Pathway-enhanced Transformer-based model for robust enumeration of cell types from the cell-free transcriptome

Shuo Yan[1,2†], Xuetao Tian[3†], Yulong Qin[1,2], Jiayi Li[1,2], Tingyu Yang[1,2], Dongyue Yu[4], Wentao Liu[1,2], Jinghua Sun[2], Chengbin Hu[2], Qing Zhou[2], Zhongzhen Liu[2] and Wen-jing Wang[2*]

[1]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China.
[2*]BGI Research, Building 11, Beishan Industrial Zone, Yantian District, Shenzhen, 518083, China.
[3]Faculty of Psychology, Beijing Normal University, Beijing, 100875, China.
[4]College of Life Sciences, Nankai University, Tianjin, 300071, China.

*Corresponding author(s). E-mail(s): wangwenjing@genomics.cn;
†These authors contributed equally: Shuo Yan and Xuetao Tian

Abstract

Analysis cell types origin of cell-free RNA can enhance the resolution of liquid biopsies, thereby deepening the understanding of molecular and cellular changes in development and disease processes. Existing deconvolution methods typically rely on meticulously curated gene expression profiles or employ deep neural network with vast and complex solution spaces that are difficult to interpret. These approaches fail to leverage the synergistic and co-expression effects among genes in biological signaling pathways, compromising their generalizability and robustness. To address this issue, we have developed 'Deconformer', a Transformer-based deconvolution model that integrates biological signaling pathways at the embedding stage. Compared to popular methods on multiple datasets, Deconformer demonstrates superior performance and robustness, and is capable of tracking the developmental process

of the placenta. Additionally, Deconformer's self-attention mechanism has identified a connection between platelet activation, dependencies with other pathways, and the severity of COVID-19. We believe that Deconformer will enable and accelerate the precise analysis of a wide range of cell-free RNA, describing disease progression and severity from the perspectives of originating cell fractions and pathway dependencies.

Keywords: Deconvolution, Transformer, Biological pathway, Interpretability, cell-free RNA

1 Introduction

Cell-free RNA (cfRNA) is a mixture of transcripts continuously released into body fluids by tissues throughout the body[1]. Previous researches have highlighted cfRNA's potential as a powerful biomarker for monitoring, prediction, diagnosis and prognosis[2–5], with its detectability in biofluids like plasma, saliva, and urine underscoring this point[6]. Analysis of the origins of cfRNA is crucial for cancer diagnosis and monitoring the development and progression of diseases and organs[2, 4, 7], aiding in understanding diseases and development in conjunction with cellular pathophysiology. Existing studies have used tissue-specific gene expression profiles (GEPs) for deconvolution of cfRNA expression profiles to trace their origins, but this is limited to a few cell types and fails to cover a full spectrum of most cells[1, 2, 4, 7]. Vorperian et al. (2022)[8] attempted to use CIBERSORTx[9] to generate cell-specific GEPs for deconvolution of plasma cfRNA (CSx-Vorperian), aiming to obtain relative fractions of cfRNA origins for a vast majority of cell types. However, their results varied significantly among individuals. This is due to genes with a high power to distinguish cell type, compared to universally expressed housekeeping genes, having lower expression levels[10], higher missing rates, and The regression-based deconvolution method highly reliant on GEP's low collinearity[11], the absence and bias of these specific genes in expression profiles greatly affect the accuracy of linear regression algorithms in deconvolution. To enhance robustness against input noise and technical biases, Menden et al. developed Scaden[12], a Deep Neural Network (DNN) method based on Multi-Layer Perceptron(MLP) for bulk deconvolution, using gene expression information to infer the cellular composition of tissues without relying on complex specific gene selection. However, the solution space of MLPs is vast and entirely data-driven, lacking an understanding of inherent correlations inside data, and optimization process is unstable. Although the hidden layers of DNNs learn higher-order latent representations of cell types, this makes interpreting the model's workings difficult. Also, most GEP-based methods require regression calculations for each sample[9], and Scaden needs separate models trained for every different expression profile[12], both incurring significant time costs. Additionally, these methods, aside from single-cell data, do not integrate

additional prior knowledge and use genes as the smallest unit, whereas gene interactions often exist in the form of networks. Single-gene analysis might miss important impacts on pathways, as physiological and cellular processes often affect groups of synergistically acting genes[13–15], with this synergy and co-expression effect being more pronounced in the release process of cfRNA. Previous research in single-cell annotation. TOSICA[16] incorporated pathway information to enhance embedding, but its computational complexity and massive parameter quantity limited the inclusion of numerous pathways or gene sets. and scBERT[17] using gene co-expression information for embedding, but encountered loss in continuous value embedding.

To address these challenges, we developed Deconformer (deconvolution transformer), a transformer-based[18] deconvolution model that integrates pathway information during the embedding process. Deconformer trained on 800,000 cfRNA samples simulated using 500,000 single-cell data from Tabula Sapiens version 5.0 (TSP)[19]. By utilizing pathway prior knowledge, Deconformer achieved lossless embedding of continuous gene expression values, and the addition of these synergistic and co-expressed pathway information endowed Deconformer with better depth representation capabilities. Moreover, its lightweight and efficient embedding design significantly increased the number of pathways or gene sets that could be included. Benefiting from the transformer's powerful self-attention mechanism and long-distance dependency capturing capability, Deconformer outperforms previous methods in terms of generalizability, noise and dropout resistance, and interpretability. Firstly, benchmark and case studies confirmed Deconformer's advantages in accuracy, robustness, and time performance. Once trained, Deconformer inferred the contribution fractions of source cells in cfRNA expression profiles from different sequencing technologies and bodily fluids, presenting variations in RNA released into the plasma by specific cells during disease progression. Secondly, after adding fetal placental data to the single-cell reference, Deconformer could infer the changes in cell types during the fetal development process in pregnancy. Lastly, by analyzing attention scores between pathways during the inference stage, we found significantly different pathways between diseased and healthy groups that could effectively cluster samples unsupervised. The differences in attention scores between pathways correlated strongly with disease, indicating that Deconformer's self-attention mechanism effectively captures pathways causing cell type differences, providing support for further biological discoveries.

Deconformer not only meets the needs of deconvolution across cfRNA expression profiles obtained through different technologies, surpassing existing methods in accuracy, but also generally reduces time costs and offers pathway-level interpretability. To our knowledge, Deconformer pioneers the application of transformer architecture in transcriptomic data deconvolution and innovatively integrates pathway information to design a lightweight, efficient, and lossless method for embedding gene expression values.

2 Results

2.1 The Deconformer algorithm

Deconformer is a cfRNA deconvolution tool based on the transformer encoder (Fig. 1). It learns to deduce cell type fractions through supervised training on simulated data, extracting insights from high-dimensional gene expression information by incorporating biological pathway and self-attention mechanisms.

