nature computational science

Article

https://doi.org/10.1038/s43588-024-00625-4

Discrete latent embedding of single-cell chromatin accessibility sequencing data for uncovering cell heterogeneity

In the format provided by the authors and unedited

Contents

Supplementary Notes

Supplementary Note 1: The validation of normality assumption of the latent space in VAE

To validate whether the latent space of the VAE model conforms to the assumption of normal distribution, we conducted a comparative experiment on the mouse Splenocyte dataset[1]. We trained a standard VAE model following the pipeline of SCALE[2] with the default parameters. Then we investigated the distributional characteristics of the latent embeddings extracted from the model from four aspects:

First, the histogram of each latent dimension can provide a visual assessment of its distribution of the latent embeddings. If the distribution appears bell-shaped and symmetric, it suggests a closer approximation to a normal distribution.

Second, a quantile-quantile (Q-Q) plot compares the quantiles of the observed latent variable with the quantiles of a normal distribution. If the points lie close to a straight line, it indicates a good fit to the normal distribution.

Third, the skewness is a measure of the asymmetry of a distribution[3]. The skewness of the normal distribution is equal to 0. Negative skewness commonly indicates that the tail is on the left side of the distribution, and positive skewness indicates that the tail is on the right.

Fourth, the kurtosis is a measure of the tailedness of a distribution[3]. The kurtosis of the normal distribution is equal to 3. The kurtosis greater than 3 commonly indicates that the peak is steep, and the kurtosis less than 3 indicates that the peak is gentle.

In the standard VAE model trained on the Splenocyte dataset, the histogram fitting curves of latent feature 4, 6, 7 and 9 display noticeable asymmetry and multiple peaks, and the Q-Q plots of these latent features reveal a notable departure from the straight line representing the standard normal distribution, which evidently deviate from the characteristics of a standard normal distribution (Supplementary Figure 1). The skewness of latent feature 2, 3, 8 and 10 is not equal to 0, and the Q-Q plots of these latent features reveal a notable departure from the straight line representing the standard normal distribution, deviating from the characteristics of a standard normal distribution (Supplementary Figure 2). The kurtosis of latent feature 1 and 5 is not equal to 3, and the Q-Q plots of these latent features reveal a notable departure from the straight line representing the standard normal distribution, deviating from the characteristics of a standard normal distribution (Supplementary Figure 3).

Supplementary Note 2: Evaluation metrics for clustering

Let **U** represent the known ground-truth cell type labels, **V** represent the predicted clustering assignments, *N* represent the total number of single cells, x_i represent the number of cells in the *i*-th cluster of **V**, y_j represent the number of cells with the *j*-th unique cell type label of **U**, and n_{ij} represent the number of cells shared between the *i*-th cluster and the *j*-th unique cell type label. The Rand Index (RI) computes a similarity measure between two clustering assignments by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clustering assignments. The ARI[4] is derived from the RI by taking into account the expected agreement due to chance:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i\binom{x_i}{2}\sum_j\binom{y_j}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_i\binom{x_i}{2} + \sum_j\binom{y_j}{2}\right] - \left[\sum_i\binom{x_i}{2}\sum_j\binom{y_j}{2}\right]/\binom{N}{2}}$$

Both AMI[5] and NMI[6] are derived from Mutual Information (MI), which is a measure of the similarity between two labels of the same data. Where N_{U_p} is the number of the samples in cluster U_p and N_{V_q} is the number of the samples in cluster V_q, the Mutual Information between clustering assignments **U** and **V** is given as:

$$MI(\mathbf{U},\mathbf{V}) = \sum_{p=1}^{N_U}\sum_{q=1}^{N_V}\frac{|U_p \cap V_q|}{N} log \frac{N|U_p \cap V_q|}{|U_p||V_q|}$$

AMI adjusts MI by considering the expected value under random clustering as follows:

$$AMI = \frac{MI(\mathbf{U},\mathbf{V}) - E[MI(\mathbf{U},\mathbf{V})]}{avg[H(\mathbf{U}), H(\mathbf{V})] - E[MI(\mathbf{U},\mathbf{V})]}$$

where $E(\cdot)$ is the expectation function and $H(\cdot)$ is the entropy function.

NMI scales MI to a range between 0 and 1, and its calculation is as follows:

$$NMI = \frac{MI(\mathbf{U},\mathbf{V})}{\sqrt{H(\mathbf{U})H(\mathbf{V})}}$$

Let TP denote the number of True Positive (i.e. the number of pairs of cells that belong to the same clusters in both **U** and **V**), FP denote the number of False Positive (i.e. the number of pairs of cells that belong to the same clusters in **V** and not in **U**) and FN denote the number of False Negative (i.e. the number of pairs of cells that belong to the same clusters in **U** and not in **V**). FMI[7] is calculated as follows:

$$FMI = \frac{\mathrm{TP}}{\sqrt{(\mathrm{TP+FP})(\mathrm{TP+FN})}}$$

Supplementary Note 3: Evaluation of performance based on kNN classification

Similar to existing research[8], we conducted experiments to evaluate the performance of cell embeddings by comparing k-nearest neighbor (kNN) classification results according to two classification metrics including classification accuracy and median F1 score for different cell types (mF1)[9, 10]. For a single-cell chromatin accessibility sequencing (scCAS) dataset, we chose 80% of the cells as a train set and the remaining 20% as a test set. Specifically, we included cell types with fewer than 10 cells entirely in the train set[9, 10]. Then we trained a kNN classifier with the cell embeddings and cell type labels of train set and used the classifier to predict the cell type labels of test set. CASTLE performs equally well in accuracy compared to cisTopic[11] and scBasset[12] (Supplementary Figure 5). Generally, compared with accuracy, F1 score is a better metric for imbalanced datasets[13-15]. CASTLE achieves better median mF1 than all baseline methods (Supplementary Figure 5), demonstrating that CASTLE excels in accurate cell type identification.

Supplementary Note 4: Stability of batch correction for CASTLE

We conducted comparative experiments to showcase the versatility of CASTLE across datasets with complex batch effects. As highlighted with red circles in Fig. 3g, CASTLE acquires better batch integration and illustrates a clearer separation between L2/3 IT (intratelencephalic neurons from layer 2 or 3 of mouse brain), L4 (layer 4 of mouse brain), L5 IT (intratelencephalic neurons from layer 5 of mouse brain) and L6 IT (intratelencephalic neurons from layer 6 of mouse brain) across two batches than baseline methods. To verify the capacity of CASTLE for treating datasets with certain abundant cell types existing only in one batch, we removed cells that belongs to one or all of L2/3 IT, L4, L5 IT and L6 IT cell types in the batch "10X" of the Brain dataset. For datasets with partially overlapping batches, CASTLE achieves highly robustness for batch correction and exhibits weak over-correction, in terms of UMAP visualization (Supplementary Figure 14a-f) and the performance of batch correction evaluated by ARI and kBET[16] (Supplementary Figure 14g). To sum up, CASTLE not only effectuates the excellent balance between removing batch variations and preserving cell heterogeneity, but also demonstrates strong robustness to datasets with complex batch effects.

Supplementary Note 5: Imbalanced datasets and rare cell types

To assess the competence of CASTLE for dealing with complex and imbalanced scCAS data, we simulated eight datasets with diverse degrees of cell type imbalance[17] based on the Stimulated Droplet dataset again (“Simulation of datasets with different imbalances” in Methods). The cell type imbalance of the original Stimulated Droplet is 0.175. To simulate datasets with higher degrees of cell type imbalance, we randomly downsampled cell types with a cell count less than 10,000 to the following fractions of their original counts: half, one-fourth, one-sixth, one-eighth, one-tenth, one-sixteenth, and one-thirtieth, and simulated the datasets with cell type imbalances of 0.281, 0.407, 0.476, 0.520, 0.551, 0.604 and 0.656. As the imbalance gradually increases, the clustering performance of all methods decreases, while CASTLE substantially outperforms all other methods at distinct cell type imbalances and displays the highest level of stability (Fig. 4c and Supplementary Figure 17).

Then we removed cell types with a cell count larger than 10,000 (Mono and CD4) and evaluated the performance over the remaining rare cell types. As the degree of cell type imbalance increases sharply, the UMAP visualization and clustering performance of CASTLE for the identification of rare cell types deteriorates slowly (Supplementary Figure 18). Specifically, CASTLE achieves the overall best clustering performance compared with four baseline methods (Supplementary Figure 18i). In summary, CASTLE exhibits strong robustness and stability for the identification of rare cell types.

Supplementary Note 6: Ablation studies for CASTLE

In CASTLE, there are seven hyperparameters including α denoting the weight of $L_{commitment}$, γ denoting the decay ratio, K denoting the size of codebook, M denoting the time of split quantization[18], ζ denoting the weight of L_{batch}, η denoting the weight of $L_{celltype}$, λ denoting the weight of $L_{classifier}$. We designed comprehensive comparative experiments about robustness analyses for these hyperparameters.

First, similar to the original studies of Vector Quantized Variational AutoEncoder (VQ-VAE)[19, 20], we aimed for the codebook to have less impact on the output of the encoder so that we set the default value of α, the weight of $L_{commitment}$, to 0.25. To validate the robustness of CASTLE with different weights of $L_{commitment}$, we trained CASTLE with different values of α (0.1, 0.25, 0.5, 1.0, 5.0 and 10.0) on 16 benchmark datasets. The results, as shown in Supplementary Figure 21a, demonstrate that CASTLE consistently obtains stable clustering performance under different values of α.

Second, we mainly followed the original studies of VQ-VAE[19, 20] to set the default value of γ, the decay ratio for updating the codebook, to 0.99. To assess the stability of CASTLE with different decay ratios, we conducted two experiments. We first trained CASTLE under a series of γ, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.96, 0.97, 0.98 and 0.99, on 16 benchmark datasets. The results, as shown in Supplementary Figure 21b, demonstrate the stability of the clustering, while the value of 0.99 is preferred for its slightly higher clustering performance. We further analyzed values of γ around 0.99. We trained CASTLE under a series values of γ, ranging from 0.991 to 0.999 on 16 benchmark datasets. As shown in Supplementary Figure 21c, it seems that the results at 0.99 is still better than those at the larger values. We again trained CASTLE under a series values of γ, ranging from 0.981 to 0.989 on 16 benchmark datasets and also obtained slightly higher clustering performance at 0.99 (Supplementary Figure 21c). All these empirical analyses, including the above results for γ ranging from 0.10 to 0.99, support that 0.99 is indeed a reasonable default value of the decay ratio γ.

Third, to evaluate the robustness of CASTLE to the size of the codebook, K, we trained CASTLE with different values of K (100, 200, 400, 600, 800 and 1000) on 16 benchmark datasets. The results (Supplementary Figure 21d) show that CASTLE is insensitive to this hyperparameter. Taking into account the balance of cell heterogeneity preservation and codebook utilization, we set the default value of K to 400.

Fourth, to evaluate the stability of CASTLE to the time of split quantization[18], we trained CASTLE with different values of M (1, 2, 5, 10, 25, 50) on 16 benchmark datasets. The results (Supplementary Figure 21e) show that CASTLE acquires relatively poor clustering performance when the time of split quantization is

1 or 2, while CASTLE attains highly stable clustering performance when the time of split quantization is higher than 5. Evidently, the lower the time of split quantization, the higher the dimension of codebook features. With the consideration that it is obviously challenging to look up the nearest neighbors for high-dimensional vectors, we set the default value of M to 10.

Fifth, to verify the robustness of CASTLE to the weight of L_{batch}, we trained CASTLE with different values of ζ ranging from 0 to 1 on 4 benchmark datasets with multiple batches. The results (Supplementary Figure 22a) show the stability of CASTLE under these values, while excessive weights of L_{batch} may lead to poor batch correction. We therefore set the default value of ζ to 0.001.

Sixth, to demonstrate the stability of CASTLE toe the weight of $L_{celltype}$, we trained CASTLE with different values of η ranging from 0 to 1 on 8 benchmark datasets when incorporating labeled reference datasets. The results (Supplementary Figure 22b) show that CASTLE yields bad results when η equals 1 on the Stimulated Droplet dataset[21], while obtaining stable clustering performance when η below 1. We therefore set the default value of η to 0.001, with which CASTLE achieves relatively better performance.

Seventh, to evaluate the sensitivity of CASTLE to the weight of $L_{classifier}$, we trained CASTLE with different values of λ ranging from 0 to 100 on 8 benchmark datasets when incorporating labeled reference datasets. The results (Supplementary Figure 22c) show that CASTLE attains the highest performance when λ equals 1, and somewhat lower performance when λ equals 0 and 100. We therefore set the default value of λ to 1 to avoid inappropriate selection of λ.

In addition, since L_{recon} and $L_{commitment}$ are basic and indispensable loss functions in CASTLE, we designed comprehensive ablation studies only for the contributions of L_{batch} on 4 benchmark datasets with multiple batches, $L_{celltype}$ and $L_{classifier}$ on 8 benchmark datasets when incorporating labeled reference datasets. Compared with CASTLE without L_{batch}, $L_{celltype}$ and $L_{classifier}$, as long as the loss weights are within a reasonable range, the incorporation of these loss functions in CASTLE consistently leads to improvements of clustering performance (Supplementary Figure 22). The above results support the effectiveness of three self-designed loss functions.

Besides, we conducted comparative experiments to validate the effectiveness of replacing $L_{codebook}$ with EMA and separating $L_{commitment}$ and $L_{codebook}$. We referred to the model that updates codebook using $L_{codebook}$ instead of EMA as CASTLE-noEMA, and the model that updates encoder and codebook using the integrated $L_{commitment}$ and $L_{codebook}$ instead of using them separately as CASTLE-integrated. Then we applied CASTLE-noEMA and CASTLE-integrated to the 16 benchmark datasets. As illustrated in Supplementary Figure 23, CASTLE obtains significantly better clustering performance than CASTLE-noEMA (one-sided paired Wilcoxon signed-rank tests p-values < 3e-3) and CASTLE-integrated (one-sided paired Wilcoxon signed-rank tests p-values < 2e-3). Evidently, CASTLE-noEMA achieves better clustering performance than CASTLE-integrated. The results above thoroughly demonstrate the effectiveness of replacing $L_{codebook}$ with EMA and separating $L_{commitment}$ and $L_{codebook}$.

Supplementary Note 7: Comparative experiments for different designs of $L_{celltype}$

As shown in the "Batch correction and reference incorporation" section of Methods, we developed $L_{celltype}$ for maximizing the distance between the cells from different cell types. We further compared CASTLE with a variant model that only minimizes the distance between the cells in the same cell types, and we referred to such a model as CASTLE-ref-labeled(min). Moreover, we referred to the model incorporating the redesigned $L_{celltype}$ for simultaneously maximizing the distance between the cells from different cell types and minimizing the distance between the cells in the same cell types as CASTLE-ref-labeled(+min). Similar to the model referred to as CASTLE-ref-labeled, we applied CASTLE-ref-labeled(min) and CASTLE-ref-labeled(+min) to the eight scCAS datasets with labeled reference datasets and obtained results for comparison.

As shown in Supplementary Figure 24b, CASTLE-ref-labeled obtained the highest clustering performance based on the cell embeddings of target datasets, consistently superior to CASTLE-ref-labeled(min) and CASTLE-ref-labeled(+min), indicating that the newly designed loss functions actually impaired the clustering performance. The results are reasonable, because we need to avoid losing potential heterogeneity of cell subtypes that may exist within cell types, especially when we are unsure about the granularity of cell type labels in the reference dataset and whether it aligns with the target dataset. All the above results demonstrate the rationality and efficiency of $L_{celltype}$ to maximize the distance between the cells from different cell types.

Supplementary Note 8: Comparative experiments with different reference datasets

Taking the Stimulated Droplet dataset[21] as an example, we further designed comprehensive comparative experiments to demonstrate the versatility of CASTLE with reference datasets under various conditions from five perspectives. First, to assess the ability of CASTLE for coping with reference datasets of different sizes, we downsampled the cells in the reference dataset (Resting Droplet dataset[21]) at different downsample rates ranging from 10% to 100%. Second, to evaluate the capacity of CASTLE for tackling reference datasets with different degrees of sparsity, we randomly dropped out the non-zero values in the reference dataset at distinct dropout rates ranging from 0 to 50%. Third, to validate the power of CASTLE for treating regular reference datasets only with major cell types, we reserved the cells belonging to major cell types of which the proportion is greater than a threshold in the reference dataset. We selected thresholds ranging from 0 to 10%. Fourth, to verify the ability of CASTLE for handling reference datasets with missing cell types present in the target dataset, we deleted one or more cell types in the reference dataset. Fifth, to assess the capacity of CASTLE for dealing with reference datasets with novel cell types, we deleted one or more cell types in the target dataset.

Supplementary Note 9: Quantitative recommendation score for reference dataset

Similar to the existing study[22], we developed a quantitative strategy that quantifies recommendation scores to guide users to select high-quality reference data. Suppose there are several different reference datasets for the target dataset, we designed a quantitative recommendation score to evaluate the quality of these reference datasets. First, we performed Principal Component Analysis (PCA)[23] on the raw cell-by-region matrix of target dataset and then acquired several sets of pseudo labels based on Louvain algorithm with different resolution parameters. Second, similar to the previous step, we performed PCA on the raw cell-by-region matrix of integrated dataset including target dataset and reference dataset, and then acquired several sets of pseudo labels for the target dataset. Third, we calculated the Silhouette Coefficient (SC)[24] for each set of pseudo labels. Generally, higher SC indicates better clustering outcomes. Finally, we calculated the recommendation score as follows:

$$Recommendation\ score = \frac{Mean\left(SC_{reference}^{target}\right)}{Mean\left(SC_{None}^{target}\right)} \cdot \frac{Std\left(SC_{None}^{target}\right)}{Std\left(SC_{reference}^{target}\right)}$$

where SC_{None}^{target} or $SC_{reference}^{target}$ represents the SCs for target dataset without or with reference dataset, $Mean(\cdot)$ and $Std(\cdot)$ are the formulas for calculating the average and the standard deviation. Intuitively, a higher recommendation score represents stronger and more consistent clustering performance.

