Predicting individualized tissue gene expression profiles with multi-task learning

Guangyi He[1,4], Maiyue Chen[2,4], Yingnan Bian[3], Ence Yang[1 *]

[1] Department of Medical Bioinformatics, School of Basic Medical Sciences, Peking University, Beijing 100191, P. R. China.

[2] School of Artificial Intelligence, Peking University, Beijing 100191, P. R. China.

[3] Enlight Medical Technologies (Shanghai) Co., Ltd, Shanghai 201318, P. R. China.

[4] These authors contributed equally: Guangyi He, Maiyue Chen.

[*] e-mail: yangence@pku.edu.cn

Abstract

Mapping biological information from peripheral 'surrogate' samples, especially transcriptomic information from blood, to tissue expression profiles has become an effective emerging alternative when invasive procedures are not ideal. However, existing approaches ignore hidden nonlinear expression relationships and disrupt tissue-shared intrinsic relevance, inevitably limiting predictive performance. Here, we propose a unified deep-learning-based multi-task learning framework, Multi-tissue Transcriptome Mapping (MTM), that enables the prediction of individualized expression profiles from any available tissue from an individual. By jointly leveraging individualized cross-tissue information through nonlinear neural networks, MTM achieves superior performance at both the sample level and the gene level with a large proportion of predictable genes. With the high accuracy of predicting the expression profiles of uncollected tissues and the ability to preserve individualized biological variations, MTM could facilitate the discovery of novel mechanisms and clinical biomarkers in biomedical research.

Gene expression profiles in diverse tissues act as molecular mediators between genotypes and phenotypes and provide a snapshot of the systemic physiological and pathological status of an individual[1]. Deciphering tissue-derived gene expression not only provides insights into the fundamental molecular mechanisms of biological processes[2-4], but also aids in clinical diagnosis, subtyping and management[5-7]. However, tissue transcriptome-based biological investigations and clinical evaluations highly depend on invasive tissue biopsy procedures, which are risky, harmful, and even unavailable in some cases. Thus, mapping biological information from peripheral 'surrogate' samples to tissue transcriptomic profiles has become an emerging solution attracting extensive research interest.

Various methods have been developed to predict tissue-specific expression for specific individuals. Early attempts utilized genotypes with the rationale that individual variance in genetically regulated expression (GReX) components of tissue-specific gene expression could be explained by genotypes to some extent, i.e., expression quantitative trait loci (eQTL). A representative work demonstrating this strategy is the PrediXcan, which takes advantage of eQTL variants to construct gene-based tests to impute GReX[8]. However, these eQTL-based methods are limited to genes with significant tissue-specific eQTL single-nucleotide polymorphisms (eSNPs) and neglect non-GReX factors in expression. As blood is involved in the circulation and bidirectional exchange of substances with organs, it is considered to reflect the in-time functional status of tissues underlying gene expression, including components attributable to feedback of traits, and to environmental and other factors[9, 10]. Recent studies have supported the advantages of predicting tissue expression profiles based on blood expression[11-14]. Nevertheless, the methods based on genotype and transcriptome usually have two common limitations: first, they are typically based on linear models or traditional machine learning methods, inevitably limiting their capability to capture complex nonlinear expression relationships in biological organisms[15, 16]; second, they build prediction models independently for a single gene in each target tissue, failing to utilise the intrinsic cross-tissue biological inherence underlying transcriptomes of multiple tissues from the same individual, which could be rescued by borrowing information from similar tissues[17-19].

Deep learning has shown its powerful capability for handling high-dimensional, complex, and nonlinear biological data[20-27]. Multi-task learning, which imposes regularity or shared representation by learning tasks simultaneously, is ideal for utilizing the intrinsic shared features of different domains (e.g., different tissues from the same individual)[28-31]. Thus, we developed MTM, a deep-learning-based multi-task learning framework to predict individualized tissue gene expression profiles using any available tissues from a person. MTM integrates mappings among multiple tissues into a single model, enabling the exploitation of the underlying complex nonlinear patterns of tissue gene expression levels within the same individual. With its outstanding predictive power and the ability to capture individualized physiological and pathological variations, MTM demonstrates its potential to accelerate transcriptome-based research and clinical applications.

Results

Framework overview

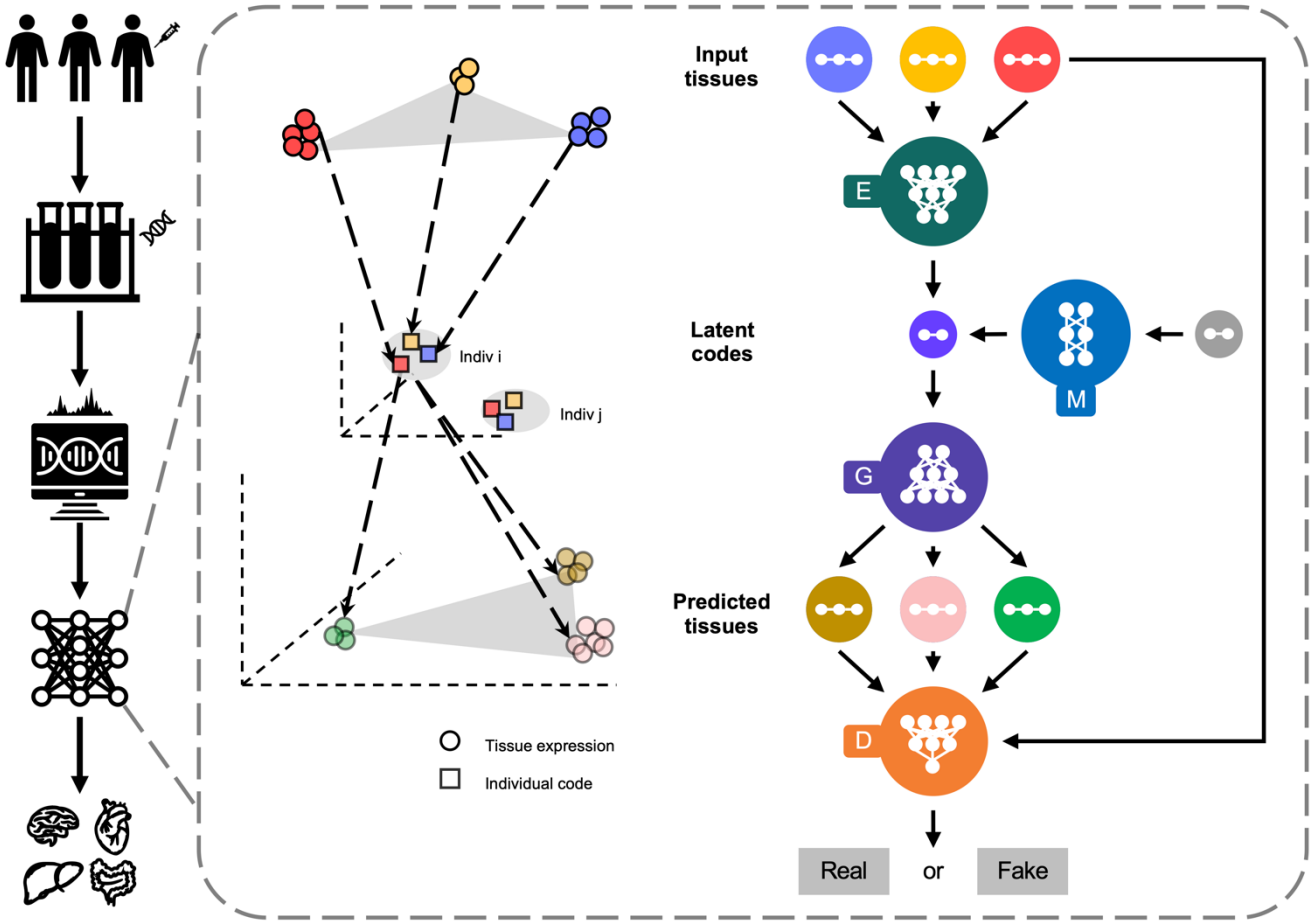

Fig. 1 Overview of MTM. MTM is based on a multi-task learning architecture and consists of four artificial neural networks, a unified encoder (*E*), a unified generator (*G*), a unified discriminator (*D*), and a mapping network (*M*). The gene expression profiles of different input tissues are embedded into a shared latent space by *E* to form individualized codes, which are then mapped to multiple tissue-specific expression spaces by *G* to realize personalized prediction of the expression profiles of different target tissues.