Deconformer consists of three parts: Pathway-based embedding, Transformer module, and fractions prediction. Pathway-based embedding transforms cfRNA expression profiles into tokens, initially using GSEA C5 gobp pathway data and the 10,000 genes with the highest variance from simulated data to create a pathway mask matrix based on Jacobi similarity. The cfRNA expression profile is row-wise replicated to match the size of the mask matrix, and a Hadamard product is calculated with the pathway mask matrix to generate a masked expression matrix. Masking the expression profile with pathway information turns the original expression profile vector into a collection of vectors for each pathway. A fully connected weight matrix subsequently processes it, acting as a linear layer that maps the masked expression matrix, which incorporates both expression and pathway information, to a lower-dimensional space. This process embeds a cfRNA expression profile into tokens representing the number of pathways used, which is lossless and incorporates biologically meaningful expert knowledge into the tokens. In this way, Deconformer efficiently formalizes information on gene expressions and pathways for Transformer, generating a pathway-specific embedding that represents the pathway-specific expression after training. A trainable [CLS] token is then concatenated to the 0th position of these tokens, used in the subsequent self-attention mechanism to capture global dependency information and predict fractions scores. These tokens, including the [CLS] token, are inputted into the transformer encoder's multi-head self-attention layer. Due to the complexity and interactivity of biological processes, there exist subtle relationships between pathways. These relationships are captured through training on a large number of simulated samples by the self-attention mechanism, yielding attention scores. The [CLS] token, in each layer of self-attention computation, carries the information of all pathway tokens and gains global dependency information, which is then transformed through a fully connected matrix into a vector of cell type probabilities. The transformer has strong interpretability, and the self-attention mechanism can compute relationships between tokens representing objects. Deconformer computes attention scores between pathways. It utilizes not only the attention scores between the [CLS] token, used for predicting cell fractions, and the pathways for analysis but also employs scores from all pathway pairs to capture the dependencies between various pathways.

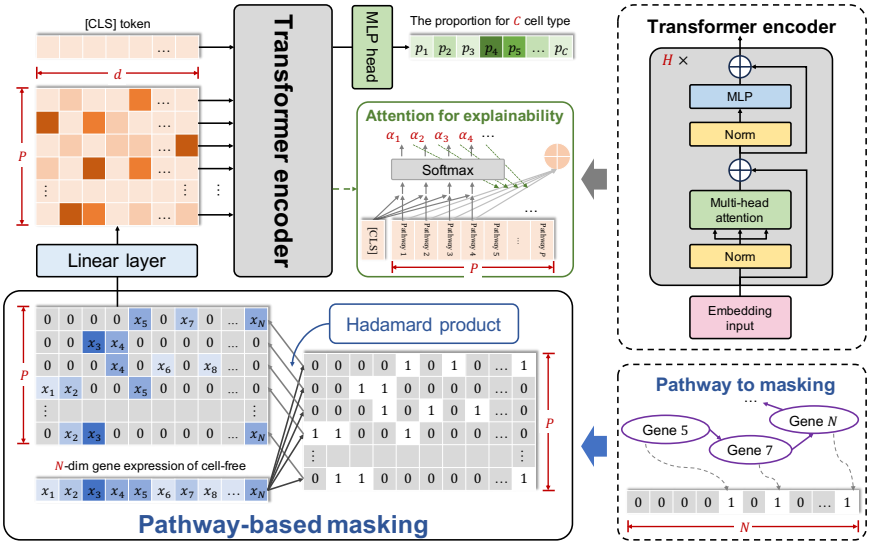

Fig. 1 The Deconformer model overview. For each input sample's expression profile, it is first extended via a row-wise replication to form a matrix of the same size as the mask matrix and then subjected to a Hadamard product with the mask matrix. The resulting matrix is then mapped to another space through a linear layer, and a [CLS] token is added at the head. After that, multi-head self-attention is calculated, and the [CLS] token, following a fully connected layer and softmax, yields the cell type scores.

2.2 Performance on simulated cfRNA and distinct body fluid cfRNA dataset

We compared Deconformer with the previously reported cfRNA deconvolution method CSx-Vorperian and Scaden, known for its superior performance on bulk data. CSx-Vorperian relies on manually curated GEP matrices but can also use single-cell data to generate corresponding GEP matrices. The Scaden method, as described in the original paper, was trained using simulated data consistent with Deconformer. We separately simulated 1000 cfRNA samples for performance testing. The results showed that Deconformer achieved the highest Lin's concordance correlation coefficient (CCC)[20], Pearson's correlation coefficient (r), and the lowest root mean square error (RMSE) on simulated data and under different dropout conditions (Fig. 2A). Given the non-negligible impact of dropout in cfRNA sequencing[21], we further tested the robustness of the method by evaluating the inference performance of the three methods under different dropouts. This was achieved by randomly dropping expression in these 1,000 simulated cfRNA profiles at rates of 0.1, 0.2, 0.3, 0.4, and 0.5. Deconformer still performed the best. Moreover, under different dropout rates, the performance degradation (loss) and its extent were smaller for Deconformer compared to Scaden (Fig. 2B, 2C). Notably, on simulated data, Deconformer's CCC reached over 99%, and even with a 50%

dropout rate, the CCC remained above 95%. As expected, the CSx-Vorperian regression-based method performed poorly on simulated data.

To validate the importance of pathway information in deconvolution tasks, we conducted a progressive perturbation experiment on the original pathway mask matrix. This experiment included two types of perturbations: reducing the non-zero values (i.e., pathway gene information) in the pathway mask matrix by 20%, representing partial loss of pathway gene information; and randomly adding an amount equivalent to 20% of the non-zero values as 1s at zero positions in the pathway mask matrix, simulating noise interference in pathway information. Using these two perturbation methods, we generated two perturbed matrices. Additionally, we created two completely random mask matrices with non-zero element ratios matching the perturbed matrices for comparison. Training Deconformer with these four different mask matrices, the results showed that the matrix with a 20% reduction in non-zero values performed best in expressions without dropout and with 50% dropout, yet was still inferior to the original pathway mask matrix, followed by the matrix with randomly added 20% 1s at zero positions. The completely random mask matrices consistently had CCC values below 0.8 across different training epochs, indicating poor performance that could not be improved by increasing the number of training iterations (Fig. 2D). This demonstrates that the potential co-expression information in the pathway mask can effectively enhance deconvolution performance, and the self-attention mechanism can better capture the true correlations between pathways.

To test Deconformer's generalization performance on cfRNA expression profiles generated from different body fluids, we conducted deconvolution on urine, saliva, seminal plasma, and blood plasma[17]. The results showed that Deconformer could identify the patterns of different body fluids, and the top five contributing cell types for each fluid source were more consistent with expectations compared to Scaden and CSx-Vorperian. For example, the primary sources of cfRNA in urine were the bladder and kidneys(Fig. 2E).