Taking the Stimulated Droplet dataset[21] as an example, we conducted multiple experiments for six different reference datasets. Overall, whether the introduced reference dataset is unlabeled or labeled, the reference data with a higher recommendation score can provide better clustering performance of the target dataset (Supplementary Figure 34). For example, when the Recommendation score of introduced reference dataset is higher than 0.95, the unlabeled incorporation will improve the clustering performance of the target dataset; conversely, if it is lower, it may be detrimental to the clustering performance. For the labeled reference dataset, the threshold of Recommendation score for enhancing or impairing clustering performance is 0.8.

Supplementary Note 10: Comparative experiments between the cell type-specific peaks identified by CASTLE and epiScanpy

We applied the function *rank_features* implemented in epiScanpy[25] to identify the basic cell type-specific peaks with the raw cell-by-region matrix and the cell type labels of the Stimulated Droplet dataset[21]. Following the same steps as the "Downstream analyses" section of Methods, we implemented downstream analysis[26, 27]using these epiScanpy-identified peaks. With regard to single-nucleotide polymorphisms (SNPs) enrichment analysis, cell type-specific peaks identified by epiScanpy also demonstrate excellent cell type specificity (Supplementary Figure 43). However, the background peaks identified by epiScanpy also exhibit strong cell type specificity (Supplementary Figure 43a), which is not evident in the background peaks identified by CASTLE. Besides, 721_B_Lymphoblasts is the tissue with most significant enrichment for pDC-specific peaks identified by epiScanpy (Supplementary Figure 43b), which is not desirable. B-specific peaks identified by epiScanpy display significant enrichment in multiple cell types including PB-CD19+B_cells, PB-BDCA4+Dentritic_cells, PB-CD56+NK_cells, PB-CD14+Monocytes and so forth (Supplementary Figure 43c). About heritability enrichment analysis, the enrichment results for blood-related phenotypes in the cell type-specific peaks identified by epiScanpy are unsatisfactory, which may be due to the inferior identification for the background peaks (Supplementary Figure 44). Overall, the enrichment results with the cell type-specific peaks identified by CASTLE (Fig. 6c-g and Supplementary Figures 41-42) are relatively stronger than that identified by epiScanpy (Supplementary Figures 43-44), especially the more significant heritability enrichment results, demonstrating the biological significance of the cell type-specific peaks identified by CASTLE.

Supplementary Note 11: Trajectory inference and motif enrichment analysis

Indeed, the discrete cell embeddings generated by CASTLE can be further employed for trajectory inference which aims to reconstruct the developmental trajectories or temporal dynamics of cells. Following the original study[28], we created the Immune dataset by selecting 18,489 CD34+ bone marrow progenitors and dendritic cells across 10 subpopulations from full hematopoiesis. We applied Slingshot[29, 30] to the trajectory inference. First, we fed the low-dimensional cell embeddings generated by CASTLE and the clustering assignments acquired from Louvain clustering to Slingshot for the clustering relationships using the function getLineages. Then we used the getCurves function to cultivate smooth curves denoting the estimated cell lineages and using the embedCurves function to map the curves to the two-dimensional UMAP space for visualization. CASTLE with Slingshot uncovers the differentiation lineage, which is clearly consistent with the true hematopoietic differentiation tree (Supplementary Figure 45a-b). For example, the differentiation path (HSC→MPP→LMPP→CLP→Pro-B) on the inferred trajectory has been proved in the original study[28]. Altogether, CASTLE facilitates trajectory inference, shedding light on the underlying regulatory dynamics and cell fate decisions during development or disease progression.

We conducted comparative experiments to assess the performance of baseline methods for trajectory inference on the Immune dataset. Similarly, we employed Slingshot with default settings to infer the smooth curves representing the estimated cell lineages from the cell embeddings of four baseline methods including SCALEX[31], SCALE[2], cisTopic[11] and scBasset[12]. We note that other baseline methods except cisTopic[11] have not been validated the ability for trajectory inference in their original papers. Obviously, the inferred trajectories from baseline methods are chaotic, redundant, and even intersect with each other (Supplementary Figure 46), which probably results from the poor performance of cell clustering (Supplementary Figure 9d) and UMAP visualization (Supplementary Figure 13b) on the Immune dataset with multiple batches. In summary, baseline methods struggled in trajectory inference over the Immune dataset, while CASTLE performed well.

We next proceeded to conduct motif enrichment analysis, which is a critical step for illuminating the context-specific regulatory mechanisms. With the Louvain clustering assignments based on cell embeddings from CASTLE, we applied scABC[32] and chromVAR[33] with default settings to the InSilico dataset[34] for the identification of the cluster-specific enriched motifs from the JASPAR database[35]. We selected the top 1000 peaks with smallest p-values for each cluster, then performed TF binding motif enrichment within these peaks using chromVAR, and subsequently visualized the top 50 most variable transcription factor (TF) binding motifs (Supplementary Figure 45c). Our analysis reveals that the most active TFs are often specific to one or two clusters, indicating their unique regulatory roles in those particular cell types. It is reported in the previous literature that some motifs are cell type-specific. For example, *FOS* is specific to human foreskin fibroblasts (BJ)[36], *GATA1::TAL1* is specific to K562 chronic myelogenous leukaemia cells (K562)[37], *SPI1* is specific to human promyeloblasts (HL-60)[38], *NFKB2* is specific to GM12878 lymphoblastoid cells (GM12878)[34, 39] and POU motifs (e.g. *POU3F2*) are specific to H1 human embryonic stem cells (H1)[32, 40] (Supplementary Figure 45d). These findings allow us to identify the active TFs in each cell type, providing insights into the specific regulatory mechanisms operating in different clusters.

Supplementary Note 12: Comparative experiments for different feature-to-peak ratios

As shown in the “Feature spectrum and cell type-specific peaks” section of Methods, we selected 50% of the cell type-specific features to identify the cell type-specific peaks. We conducted comparative experiments to elucidate the rationale behind our choice of a 50% feature-to-peak ratio. Following the steps in the “Feature spectrum and cell type-specific peaks” section of Methods, we identified the cell type-specific peaks with four different feature-to-peak ratios including 25%, 50%, 75%, and 100% and implemented downstream analysis[26, 27] on the Stimulated Droplet dataset[21]. With regard to single-nucleotide polymorphisms (SNPs) enrichment analysis, cell type-specific peaks identified by CASTLE with the ratio of 50% (Fig. 6c-e and Supplementary Figure 41) show more remarkable cell type specificity than that with other ratios (Supplementary Figures 47-50). As shown in Supplementary Figure 47a-b, at the feature-to-peak ratio of 50%, in 15 out of 28 cases (53.6%), the SNP enrichment folds between the number of significantly enriched tissues or significantly enriched blood-related tissues on the cell type-specific peaks versus background peaks are the highest. In contrast, considering ties, at the feature-to-peak ratio of 25%, 75% and 100%, the highest fold changes only appear 8 (28.6%), 2 (7.1%) and 9 (32.1%) cases, respectively. For example, the background peaks identified with the feature-to-peak ratios of 25% and 75% also exhibit strong cell type specificity (Supplementary Figure 48a and Supplementary Figure 49a), which is not desirable. PB-CD19+B_cells is the tissue with the most significant enrichment for pDC-specific peaks identified with the feature-to-peak ratios of 75% and 100% (Supplementary Figure 49b and Supplementary Figure 50b). About heritability enrichment analysis, the enrichment results with the feature-to-peak ratio of 50% (Fig. 6f-g and Supplementary Figure 42) are relatively stronger than that with other ratios (Supplementary Figure 47 and Supplementary Figures 51-53). Taking all of the 10 traits into consideration (Supplementary Figure 47c-l), in 118 out of 140 cases (84.3%), the highest fold changes between the heritability enrichments of cell type-specific peaks and background peaks for blood-related traits appear at the feature-to-peak ratio of 50%. In contrast, at the feature-to-peak ratio of 25%, 75% and 100%, the highest fold changes only appear in 5 (3.6%), 15 (10.7%) and 2 (1.4%) cases, respectively. Specifically, the background peaks identified with the feature-to-peak ratio of 100% also display high heritability enrichment (Supplementary Figure 53), which is not desirable. The CD8-specific peaks identified with the feature-to-peak ratios of 25%, 75% and 100% show lower heritability enrichment than the background peaks (Supplementary Figure 51a-b, Supplementary Figure 52a and Supplementary Figure 53a-b), which is not evident in the CD8-specific peaks identified with the feature-to-peak ratio of 50% (Fig. 6f-g). Overall, we first theorized that 50% might be better feature-to-peak ratio in principle, and then demonstrated by the comparative experiments that 50% is indeed a better feature-to-peak ratio compared with 25%, 75% and

100% on the Stimulated Droplet dataset.

Moreover, we conducted SNP enrichment analysis and heritability enrichment analysis with different feature-to-peak ratios on 4 additional real datasets, including the Resting Droplet dataset[21], the Immune dataset[28], the Fetal Lung dataset[41] and the Fetal Liver dataset[41]. For the Resting Droplet dataset, as shown in Supplementary Figure 54a-b, at the feature-to-peak ratio of 50%, in 20 out of 28 cases (71.4%), the SNP enrichment folds between the number of significantly enriched tissues or significantly enriched dataset-related tissues on the cell type-specific peaks versus background peaks are the highest. In contrast, considering ties, at the feature-to-peak ratio of 25%, 75% and 100%, the highest fold changes only appear 5 (17.9%), 3 (10.7%) and 5 (17.9%) cases, respectively. Similarly, taking all of the 4 traits into consideration (Supplementary Figure 54c-f), in 27 out of 56 cases (48.2%), the highest fold changes between the heritability enrichments of cell type-specific peaks and background peaks for dataset-related traits appear at the feature-to-peak ratio of 50%. In contrast, considering ties, at the feature-to-peak ratio of 25%, 75% and 100%, the highest fold changes only appear in 8 (14.3%), 14 (25.0%) and 7 (12.5%) cases, respectively. Specifically, at the feature-to-peak ratio of 50%, we identify PB-BDCA4+Dentritic_cells as the tissue with most significant enrichment (p-value < 5e-5) for pDC-specific peaks (Supplementary Figure 55b) and 721_B_Lymphoblasts as the tissue with the most significant enrichment (p-value < 2e-3) for B-specific peaks (Supplementary Figure 55c). The enrichment of heritability for the HbA1c in the cell type-specific peaks is higher than that in the background peaks (Supplementary Figure 56a).

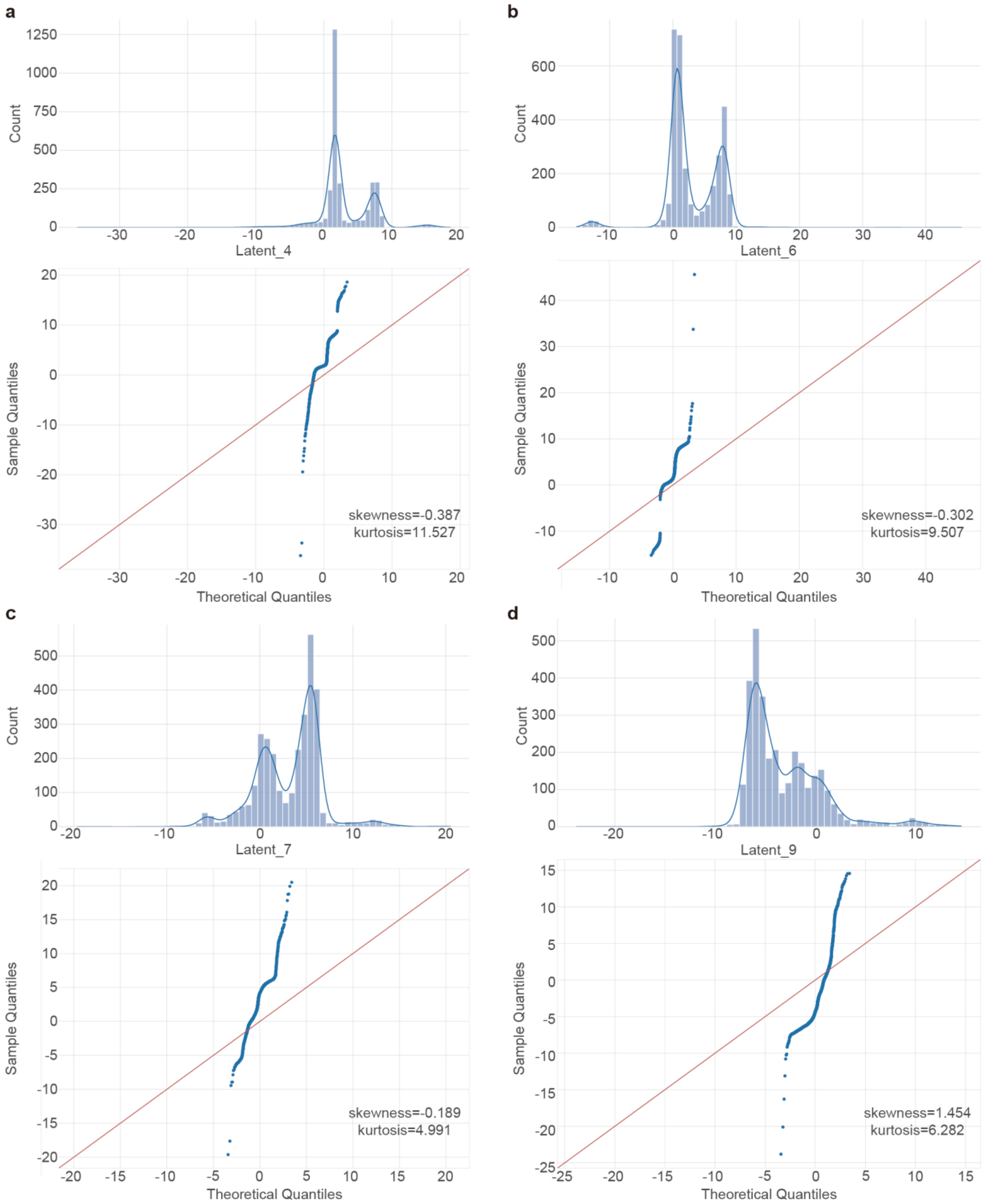
For the Immune dataset, as shown in Supplementary Figure 57a-b, at the feature-to-peak ratio of 50%, in 12 out of 20 cases (60.0%), the SNP enrichment folds between the number of significantly enriched tissues or significantly enriched dataset-related tissues on the cell type-specific peaks versus background peaks are the highest. In contrast, considering ties, at the feature-to-peak ratio of 25%, 75% and 100%, the highest fold changes only appear 2 (10.0%), 5 (25.0%) and 7 (35.0%) cases, respectively. Similarly, taking all of the 4 traits into consideration (Supplementary Figure 57c-f), in 31 out of 40 cases (77.5%), the highest fold changes between the heritability enrichments of cell type-specific peaks and background peaks for dataset-related traits appear at the feature-to-peak ratio of 50%. In contrast, considering ties, at the feature-to-peak ratio of 25%, 75% and 100%, the highest fold changes only appear in 11 (27.5%), 0 (0.0%) and 14 (35.0%) cases, respectively. Specifically, at the feature-to-peak ratio of 50%, PB-BDCA4+Dentritic_cells is the tissue with the most significant enrichment (p-value < 5e-6) for pDC-specific peaks (Supplementary Figure 58k). The enrichment of heritability for the Monocyte count in the cell type-specific peaks is higher than that in the background peaks except CLP cells (Supplementary Figure 59d).

For the Fetal Lung dataset, as shown in Supplementary Figure 60a-b, at the feature-to-peak ratio of 50%, in 9 out of 18 cases (50.0%), the SNP enrichment folds between the number of significantly enriched tissues or significantly enriched dataset-related tissues on the cell type-specific peaks versus background peaks are the highest. In contrast, considering ties, at the feature-to-peak ratio of 25%, 75% and 100%, the highest fold changes only appear 6 (33.3%), 5 (27.8%) and 5 (27.8%) cases, respectively. Similarly, taking all of the 4 traits into consideration (Supplementary Figure 60c-f), in 32 out of 36 cases (88.9%), the highest fold changes between the heritability enrichments of cell type-specific peaks and background peaks for dataset-related traits appear at the feature-to-peak ratio of 50%. In contrast, considering ties, at the feature-to-peak ratio of 25%, 75% and 100%, the highest fold changes only appear in 9 (25.0%), 2 (5.6%) and 0 (0.0%) cases, respectively. Specifically, at the feature-to-peak ratio of 50%, we identify lung and fetal lung as the tissues with significant enrichment (p-value < 2e-2) for cell type-specific peaks in bronchiolar and alveolar epithelial cells (Supplementary Figure 61b). The enrichment of heritability for the lung FVC smoke in the cell type-specific peaks is higher than that in the background peaks except ciliated epithelial cells and lymphatic endothelial cells (Supplementary Figure 62a).

