted: 198, 268, and 502 samples with age from GSE7390, GSE20194, and GSE25066; 198 samples with tumor size from GSE7390; 198 and 502 samples with tumor grade from GSE7390 and GSE25066; and 265 samples with race from GSE20194. These were used to assess disentangled covariates, not included in training.

Log2-transformed data were processed with k-means++ clustering [43] to reduce dimensionality and prevent overfitting.

Simulated data for eQTL analysis

We simulated multi-tissue eQTL data using PigGTEx chromosome 1 muscle genotypes, modeling 2000 genes across 3 tissues with tissue-specific, shared, and unique eGenes. *Cis*-regulatory effects used 5–15 SNPs per gene within 1 Mb, with effects scaled by MAF and heritability (0.5%–80%). Nonlinear confounders included batch, age, and sex effects. Expression combined genetic and confounding effects with tissue-specific noise. Genotype data included 10 000 SNPs (MAF > 1%) and replicated IDs for paired designs. Details are in the Supplementary Information.

PigGTEx datasets

To evaluate the capability of USADAE in disentangling covariates across multi-group datasets, we integrated multi-tissue transcriptomic and genotypic data from PigGTEx project [13]. We analyzed 11 tissues (sample size >150 each), including muscle, adipose, and others. Transcriptomic data were preprocessed by dual-threshold filtering (TPM > 0.1 and raw count >6 in ≥20% samples per tissue), from which the top 10 000 high-variance genes were selected and aggregated across tissues. Prior to modeling, data were log2-transformed and dimension-reduced to 1500 via k-means++ clustering [43]. Genotype data and PCA covariates for *cis*-eQTL mapping were sourced directly from the PigGTEx database, retaining only SNPs with MAF ≥ 5%.

Covariates evaluation

The accuracy of extracted covariates was evaluated by training machine learning models to predict known covariates withheld during training. For continuous and categorical variables, we employed regression and classification models, respectively, and assessed performance using a 5-fold cross-validation protocol. Model generalizability was quantified by R^2 for regression tasks, and by AUC, F1-score, precision, and recall for classification tasks against the ground-truth labels (Supplementary Information).

Differential expression analyses and evaluation

DEA were conducted using DESeq2 [44] for simulated data, a limma-voom [45] pipeline for public GEO datasets, with covariates (e.g. batch, sex, or those extracted by USADAE) integrated into the models. For simulated data, algorithm performance was evaluated by comparing predicted differentially expressed genes (DEGs) against a ground-truth list using ROC curves, with AUC measuring classification accuracy. True positives, false positives, false negatives, and true negatives were defined based on dual-threshold (adjusted P-values <.05 and $|\text{log2FC}| > 0.263$) and ground-truth status. Precision, recall, and F1-score assessed DEA improvement with covariates integration. For GEO datasets, where true DEGs are unknown, KEGG enrichment analysis on predicted DEGs was performed using clusterProfiler's enrichKEGG functions [46].

Cis-eQTL mapping

Similar to the PEER and PCA algorithm, USADAE was applied solely to RNA-seq data to infer hidden covariates. The model utilized the complete set of RNA-seq data from all tissues as input, with tissue identity serving as the grouping variable. Subsequently, the inferred covariates were applied as adjustments in *cis*-eQLT mapping to control for hidden confounders for each tissue. The analytical workflow in this analysis closely mirrors the MolQTL mapping approach outlined in the PigGTEx [13]. Key software or packages utilized include edgeR [47] and TensorQTL (1.0.10) [48].

Replication rate (π) analysis

To assess the robustness of eQTL discoveries across methods, we computed replication rates (π statistics) via 5-fold cross-validation. In each fold, one subset served as the discovery set to identify significant eGene-eQTL pairs, while the remaining four were replication sets. The replication rate for a given discovery set was calculated as the proportion of its significant signals replicated in at least k (ranging from 1 to 4) of the replication datasets:

$$\pi_k = \frac{\text{Number of signals replicated at least } k \text{ times (n)}}{\text{Total number of signals in the main dataset (N)}}$$

Key Points

- USADAE overcomes the limitations of linear methods by capturing nonlinear confounding patterns in RNA-seq data.
- USADAE optimizes Jensen–Shannon divergence to encode data into statistically independent biological and confounder latent spaces.
- USADAE robustly corrects covariates, enhancing differential expression and eQTL analyses with better performance across simulated and real-world datasets.

Acknowledgements

This work was supported by the Selection and Breeding of New Local Pig Breeds and Promotion of Industrialization (2024-XPY-00-001), Selection and Breeding of Guangdong Small-ear Spotted Pig (2024-XPY-00-005), Modern Agricultural Industrial Technology System Innovation Team of Guangdong Province (2024CXTD22), and the ear-marked fund for CARS-35.