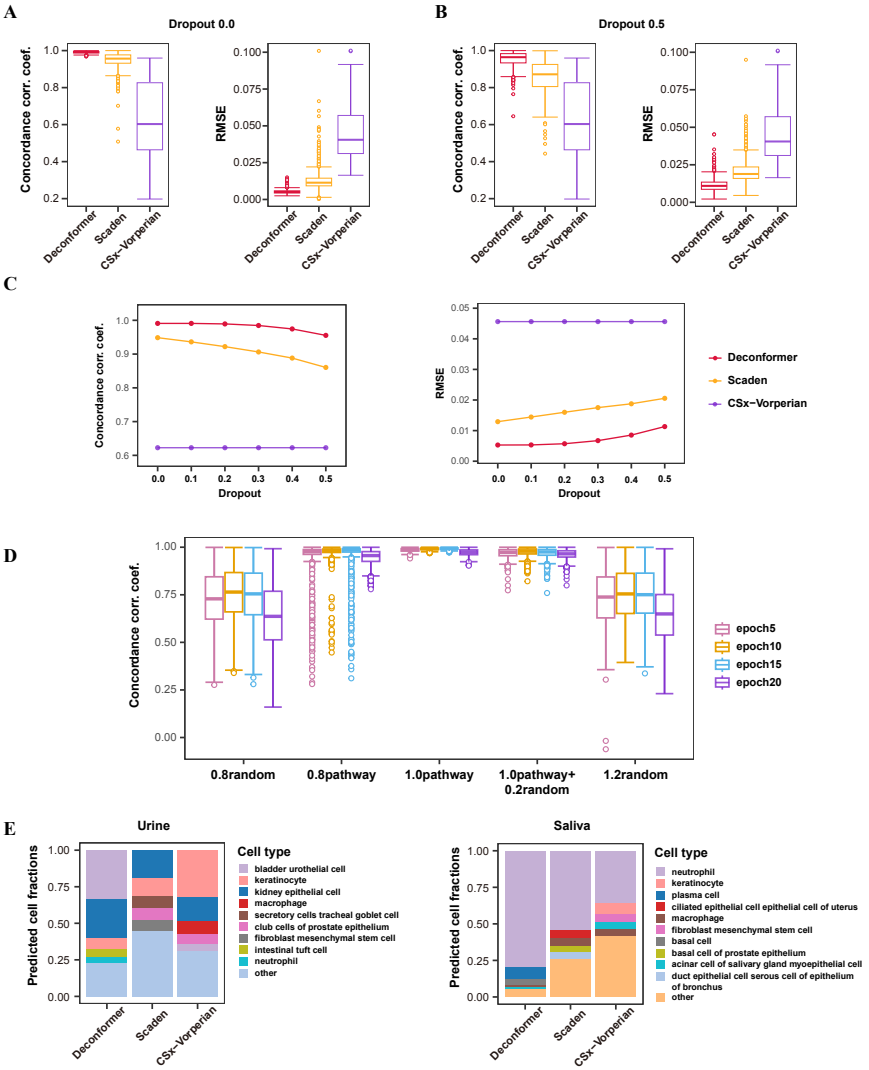

Fig. 2 Evaluation of accuracy and generalization ability on simulated cfRNA, as well as the importance of pathway information. A: Deconvolution performance of Deconformer, Scaden, and CSx-Vorperian on simulated data. B: Deconvolution performance of the three methods after dropping 50% of the simulated data expression profile. C: Deconvolution performance of the three methods on simulated data under different dropout conditions. D: Performance of Deconformer models trained with different perturbed and completely random mask matrices on simulated data. E: top-5 cell types and the sum of 'other' for different body fluids, urine(left), and saliva(right), across the three methods.

2.3 Accurate identification of cancer lesion differential signals

cfRNA shows high specificity in early cancer screening, presenting promising prospects[5,21]. Differentiating transcripts of lesion-specific cells from high-background transcripts originating from blood cells is crucial for early cancer screening and detection. We conducted tests on hepatocellular carcinoma, colorectal cancer, gastric cancer, multiple myeloma, and their precancerous states to demonstrate that Deconformer accurately identifies changes in cfRNA contribution fractions from tissues and organs directly affected by the disease or involved in solid and non-solid tumors, which are significant and stable. Generally, lesions affect the surrounding tissues and cell types, as well as the immune system, These cell types, anticipated to undergo changes in source fractions, are listed in the supplementary materials along with another category of cell types that show definite changes and have related cell counts. To emphasize the advantages and potential limitations of Deconformer, we compared its conclusive results only with CSx-Vorperian and Scaden. Firstly, we tested solid tumors; in hepatocellular carcinoma, using Deconformer to infer datasets from Chen et al. (2022)[22] and Roskams-Hieter et al. (2022)[23], we observed a significant increase in the liver fractions in the patient group across both datasets, a finding not detected by Scaden in the Roskams-Hieter et al. (2022) dataset, and although CSx-Vorperian showed a difference, the liver fractions for the majority of individuals in the control group were zero (Fig. 1A). In colorectal and gastric cancers, deconvolving datasets from Chen et al. (2022) and Tao et al. (2023)[24], we observed increased platelet counts in both cancer groups, consistent with thrombocytosis common in colorectal cancer patients; a decrease in B cells, stromal cells, immature intestinal epithelial cells, intestinal stem cells, and transitional proliferating cells of the colon, all closely related to the lifecycle of intestinal epithelia and playing critical roles in the self-renewal and repair processes of the intestines, proven to be directly related to colorectal cancer[25, 26]. In gastric cancer, both datasets showed a decrease in stromal cells, NK cells, monocytes, B cells, neutrophils, and duodenum glandular cells[27]. To further validate the cell pathophysiological significance and disease discrimination ability of the ability of the cell type fractions, we used the deconvolution results from Tao et al. (2023)'s data for classification modeling of two types of cancer. The results showed a performance comparable to the best multi-omics results in intestinal cancer and better than cfRNA abundance models in gastric cancer, possibly due to the reference cell types included.

In the multiple myeloma (MM) data of Roskams-Hieter et al. (2022)[23], we observed a significant increase in plasma cells and a marked rise in erythrocyte and erythroid progenitor in monoclonal gammopathy of undetermined significance (MGUS) group. In the MM group, we did not observe a significant increase in plasma cells, but we also noticed an increase in erythrocyte and erythroid progenitor (Fig. 3B). MM and MUGS generally present with anemia, but our observation of an increase could be due to the uncontrolled

abnormal proliferation of plasma cells and cancerous plasma cells in the bone marrow, compressing normal erythrocyte and erythroid progenitor, causing damage and apoptosis of these cells, leading to anemia and a corresponding increase in cfRNA signals in plasma. The increase in plasma cells in MUGS can well explain the disease's onset and progression. The absence of increased plasma cells in the MM led us to speculate significant changes in RNA expression between cancerous and normal plasma cells in MM patients, causing Deconformer to fail to recognize cancerous plasma cells.

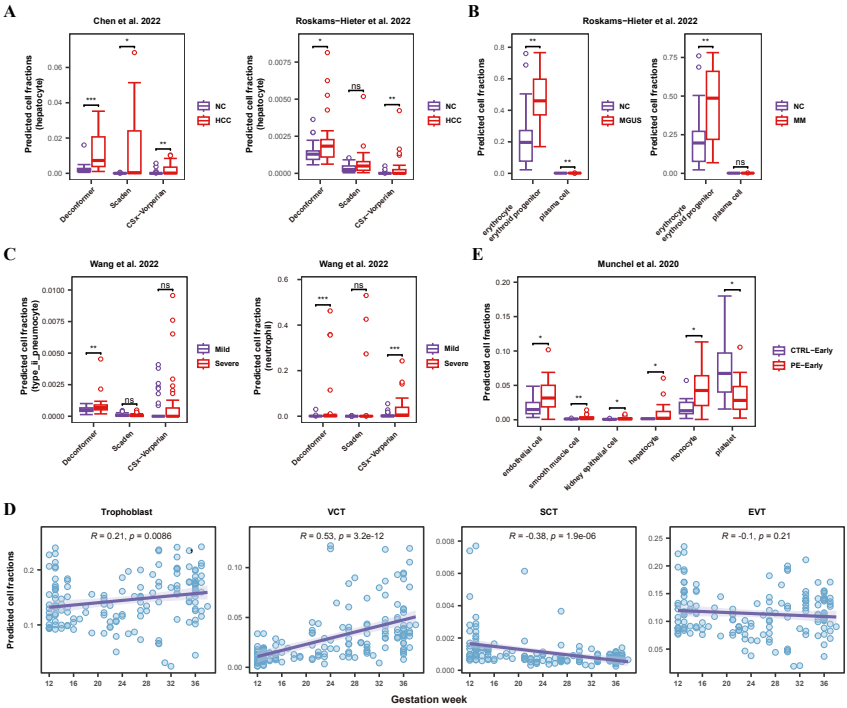

Fig. 3 Comparison of the ability to differentiate cell types on disease differences and Deconformer can track organ development. A: Capture of liver signal differences between liver cancer and healthy groups by Deconformer, Scaden, and CSx-Vorperian in the two liver cancer datasets of Chen et al. 2022 and Roskams-Hieter et al. 2022. B: Differences in cell type fractions scores in MGUS and MM as captured by Deconformer. C: Differences in type_ii_pneumocyte and neutrophils between severe and mild COVID-19 patients by Deconformer, Scaden, and CSx-Vorperian in Wang et al. 2022 data. D: The changes in fractions of trophoblast and VCT, EVT, SCT over gestational weeks. E: Differences in cell type fractions between early-onset preeclampsia and healthy individuals in Munchel et al. 2020 pregnancy data as captured by Deconformer.