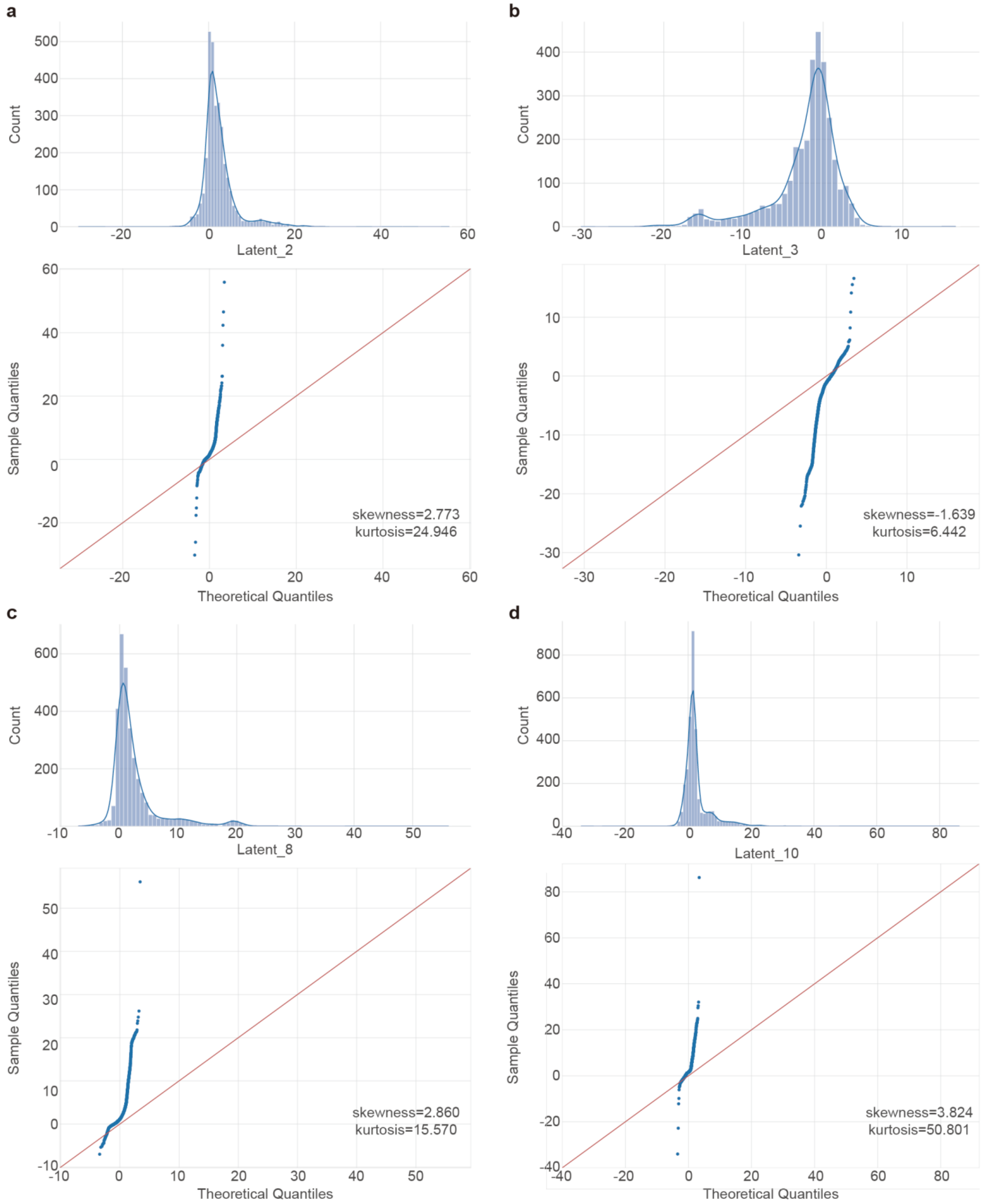
For the Fetal Liver dataset, as shown in Supplementary Figure 63a-b, at the feature-to-peak ratio of 50%, in 10 out of 16 cases (62.5%), the SNP enrichment folds between the number of significantly enriched tissues or significantly enriched dataset-related tissues on the cell type-specific peaks versus background peaks are the highest. In contrast, at the feature-to-peak ratio of 25%, 75% and 100%, the highest fold changes only appear 2 (12.5%), 3 (18.8%) and 1 (6.3%) cases, respectively. Similarly, taking all of the 4 traits into consideration (Supplementary Figure 63c-f), in 26 out of 32 cases (81.3%), the highest fold changes between the heritability enrichments of cell type-specific peaks and background peaks for dataset-related traits appear at the feature-to-peak ratio of 50%. In contrast, at the feature-to-peak ratio of 25%, 75% and 100%, the highest fold changes only appear in 6 (18.8%), 0 (0.0%) and 0 (0.0%) cases, respectively. Specifically, at the feature-to-peak ratio of 50%, we identify liver as the tissue with the most significant enrichment (p-value < 2e-3) for cell type-specific peaks in hematopoietic stem cells (Supplementary Figure 64c). The enrichment of heritability for the HDL cholesterol in the cell type-specific peaks is higher than that in the background peaks (Supplementary Figure 65b).

To sum up, we demonstrated by the comparative experiments that 50% is indeed a better feature-to-peak ratio compared with 25%, 75% and 100% (Supplementary Figures 47, 54, 57, 60, 63). At the feature-to-peak ratio of 50%, the identified cell type-specific peaks on 5 real datasets show remarkable tissue specificity (Supplementary Figures 41, 55, 58, 61, 64) and are functionally relevant and potentially involved in the regulation of the studied phenotypes (Supplementary Figures 42, 56, 59, 62, 65).

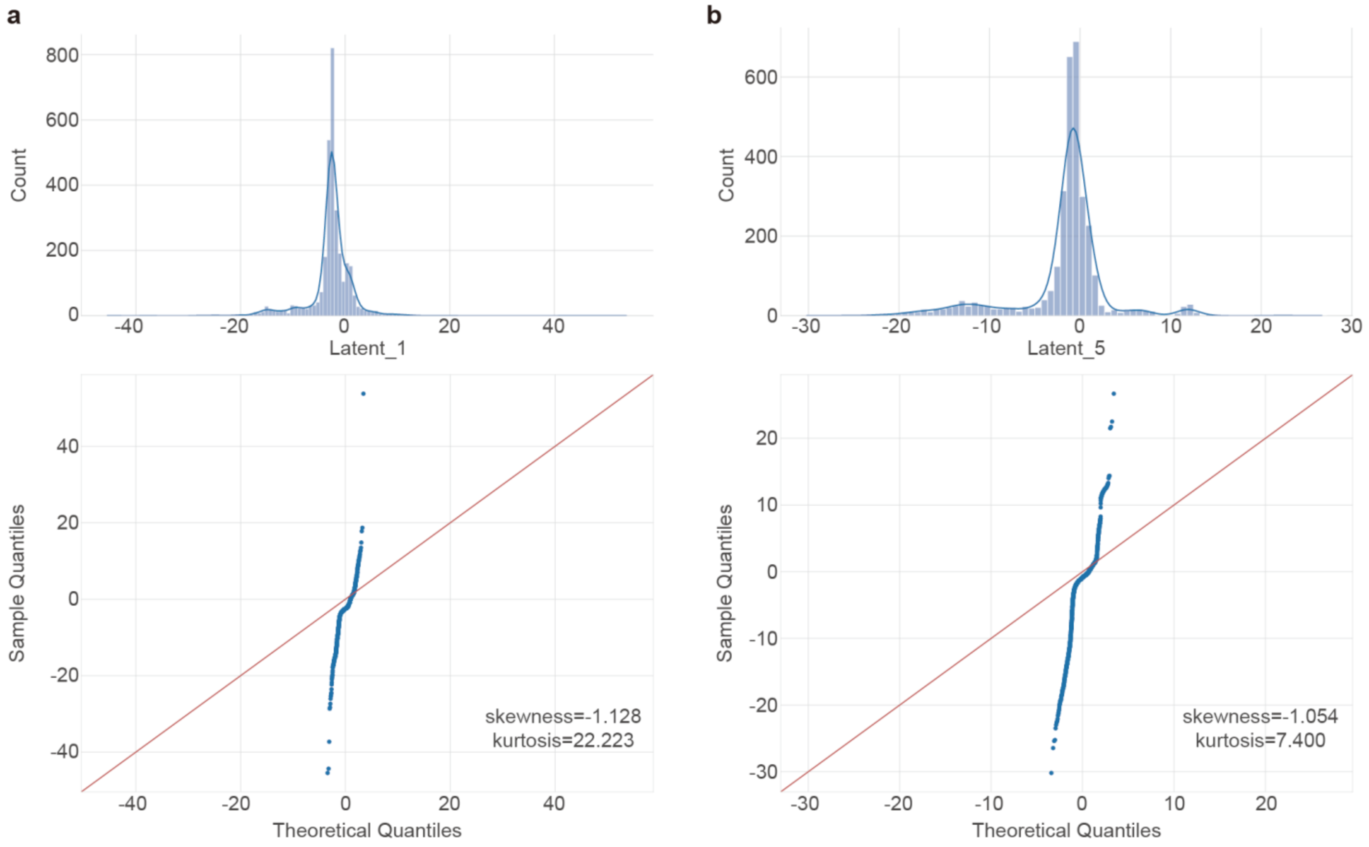
Supplementary Figures



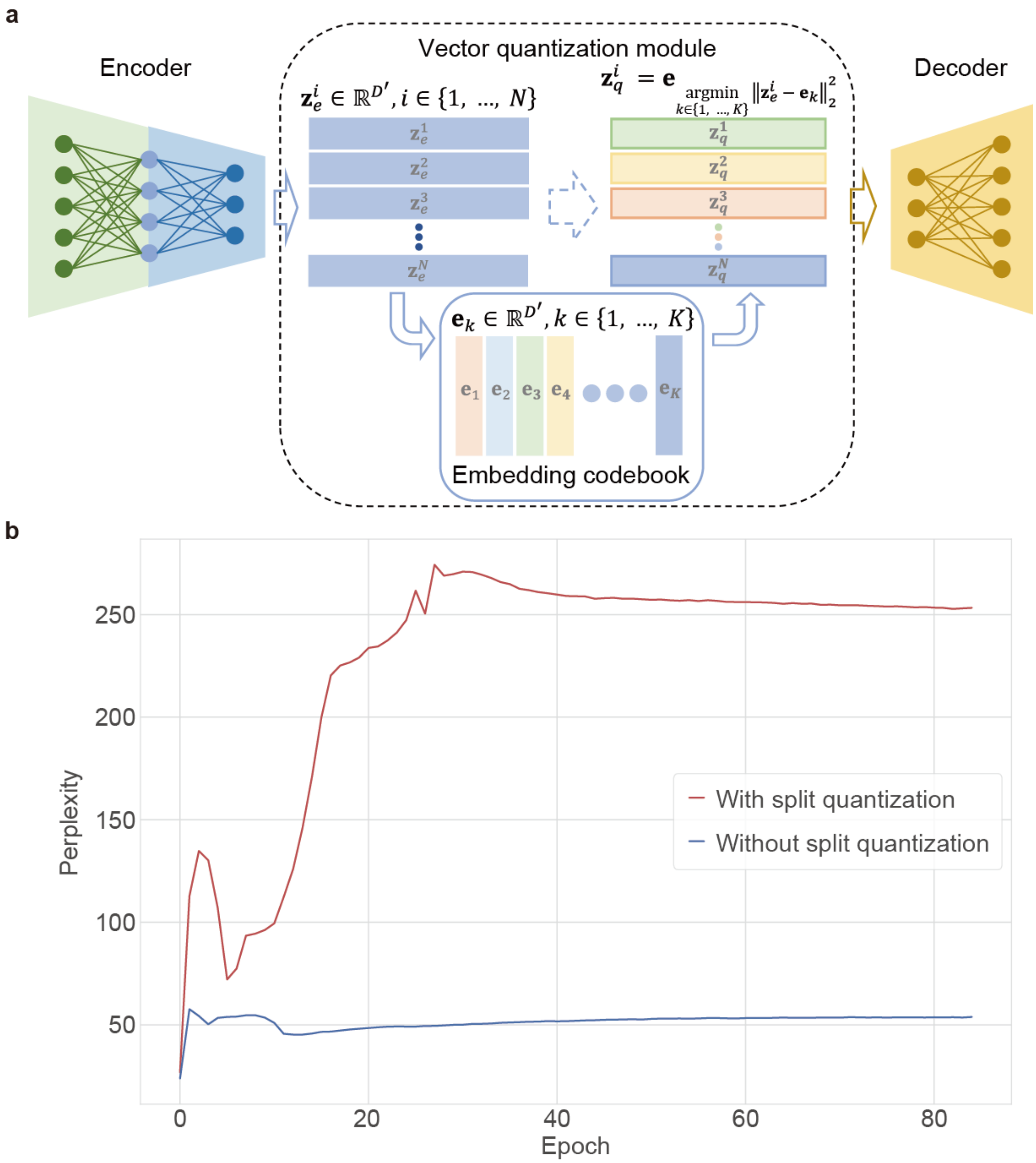
Supplementary Figure 1. Multimodal behavior of latent features in VAE. **a-d**, In the standard VAE model trained on the Splenocyte dataset, latent feature 4 (**a**), 6 (**b**), 7 (**c**) and 9 (**d**) exhibit multimodal behavior by the frequency distribution histogram (top) and Q-Q plot (bottom).



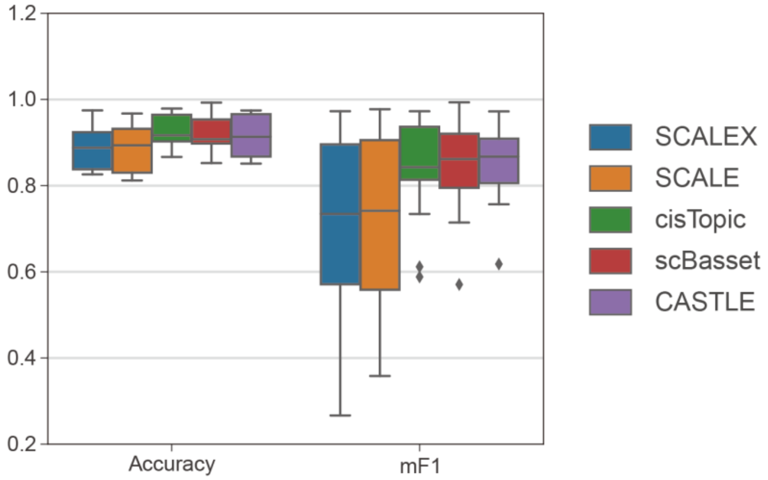
Supplementary Figure 2. Unexpected skewness of latent features in VAE. **a-d**, In the standard VAE model trained on the Splenocyte dataset, latent feature 2 (**a**), 3 (**b**), 8 (**c**) and 10 (**d**) exhibit unexpected skewness (equal to 0 in standard normal distribution) by the frequency distribution histogram (top) and Q-Q plot (bottom).



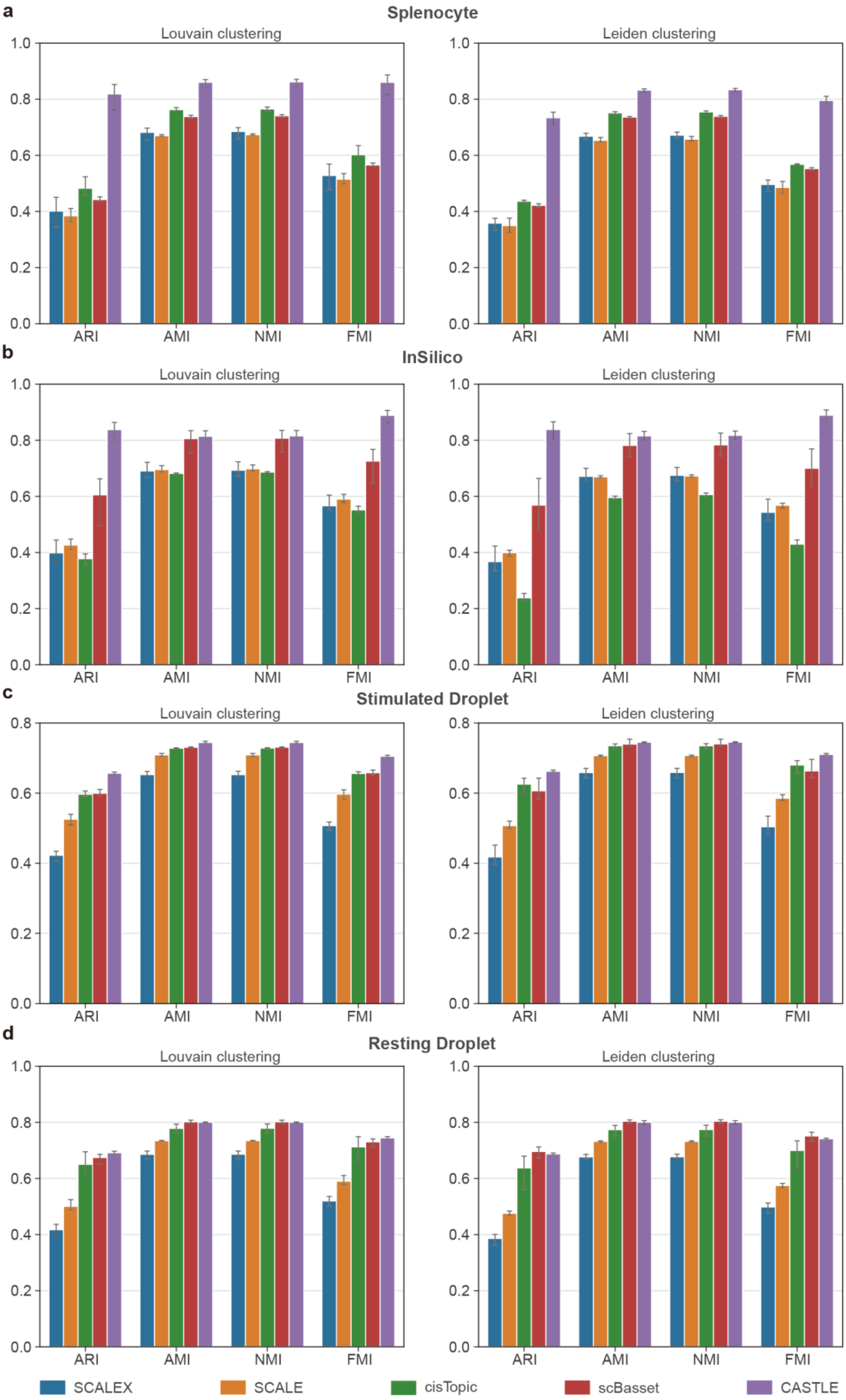
Supplementary Figure 3. Unexpected kurtosis of latent features in VAE. **a-b**, In the standard VAE model trained on the Splenocyte dataset, latent feature 1 (**a**) and 5 (**b**) exhibit unexpected kurtosis (equal to 3 in standard normal distribution) by the frequency distribution histogram (top) and Q-Q plot (bottom).