To achieve individualized gene expression prediction, we proposed a deep learning framework, Multi-tissue Transcriptome Mapping (MTM), a unified multi-task learning approach capable of predicting tissue-specific gene expression profiles using any available tissues from an individual (Fig. 1). Gene expression levels of 19,291 protein-coding

genes across 49 tissues from 948 individuals in the Genotype-Tissue Expression (GTEx) datasets[32] were used to develop our model. The expression levels of input tissue samples across different tissues were processed by a unified encoder network equipped with the tissue conditioning module (TCM), to encourage personalized representations in a shared latent space. Then, the latent codes were mapped to multiple tissue-specific expression spaces through a prediction task conditioning generator to achieve personalized prediction of tissue expression profiles. Two auxiliary networks, a mapping network and a discriminator network, were employed to smooth the latent space and to refine the prediction results, respectively (see details in the Methods). With the multi-task learning architecture, MTM is not limited to predicting the expressions of a single target tissue based on a single input tissue, but rather supports arbitrary predefined inputs and predicting targets in a unified model, which introduces more flexibility along with potentially improved performance.

MTM outperforms the baseline methods

The prediction quality of MTM at the sample level and the gene level was evaluated on the task of predicting individualized gene expression profiles of the other 48 tissues within the development set (GTEx) based on the expression profile of whole blood. Two representative single-tissue approaches, B-GEX (method S1) and TEEBoT (method S2), were used as baselines in the evaluation for comparison. At the sample level, the sample-wise prediction accuracy was measured by Pearson's correlation coefficient, which was 0.21±0.07 across all genes among the 48 tissues (ranging from 0.04 to 0.34). Compared with 0.00±0.02 for S1 and 0.08±0.09 for S2, MTM showed significantly higher levels of accuracy (Wilcoxon rank sum test, $P < 0.05$) than S1 or S2 did in all 48 tissues. On the gene-scale, MTM (0.23±0.07, ranging from 0.04 to 0.34) also demonstrated significantly higher gene-wise accuracy (Wilcoxon rank sum test, $P < 0.05$) than S1 (0.00±0.03) and S2 (0.07±0.11) did in 48 and 47 tissues out of the 48 tissues, respectively (Fig. 2a and Supplementary Table 1).

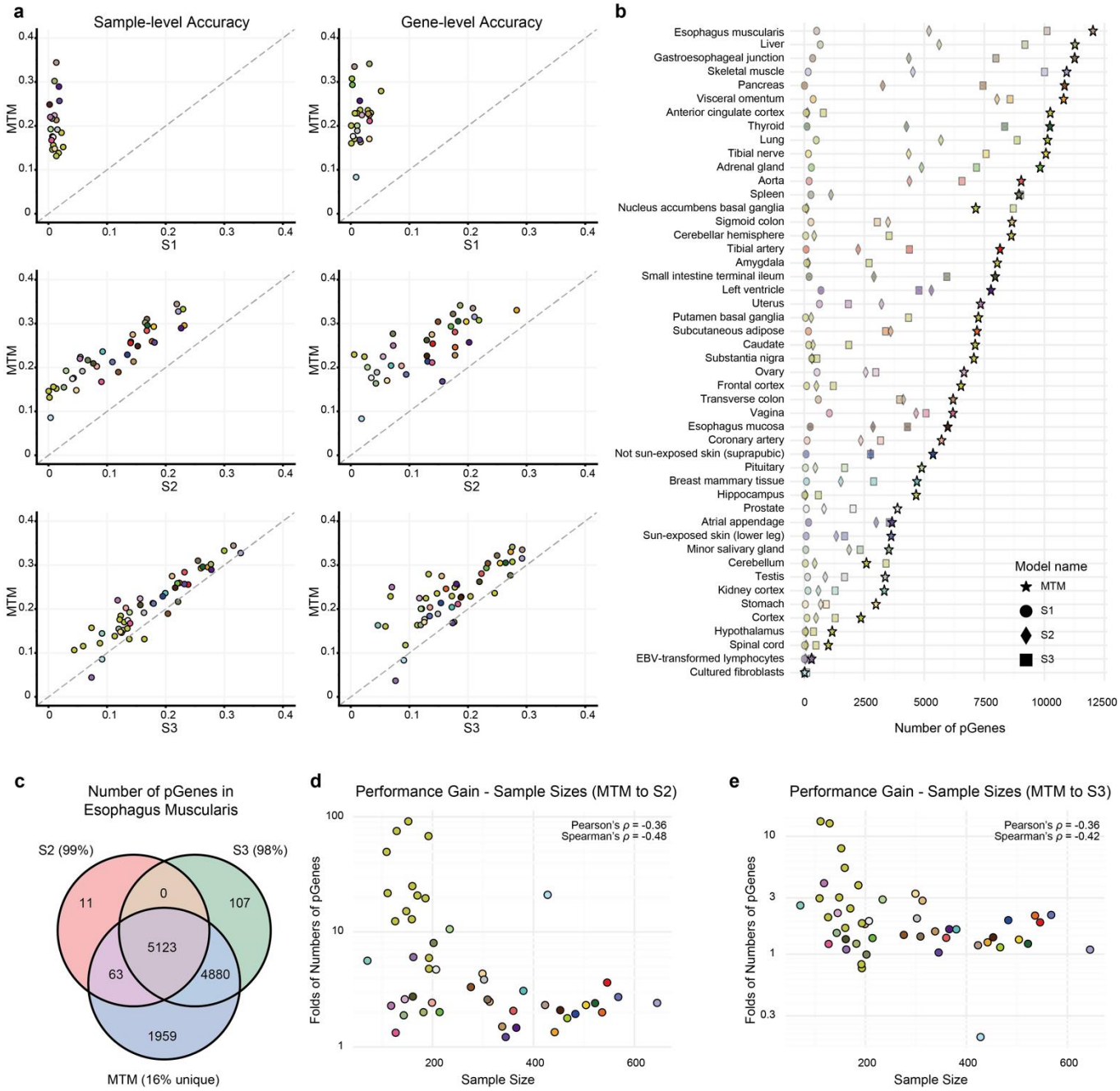

Fig. 2 Comparison of the performance for predicting tissue-specific gene expression. a The sample-wise and gene-wise accuracy (Pearson's correlation coefficient) of MTM against single-tissue approaches, including linear (S1 and S2) and nonlinear (S3) models. **b** The number of pGenes (Pearson's $\rho > 0.3$) of different models in different target tissues. **c** Venn diagram showing the overlap of the pGenes from S2, S3 and MTM in the oesophagus muscularis. Approximately 99% and 98% of pGenes from S2 and S3 were included in the MTM-derived pGenes, respectively. **d**, **e** The relationship between the fold changes in the number of pGenes and the sample sizes of different tissues for MTM to S2 (**d**) and MTM

to S3 (**e**). Each scatter point colour represents a specific tissue as defined by the GTEx Consortium[32].