2.4 Changes in affected organs and peripheral immunity due to viral infection

To further demonstrate Deconformer's ability to capture differences in disease progression at lesion sites and in peripheral immunity for various diseases, we applied Deconformer to datasets on COVID-19 infection by Wang et al. (2022) [28]and HBV infection by Sun et al. (2022)[29]. COVID-19, an infectious disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), primarily exhibits diffuse alveolar damage in severe cases[30]. Deconformer's deconvolution results indicated a higher fractions of type ii pneumocyte cells in severely ill patients compared to those with mild symptoms, a distinction not apparent in results from Scaden and CSx-Vorperian. Regarding changes in peripheral immunity, the referenced dataset[28] showed significantly higher neutrophil counts in severely ill COVID-19 patients than in those with mild symptoms, a finding consistent with both Deconformer and CSx-Vorperian results, whereas Scaden failed to detect any differences in neutrophils(Fig. 2C). Additionally, we observed an increase in monocytes[31] in severe cases in Deconformer's results, corroborating previous studies. The rise in T cell and NK cells, contrary to the previously reported decline in circulating T and NK cells in severe COVID-19[32–35], suggests that the depletion of immune cell subgroups may release more signals into the plasma. Besides these immune cells, we also noted a decrease in platelets in severely ill patients, aligning with the differences in blood cell counts reported in the corresponding studies. By deconvoluting the sample data from the study by Sun et al. [29] we observed the progressive changes in liver fractions across different serological phenotypes of HBV infection.Compared to normal individuals, patients with the negative HBeAg serological pattern showed increased liver fractions, but this increase was even more pronounced and significant in patients with the positive HBeAg pattern[29]. Additionally, in the positive HBeAg patient group, we observed a decrease in the fractions of NK cells and T cells. Existing research has suggested that these differences could be due to a reduction in NK cell fractions[36] and T cell exhaustion[37, 38], reflecting more severe liver damage and immune characteristics in patients with chronic infection, particularly those with the positive HBeAg serological pattern. Based on the above, we hypothesize that changes in peripheral immunity due to acute and chronic viral infections manifest differently in cfRNA expression profiles.

2.5 Detecting and tracking fetal organ and placenta development

The deconvolution of cfRNA offers a comprehensive means to monitor maternal and fetal organ health and to track fetal development. Moufarrej et al. [39] qualitatively observed changes in the mother and placenta during fetal development through cell and tissue-specific scores, but these results do not reflect the relative developmental changes between different organs of the individual. We included cells from the placental trophoblast layers SCT, EVT,

VCT[40], and the four main fetal organs—heart, liver, brain, kidney[41]—in the target cell types for Deconformer deconvolution. Considering the potential high collinearity between homologous fetal and maternal cells, detectability, and batch effect interference, we reduced the number of maternal cell types included, retaining only the top 25 types from healthy individuals in the pregnant women's dataset, along with the cell types encompassed by the fetal cell types, totaling 27 types. Finally, there were 4 fetal cell types, 3 placental trophoblast cell types, and 27 adult cell types, summing up to 34 types. Using Deconformer to deconvolve healthy pregnant women's plasma cfRNA profile [42], We observed that the fractions of VCT most notably increases with gestation, EVT does not show significant changes, SCT is on a downward trend. The overall fractions of the trophoblast (the sum of SCT, VCT, EVT) increase with gestation (Fig. 3D). Among fetal cells, the fractions of mesencephalon and hepatocyte cells increased significantly with gestation, while cardiomyocyte cells significantly decreased, and there was no significant change in the fractions of kidney epithelial cells. In the early-onset pre-eclampsia (PE) group, Deconformer showed significantly higher contributions of endothelial_cell, kidney_epithelial_cell, smooth_muscle_cell, hepatocyte, monocyte[43] compared to the control group, consistent with the symptoms of proteinuria, impaired liver function, renal insufficiency, and increased blood pressure in early-onset PE patients [39] (Fig. 3E), and an increased contribution of platelets, possibly related to thrombocytopenia in early-onset PE patients. But these changes and differences were not observed in the late-onset PE (Preeclampsia) group, which also corroborates, from the changes in affected organs, that early-onset PE is more severe than late-onset PE. It is noteworthy that in both early-onset and late-onset PE, the SCT fractions significantly increased, even though the absolute values of the SCT scores are very small. For preICP patients in the Sun et al. (2022)[29] dataset, using Deconformer with added fetal and placental cells, we found a significant increase in the liver fractions, consistent with reports. The Deconformer with adult reference cells yielded the same result, demonstrating Deconformer's robustness.

2.6 Deconformer can identify dependencies between pathways in cfRNA

Previous deconvolution methods struggled to adequately explain or provide insights into differences in cell type origin fractions. We obtained the last layer's attention scores from the inference stage of COVID and HCC samples. Our interpretability analysis is in two parts: first, we analyzed the attention scores between the [CLS] token used for calculating cell fractions scores and other pathways. Wang et al. (2022) [28] found that neutrophil activation–related processes are primarily enriched in severe patients. Deconformer includes seven such related pathways, among which five show significant differences under the Wilcoxon rank-sum test. This demonstrates that the attention score of the [CLS] token can effectively distinguish between severe and mild COVID-19 patients(Fig. 4A,4B). Next, we calculated the Cohen's d Effect of [CLS]

token attention scores for different groups, using pathways with an absolute value greater than 0.8 for self-clustering(Fig 4C). This method could better distinguish between severe and mild cases compared to similarly calculated differential genes, reflecting Deconformer's noise reduction capability to a certain extent.

To further explore interactions between pathways in plasma cfRNA, we considered the overall correlations and dependencies among 5,000 pathways. We performed Cohen's d Effect statistics on 25 million pathway pairs, selecting the top 100 pairs with the highest absolute Cohen's d Effect for pathway network analysis. We found that Regulation of Platelet Activation, which appeared most frequently among these differential pathway pairs, might influence many physiological processes related to severe COVID-19, consistent with previous reports[44–46]. Additionally, pathways such as Leukocyte Aggregation[47], Ras protein signal Transduction, and Positive Regulation of Myoblast Differentiation, which have been proven related to severe COVID-19, also showed high connectivity(Fig 4D). In the liver cancer data, we found interesting results, such as in liver cancer, pathways like Negative Regulation of Lipase Activation[48], Smooth Muscle Cell Apoptotic Process, and Serine Family Amino Acid Catabolic Process[49] showed the most dependencies with other pathways (Fig 4E). Then, we analyzed the differences in embeddings of pathways strongly related to diseases, finding that highly background-dependent pathways showed greater variability in embeddings across different sample states. These findings indicate that Deconformer can learn high background-dependent information at the embedding stage. The self-attention mechanism effectively learns the correlations between pathways. Additionally, for pathways and genes not identified in the expression profile, Deconformer can still effectively identify potential key pathways, offering a new perspective on the interactions between biological pathways in plasma cfRNA for diseases and discovering pathways of importance.