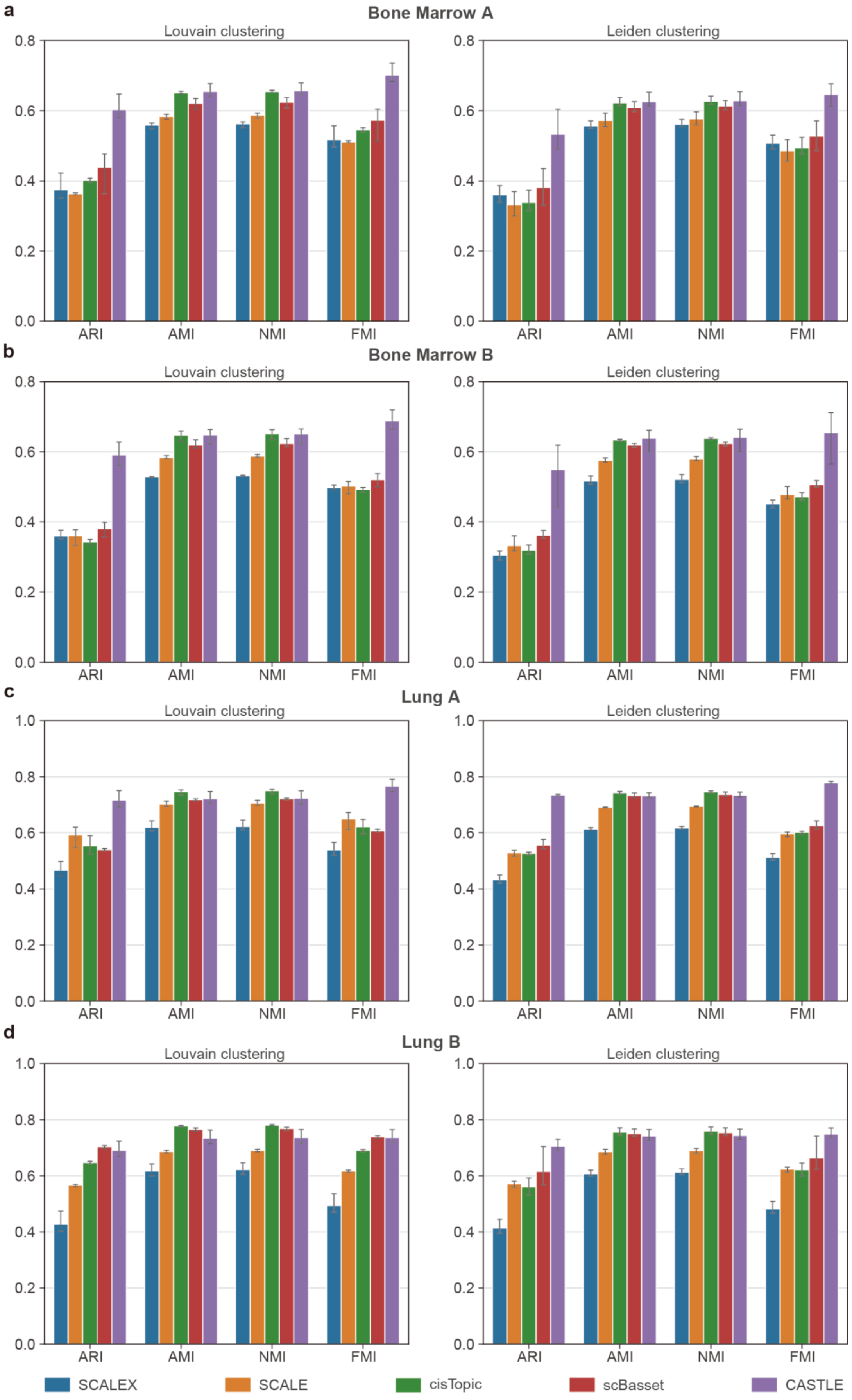
Supplementary Figure 4. Comparison between the models with and without split quantization. a, The model architecture without split quantization. **b**, Utilization of the codebook (that is, perplexity) increases when applying split quantization on the Stimulated Droplet dataset.



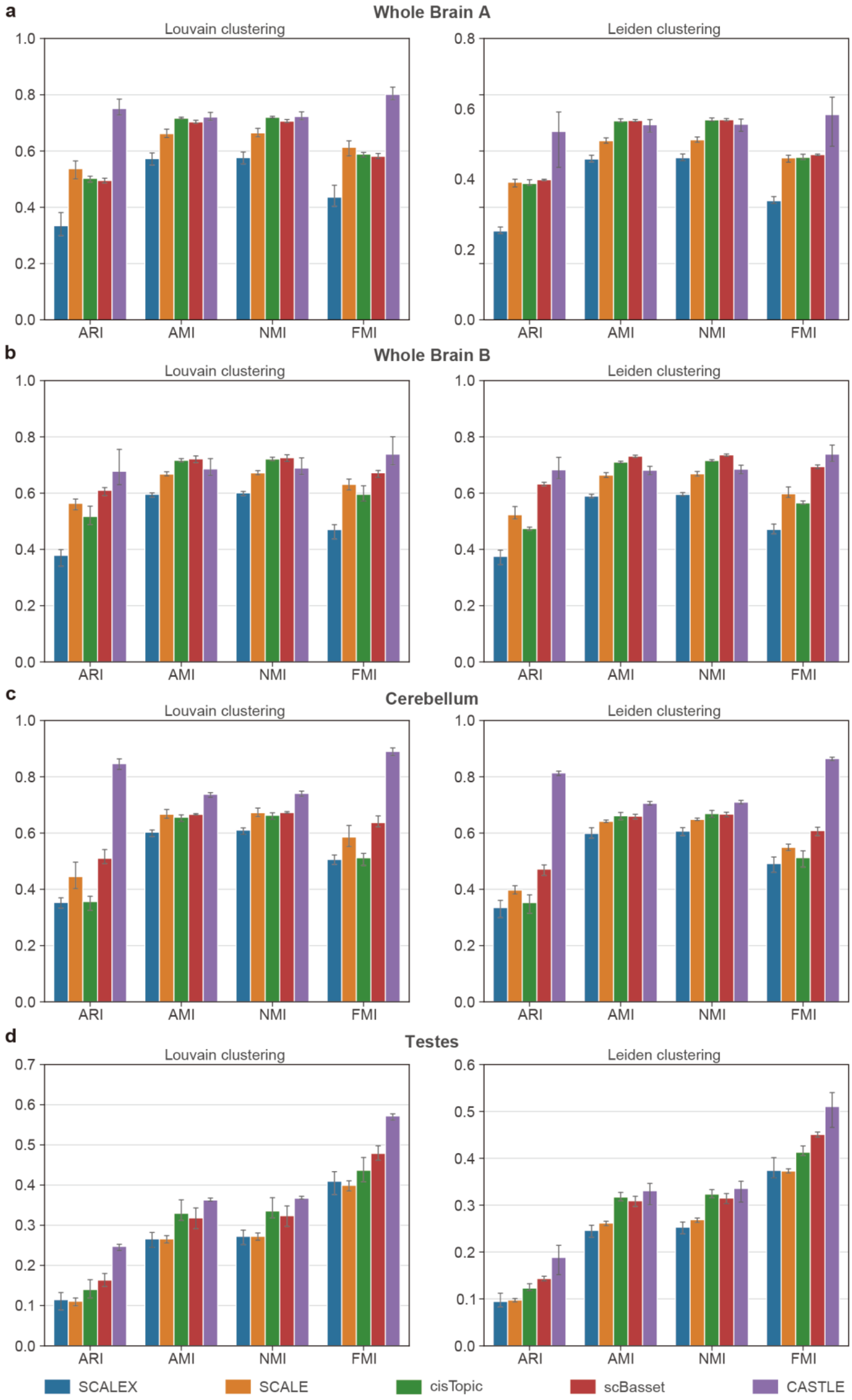
Supplementary Figure 5. Evaluation of performance based on kNN classification. Boxplot of kNN classification performance for CASTLE and four baseline methods including SCALEX, SCALE, cisTopic and scBasset, evaluated by Accuracy and mF1 on 16 benchmark datasets. Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range and points represent outliers.



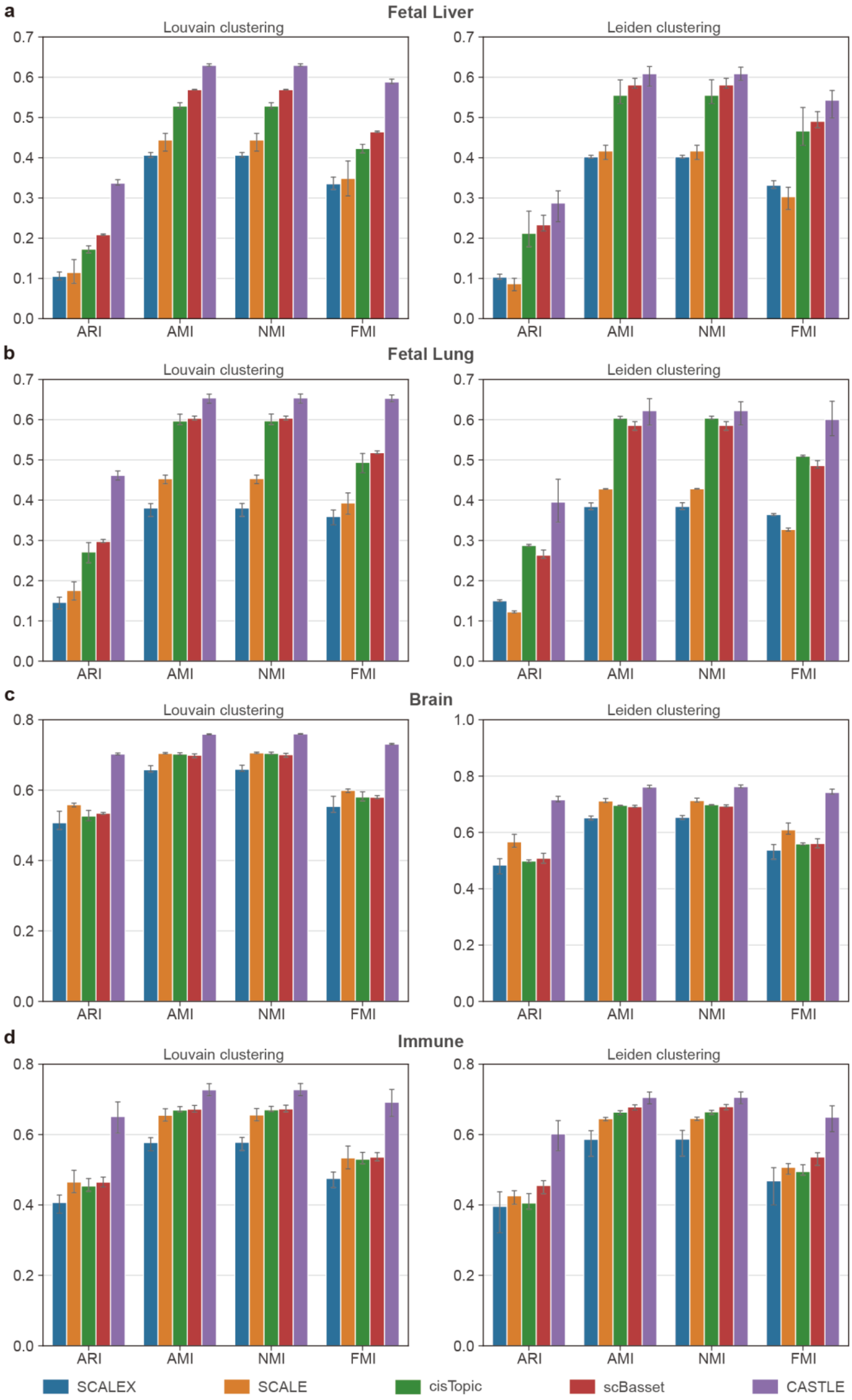
Supplementary Figure 6. Clustering performance of CASTLE compared with baseline methods. a-d, Performance evaluated by ARI, AMI, NMI and FMI based on the Louvain clustering and Leiden clustering on the Splenocyte (**a**), InSilico (**b**), Stimulated Droplet (**c**) and Resting Droplet (**d**) datasets compared with baseline methods. Each method was run three times with different random seeds. The error bars denote the 95% confidence interval, and the centers of error bars denote the average value.



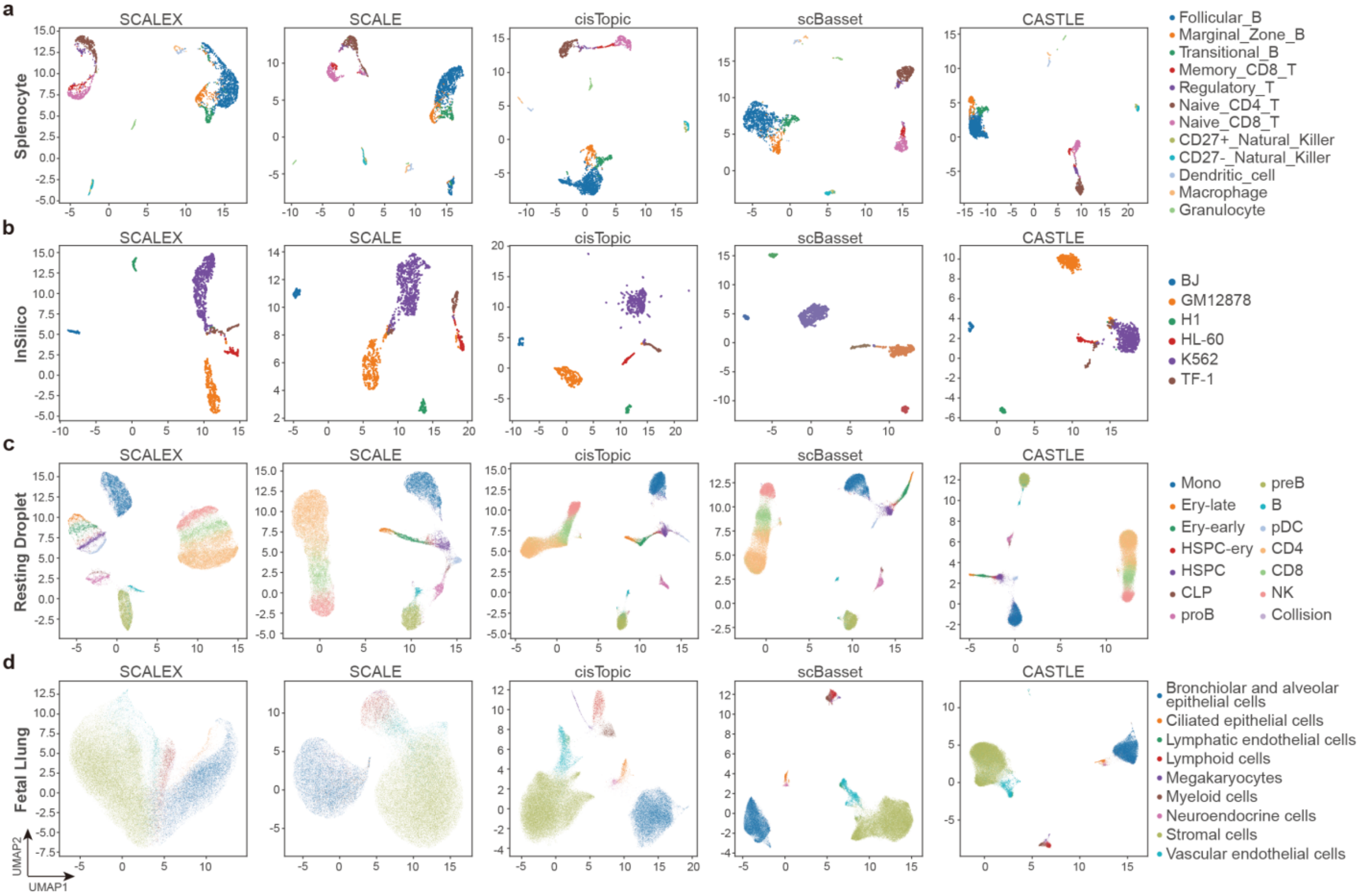
Supplementary Figure 7. Clustering performance of CASTLE compared with baseline methods. a-d, Performance evaluated by ARI, AMI, NMI and FMI based on the Louvain clustering and Leiden clustering on the Bone Marrow A (**a**), Bone Marrow B (**b**), Lung A (**c**) and Lung B (**d**) datasets compared with baseline methods. Each method was run three times with different random seeds. The error bars denote the 95% confidence interval, and the centers of error bars denote the average value.