In terms of predictable genes (pGenes; defined as genes with Pearson's $\rho > 0.3$ between predicted and observed expression), MTM on average identified 6,597 pGenes across the 48 tissues (ranging from 21 to 12,025), which was substantially higher than the 218 (ranging from 0 to 1,048) of S1 and the 2,301 (ranging from 1 to 8,031) of S2 (Fig. 2b and Supplementary Table 2). For the S1- and S2-derived pGenes, a median of 94.2% (7.7% – 100.0%) and 94.5% (0.0% – 99.8%) were identified by MTM, respectively. In contrast, a median of 2.2% (0.0% – 15.9%) and 32.6% (0.0% – 73.1%) of MTM-derived pGenes were captured by S1 and S2, respectively. When comparing the fold changes in the number of pGenes of MTM to S2 (ranging from 1.22- to 91.37-fold, mean = 11.39-fold, median = 2.91-fold), we found that the performance gain was negatively correlated with sample size (Spearman's $\rho = -0.48$, $P = 5.26\times10^{-4}$, Fig. 2d), implying that the missing information caused by small sample sizes could be partially recovered from other tissues with the application of MTM.

Both the nonlinearity and multi-task framework contribute to performance

To investigate the independent contribution of the nonlinear and multi-task components, we performed ablative experiments and assessed their influence on the overall performance. We constructed an intermediate model (S3) for comparison, which consisted of a set of nonlinear neural networks for each tissue based on simple multilayer perceptrons.

For the nonlinearity component, we compared the performance of the linear models (S2) with the constructed nonlinear models (S3) for each tissue. As expected, the accuracy of S3 outperformed S2 both at the sample (0.17±0.07) and gene levels (0.17±0.07). The sample-wise accuracies and gene-wise accuracies of S3 were significantly ($P < 0.05$) higher than those of S2 in 46 tissues and 43 tissues, respectively. Compared with S2, the average number of pGenes resulting from S3 (on average 4,057 pGenes, ranging from 106 to 10,110, Supplementary Table 2) was substantially higher, with fold changes

ranging from 0.57- to 106.00-fold (mean = 7.11-fold, median = 1.93-fold).

For the multi-task learning component, MTM was compared with the constructed nonlinear single-tissue models (S3). At the sample level, the accuracy of MTM outperformed S3 in 41 tissues, with the performance in 24 tissues showing statistical significance ($P < 0.05$). In addition, at the gene level, MTM significantly ($P < 0.05$) outperformed S3 in 42 tissues (Fig. 2a and Supplementary Table 1). Furthermore, the average number of pGenes resulting from MTM (6,597) was substantially higher than that of S3 (4,057) (Fig. 2b and Supplementary Table 2), with fold changes ranging from 0.20- to 13.38-fold (mean = 2.46-fold, median = 1.65-fold). Approximately 92.4% of the S3-derived pGenes were still captured by MTM (Fig. 2c and Supplementary Fig. 1). The above results suggested that apart from advantages introduced by nonlinear neural networks, jointly leveraging the individualized cross-tissue information provides extra prediction improvements perpendicularly.

Intermediate features and outputs of MTM provided insights into the performance basis

To determine how MTM integrated different tissues from the same individual as an entirety, we next explored the characteristics of the individualized representations learned by MTM. Intraindividual latent codes derived from the same individual showed significantly ($P < 2.2\times10^{-16}$) higher similarities than interindividual latent codes did (Fig. 3b and 3c), indicating that individualized properties of different tissues from the same individual were captured well by MTM. Then, we explored the characteristics of the decoding paths from the same individual representation to the tissue expression profiles. We found that the pairwise similarities between the decoding paths of the data flow (i.e., activations of different layers) were highly consistent (Spearman's $P < 0.05$) with those between tissue expression profiles (Fig. 3d), suggesting that the biological similarities between tissues were reflected in the decoding process. Together, the integration of multiple tissues into a single model by our multi-task learning architecture resulted in individualized representations and decoding

rules with biological patterns.

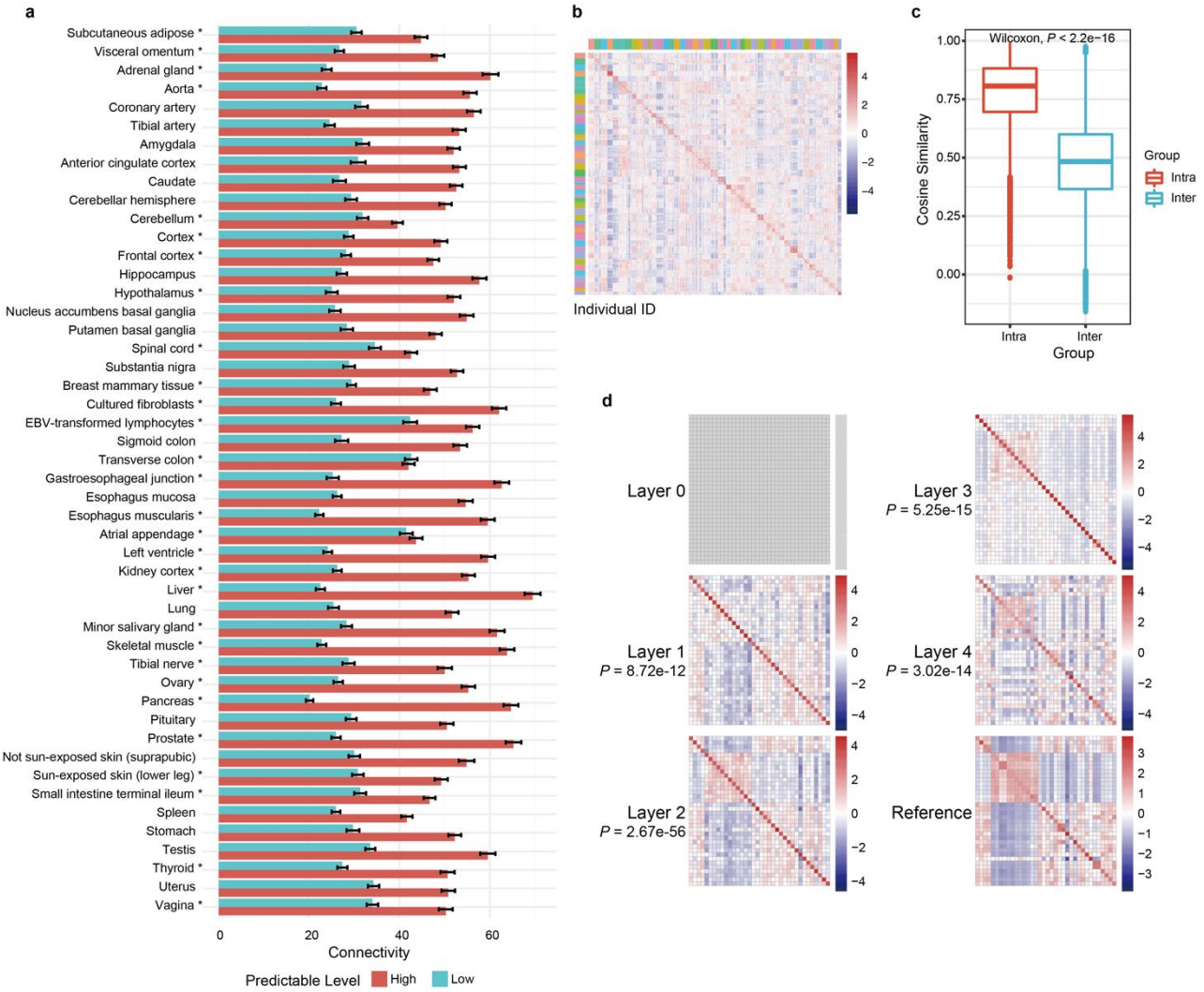

Fig. 3 Characteristics of predicted and intermediate data derived from MTM. a Comparison of the connectivity levels (average degrees) between highly predictable (top 25 percentiles of predictable genes) and unpredictable genes (bottom 25 percentiles of predictable genes) in different tissues. **b** Heatmap displaying the pairwise similarities of codes of different tissues from 50 random individuals. **c** Comparison of similarities of intraindividual and interindividual latent codes. **d** Pairwise similarities between different intermediate data flow towards different target tissues, which started from the same latent code of the blood expression from one specific individual. Each grid is coloured to indicate a similarity level between intermediate data of the corresponding two tissues. 'Reference' refers to pairwise similarities between the expression

profiles of different target tissues.