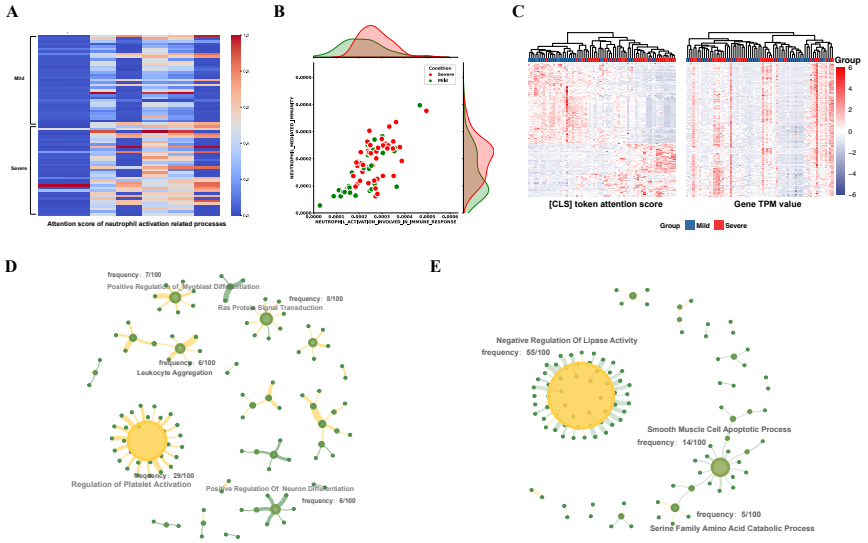

Fig. 4 The dependency relationships of pathways in plasma cfRNA have biological significance. A: Differences in attention scores between [CLS] token and neutrophil activation pathway in mild and severe COVID-19 cases. B: The attention scores of two neutrophil activation-related pathways with the [CLS] token can differentiate between severe and mild COVID-19 cases. C: Heatmaps of unsupervised clustering of significant differences in attention scores and gene expression profiles. D: Interactions and dependencies between pathways in plasma cfRNA of COVID-19 patients as identified by attention score between all pathway pairs. E: Interactions and dependencies between pathways in plasma cfRNA of liver cancer patients as identified by attention score between all pathway pairs.

3 Discussion

To address the robustness and accuracy of cfRNA deconvolution and the interpretability of critical pathways, we developed Deconformer. It learns patterns between pathway interactions and cell type fractions in simulated cfRNA gene expression profiles for inferring cell type fractions in real cfRNA expression spectra. Deconformer is the first to apply the transformer to transcriptomic data deconvolution. It explicitly incorporates prior knowledge into the model's embedding module, using gene groups with synergistic and co-expression effects as the smallest unit for modeling, rather than individual genes susceptible to random noise. Moreover, its embedding design is lightweight and efficient, accommodating nearly ten thousand pathways or gene sets, enhancing the model's comprehensive understanding of inter-pathway interactions.

Deconformer demonstrates a particularly notable robustness to noise and bias. With a single training session, it can infer expression profiles obtained from different bodily fluids and sequencing technologies, surpassing existing advanced methods in both simulated and multiple real datasets. Additionally, Deconformer can flexibly track placental and fetal development by adjusting single-cell reference datasets.

We hope Deconformer advances the understanding of pathway-pathway interactions in cfRNA from plasma, directly modeling small sample data, which is difficult for complex models capturing pathway interactions in bodily fluid cfRNA. Through Deconformer's deconvolution process, this objective is also achieved. In COVID-19 data, using Deconformer's attention scores, we mapped the dependency network between pathways. The most significant differences in pathway-pair attention levels and network hub pathways were in platelet activation, directly related to severe cases. This offers important pathways for studying disease mechanisms and treatments in the peripheral immune system and supports further biological discoveries.

Although Deconformer achieved good performance on the samples and data used in this study, it is important to note that cell type similarities, sample heterogeneity and complexity, and experimental noise and bias significantly limit deconvolution accuracy. Interpretation of individual results should be approached with caution. In the future, as single-cell datasets become more comprehensive, more detailed cell types like immune cells residing in different tissues and cancer cells can be added to the reference data to obtain relative fractions of circulating and resident immune and cancer cells. Currently, Deconformer's pathway-related genes consist of mRNA. However, non-coding RNAs are also crucial in disease research, and integrating functional pathways of non-coding RNAs into Deconformer is equally important. Also, grouping pathways by function and adding positional encoding to tokens might enhance the model's representational power and training speed. Of course, Deconformer's structure can be easily adapted to other tasks, and we hope to explore its flexibility and versatility in various downstream tasks, such as spatial transcriptomics segmentation and disease differential studies.

4 Methods

4.1 Single-cell datasets and preprocessing

The adult single-cell dataset for simulating cfRNA was sourced from the Tabula Sapiens project. The placental single-cell dataset was obtained from Vento-Tormo et al. (2018), and the fetal single-cell data from Xu et al. (2023). Only cells from droplet sequencing ('10x') were used for analysis in these datasets. For the adult single-cell data, following Sevahn's deconvolution method, cell types with similar expressions were merged and removed to minimize interference in the subsequent design of deconvolution models. We retained only cell types with more than 10 cells, considering the need for diversity in the single-cell data selected for simulation to accurately represent each cell type's characteristics. The final target cell types for deconvolution were 60 kinds, covering major cell types and those with higher specificity. Raw count values from the DecontX adjusted layer were used to minimize signal spread contamination in individual cells. For placental and fetal data, we included placental trophoblast (SCT, EVT, VCT) and cells from four main

fetal organs—heart, liver, brain, kidney—in the target cell types for Deconformer deconvolution. The expression values of individual cells were normalized to 10,000 using the "normalize_total" function from Scanpy. Since the genes in the pathways used are predominantly mRNA, only mRNA was retained for subsequent analysis. These values were used as input vectors for generating simulated cfRNA.

4.2 Simulated cfRNA data

Deconformer was trained on simulated cfRNA data, generated as follows: 1. Randomly sample 2 to 20 (1/3 of the total) cell types from a pool of 60 original types with equal probability. 2. For each cell type, randomly sample 200 to 800 cells for the mixture. If the cell number of a particular type is less than 200, all available cells were included in the mixture. 3. Specify the mixture fractions for the cell types by assigning a random ratio to each cell type with the sum of all ratios being 1. 4. Accumulate the expression data for the cell types, where the expression value of each cell type was first averaged and then aggregated into a weighted sum based on the mixture fractions. Based on these phases, we simulated 800,000 cfRNA samples, as well as an additional 1,000 cfRNA samples for validation and performance testing. When the simulation data was generated based on placental and fetal cell, the pool of original cell types was 34, and the sampling number is from 2 to 17 (1/2 of total), while retaining consistency in other sampling parameters.