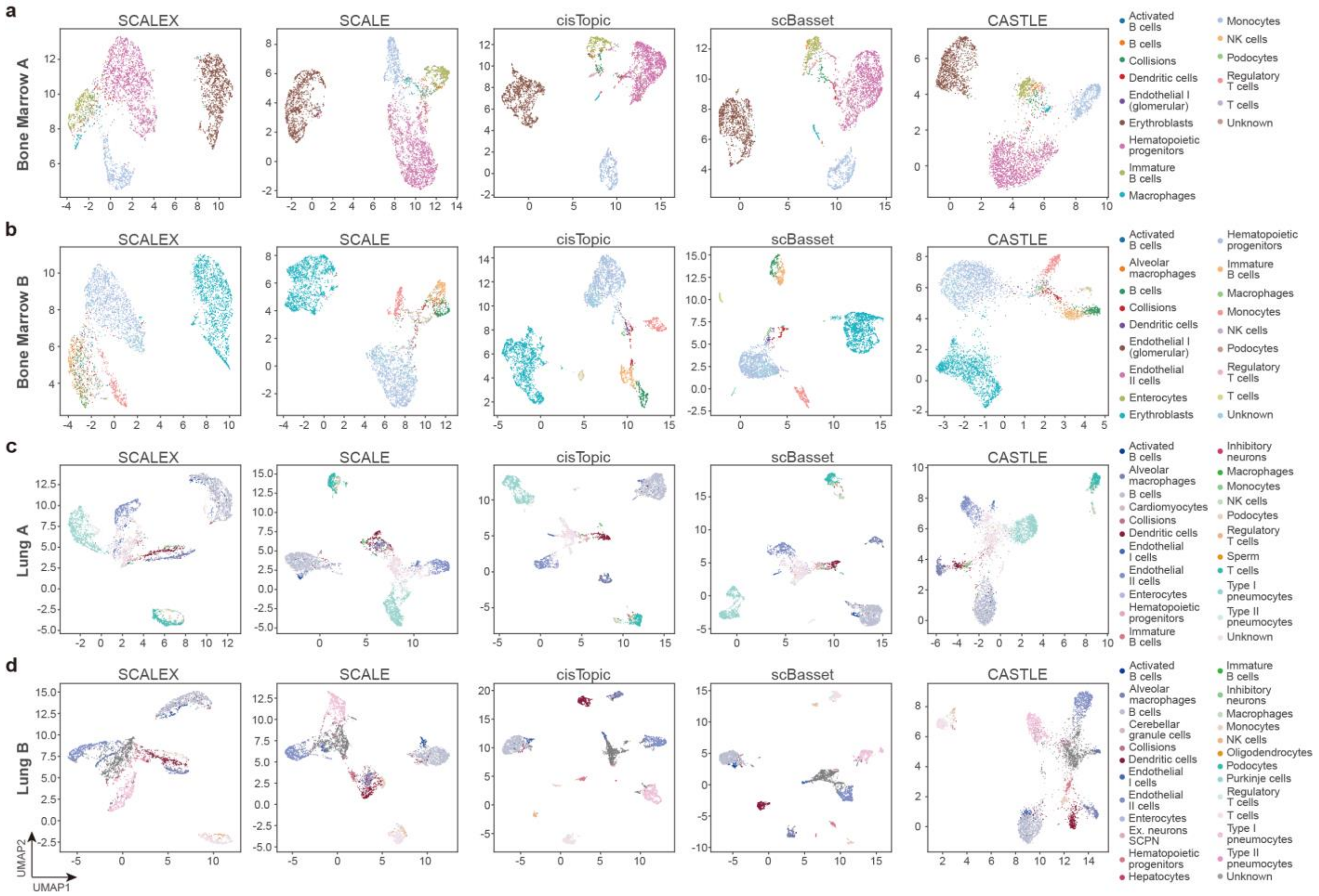
Supplementary Figure 8. Clustering performance of CASTLE compared with baseline methods. a-d, Performance evaluated by ARI, AMI, NMI and FMI based on the Louvain clustering and Leiden clustering on the Whole Brain A (**a**), Whole Brain B (**b**), Cerebellum (**c**) and Testes (**d**) datasets compared with baseline methods. Each method was run three times with different random seeds. The error bars denote the 95% confidence interval, and the centers of error bars denote the average value.



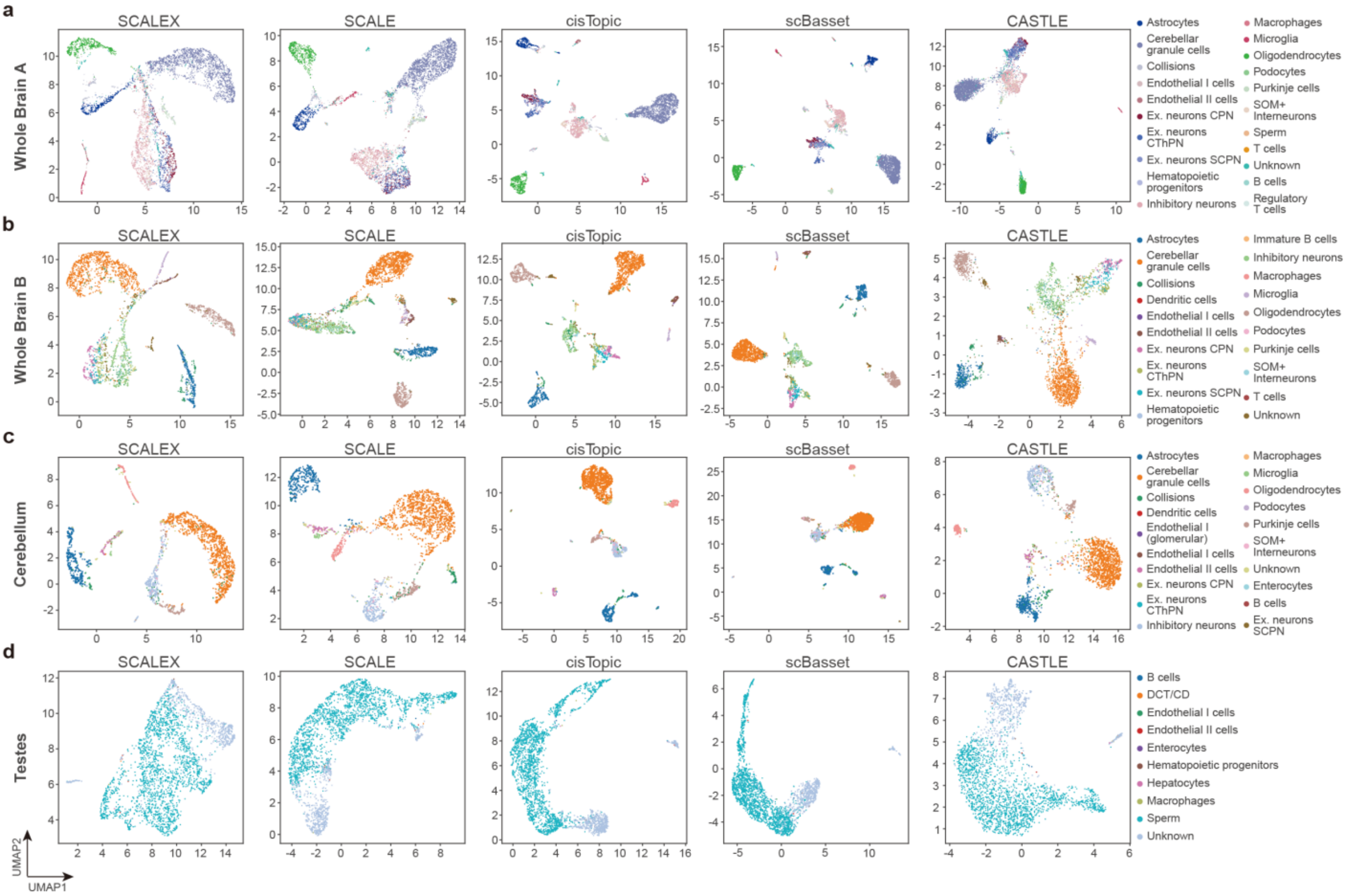
Supplementary Figure 9. Clustering performance of CASTLE compared with baseline methods. a-d, Performance evaluated by ARI, AMI, NMI and FMI based on the Louvain clustering and Leiden clustering on the Fetal Liver (**a**), Fetal Lung (**b**), Brain (**c**) and Immune (**d**) datasets compared with baseline methods. Each method was run three times with different random seeds. The error bars denote the 95% confidence interval, and the centers of error bars denote the average value.



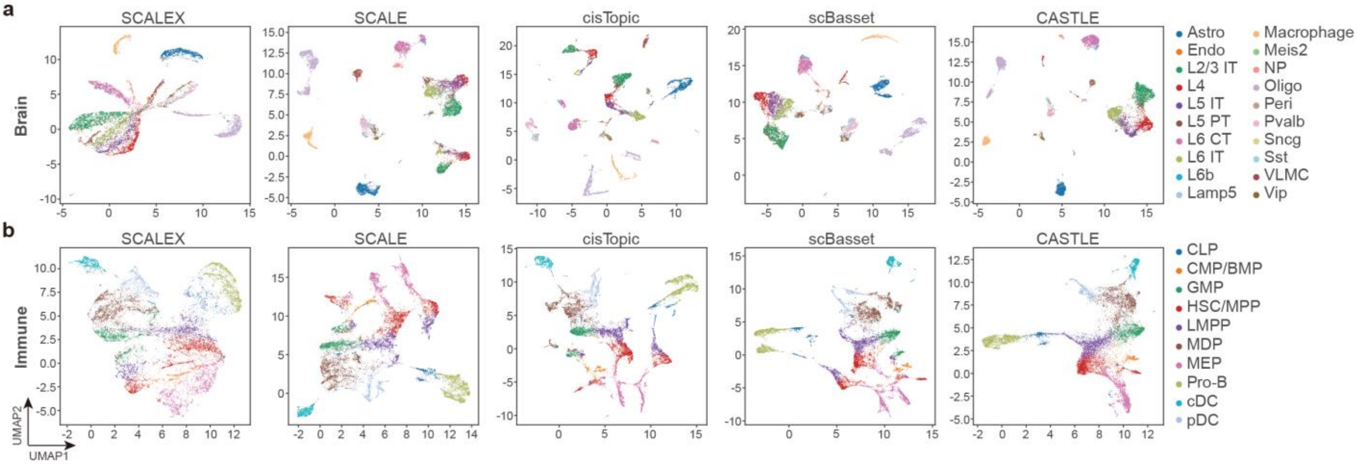
Supplementary Figure 10. UMAP visualization of CASTLE compared with baseline methods. a-d, UMAP visualizations of the cell embeddings from SCALEX, SCALE, cisTopic, scBasset and CASTLE on the Splenocyte (**a**), InSilico (**b**), Resting Droplet (**c**) and Fetal Lung (**d**) datasets.



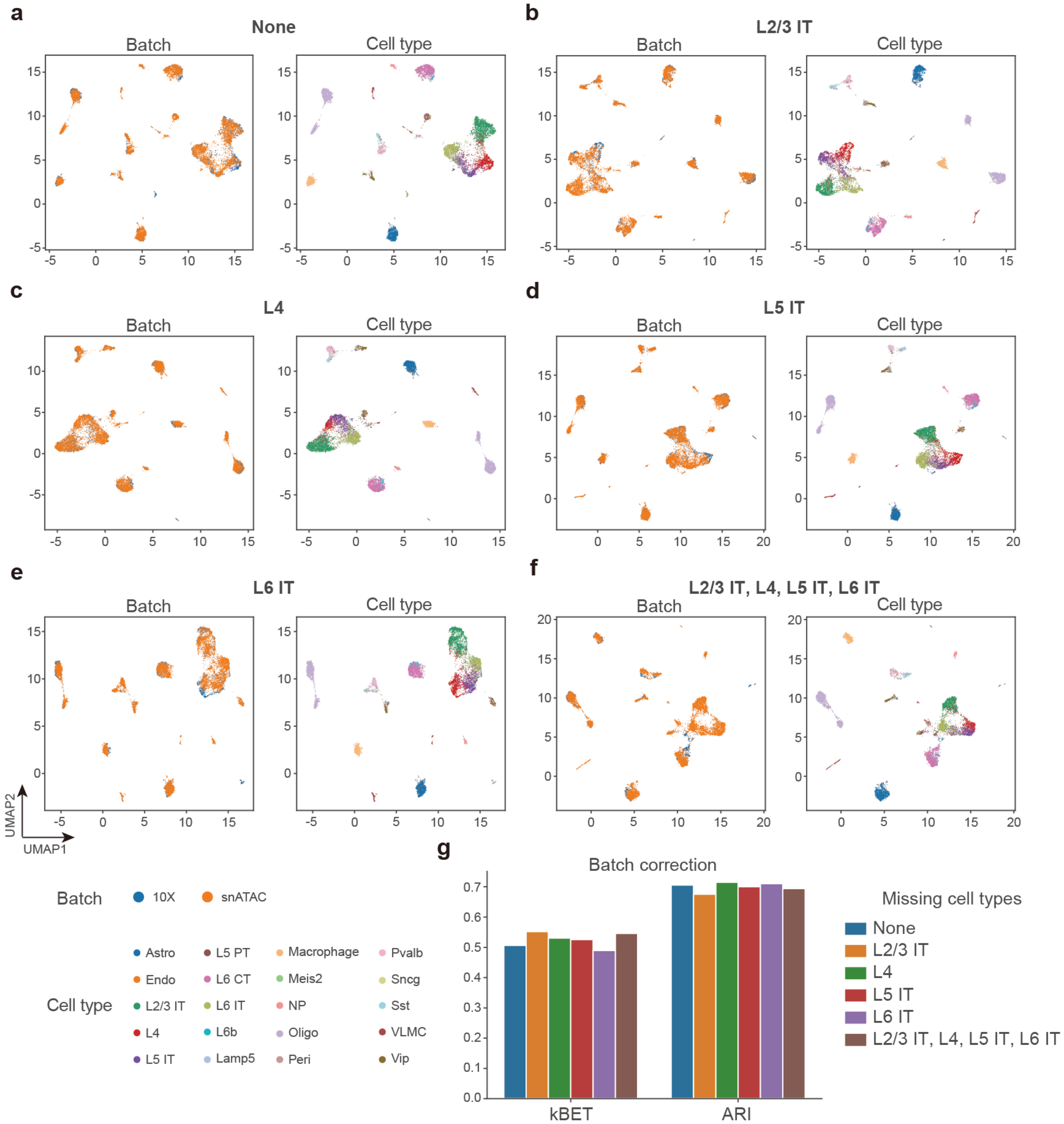
Supplementary Figure 11. UMAP visualization of CASTLE compared with baseline methods.a-d, UMAP visualizations of the cell embeddings from SCALEX, SCALE, cisTopic, scBasset and CASTLE on the Bone Marrow A (**a**), Bone Marrow B (**b**), Lung A (**c**) and Lung B (**d**) datasets.



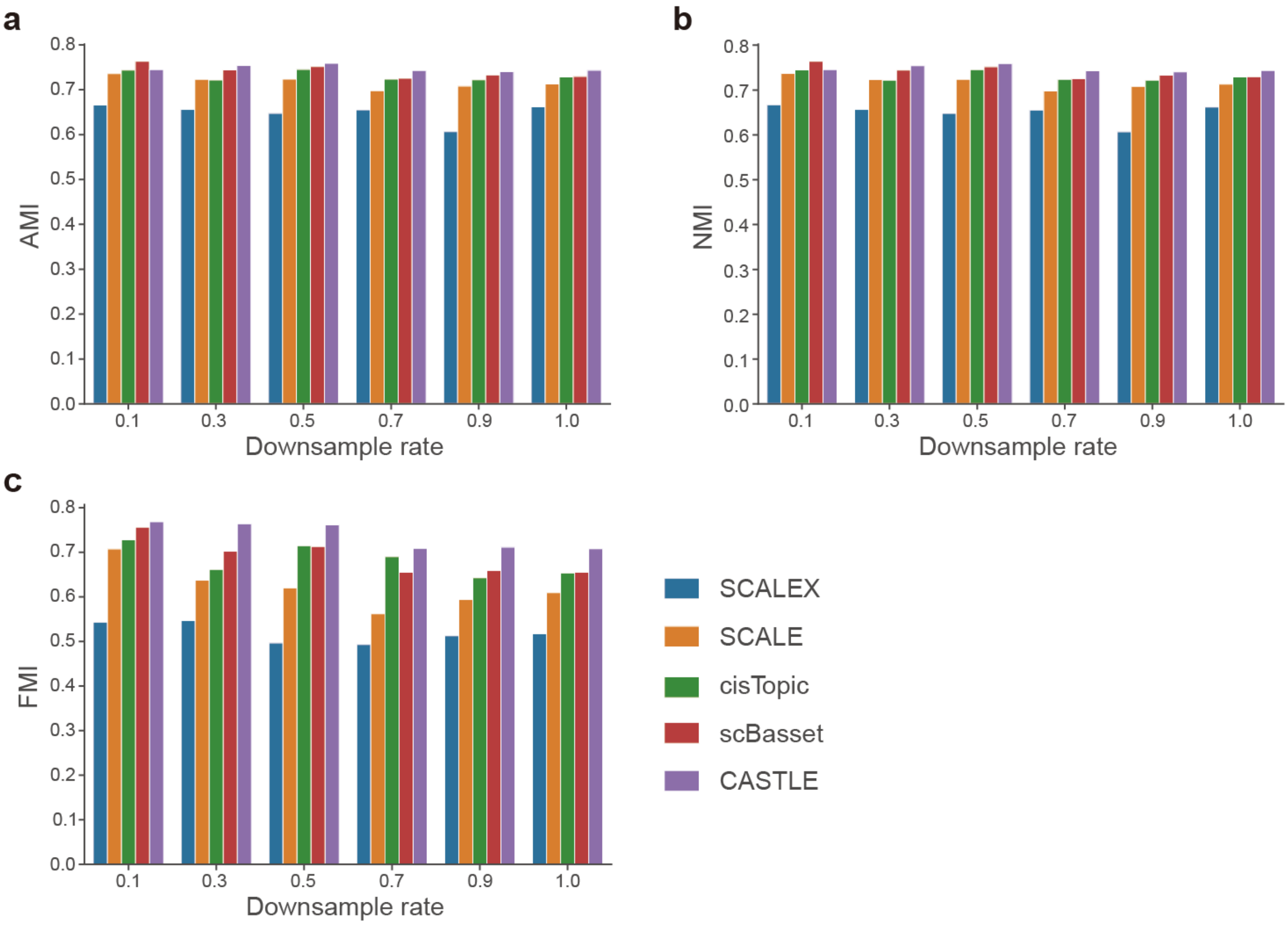
Supplementary Figure 12. UMAP visualization of CASTLE compared with baseline methods.a-d, UMAP visualizations of the cell embeddings from SCALEX, SCALE, cisTopic, scBasset and CASTLE on the Whole Brain A (**a**), Whole Brain B (**b**), Cerebellum (**c**) and Testes (**d**) datasets.