Given that MTM is able to identify more pGenes, distinctive features of the pGenes may also aid in the interpretation of the overall prediction strength of MTM from a different perspective. We found that tissues with similar expression profiles tended to share similar pGenes (Spearman's ρ = 0.56, P = 1.40×10^{-92}), which was consistent with previous discoveries. The pGenes tended to have significantly (Wilcoxon rank sum test, $P < 0.05$) higher expression levels than the unpredictable genes in all 48 tissues (Supplementary Fig. 3). The pGenes showed significantly (Wilcoxon rank sum test, $P < 0.05$) higher conservation than the unpredictable genes in 9 tissues but significantly lower conservation in 35 out of the 48 tissues (Supplementary Fig. 4), suggesting that there was no obvious preference for conservation in pGenes. Notably, there were significantly higher levels of connectivity (average degrees) in the protein interaction network (PIN) for pGenes than for unpredictable genes in a large portion of tissues (higher in 47 tissues, with 28 tissues showing statistical significance, while significantly lower in one tissue, Wilcoxon rank sum test) (Fig. 3a). By comparison, the S2-derived pGenes showed significantly lower levels of connectivity in 11 tissues, while exhibiting significantly higher levels in 26 tissues. The difference between MTM and S2 suggested that MTM might be more capable of capturing the complex interactions in the PIN.

MTM could preserve personalized biological information

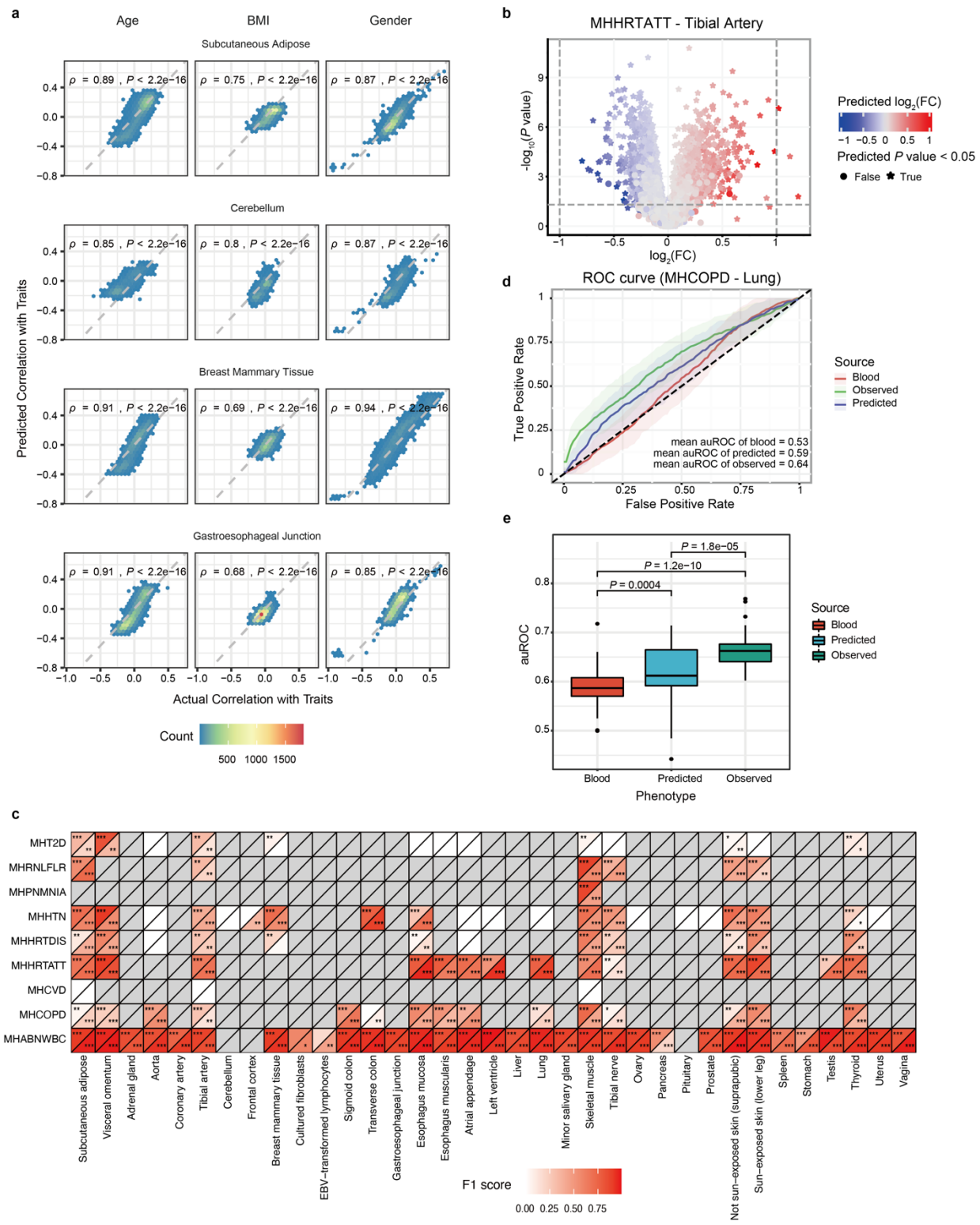

Fig. 4 Predicted data preserve personalized biological variations. a Hexagonal heatmaps show the relations between the associations of genes with traits (age, gender and BMI) using predicted expression and those using the actual expression

167 within GTEx. Concordance in terms of Spearman's correlations (Spearman's ρ) between association levels (Pearson's ρ)
168 based on the predicted and actual gene expression are displayed in the plots. Hexagonal cells are coloured according to
169 numbers of genes (count). **b** Comparison of results in differential expression (DE) analysis between the predicted
170 expression and the observed tibial artery expression with heart attack. The position of each point represents the observed
171 results, while the colour indicates the predicted $\log_2$ fold change. **c** Performance of predicting differentially expressed
172 genes (DEGs) of different diseases in different tissues using predicted data. The number of stars in the grids indicates the
173 overlap significance level of up- or downregulated genes between observed and predicted data (*=significant at 0.05, **=
174 significant at 1×10^{-10} and ***=significant at 1×10^{-100}). The colour represents the level of the F1 score for predicting DE
175 genes. **d** A specific case of receiver operating characteristic (ROC) curves for predicting chronic obstructive pulmonary
176 disease (COPD) status in the observed lung expression, the predicted lung expression, and the input blood expression. **e**
177 Performance of predicting disease status (auROC) of different cases using observed tissue data, predicted tissue data, and
178 input blood data.
179
180
181 As MTM resulted in highly personalized representations of individuals, we assessed the extent to which the predicted
182 tissue expression profiles could reveal trait-associated information. Correlation analysis of gene expression levels of
183 pGenes was conducted against individuals' traits (including age, gender and BMI) in the predicted data and compared with
184 the actual data (observed data). The age-specific gene expression associations were well preserved (Pearson's $\rho > 0.3$, $P <$
185 0.05) in the predicted expression data in 46 out of the 48 tissues, with a ρ_{median} of 0.81. Similarly, the test statistics of gender
186 ($\rho_{median} = 0.76$) and BMI ($\rho_{median} = 0.56$) of the predicted data showed significant positive (Pearson's $\rho > 0.3$, $P < 0.05$)
187 correlations with those of the actual data in all 43 gender-independent tissues and 39 out of the 48 tissues, respectively
188 (Fig. 4a and Supplementary Fig. 5–7). The above results suggested that the individualized tissue expression profiles

predicted by MTM preserved the trait-related expression changes well.