4.3 The proposed Deconformer model

Gene expression of cell-free RNA (cfRNA) is the accumulation of genes in all cells and can be represented as a N-dimensionality vector $x = [x_i]_{i=1}^{N}$, where N denotes the number of genes and x_i denotes the expression value of i-th gene. Taking the gene expression as input, a deconvolution model will output the fractions of each type of cell to all cells, represented by a C-dimensionality vector $p = [p_j]_{j=1}^{C}$, where C is the number of cell types and p_j is the fractions of j-th type of cell. In this article, we design a novel Deconformer model, which leverages prior knowledge of pathways, adopts the advanced paradigm of Transformer and tailors its architecture specifically for the task of cfRNA deconvolution. On one hand, pathway information indicates which genes have certain associations and is beneficial for the model to solve faster and more accurately. On the other hand, Transformer has a global receptive field, enabling it to effectively learn global representations and long-range dependencies, and achieves excellent performance on non-sequential data. The details of Deconformer model are presented from the following three aspects:

4.3.1 Pathway selection

The pathways in this study were constructed using the C5 GOBP dataset (Gene sets derived from the GO Biological Process ontology) from the GSEA database. Each pathway in C5 GOBP consists of several genes and thus we

evaluated its value by whether it covers more useful genes. Specifically, the top-G genes having the highest between-subject variance are selected and used as evaluation criteria. The Jaccard similarity between each pathway and the G genes is calculated, and the top-P pathways having the biggest similarity are selected for further modeling. In this study, G and P are set as $10,000$ and $5,000$ respectively.

4.3.2 Pathway-based embedding

As above-mentioned, the input of the model is N-dimensionality gene expression $x = [x_i]_{i=1}^{N}$. To model its different representations over each pathway, x is first extended via a row-wise replication and the obtained matrix is denoted as $M_X \in \mathbb{R}^{P \times N}$. Meanwhile, according to the involved genes of the selected P pathways, a pathway-based mask matrix is established, denoted as $M_P \in \{0,1\}^{P \times N}$, where each element is valued as 0 or 1, representing whether a pathway covers one specific gene. As Equation (1), the mask matrix M_P act on M_X with Hadamard product to achieve a new representations $M_E \in \mathbb{R}^{P \times N}$.

$$M_E = M_P \odot M_X \tag{1}$$

Then, the matrix M_E is transformed into a deeper d-dimensionality space through a linear layer as Equation (2), where $W^E \in \mathbb{R}^{N \times d}$ is the trainable parameters and $M_T \in \mathbb{R}^{P \times d}$ can be seen as the representations of the input gene expression over different pathways. In this study, d is set as 128.

$$M_T = M_E W^E \tag{2}$$

4.3.3 Model building

On the basis, Transformer module is leveraged to capture the intrinsic correlation among the representations over different pathways. Concretely, an extra trainable d-dimensionality [CLS] token is randomly initialized using a normal distribution and added to the header of M_T, resulting in a token list $T \in \mathbb{R}^{(P+1) \times d}$ as the input of Transformer. The token list T is encoded by four consecutive layers of Transformer, where the original positional embedding of Transformer is not utilized because there is no positional relationship between the pathways. Each layer of Transformer consists of multi-head attention layer and the following residual connection, layer normalization, as well as feedforward neural network layers. The process with input token list being $X \in \mathbb{R}^{(P+1) \times d}$ is implemented by the following steps.

(a) Multi-head attention. The layer performs H-time self-attention operations. The h-th attention is performed as Equation (3), (4) and (5). In this study, H is set as 4.

$$Q_h = XW_h^Q, K_h = XW_h^K, V_h = XW_h^V \tag{3}$$

$$H_h = Attention\left(Q_h, K_h, V_h\right) = softmax\left(\frac{Q_h K_h^{\mathrm{T}}}{\sqrt{d}}\right) V_h \tag{4}$$

$$MH = Concat\left(H_1, H_2, \cdots, H_H\right) W^{MH},\ MH \in \mathbb{R}^{(P+1)\times d} \tag{5}$$

where $\left\{W_h^Q\right\}_{h=1}^H$, $\left\{W_h^K\right\}_{h=1}^H$, $\left\{W_h^V\right\}_{h=1}^H \in \mathbb{R}^{d\times d}$ and $W^{MH} \in \mathbb{R}^{Hd\times d}$ are trainable parameters.

(b) Residual connection and layer normalization. This layer adds the original input X and the multi-head attention output MH together, and applies layer normalization over this as Equation (6).

$$R^1 = LayerNorm\left(X + MH\right),\ R^1 \in \mathbb{R}^{(P+1)\times d} \tag{6}$$

(c) Feedforward neural network. In this layer, for each token R_i^1 in R^1, its representation is updated by a two-layer neural network as Equation (7).

$$FNN_i = GELU\left(R_i^1 W^1 + b^1\right) W^2 + b^2, FNN \in \mathbb{R}^{(P+1)\times d} \tag{7}$$

where $W^1 \in \mathbb{R}^{d\times d_1}$, $W^2 \in \mathbb{R}^{d_1\times d}$, $b^1 \in \mathbb{R}^{1\times d_1}$ and $b^2 \in \mathbb{R}^{1\times d}$ are trainable parameters, and $GELU(\bullet)$ is the activation function with Gaussian error linear unit. In this study, d_1 is set as 512.

(d) Residual connection and layer normalization again. This layer adds R^1 and FNN together, and execute layer normalization again as Equation (8). Finally, one Transformer layer outputs a token list of the same size as the input, i.e. R^2.

$$R^2 = LayerNorm\left(R^1 + FNN\right),\ R^2 \in \mathbb{R}^{(P+1)\times d} \tag{8}$$

Through these four Transformer layers, the token list T is transformed into $\tilde{T} \in \mathbb{R}^{(P+1)\times d}$, which embeds rich correlation information between pathways. Especially, the corresponding token at the [CLS] position, $\tilde{T}_0 \in \mathbb{R}^{1\times d}$, is utilized to further predict the fractions of each cell type as Equation (9).

$$p = softmax(\tilde{T}_0 W^P + b^P) \tag{9}$$

where $W^P \in \mathbb{R}^{d\times c}$ and $b^P \in \mathbb{R}^{1\times c}$ are trainable parameters. Each element of $p = [p_j]_{j=1}^C$ represents the fractions of corresponding cell type, and their sum is 1.0.

4.3.4 Model training

To optimize the proposed Deconformer model, we utilize a mean squared error (MSE) loss between the prediction p and ground-truth $\hat{p}$ as Equation (10), where $I(\bullet)$ is indicator function, and train it by mini-batch Adam optimizer with an end-to-end strategy. Training hyperparameters were tuned to the following values: max learning rate, 0.0005; learning scheduler, linear with warmup; warmup steps, 10000; batch size, 128. Typically, Deconformer

converges within 15 epochs. Training for more epochs does not significantly increase the CCC on simulated data and poses a risk of overfitting. Fig 2D

$$\min I(\hat{p} \neq 0) \parallel p - \hat{p} \parallel_2^2 \tag{10}$$

4.4 Model interpretability

The Deconformer model is based on a self-attention mechanism, and we conducted interpretability analysis to explore key pathways in decision-making. Attention weights reflect each pathway's contribution and the interactions between pathways. Attention weights can be obtained from Equation (4), and we averaged the attention scores from multiple heads, integrating attention matrices from multiple heads into one matrix. In this average attention matrix, each value represents the degree of attention from pathway i to pathway j. We first analyzed the attention scores between the [CLS] token and other pathway tokens, which is a vector of the length of the number of pathways. This way, we can obtain the degree of attention of the [CLS] token to other pathways in different states of samples used for classification. Secondly, to obtain differences in attention degrees between each pathway, we also analyzed the attention between 5000 pathways, i.e., 25 million pathway pairs. The calculation of differences in pathways used the Cohen's d statistic, and the pathways were ranked by the absolute value of their differences.