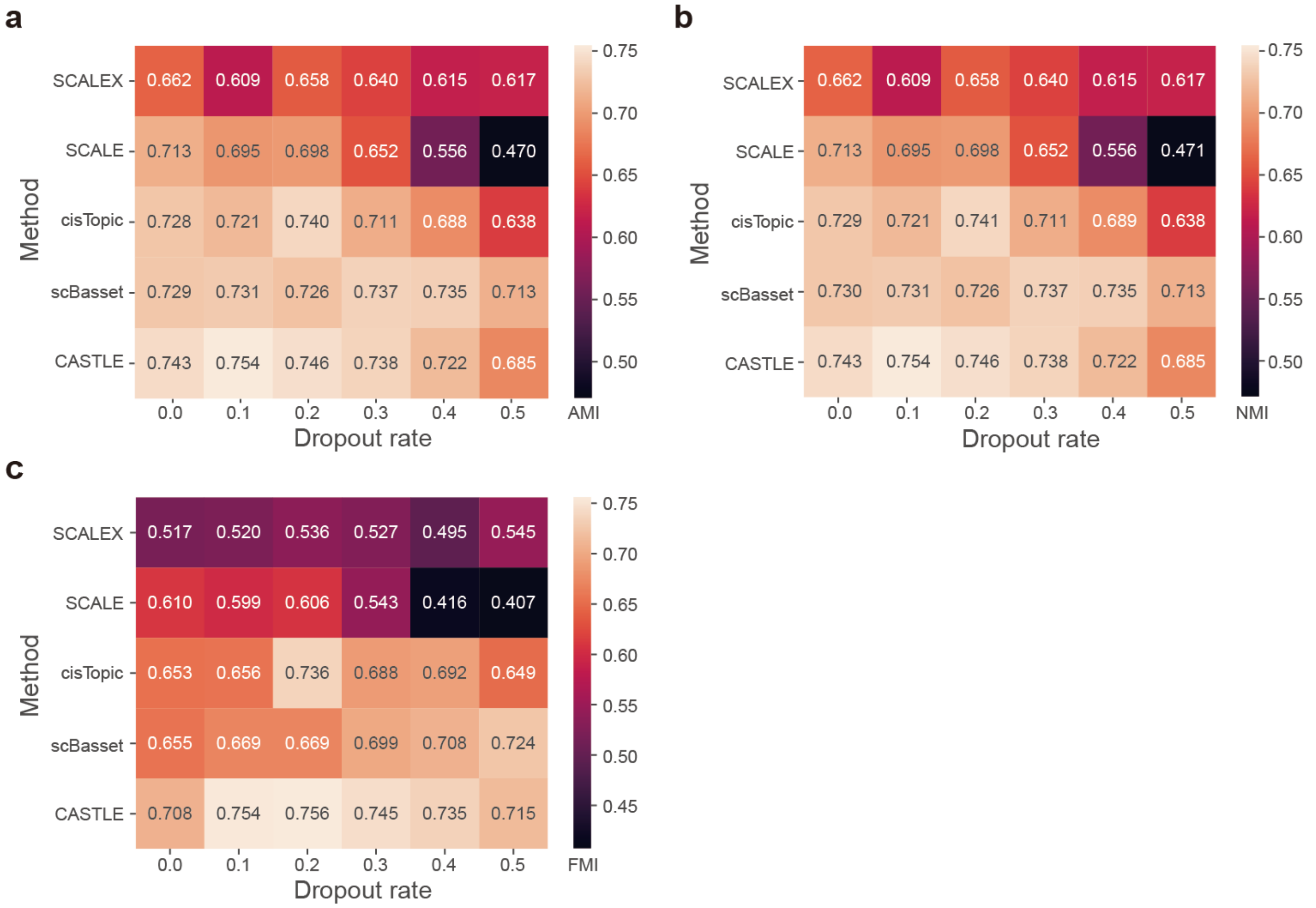
Supplementary Figure 13. UMAP visualization of CASTLE compared with baseline methods.a-b, UMAP visualizations of the cell embeddings from SCALEX, SCALE, cisTopic, scBasset and CASTLE on the Brain (**a**) and Immune (**b**) datasets.



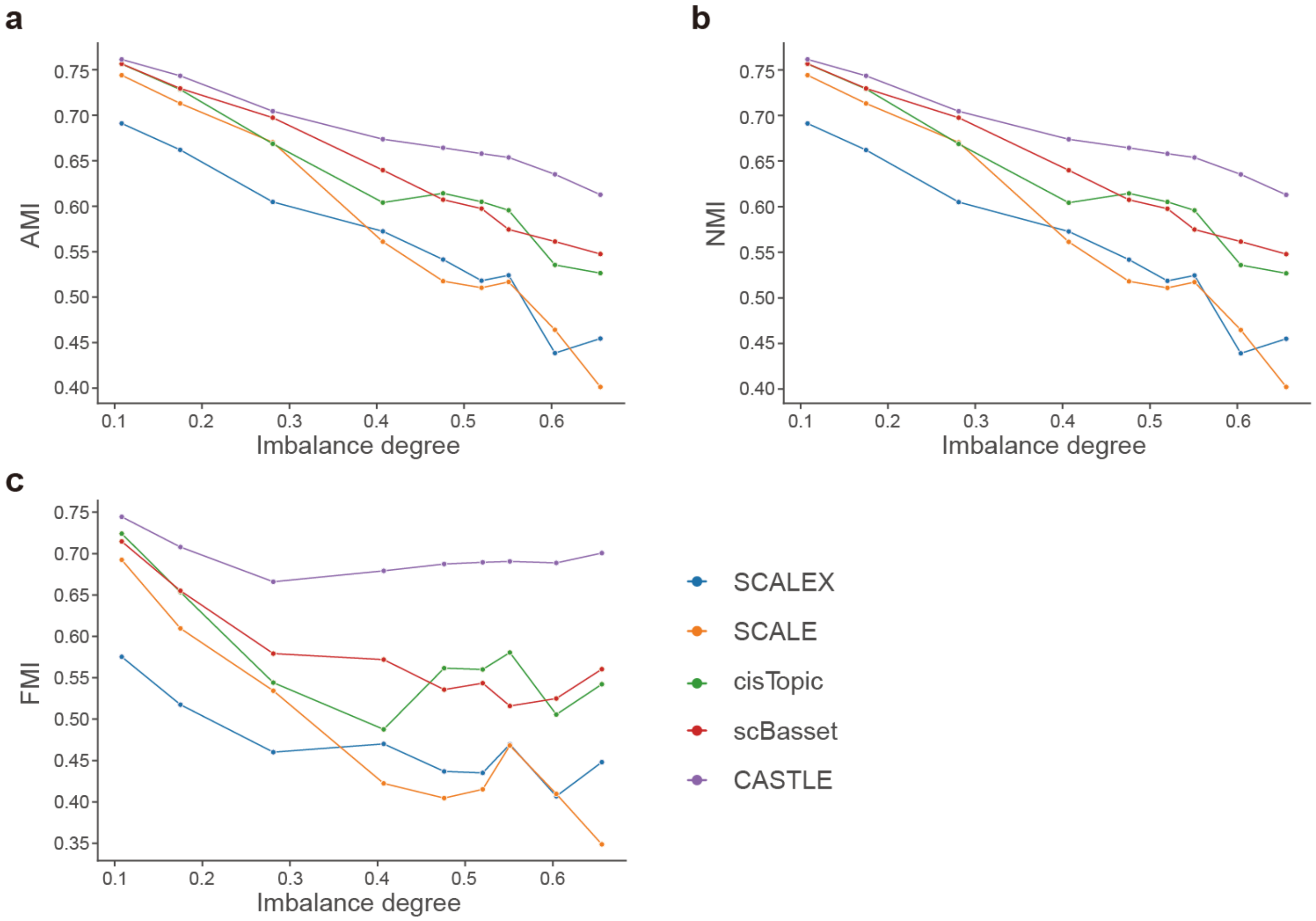
Supplementary Figure 14. Performance of batch correction for CASTLE over datasets with complex batch effects. a-f, UMAP visualizations of cell embeddings for the Brain dataset with None (**a**), L2/3 IT (**b**), L4 (**c**), L5 IT (**d**), L6 IT (**e**), and L2/3 IT, L4, L5 IT, L6 IT (**f**) missing cell types in the batch "10X". Cells are colored by data batch (left) and cell type (right). **g**, Performance comparison of batch correction for CASTLE on the Brain dataset with None, L2/3 IT, L4, L5 IT, L6 IT, and L2/3 IT, L4, L5 IT, L6 IT missing cell types in the batch "10X".



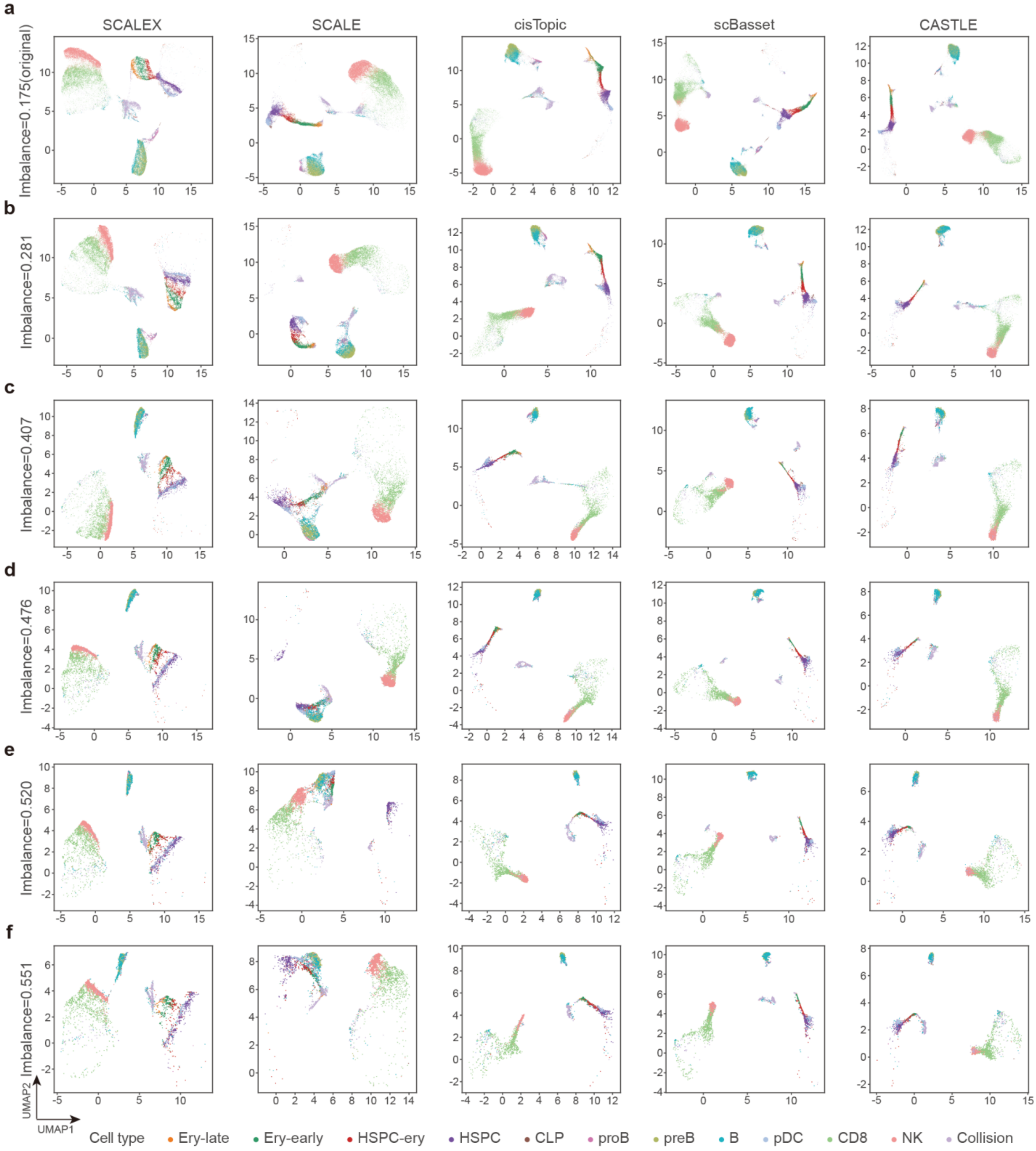
Supplementary Figure 15. Clustering performance under different downsample rates. a-c, AMI (**a**), NMI (**b**), FMI (**c**) of different methods under different downsample rates on the Stimulated Droplet dataset.



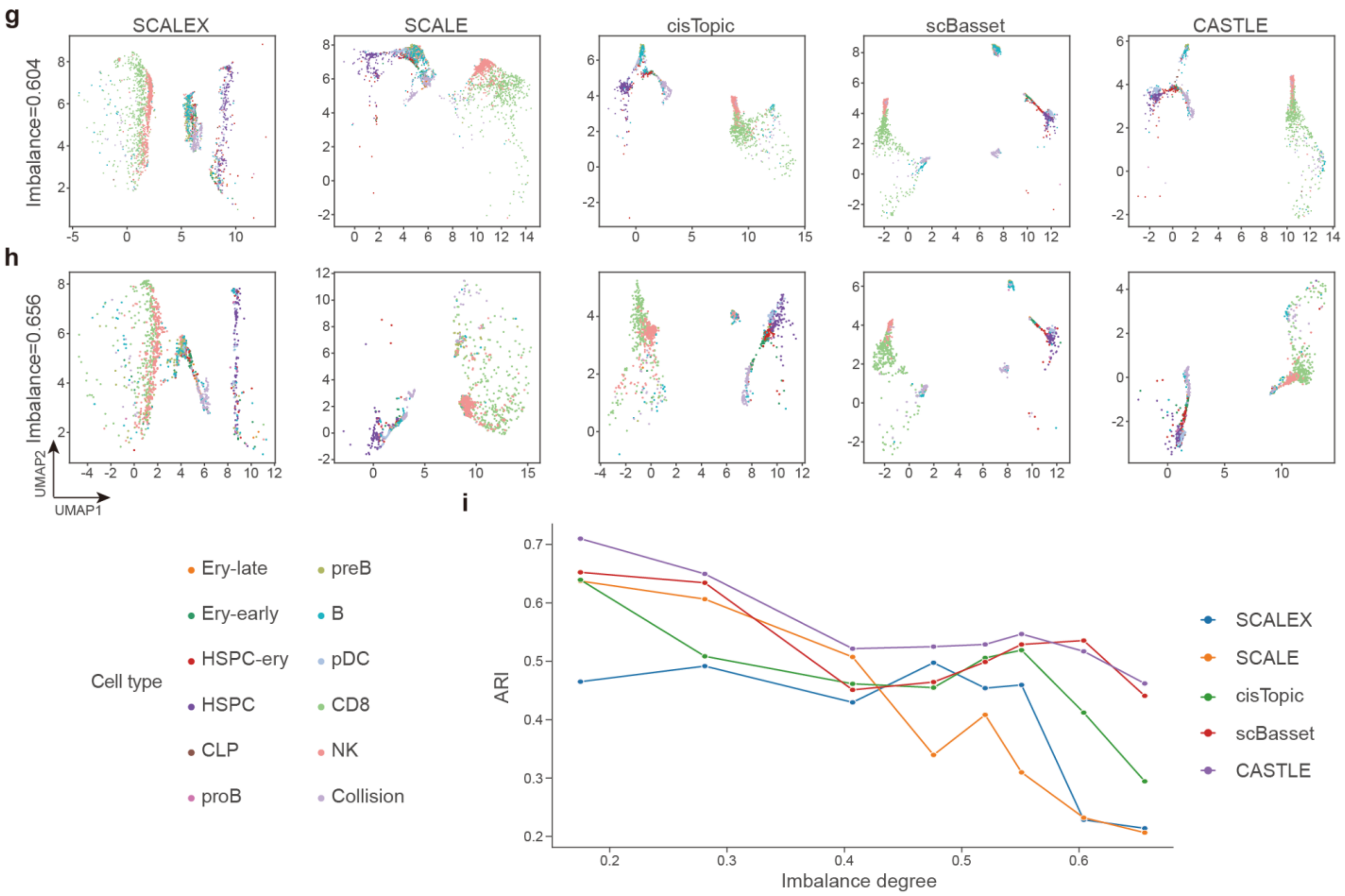
Supplementary Figure 16. Clustering performance under different dropout rates. a-c, AMI (**a**), NMI (**b**), FMI (**c**) of different methods under different dropout rates on the Stimulated Droplet dataset.