Apart from trait-associated changes, we investigated the consistency between tissue-specific disease-related dysregulations in the predicted data and those of the actual data (GTEx). Differential expression analysis (DE analysis) in pGenes was conducted for the combination of each tissue and each disease with at least 50 cases and 50 controls annotated in GTEx, resulting in 117 disease-related tissue-disease pairs across 35 tissues and 9 diseases (with at least 10 differentially expressed genes (DEGs), FDR < 0.05). In these 117 disease-related pairs, 112 pairs exhibited both high sign consistency (sign consistency > 0.7) and high correlation (Pearson's $\rho > 0.3$) of $\log_2$ fold changes of pGenes, among which DEGs showed even higher concordance (Supplementary Table 3), suggesting that MTM captured both the direction and the level of the disease-related expression changes. Among 109 pairs with at least 10 observed upregulated DEGs, the predicted DEGs significantly overlapped (hypergeometric test, $P < 0.05$) with the reference DEGs in 93 pairs ($F1_{median} = 0.65$, F1 score > 0.7 in 44 pairs). For downregulated DEGs, the predicted DEGs significantly overlapped with reference DEGs in 94 pairs ($F1_{median} = 0.60$, F1 score > 0.7 in 41 pairs) among 112 affected pairs (with at least 10 observed downregulated DEGs, Fig. 4c).

To assess whether the predicted tissue expression profiles could be employed in indicating disease status, we next compared the prediction performance of the predicted tissue expression, the actual tissue expression and the input blood expression profiles. Among the 117 disease-related pairs mentioned above, 53 pairs were focused where the actual tissue expression profiles were more informative than those of blood for disease status prediction with *areas under the ROC curve* (auROCs) of at least 0.6. Although less predictive (Wilcoxon paired test, $P = 1.82\times10^{-5}$) than the actual tissue expression ($auROC_{mean} = 0.66$), the mean auROCs of the predicted tissue expression of pGenes across the 53 pairs ($auROC_{mean} = 0.62$) were significantly (Wilcoxon paired test, $P = 4.03\times10^{-4}$) higher than those of the original blood expression ($auROC_{mean} = 0.59$), suggesting that predicted tissue expression profiles were informative for disease status (and Supplementary Fig. 8).

MTM facilitates the identification of tissue-specific disease-associated dysregulations

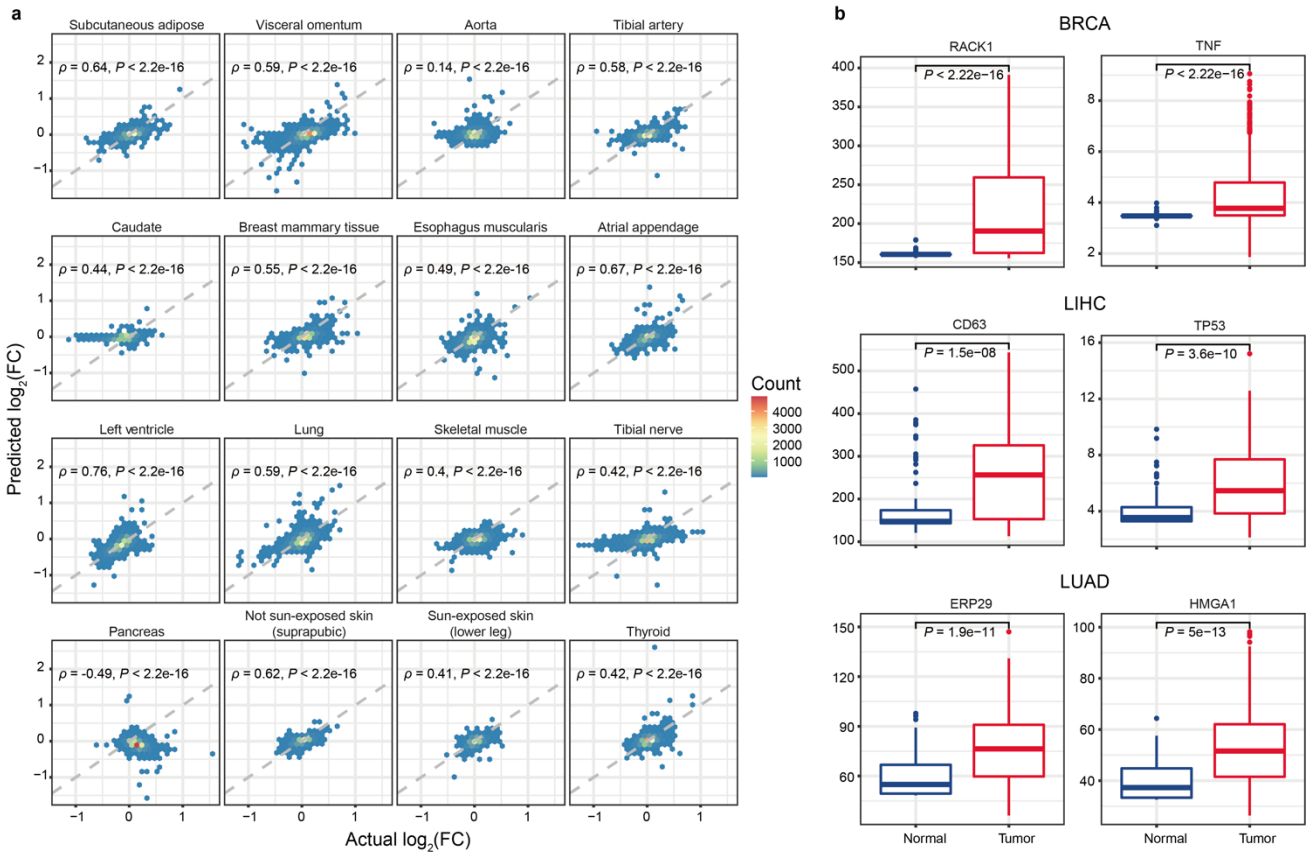

Fig. 5 Potential utilities in analysing disease-related gene signatures in external data. a Hexagonal heatmaps show the relations between the $\log_2$ fold changes of the predicted tissue expression (using blood from external T2D patients) and the reference $\log_2$ fold changes (GTEx T2D subsets). Hexagonal cells are coloured according to numbers of genes (count). **b** Examples of the expression levels of the reported blood biomarkers in the predicted BRCA (breast cancer biomarkers), LIHC (liver cancer biomarkers) and LUAD (lung cancer biomarkers) blood expression profiles.

Since MTM could capture individualized biological information from any input tissue expression profiles, we further explored the potential application scenarios for MTM. We first investigated whether MTM could predict disease-related dysregulations of tissues from blood expression on an external dataset. The external blood expression data were input into MTM to predict the expression profiles of other tissues, which were then used to perform DE analysis and were compared with the GTEx reference (with at least 50 cases and 50 controls). The dataset includes 25 type 2 diabetes (T2D) patients

and 33 normal subjects [33]. Among 16 affected tissues (with at least 10 observed up- or downregulated DEGs under FDR < 0.05 within GTEx), the predicted T2D-related expression changes (fold changes) exhibited high concordance with the reference data (Pearson's ρ ranged from -0.49 to 0.76, median = 0.52, Fig. 5a), which were significantly higher (Wilcoxon rank sum test, $P = 0.03$) than those of 11 unaffected tissues (Pearson's ρ ranged from -0.54 to 0.65, median = 0.32).

Another potential application of MTM is to predict blood expressions from other tissues, which may offer opportunities for data reanalysis or screening liquid biopsy markers when blood samples are temporarily unavailable. We used tumour and control tissue samples of tumours and controls from the Cancer Genome Atlas (TCGA)[34] to predict the corresponding blood expression levels. Three cancer types with at least 50 cases and 50 controls, including BRCA, LIHC, and LUAD, were included in the following analysis. The expression levels of reported blood biomarkers of these cancers[35] were estimated in the predicted blood expression of pGenes. On average, 83.5% of biomarkers were successfully validated (FDR < 0.05, $\log_2$FC > 0) in the three cancers. In the predicted blood expression of BRCA and LUAD subgroups, 13 out of 18 and 18 out of 23 reported blood biomarkers were significantly upregulated, respectively. Notably, among the 11 reported blood-based biomarkers of LIHC, all of them were significantly upregulated as expected in DE analysis (Fig. 5b and Supplementary Fig. 9–11).