Author contributions

X.C.: Conceptualization, Formal analysis, Writing – original draft; L.G.: Visualization, Writing – review & editing; Y.C.: Writing – review & editing; D.M.: Supervision, Writing – review & editing; X.L.: Supervision, Writing – review & editing. All authors approved the final version.

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Funding

None declared.

Data availability

Publicly available resources include: (i) the GEO datasets from the NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) and (ii) the RNA-seq data from the PigGTEx datasets (https://piggtex.farmgtex.org/). The code for the USADAE algorithm is available at https://github.com/chenxuya/USADAE.

References

1. Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63. https://doi.org/10.1038/nrg2484
2. Kukurba KR, Montgomery SB. RNA sequencing and analysis. *Cold Spring Harb Protoc* 2015;**2015**:951–69. https://doi.org/10.1101/pdb.top084970
3. Conesa A, Madrigal P, Tarazona S *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;**17**:13. https://doi.org/10.1186/s13059-016-0881-8
4. Corchete LA, Rojas EA, Alonso-López D *et al.* Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci Rep* 2020;**10**:19737. https://doi.org/10.1038/s41598-020-76881-x
5. Chen W, Zhang S, Williams J *et al.* A comparison of methods accounting for batch effects in differential expression analysis of UMI count based single cell RNA sequencing. *Comput Struct Biotechnol J* 2020;**18**:861–73. https://doi.org/10.1016/j.csbj.2020.03.026
6. Molania R, Foroutan M, Gagnon-Bartsch JA *et al.* Removing unwanted variation from large-scale RNA sequencing data with PRPS. *Nat Biotechnol* 2023;**41**:82–95. https://doi.org/10.1038/s41587-022-01440-w
7. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* 2020;**2**:lqaa078. https://doi.org/10.1093/nargab/lqaa078
8. Stegle O, Parts L, Piipari M *et al.* Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* 2012;**7**:500–7. https://doi.org/10.1038/nprot.2011.457
9. Lazar C, Meganck S, Taminau J *et al.* Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform* 2013;**14**:469–90. https://doi.org/10.1093/bib/bbs037
10. Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res* 2014;**42**:e161. https://doi.org/10.1093/nar/gku864
11. Risso D, Ngai J, Speed TP *et al.* Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 2014;**32**:896–902. https://doi.org/10.1038/nbt.2931
12. Liu S, Gao Y, Canela-Xandri O *et al.* A multi-tissue atlas of regulatory variants in cattle. *Nat Genet* 2022;**54**:1438–47. https://doi.org/10.1038/s41588-022-01153-5
13. Teng J, Gao Y, Yin H *et al.* A compendium of genetic regulatory effects across pig tissues. *Nat Genet* 2024;**56**:112–23. https://doi.org/10.1038/s41588-023-01585-7
14. The GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;**369**:1318–30. https://doi.org/10.1126/science.aaz1776