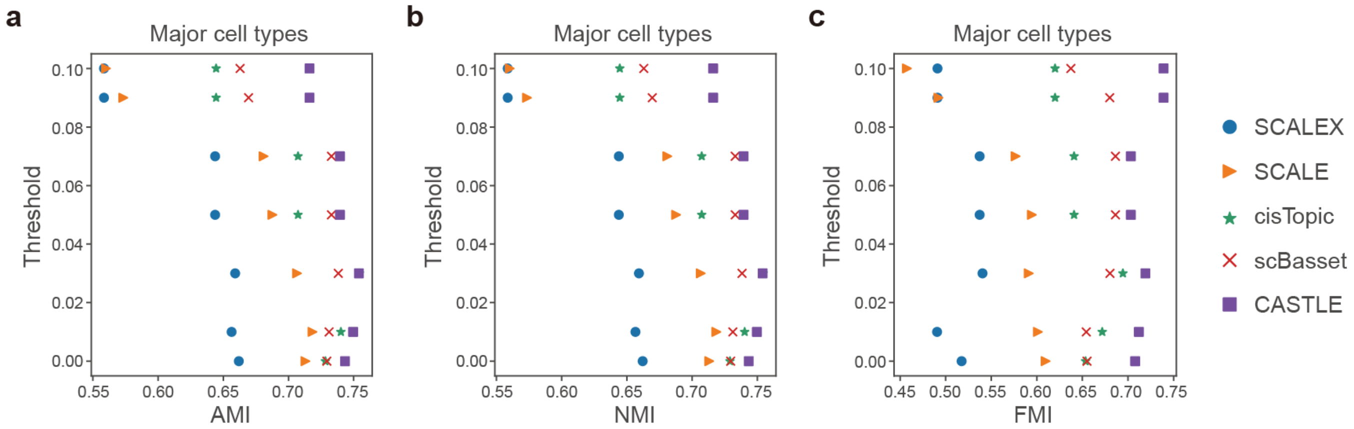
Supplementary Figure 17. Clustering performance under different imbalance degrees. a-c, AMI (**a**), NMI (**b**), FMI (**c**) of different methods under different degrees of cell type imbalance on the Stimulated Droplet dataset.



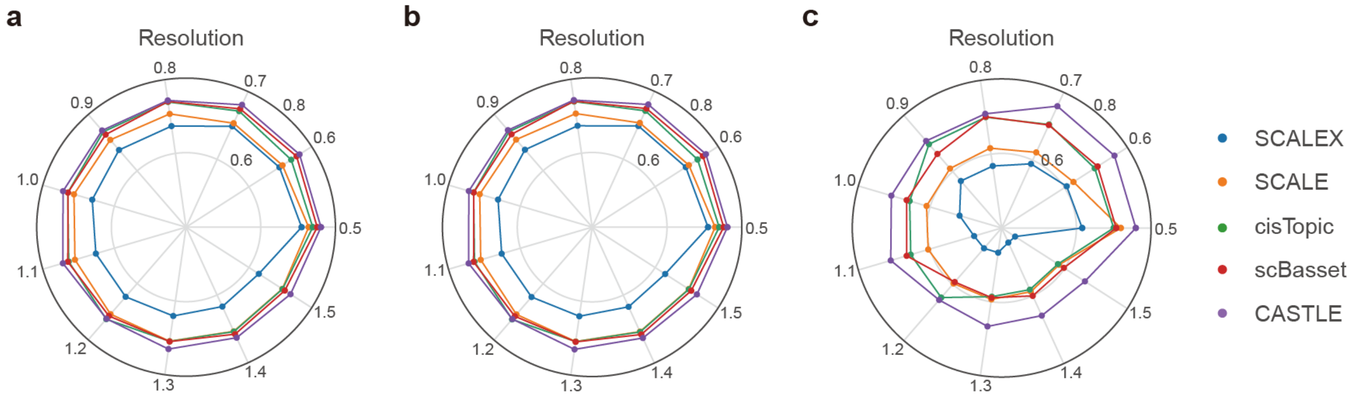
Supplementary Figure 18. Performance of CASTLE for the identification of rare cell types. a-h, UMAP visualizations of cell embeddings from SCALEX, SCALE, cisTopic, scBasset and CASTLE on the Stimulated Droplet dataset with degree of cell type imbalance of 0.175 (original) (**a**), 0.281 (**b**), 0.407 (**c**), 0.476 (**d**), 0.520 (**e**), 0.551 (**f**), 0.604 (**g**) and 0.656 (**h**). **i**, Clustering performance of different methods for recovering the rare cell types on the Stimulated Droplet datasets with different degrees of cell type imbalance.



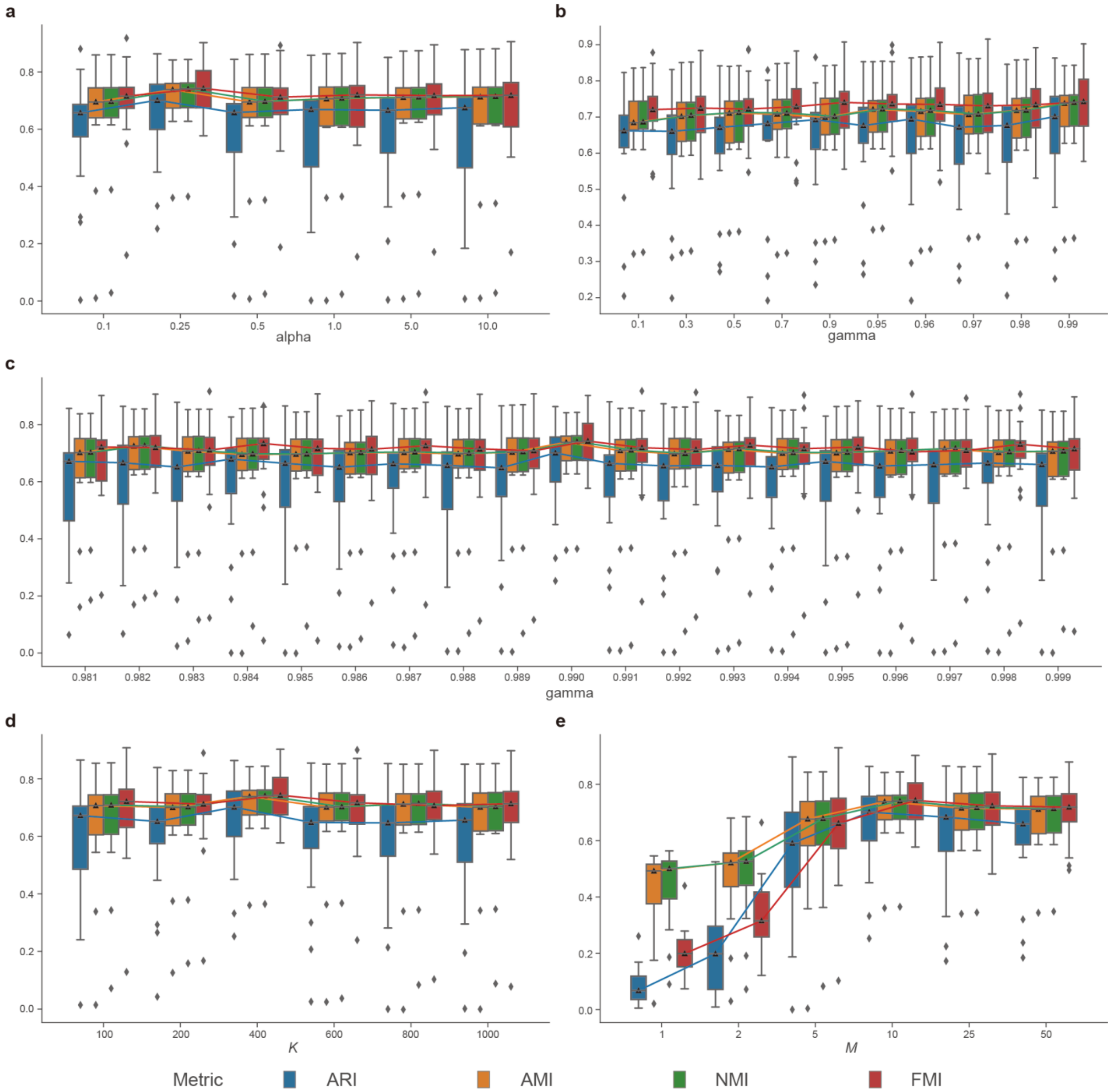
Supplementary Figure 18 (continue). Performance of CASTLE for the identification of rare cell types. a-h, UMAP visualizations of cell embeddings from SCALEX, SCALE, cisTopic, scBasset and CASTLE on the Stimulated Droplet dataset with degree of cell type imbalance of 0.175 (original) (**a**), 0.281 (**b**), 0.407 (**c**), 0.476 (**d**), 0.520 (**e**), 0.551 (**f**), 0.604 (**g**) and 0.656 (**h**). **i**, Clustering performance of different methods for recovering the rare cell types on the Stimulated Droplet datasets with different degrees of cell type imbalance.



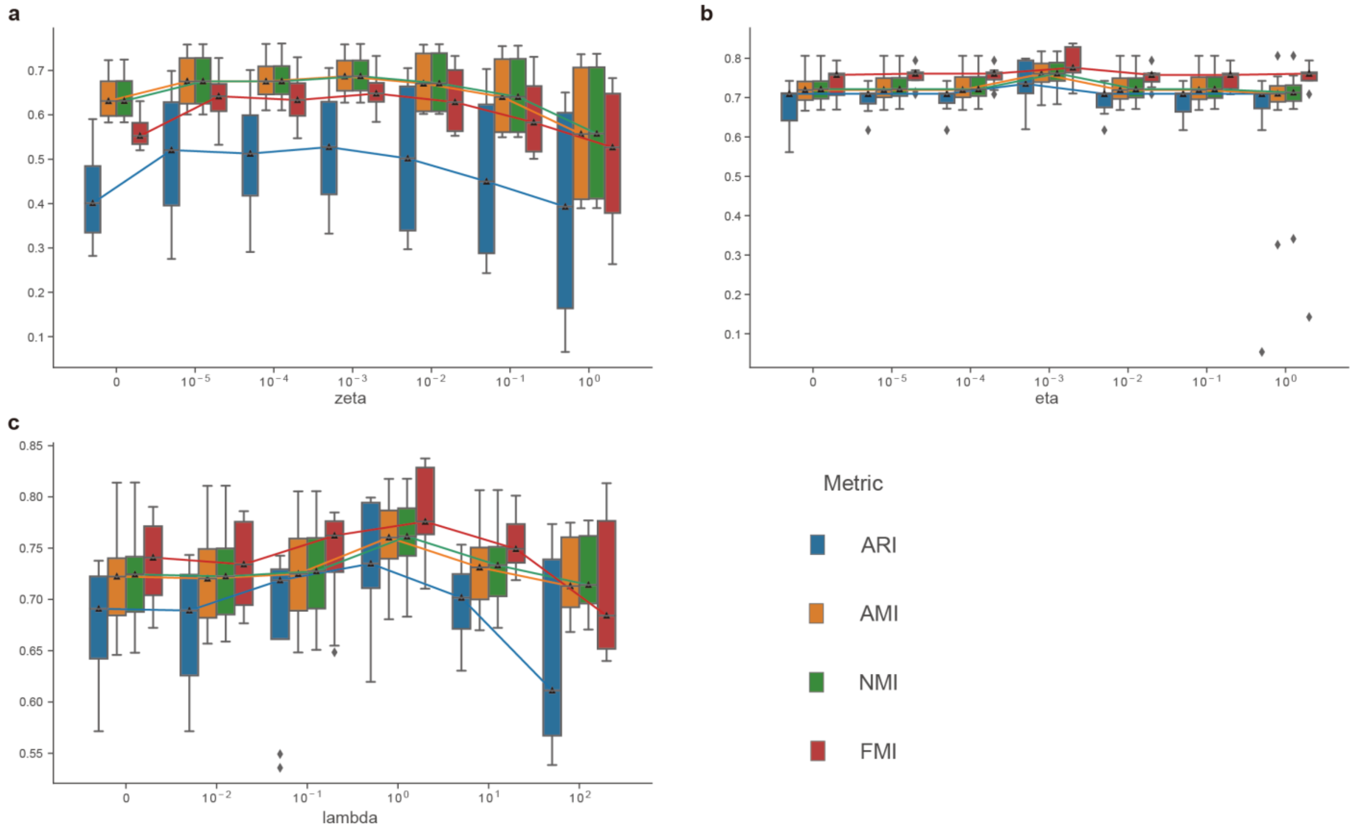
Supplementary Figure 19. Clustering performance with major cell types. a-c, AMI (**a**), NMI (**b**), FMI (**c**) of different methods under different thresholds for reserving major cell types on the Stimulated Droplet dataset.



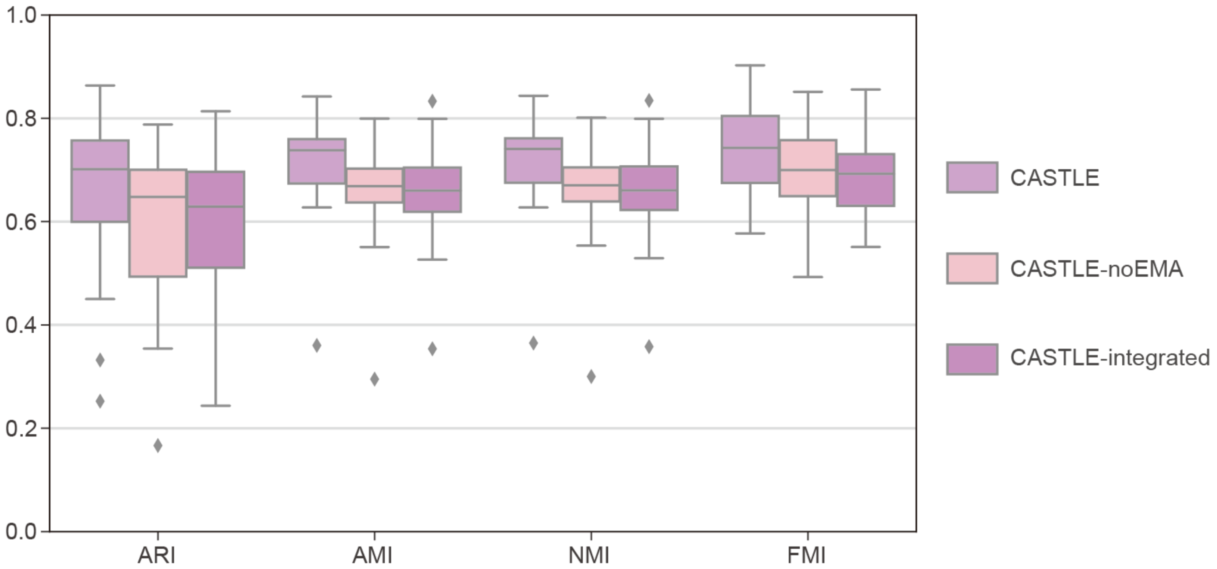
Supplementary Figure 20. Clustering performance under different resolutions. a-c, AMI (**a**), NMI (**b**), FMI (**c**) of different methods under different resolutions in Louvain clustering on the Stimulated Droplet dataset.