Discussion

Mapping biological information from peripheral 'surrogate' samples to tissue transcriptome profiles has emerged as an effective alternative with potential utility in areas of biomedical science, especially for circumstances when tissue samples are unavailable. In this work, we have proposed MTM, a deep-learning-based multi-task learning framework that enables the prediction of individualized tissue gene expression profiles from any available gene expression of the same individual using a single unified model. Through the nonlinear representation capability of neural networks and by fully leveraging information from all tissues of the same individual via the multi-task learning architecture, MTM not only bypasses the need to establish an independent model for one mapping direction (such as blood to one brain area), but also substantially enhances the performance on both the overall sample levels and fine-grained gene levels. In particular, MTM identifies more predictable genes with well-preserved personalized physiological and pathological characteristics, providing a novel and valuable tool for tissue transcriptome-based biomedical investigations from a more comprehensive and systemic perspective.

The substantial improvement in performance for MTM may be attributed to several factors. First, it is commonly recognized that an individual's gene expression levels in different tissues are inherently linked by shared biological foundations, which could be decomposed into the static part that shares the same genetic material across tissues (i.e., GReX) and the dynamic part that is related to the temporal and spatial factors of the individual, including physiological states as well as environmental and other factors. The multi-task learning design of MTM encourages the exploitation of both the static and dynamic intrinsic relevance in all tissues from the same individual, enabling the model to extract and refine decoding rules with biological patterns and to produce better individualized representations instead of solely twisting data, thus leading to improved predictive performance, especially in tissues with smaller sample sizes. Interestingly, MTM performed worst in the two cell lines, with the lowest numbers of pGenes being captured, indicating that the integrity of internal information for the individual was lost in the cell lines that were cultured or transformed *in vitro*. Second, MTM

may be more capable of capturing complex gene–gene interactions with nonlinearity and multi-task components. Last but not least, predictive information, including subtle signals, is preserved to the greatest extent in MTM, since dimension reduction is not required for transcriptomic data in large-scale neural networks driven by the backpropagation algorithm[36]. With these properties, the framework of MTM may serve as an effective paradigm in similar biomedical scenarios where complex intrinsic relationships underlie multiple entities, such as other large-scale omics data of different tissues or cells from the same individuals.

The high accuracy of MTM for cross-tissue expression prediction with well-preserved individualized biological information supports its potential for biomedical applications, especially in tissue-specific biomarker discovery in liquid biopsy. However, there are still gaps that remain to be filled. A major challenge is to remove the heterogeneity in real data to align with the fixed develop sets for deployed machine learning models in practice, which imposes difficulties on existing data integration methods that cannot keep the develop sets intact. Another challenge is that the multi-task learning architecture of MTM highly depends on large-scale computational resources, which limits the incorporation of additional omics data. With the future advances of computational resources and data integration algorithms, as well as new incoming data to support continuous learning, MTM will promote real-world translational applications, such as decoding molecular mechanisms and mining clinical biomarkers.

Methods

Data preparation for MTM

The gene expression data from human tissues were downloaded from the v8 release of the Genotype-Tissue Expression (GTEx) portal (https://www.gtexportal.org/)[32]. The metadata, including the subject-level and sample-level information, were obtained through the database of Genotypes and Phenotypes (dbGaP) (https://www.ncbi.nlm.nih.gov/gap/). Individuals whose blood RNA-seq and genotype data were available were included in our study. The expression values (transcripts per million, TPM) of 19,291 protein-coding genes from 17,329 samples

(49 tissues with sample sizes greater than 50) across 948 individuals were used for model development. The tissue expression profiles were randomly split into the training set (80%) and the validation set (20%) based on individual labels. Z-score standardization was first conducted within each tissue for the training data, and then the recorded scaling factors were used for the validation data to prevent information leakage.

Description of MTM

Key elements. The MTM architecture consists of an encoder, a generator, a discriminator, and a mapping network. The encoder (*E*) embeds expression profiles into a shared latent space, resulting in personalized representations, i.e., latent codes. The generator (*G*) is symmetrical to the structure of the encoder, through which the latent codes were mapped into tissue-specific expression spaces in the form of individualized tissue expression profiles. The discriminator (*D*) is employed to identify the mismatches of the predicted distributions and the actual distributions to send feedback signals for the generator to improve the prediction in an adversarial manner[37]. The auxiliary mapping network (*M*) assists *E* in learning a smooth space[38]. All *E*, *G* and *D* networks are multi-task networks equipped with TCMs, which were designed to route and transform the data through tissue-specific or shared connections. The TCM consists of three parts: a shared fully connected layer and a learnable instance-level affine transformation conditioned on tissue labels followed by a leaky rectified linear unit (Leaky ReLU) to introduce nonlinearity. The detailed structure of the MTM is described in Supplementary Fig. 12–16.

Objectives. The generator (*G*) translates latent codes from a shared latent space ($c \in \mathcal{W}$) to predict the expression profiles conditioned on the target tissue *t*, $G(c|t) \rightarrow x_t$. The latent codes could either be reference-guided by *E* ($E(x|s) \rightarrow c$) or transformed from random Gaussian noise $z \sim \mathcal{N}(0,1)$ by *M* ($M(z) \rightarrow \tilde{c}$). To enforce *G* to utilize the latent code $\tilde{c}$ when predicting the expression profile $G(\tilde{c}|t)$, we employed the L1 reconstruction loss:

$$\mathcal{L}_R = \mathbb{E}[\|E(G(\tilde{c}|t)|t) - \tilde{c}\|_1]. \quad (1)$$

To minimize the divergence between the distributions of the actual expression profiles and of the predicted ones ($G(\cdot$

)), a training strategy from a generative adversarial net (GAN) is used, where *D* tries to distinguish the generated samples from real samples, while *G* tries to fool *D*. Instead of the original GAN, which minimizes the Jensen–Shannon divergence [37], we used an adversarial objective function in the form of the hinge loss for robust training[39, 40]:

$$\mathcal{L}_{adv}(D) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\max(0, 1 - D(\boldsymbol{x}|\boldsymbol{t}))] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\max(0,1 + D(G(\boldsymbol{z}|\boldsymbol{t})))], \tag{2}$$

$$\mathcal{L}_{adv}(G) = -\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[D\left(G(\boldsymbol{z}|\boldsymbol{t})\right)\right]. \tag{3}$$

To achieve individualized tissue expression prediction, given the expression profile of the source tissue x_s from a specific individual, we use the L1 loss to penalize the difference between the predicted expression profiles $G(E(x_s|s)|t)$ and the actual expression profile of the target tissue x_t:

$$\mathcal{L}_I = \mathbb{E}[\|G(E(x_s|s)|t) - x_t\|_1]. \tag{4}$$

To ensure that the predicted expression profile properly preserves the individual invariant characteristics, a cycle consistency loss is employed:

$$\mathcal{L}_C = \mathbb{E}\left[\left\|G\left(E\left(G(E(x_s|s)|t)\right)\middle|s\right) - x_s\right\|_1\right]. \tag{5}$$

The final objective function is as follows:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_R\mathcal{L}_R + \lambda_I\mathcal{L}_I + \lambda_C\mathcal{L}_C.$$

Training hyperparameters. The data were split into mini-batches of 256 samples for training. The Adam optimizer[41] was employed to optimize the network parameters using the following hyperparameters: learning rate = 0.001, β1 = 0.5 and β2 = 0.9. The number of maximum training epochs was set to 200. All models were trained using the PyTorch framework (version 1.10.2)[42] on Tesla V100S graphics processing units (NVIDIA).

Description of intermediate model S3 for ablative experiments

The S3 model consisted of a set of nonlinear neural networks for each tissue based on simple multilayer perceptrons.