4.5 Comparison methods

For benchmarking, we selected two representative tools, Scaden and CSx-Vorperian, which are regression-based algorithms on GEP and based on DNN, respectively. In real data comparison, we only conducted comparisons on data with clear evidence support or corresponding blood cell counts.

CSx-Vorperian

CSx-Vorperian was tested following its provided instructions and tutorials. Due to its high time complexity, approximately 14 to 15 hours per sample, for simulated data comparison, we randomly sampled 100 out of 1000 validation data for performance testing. The GEP provided by Vorperian et al. (2022) includes 62 cell types, and the 60 cell types in the simulated adult cfRNA are included, so only the 60 cell types present in the simulated cfRNA were considered during performance testing.

Scaden

Given the significantly larger variety of single-cell sources in the simulated data compared to those tested in the Scaden paper and the high dropout in cfRNA, the Scaden model was unable to filter out enough genes with a variance greater than 0.1 as in the original paper. Therefore, based on the instructions and source code from the original Scaden paper, we rewrote a version of Scaden using PyTorch2.0. The only change we made was in the gene selection part, where we selected the top 10,000 genes with the highest variance instead of those with a variance greater than 0.1. We then tested the impact of different

training set sizes on the performance of Scaden on a validation machine and ultimately determined that performance does not improve further with training sets larger than 9,000. For performance comparison using Scaden, we used a model trained on a training set based on 9,000.

4.6 Performance evaluation and statistical analysis

The performance evaluation metrics for the algorithm include CCC (Concordance Correlation Coefficient) and RMSE (Root Mean Square Error). Relying solely on a single metric can make it challenging to evaluate performance reasonably in all scenarios. The CCC value is prone to distortion when the cell types in the expression spectrum are few. Given that the loss functions for the Deconformer and Scaden models are sparse MSE (Mean Squared Error) and MAE (Mean Absolute Error) respectively, the RMSE can also be low when the model is overfitting. Therefore, it is necessary to use both CCC and RMSE for performance assessment. Statistical differences are evaluated using the Wilcoxon rank-sum (Mann–Whitney-U) test.

4.7 cfRNA dataset and processing

The datasets for cancer-related diseases are from Chen et al. (2022), Roskams-Hieter et al. (2022), Tao et al. (2023). For infection-related diseases, COVID-19 research data is from Wang et al. (2022), HBV infection data from Sun et al. (2023). Placental development and pregnancy-related diseases are from Munchel et al. (2020), Sun et al. (2023). Bodily fluid data is from Liu et al. (2022). These data were processed to TPM, with the quality control criteria for the samples based on the rules outlined in the original articles.

4.8 Software and hardware environment

Deconformer utilizes Python 3.10.12 and is primarily built using torch (version 2.0.0+cu117) and scanpy (version 1.9.3). This study used the Nvidia A100 40GB to train models. If only using pre-trained model parameters for inference, a laptop without GPU is sufficient. Downstream analysis and plotting employed joblib (version 1.3.2), as well as R 4.2.0, ggplot2 (3.4.4), pheatmap (1.0.12), igraph (1.5.1), ggraph (2.1.0), and their dependent packages.

5 Data Availability

All data used in the article are given in the method part, and these data are publicly accessible, except for Liu et al. , which was obtained from the author. Samples from Chen et al. (GSE174302), Toden et al. (GSE186607), Roskams-Hieter et al. (GSE182824) were downloaded from the Sequence Read Archive with the respective accession numbers. samples from Munchel et al. were downloaded from the Resources of the article. Samples from Wang et al. (CNP0001306) were downloaded from CNGB Sequence Archive (CNSA; https://db.cngb.org/cnsa/) of the China National GeneBank

DataBase (CNGBdb) under accession number CNP0001306. Samples from Sun et al. were downloaded from Genome Sequence Archive for Human Database with accession numbers HRA003387. The code of cfRNA alignment and quantification is available at https://github.com/wonderful1/PALM-Seq-cfRNA . Tabula Sapiens was downloaded from the Chan Zuckerberg Biohub (https://tabula-sapiens-portal.ds.czbiohub.org, version 5.0). Placental single-cell data (Vento-Tormo et al.) were downloaded from ArrayExpress with experiment codes E-MTAB-6701. Fetal single-cell data (Xu et al.) were downloaded from the Gene Expression Omnibus (GEO) under accession code GSE157329.

6 Code Availability

The code for data preprocessing, Deconformer modeling, and downstream analysis, as well as the PyTorch2.0 version of Scaden, are available on https://github.com/findys/Deconformer. Detailed instructions are provided there.

7 Acknowledgement

This work was supported by grants from National Natural Science Foundation of China (62207002), and Science, Technology and Innovation Commission of Shenzhen Municipality (JCYJ20170412152854656).

8 Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Koh, W. *et al.* Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proceedings of the National Academy of Sciences* **111** (20), 7361–7366 (2014). https://doi.org/10.1073/pnas.1405528111 .

[2] Ngo, T. T. M. *et al.* Noninvasive blood tests for fetal development predict gestational age and preterm delivery. *Science* **360** (6393), 1133–1136 (2018). https://doi.org/10.1126/science.aar3819 .

[3] Ibarra, A. *et al.* Non-invasive characterization of human bone marrow stimulation and reconstitution by cell-free messenger RNA sequencing. *Nature Communications* **11** (1), 400 (2020). https://doi.org/10.1038/s41467-019-14253-4 .

[4] Toden, S. *et al.* Noninvasive characterization of Alzheimer's disease by circulating, cell-free messenger RNA next-generation sequencing. *Science Advances* **6** (50), eabb1654 (2020). https://doi.org/10.1126/sciadv.abb1654 .

[5] Larson, M. H. *et al.* A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. *Nature Communications* **12** (1), 2357 (2021). https://doi.org/10.1038/s41467-021-22444-1 .

[6] Hulstaert, E. *et al.* Charting Extracellular Transcriptomes in The Human Biofluid RNA Atlas. *Cell Reports* **33** (13), 108552 (2020). https://doi.org/10.1016/j.celrep.2020.108552 .

[7] Boiarsky, R. *et al.* Single cell characterization of myeloma and its precursor conditions reveals transcriptional signatures of early tumorigenesis. *Nature Communications* **13** (1), 7040 (2022). https://doi.org/10.1038/s41467-022-33944-z .

[8] Vorperian, S. K. *et al.* Cell types of origin of the cell-free transcriptome. *Nature Biotechnology* **40** (6), 855–861 (2022). https://doi.org/10.1038/s41587-021-01188-9 .

[9] Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* **37** (7), 773–782 (2019). https://doi.org/10.1038/s41587-019-0114-2 .

[10] Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618** (7965), 616–624 (2023). https://doi.org/10.1038/s41586-023-06139-9 .

[11] Aliee, H. & Theis, F. J. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. *Cell Systems* **12** (7), 706–715.e4 (2021). https://doi.org/10.1016/j.cels.2021.05.006 .

[12] Menden, K. *et al.* Deep learning–based cell composition analysis from tissue expression profiles. *Science Advances* **6** (30), eaba2619 (2020). https://doi.org/10.1126/sciadv.aba2619 .

[13] Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362** (6416), eaau5324 (2018). https://doi.org/10.1126/science.aau5324 .

[14] Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102** (43), 15545–15550 (2005). https://doi.org/10.1073/pnas.0506580102 .

[15] Hwang, S. *et al.* HumanNet v2: Human gene networks for disease research. *Nucleic Acids Research* **47** (D1), D573–D580 (2019). https://doi.org/10.1093/nar/gky1126 .