15. Ren X, Kuan P-F. Negative binomial additive model for RNA-seq data analysis. *BMC Bioinformatics* 2020;**21**:171. https://doi.org/10.1186/s12859-020-3506-x
16. Li Z, Katz S, Saccenti E *et al.* Novel multi-omics deconfounding variational autoencoders can obtain meaningful disease subtyping. *Brief Bioinform* 2024;**25**:bbae512. https://doi.org/10.1093/bib/bbae512
17. Wang T, Johnson TS, Shao W *et al.* BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome Biol* 2019;**20**:165. https://doi.org/10.1186/s13059-019-1764-6
18. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv:1312.6114. 2013. https://doi.org/10.48550/arXiv.1312.6114
19. Lopez R, Regier J, Cole MB *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8. https://doi.org/10.1038/s41592-018-0229-2
20. Li X, Wang K, Lyu Y *et al.* Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat Commun* 2020;**11**:2338. https://doi.org/10.1038/s41467-020-15851-3
21. Xu C, Lopez R, Mehlman E *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol Syst Biol* 2021;**17**:e9620. https://doi.org/10.15252/msb.20209620
22. Shen X, Shen H, Wu D *et al.* Scalable batch-correction approach for integrating large-scale single-cell transcriptomes. *Brief Bioinform* 2022;**23**:bbac327. https://doi.org/10.1093/bib/bbac327
23. Zou B, Zhang T, Zhou R *et al.* deepMNN: deep learning-based single-cell RNA sequencing data batch correction using mutual nearest neighbors. *Front Genet* 2021;**12**:708981. https://doi.org/10.3389/fgene.2021.708981
24. Dincer AB, Janizek JD, Lee S-I. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* 2020;**36**:i573–82. https://doi.org/10.1093/bioinformatics/btaa796
25. Wang Y, Klijn JGM, Zhang Y *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;**365**:671–9. https://doi.org/10.1016/S0140-6736(05)17947-1
26. Desmedt C, Piette F, Loi S *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res* 2007;**13**:3207–14. https://doi.org/10.1158/1078-0432.CCR-06-2765
27. Shi L, Campbell G, Jones WD *et al.* The MicroArray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol* 2010;**28**:827–38. https://doi.org/10.1038/nbt.1665
28. Popovici V, Chen W, Gallas BD *et al.* Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res* 2010;**12**:R5. https://doi.org/10.1186/bcr2468
29. Hatzis C, Pusztai L, Valero V *et al.* A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* 2011;**305**:1873–81. https://doi.org/10.1001/jama.2011.593
30. Chipumuro E, Marco E, Christensen CL *et al.* CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer. *Cell* 2014;**159**:1126–39. https://doi.org/10.1016/j.cell.2014.10.024
31. Tabchy A, Valero V, Vidaurre T *et al.* Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. *Clin Cancer Res* 2010;**16**:5351–61. https://doi.org/10.1158/1078-0432.CCR-10-1265
32. Cochaud S, Giustiniani J, Thomas C *et al.* IL-17A is produced by breast cancer TILs and promotes chemoresistance and proliferation through ERK1/2. *Sci Rep* 2013;**3**:3456. https://doi.org/10.1038/srep03456
33. Clusan L, Ferrière F, Flouriot G *et al.* A basic review on estrogen receptor signaling pathways in breast cancer. *Int J Mol Sci* 2023;**24**:6834. https://doi.org/10.3390/ijms24076834
34. Brufsky AM, Dickler MN. Estrogen receptor-positive breast cancer: exploiting signaling pathways implicated in endocrine resistance. *Oncologist* 2018;**23**:528–39. https://doi.org/10.1634/theoncologist.2017-0423
35. Esquivel-Velázquez M, Ostoa-Saloma P, Palacios-Arreola MI *et al.* The role of cytokines in breast cancer development and progression. *J Interf Cytokine Res* 2015;**35**:1–16. https://doi.org/10.1089/jir.2014.0026
36. Kawaguchi K, Sakurai M, Yamamoto Y *et al.* Alteration of specific cytokine expression patterns in patients with breast cancer. *Sci Rep* 2019;**9**:2924. https://doi.org/10.1038/s41598-019-39476-9
37. Zhou HJ, Li L, Li Y *et al.* PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol* 2022;**23**:210. https://doi.org/10.1186/s13059-022-02761-4
38. Goodfellow I, Pouget-Abadie J, Mirza M *et al.* Generative adversarial networks. *Commun ACM* 2020;**63**:139–44. https://doi.org/10.1145/3422622
39. Kim H, Mnih A. Disentangling by Factorising. In: Dy JG, Krause A (eds.), *Proceedings of the 35th International Conference on Machine Learning*, pp. 2649–58. Cambridge, MA: PMLR, 2018.
40. Belghazi MI, Rajeswar S, Baratin A *et al.* MINE: Mutual information neural estimation. In: Dy J, Krause A (eds.), *Proceedings of the 35th International Conference on Machine Learning*, vol. **80**, pp. 531–40. Stockholm, Sweden: PMLR, 2018.
41. Menéndez ML, Pardo JA, Pardo L *et al.* The Jensen-Shannon divergence. *J Frankl Inst* 1997;**334**:307–18. https://doi.org/10.1016/S0016-0032(96)00063-4
42. Srivastava N, Hinton G, Krizhevsky A *et al.* Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929–58.
43. Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. In: Bansal N, Pruhs K, Stein C (eds.), *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pp. 1027–35. Philadelphia, PA: SIAM, 2007.
44. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8
45. Ritchie ME, Phipson B, Wu D *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47. https://doi.org/10.1093/nar/gkv007
46. Yu G, Wang L-G, Han Y *et al.* clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;**16**:284–7. https://doi.org/10.1089/omi.2011.0118
47. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40. https://doi.org/10.1093/bioinformatics/btp616
48. Taylor-Weiner A, Aguet F, Haradhvala NJ *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol* 2019;**20**:228. https://doi.org/10.1186/s13059-019-1836-7

Briefings in Bioinformatics, 2026, **27**, bbag261
https://doi.org/10.1093/bib/bbag261
Problem Solving Protocol