For each tissue, a fully connected neural network with Leaky ReLU activations was built to predict the expression profiles of the target tissue from the expression of a specific source (blood). The number of maximum training epochs was set to 1000. Other training hyperparameters were kept the same as those of MTM.

Performance evaluation

The prediction quality was indicated by the Pearson correlation coefficient (Pearson's ρ) between the predicted and observed expression data. The sample-wise accuracies across all genes were calculated on Z-score standardized data to highlight the individual variance in each tissue. The gene-wise accuracies refer to Pearson's ρ of each gene across individuals for each tissue. Predictable genes were defined as genes with gene-wise $\rho > 0.3$ between predicted and observed expression in each tissue. Accuracy indicators in this work were reported in a fivefold cross-validation (CV) manner. For fair comparison, the performance of S1, S2 and S3 was evaluated in a CV manner using the same inputs as were used with MTM (blood expression data from the same training and validation samples).

Exploration of the characteristics of predictable genes

Tissue similarity. Similarities between tissues were measured with Pearson's ρ between the mean expression levels across genes (log2-scaled) of each tissue. When comparing certain pairwise statistics (x1) of tissues with the tissue similarities (x2), element-wise values were flattened to two arrays (X1 and X2), and then Spearman's correlation coefficient between the two was calculated.

Expression levels. In each tissue, the Wilcoxon rank sum test was used to compare the mean expression levels of the top 25 percentiles of predictable genes (highly predictable genes) with those of the control groups, with the control groups defined as the bottom 25 percentiles of predictable genes (unpredictable genes).

Conservation. The conservation distributions of genes in terms of PhastCons scores were extracted with DeepTools[43]

from the 100-way vertebrate species alignment downloaded from the UCSC (University of California Santa Cruz) genome browser[44, 45]. The conservation levels of the highly predictable genes were compared with those of the unpredictable genes in each tissue (Wilcoxon rank sum test).

Connectivity. The degree of connectivity distribution for each gene was extracted from the downloaded PIN. Then, the average degrees of the two groups of genes (highly predictable genes and unpredictable genes) were compared. The average degrees of the highly predictable genes were compared with those of the unpredictable genes in each tissue (Wilcoxon rank sum test).

Exploration of the characteristics of the intermediate features of MTM

Personalized representations from the latent space of the MTM encoder were extracted across different individuals and tissues and then used to calculate pairwise similarities (Pearson's correlation coefficients). Intraindividual codes refer to latent codes of different tissues from the same individual, while interindividual codes refer to latent codes from different individuals.

Decoding paths of the data flow from latent codes to tissue expression spaces refers to activations of different layers through the MTM generator. The decoding paths of each layer from an individual's code from blood to different tissues were used to calculate pairwise similarities (Pearson's correlation coefficients). Spearman's correlation coefficients between the pairwise similarities of the decoding paths and the pairwise tissue similarities were calculated to determine the extent to which the decoding rules reflected biological patterns.

Phenotype-related analysis

For traits including age, gender and BMI, Pearson's correlation coefficients were used to measure the associations of each gene with traits in each tissue. Then, Spearman's correlation coefficients between the predicted and actual (GTEx)

369 associations with traits were calculated to determine the extent to which the predicted tissue expression preserved the trait-
370 related expression changes.

371 For disease-associated dysregulations, the sign consistency and Pearson's correlation coefficient between $\log_2$ fold
372 changes across genes in predicted and actual data were used to measure the overall concordance of the prediction.
373 Differential expression analysis was conducted using the nonparametric Wilcoxon rank sum test. When comparison of
374 differentially expressed genes (DEGs, under FDR < 0.05) between predicted and actual data was applicable,
375 hypergeometric tests were conducted to examine the overlap significance, and the F1 scores, which convey the balance
376 between precision and recall, were used to measure the performance of identifying DEGs.

377 For disease status prediction, we built least absolute shrinkage and selection operator (LASSO) models with scikit-
378 learn (version 0.24.2)[46] for each disease-tissue pair in the predicted expression, the actual expression and the input blood
379 expression. The values of the *area under the ROC curve* (auROC) were used to indicate performance in a fivefold CV
380 manner with 50 independent repetitions.

381

382 Exploration of MTM applications on external datasets

383 For blood expression profiles, the raw sequencing data (fastq files) of GSE184050[33] were downloaded, and the gene
384 expression values of the 19,291 protein-coding genes were quantified according to the pipeline provided by the GTEx
385 Consortium[32]. For tissue expression profiles, the expression levels of TCGA samples were downloaded from the GDC
386 Data Portal (https://portal.gdc.cancer.gov/)[34, 47]. Uncovered genes were filled with the average values for each tissue. To
387 reduce batch effects, we simply aligned the mean and standard deviation of each gene in each tissue to those of the GTEx
388 data. For external datasets, the MTM was trained in a hold-out manner, with 80% of the data used for training and the
389 remaining 20% for testing.

390 The standardized blood expression profiles of GSE184050 (type 2 diabetes, T2D) were input into MTM to predict

the tissue expression profiles of corresponding individuals. Then, Pearson's correlation coefficients between $\log_2$ fold changes across genes in predicted data and reference data (case and control subjects for T2D in GTEx) were used to measure the overall concordance of the predicted disease-related dysregulations.

TCGA tissue samples for breast cancer (BRCA), liver and intrahepatic bile ducts (LIHC) and bronchus and lung (LUAD) with sample sizes of both case and control subjects greater than 50 were input into the MTM to predict the blood expression profiles of corresponding individuals. Then, the known blood-based biomarkers for breast, liver and lung cancer that were collected from[35, 48] were extracted for differential expression analysis.

References

1. F. S. Collins and H. Varmus. A new initiative on precision medicine. *N. Engl. J. Med.* 372, 793-5 (2015).
2. W. Ji, W. Zhou, K. Gregg, N. Yu, S. Davis and S. Davis. A method for cross-species gene expression analysis with high-density oligonucleotide arrays. *Nucleic Acids Res.* 32, e93 (2004).
3. K. Lage, N. T. Hansen, E. O. Karlberg, A. C. Eklund, F. S. Roque, P. K. Donahoe, Z. Szallasi, T. S. Jensen and S. Brunak. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20870-5 (2008).
4. A. Kutsenko, T. Svensson, B. Nystedt, J. Lundeberg, P. Bjork, E. Sonnhammer, S. Giacomello, N. Visa and L. Wieslander. The Chironomus tentans genome sequence and the organization of the Balbiani ring genes. *BMC Genom.* 15, 819 (2014).
5. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-7 (1999).
6. J. Irgon, C. C. Huang, Y. Zhang, D. Talantov, G. Bhanot and S. Szalma. Robust multi-tissue gene panel for cancer detection. *BMC Cancer* 10, 319 (2010).
7. K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. M. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov, J. Zhang, C. Kandoth, R. Akbani, H. Shen, L. Omberg, A. Chu, A. A. Margolin, L. J. Van't Veer, N. Lopez-Bigas, P. W. Laird, B. J. Raphael, L. Ding, A. G. Robertson, L. A. Byers, G. B. Mills, J. N. Weinstein, C. Van Waes, Z. Chen, E. A. Collisson, N. Cancer Genome Atlas Research, C. C. Benz, C. M. Perou and J. M. Stuart. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158, 929-944 (2014).
8. E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E. Eyler, J. C. Denny, G. T. Consortium, D. L. Nicolae, N. J. Cox and H. K. Im. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091-8 (2015).
9. S. Mohr and C. C. Liew. The peripheral-blood transcriptome: new insights into disease and risk assessment. *Thends Mol. Med.* 13, 422-32 (2007).
10. C. C. Liew, J. Ma, H. C. Tang, R. Zheng and A. A. Dempsey. The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. *J. Lab. Clin. Med.* 147, 126-32 (2006).
11. J. W. Halloran, D. Zhu, D. C. Qian, J. Byun, O. Y. Gorlova, C. I. Amos and I. P. Gorlov. Prediction of the gene expression

in normal lung tissue by the gene expression in blood. *BMC Med. Genom.* 8, 77 (2015).