[16] Chen, J. *et al.* Transformer for one stop interpretable cell type annotation. *Nature Communications* **14** (1), 223 (2023). https://doi.org/10.1038/s41467-023-35923-4 .

[17] Yang, F. *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **4** (10), 852–866 (2022). https://doi.org/10.1038/s42256-022-00534-z .

[18] Vaswani, A. *et al.* Guyon, I. *et al.* (eds) *Attention is all you need.* (eds Guyon, I. *et al.*) *Advances in Neural Information Processing Systems*, Vol. 30 (Curran Associates, Inc., 2017). URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[19] THE TABULA SAPIENS CONSORTIUM. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376** (6594), eabl4896 (2022). https://doi.org/10.1126/science.abl4896 .

[20] Lin, L. I.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **45** (1), 255 (1989). https://doi.org/10.2307/2532051, 2532051 .

[21] Safrastyan, A., Zu Siederdissen, C. H. & Wollny, D. Decoding cell-type contributions to the cfRNA transcriptomic landscape of liver cancer. *Human Genomics* **17** (1), 90 (2023). https://doi.org/10.1186/s40246-023-00537-w .

[22] Chen, S. *et al.* Cancer type classification using plasma cell-free RNAs derived from human and microbes. *eLife* **11**, e75181 (2022). https://doi.org/10.7554/eLife.75181 .

[23] Roskams-Hieter, B. *et al.* Plasma cell-free RNA profiling distinguishes cancers from pre-malignant conditions in solid and hematologic malignancies. *npj Precision Oncology* **6** (1), 28 (2022). https://doi.org/10.1038/s41698-022-00270-y .

[24] Tao, Y. *et al.* Cell-free multi-omics analysis reveals potential biomarkers in gastrointestinal cancer patients' blood. *Cell Reports Medicine* **4** (11), 101281 (2023). https://doi.org/10.1016/j.xcrm.2023.101281 .

[25] Humphries, A. & Wright, N. A. Colonic crypt organization and tumorigenesis. *Nature Reviews Cancer* **8** (6), 415–424 (2008). https://doi.org/10.1038/nrc2392 .

[26] Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457** (7229), 608–611 (2009). https://doi.org/10.1038/nature07602

.

[27] Bhowmick, N. A., Neilson, E. G. & Moses, H. L. Stromal fibroblasts in cancer initiation and progression. *Nature* **432** (7015), 332–337 (2004). https://doi.org/10.1038/nature03096 .

[28] Wang, Y. *et al.* Plasma cell-free RNA characteristics in COVID-19 patients. *Genome Research* **32** (2), 228–241 (2022). https://doi.org/10.1101/gr.276175.121 .

[29] Sun, J. *et al.* Early characterisation and prediction of liver diseases in pregnancy by plasma cell-free RNAs. *Clinical and Translational Medicine* **13** (10), e1439 (2023). https://doi.org/10.1002/ctm2.1439 .

[30] Carsana, L. *et al.* Pulmonary post-mortem findings in a series of COVID-19 cases from northern Italy: A two-centre descriptive study. *The Lancet Infectious Diseases* **20** (10), 1135–1140 (2020). https://doi.org/10.1016/S1473-3099(20)30434-5 .

[31] Rajamanickam, A. *et al.* Dynamic alterations in monocyte numbers, subset frequencies and activation markers in acute and convalescent COVID-19 individuals. *Scientific Reports* **11** (1), 20254 (2021). https://doi.org/10.1038/s41598-021-99705-y .

[32] Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nature Medicine* **26** (7), 1070–1076 (2020). https://doi.org/10.1038/s41591-020-0944-y .

[33] Venet, F. *et al.* T cell response against SARS-CoV-2 persists after one year in patients surviving severe COVID-19. *eBioMedicine* **78**, 103967 (2022). https://doi.org/10.1016/j.ebiom.2022.103967 .

[34] Xu, G. *et al.* The differential immune responses to COVID-19 in peripheral and lung revealed by single-cell RNA sequencing. *Cell Discovery* **6** (1), 73 (2020). https://doi.org/10.1038/s41421-020-00225-2 .

[35] Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine* **27** (5), 904–916 (2021). https://doi.org/10.1038/s41591-021-01329-2 .

[36] Sun, C., Sun, H., Zhang, C. & Tian, Z. NK cell receptor imbalance and NK cell dysfunction in HBV infection and hepatocellular carcinoma. *Cellular & Molecular Immunology* **12** (3), 292–302 (2015). https://doi.org/10.1038/cmi.2014.91 .

[37] Cully, M. Reinvigorating exhausted T cells in hepatitis B infection. *Nature Reviews Immunology* **17** (4), 218–218 (2017). https://doi.org/10.1038/

nri.2017.35 .

[38] Ye, B. *et al.* T-cell exhaustion in chronic hepatitis B infection: Current knowledge and clinical significance. *Cell Death & Disease* **6** (3), e1694–e1694 (2015). https://doi.org/10.1038/cddis.2015.42 .

[39] Moufarrej, M. N. *et al.* Early prediction of preeclampsia in pregnancy with cell-free RNA. *Nature* **602** (7898), 689–694 (2022). https://doi.org/10.1038/s41586-022-04410-z .

[40] Vento-Tormo, R. *et al.* Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563** (7731), 347–353 (2018). https://doi.org/10.1038/s41586-018-0698-6 .

[41] Xu, Y. *et al.* A single-cell transcriptome atlas profiles early organogenesis in human embryos. *Nature Cell Biology* **25** (4), 604–615 (2023). https://doi.org/10.1038/s41556-023-01108-w .

[42] Munchel, S. *et al.* Circulating transcripts in maternal blood reflect a molecular signature of early-onset preeclampsia. *Science Translational Medicine* **12** (550), eaaz0131 (2020). https://doi.org/10.1126/scitranslmed.aaz0131 .

[43] Castleman, J. S., Lip, G. Y. H. & Shantsila, E. Monocytes are increased in pregnancy after gestational hypertensive disease. *Scientific Reports* **12** (1), 10358 (2022). https://doi.org/10.1038/s41598-022-13606-2 .

[44] Tang, Z. *et al.* CD36 mediates SARS-CoV-2-envelope-protein-induced platelet activation and thrombosis. *Nature Communications* **14** (1), 5077 (2023). https://doi.org/10.1038/s41467-023-40824-7 .

[45] Nishikawa, M. *et al.* Massive image-based single-cell profiling reveals high levels of circulating platelet aggregates in patients with COVID-19. *Nature Communications* **12** (1), 7135 (2021). https://doi.org/10.1038/s41467-021-27378-2 .

[46] Yang, L. *et al.* The signal pathways and treatment of cytokine storm in COVID-19. *Signal Transduction and Targeted Therapy* **6** (1), 255 (2021). https://doi.org/10.1038/s41392-021-00679-0 .

[47] Klenk, C. *et al.* Platelet aggregates detected using quantitative phase imaging associate with COVID-19 severity. *Communications Medicine* **3** (1), 161 (2023). https://doi.org/10.1038/s43856-023-00395-6 .

[48] Zhu, W. *et al.* Monoacylglycerol lipase promotes progression of hepatocellular carcinoma via NF-κB-mediated epithelial-mesenchymal transition. *Journal of Hematology & Oncology* **9** (1), 127 (2016). https://doi.org/10.

1186/s13045-016-0361-3 .

[49] Zhang, Z. *et al.* Serine catabolism generates liver NADPH and supports hepatic lipogenesis. *Nature Metabolism* **3** (12), 1608–1620 (2021). https://doi.org/10.1038/s42255-021-00487-4 .