12. J. Wang, E. R. Gamazon, B. L. Pierce, B. E. Stranger, H. K. Im, R. D. Gibbons, N. J. Cox, D. L. Nicolae and L. S. Chen. Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx. *Am. J. Hum. Genet.* 98, 697-708 (2016).
13. W. Xu, X. Liu, F. Leng and W. Li. Blood-based multi-tissue gene expression inference with Bayesian ridge regression. *Bioinformatics* 36, 3788-3794 (2020).
14. M. Basu, K. Wang, E. Ruppin and S. Hannenhalli. Predicting tissue-specific gene expression from whole blood transcriptome. *Sci. Adv.* 7, eabd6991 (2021).
15. L. Von Bertalanffy. Quantitative laws in metabolism and growth. *Q. Rev. Biol.* 32, 217-31 (1957).
16. J. Hasty, D. McMillen, F. Isaacs and J. J. Collins. Computational studies of gene regulatory networks: in numero molecular biology. *Nat. Rev. Genet.* 2, 268-79 (2001).
17. Y. Hu, M. Li, Q. Lu, H. Weng, J. Wang, S. M. Zekavat, Z. Yu, B. Li, J. Gu, S. Muchnik, Y. Shi, B. W. Kunkle, S. Mukherjee, P. Natarajan, A. Naj, A. Kuzma, Y. Zhao, P. K. Crane, H. Lu, H. Zhao and C. Alzheimer's Disease Genetics. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.* 51, 568-576 (2019).
18. A. N. Barbeira, M. Pividori, J. Zheng, H. E. Wheeler, D. L. Nicolae and H. K. Im. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* 15, e1007889 (2019).
19. D. Zhou, Y. Jiang, X. Zhong, N. J. Cox, C. Liu and E. R. Gamazon. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nat Genet* 52, 1239-1246 (2020).
20. R. Lopez, J. Regier, M. B. Cole, M. I. Jordan and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* 15, 1053-1058 (2018).
21. C. Maj, T. Azevedo, V. Giansanti, O. Borisov, G. M. Dimitri, S. Spasov, I. Alzheimer's Disease Neuroimaging, P. Lio and I. Merelli. Integration of Machine Learning Methods to Dissect Genetically Imputed Transcriptomic Profiles in Alzheimer's Disease. *Front. Genet.* 10, 726 (2019).
22. M. Marouf, P. Machart, V. Bansal, C. Kilian, D. S. Magruder, C. F. Krebs and S. Bonn. Realistic *in silico* generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat. Commun.* 11, 166 (2020).
23. J. Park, H. Kim, J. Kim and M. Cheon. A practical application of generative adversarial networks for RNA-seq analysis to predict the molecular progress of Alzheimer's disease. *PLoS Comput. Biol.* 16, e1008099 (2020).
24. T. Azevedo, G. M. Dimitri, P. Lio and E. R. Gamazon. Multilayer modelling of the human transcriptome and biological mechanisms of complex diseases and traits. *NPJ Syst. Biol. Appl.* 7, 24 (2021).
25. R. Vinas, H. Andres-Terre, P. Lio and K. Bryson. Adversarial generation of gene expression data. *Bioinformatics* 38, 730-737 (2021).
26. R. Viñas, T. Azevedo, E. R. Gamazon and P. Liò. Deep Learning Enables Fast and Accurate Imputation of Gene Expression. *Front. Genet.* 12, 624128 (2021).
27. H. A. Elmarakeby, J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S. H. AlDubayan, K. Salari, S. Kregel, C. Richter, T. E. Arnoff, J. Park, W. C. Hahn and E. M. Van Allen. Biologically informed deep neural network for prostate cancer discovery. *Nature* 598, 348-352 (2021).
28. R. Caruana. Multitask Learning. *Mach. Learn.* 28, 41-75 (1997).
29. S. Ruder. An Overview of Multi-Task Learning in Deep Neural Networks. Preprint at https://arxiv.org/abs/1706.05098 (2017).
30. Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. Preprint at Preprint at https://doi.org/10.48550/arXiv.1711.09020 (2017).
31. M. Yang, W. Zhao, W. Xu, Y. Feng, Z. Zhao, X. Chen and K. Lei. Multitask Learning for Cross-Domain Image Captioning. *IEEE Trans. Multimedia* 21, 1047-1061 (2019).
32. G. T. Consortium, D. A. Laboratory, G. Coordinating Center -Analysis Working, G. Statistical Methods groups-Analysis Working, G. g. Enhancing, N. I. H. C. Fund, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, N. Biospecimen Collection Source

Site, R. Biospecimen Collection Source Site, V. Biospecimen Core Resource, B. Brain Bank Repository-University of Miami Brain Endowment, M. Leidos Biomedical-Project, E. Study, I. Genome Browser Data, E. B. I. Visualization, I. Genome Browser Data, U. o. C. S. C. Visualization-Ucsc Genomics Institute, a. Lead, D. A. Laboratory, C. Coordinating, N. I. H. p. management, c. Biospecimen, Pathology, Q. T. L. m. w. g. e, A. Battle, C. D. Brown, B. E. Engelhardt and S. B. Montgomery. Genetic effects on gene expression across human tissues. *Nature* 550, 204-213 (2017).

33. H. H. Chen, L. E. Petty, K. E. North, J. B. McCormick, S. P. Fisher-Hoch, E. R. Gamazon and J. E. Below. Novel diabetes gene discovery through comprehensive characterization and integrative analysis of longitudinal gene expression changes. *Hum. Mol. Genet.* 31, 3191-3205 (2022).
34. N. Cancer Genome Atlas Research, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander and J. M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45, 1113-20 (2013).
35. L. Uttley, B. L. Whiteman, H. B. Woods, S. Harnan, S. T. Philips, I. A. Cree and C. Early Cancer Detection. Building the Evidence Base of Blood-Based Biomarkers for Early Detection of Cancer: A Rapid Systematic Mapping Review. *EBioMedicine* 10, 164-73 (2016).
36. Y. LeCun, Y. Bengio and G. Hinton. Deep learning. *Nature* 521, 436-444 (2015).
37. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio. Generative adversarial networks. *Commun. ACM* 63, 139–144 (2020).
38. T. Karras, S. Laine and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. Preprint at https://arxiv.org/abs/1812.04948 (2018).
39. J. H. Lim and J. C. Ye. Geometric GAN. Preprint at https://arxiv.org/abs/1705.02894 (2017).
40. J. Gui, Z. Sun, Y. Wen, D. Tao and J. Ye. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Trans. Knowl. Data Eng.* 1-1 (2021).
41. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
42. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala. PyTorch: an imperative style, high-performance deep learning library. In Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS, 2019).
43. F. Ramirez, D. P. Ryan, B. Gruning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dundar and T. Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 44, W160-5 (2016).
44. A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034-50 (2005).
45. M. Haeussler, A. S. Zweig, C. Tyner, M. L. Speir, K. R. Rosenbloom, B. J. Raney, C. M. Lee, B. T. Lee, A. S. Hinrichs, J. N. Gonzalez, D. Gibson, M. Diekhans, H. Clawson, J. Casper, G. P. Barber, D. Haussler, R. M. Kuhn and W. J. Kent. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* 47, D853-D858 (2019).
46. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, pp 2825–2830 (2011).
47. R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe and L. M. Staudt. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* 375, 1109-12 (2016).
48. Z. Zuo, H. Hu, Q. Xu, X. Luo, D. Peng, K. Zhu, Q. Zhao, Y. Xie and J. Ren. BBCancer: an expression atlas of blood-based biomarkers in the early diagnosis of cancers. *Nucleic Acids Res.* 48, D789-D796 (2020).