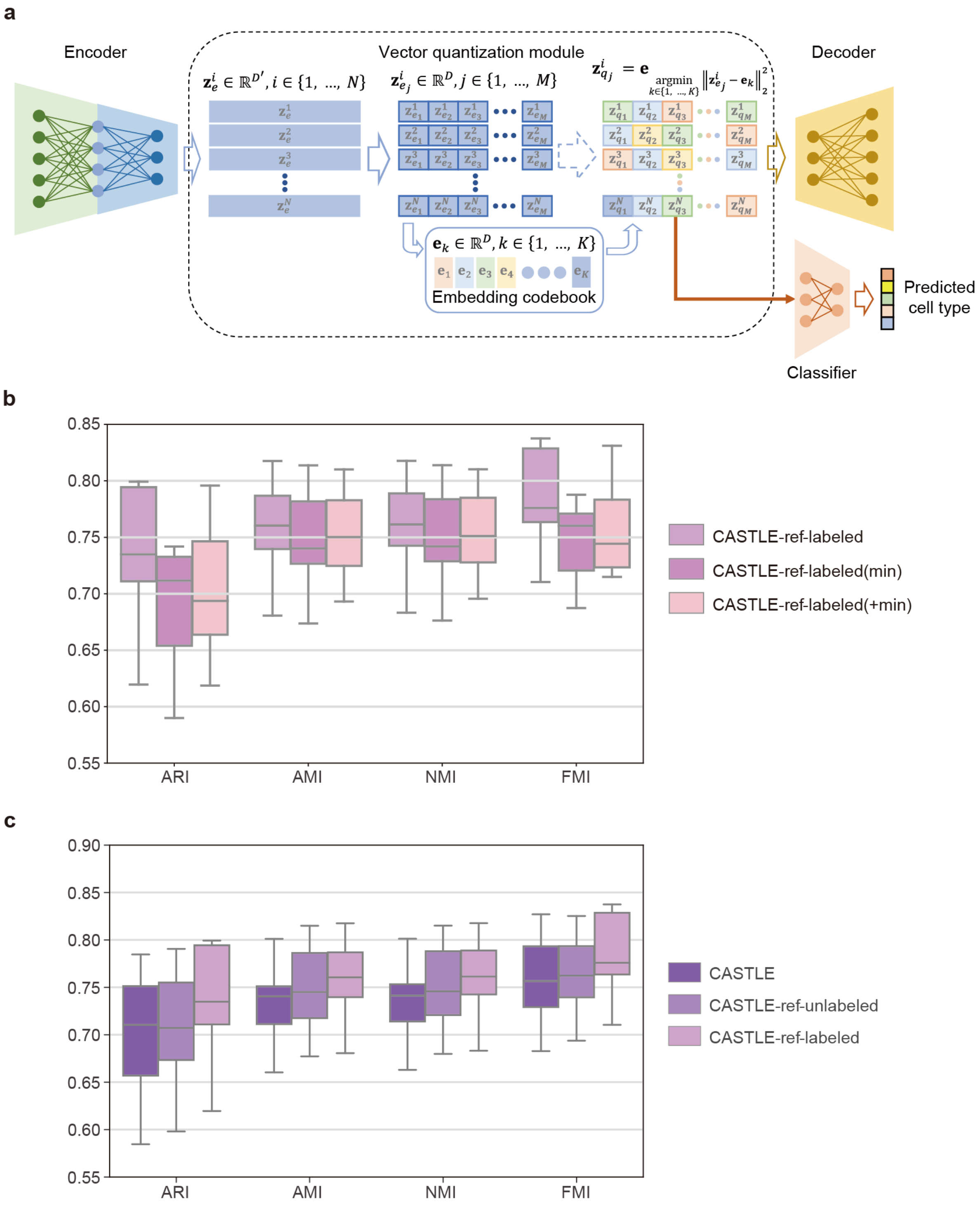
Supplementary Figure 21. Performance of CASTLE with different values of hyperparameters. a, Clustering performance of CASTLE under different values of alpha (weight of commitment loss) on 16 benchmark datasets. **b**, Clustering performance of CASTLE under different values of gamma (decay ratio) with a precision of 0.01 on 16 benchmark datasets. **c**, Clustering performance of CASTLE under different values of gamma with a precision of 0.001 on 16 benchmark datasets. **d**, Clustering performance of CASTLE under different values of *K* (size of codebook) on 16 benchmark datasets. **e**, Clustering performance of CASTLE under different values of *M* (time of split quantization) on 16 benchmark datasets. Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range and points represent outliers.



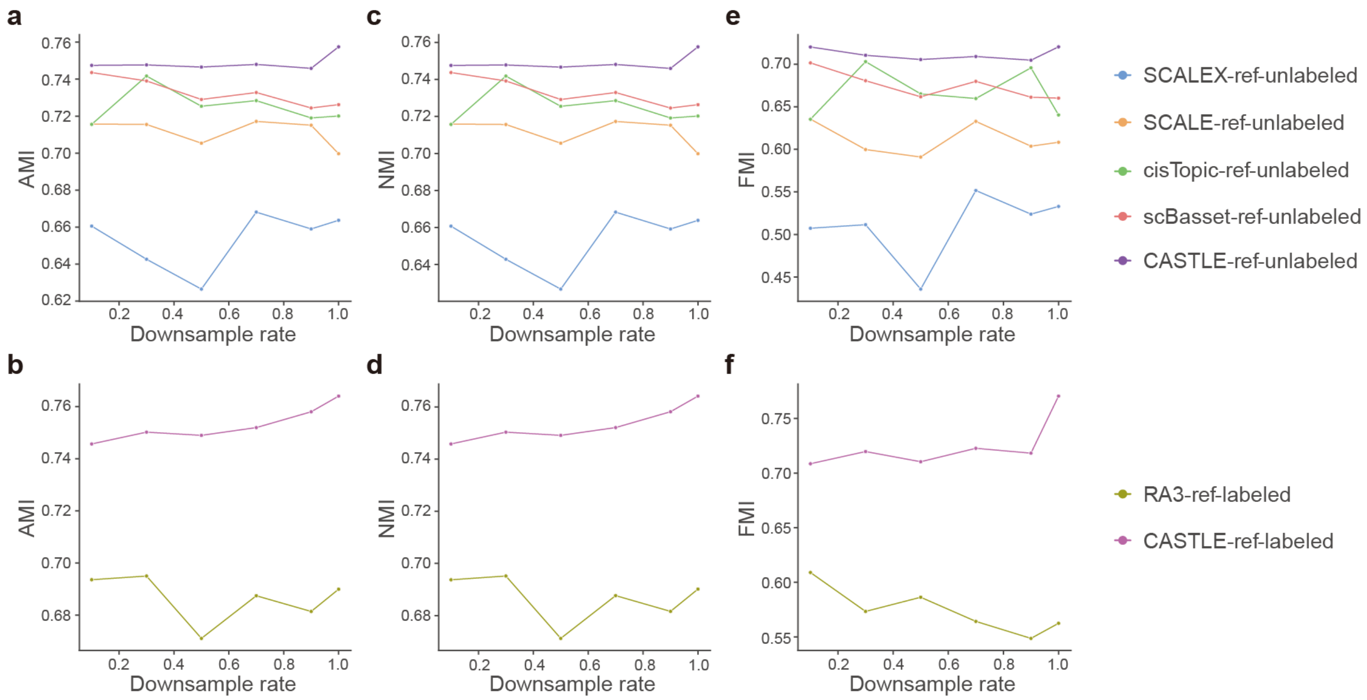
Supplementary Figure 22. Performance of CASTLE with different weights of self-designed loss functions. a, Clustering performance of CASTLE under different values of zeta (weight of batch loss) on 4 benchmark datasets with multiple batches. **b**, Clustering performance of CASTLE under different values of eta (weight of cell type loss) on 8 benchmark datasets when incorporating labeled reference datasets. **c**, Clustering performance of CASTLE under different values of lambda (weight of classifier loss) on 8 benchmark datasets when incorporating labeled reference datasets. Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range and points represent outliers.



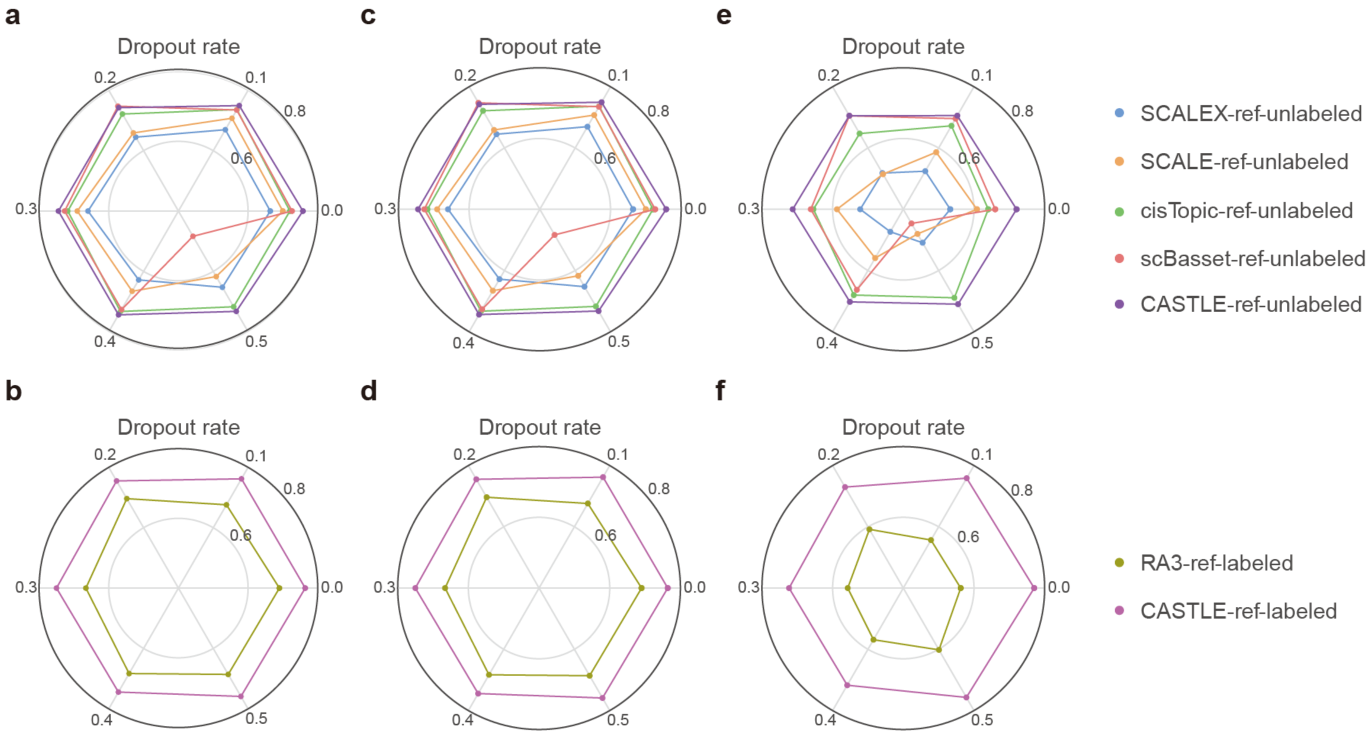
Supplementary Figure 23. Clustering performance with different designs of loss functions. Clustering performance evaluated by ARI, AMI, NMI and FMI on 16 benchmark datasets for CASTLE, CASTLE-noEMA and CASTLE-integrated. We referred to the model that updates codebook using $L_{codebook}$ instead of EMA as CASTLE-noEMA, and the model that updates encoder and codebook using the integrated $L_{commitment}$ and $L_{codebook}$ instead of using them separately as CASTLE-integrated. Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range and points represent outliers.



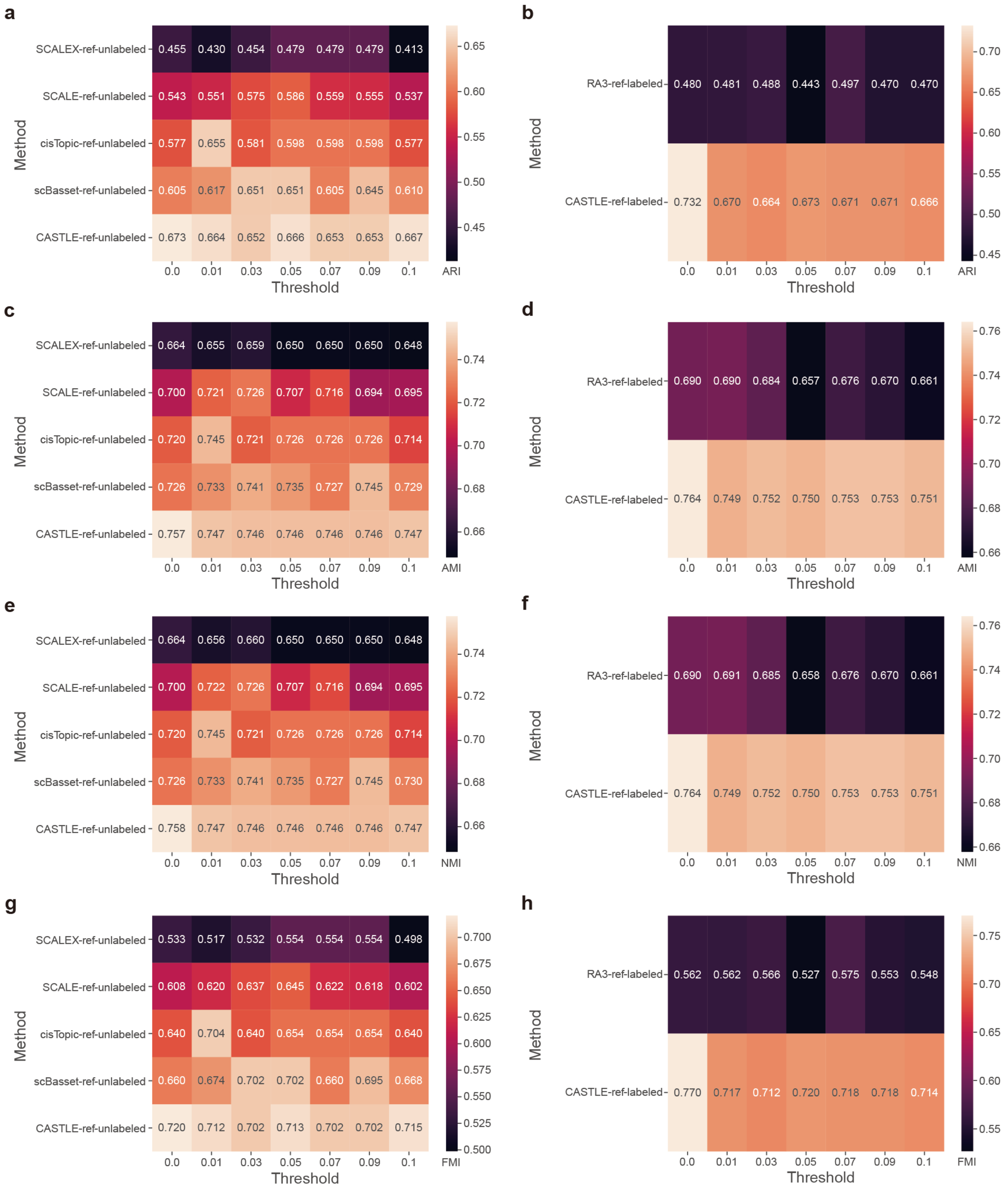
Supplementary Figure 24. Model architecture and performance of CASTLE with reference data. a, Model architecture of CASTLE when incorporating reference dataset with known cell type label. **b**, Clustering performance evaluated by ARI, AMI, NMI and FMI on 8 benchmark datasets when incorporating labeled reference datasets with $L_{celltype}$, redesigned $L_{celltype}$ for only minimizing the distance between the cells from the same cell types, or redesigned $L_{celltype}$ for simultaneously maximizing the distance between the cells from different cell types and minimizing the distance between the cells from the same cell types. **c**, Clustering performance evaluated by ARI, AMI, NMI and FMI on 8 benchmark datasets when incorporating unlabeled or labeled reference datasets. Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range and points represent outliers.



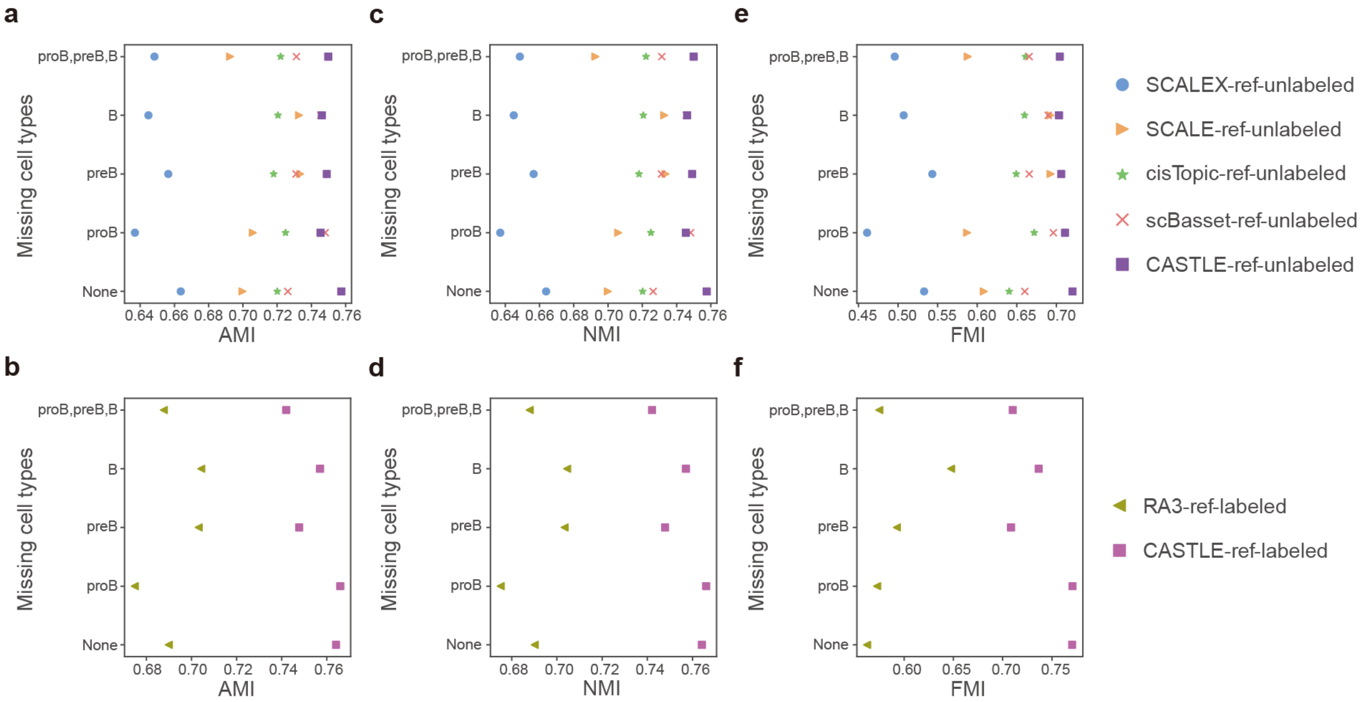
Supplementary Figure 25. Clustering performance under different downsample rates. a-f, AMI (**a**, **b**), NMI (**c**, **d**), FMI (**e**, **f**) of different methods on the Stimulated Droplet dataset when incorporating unlabeled (**a**, **c**, **e**) or labeled (**b**, **d**, **f**) reference dataset (Resting Droplet dataset) under different downsample rates.



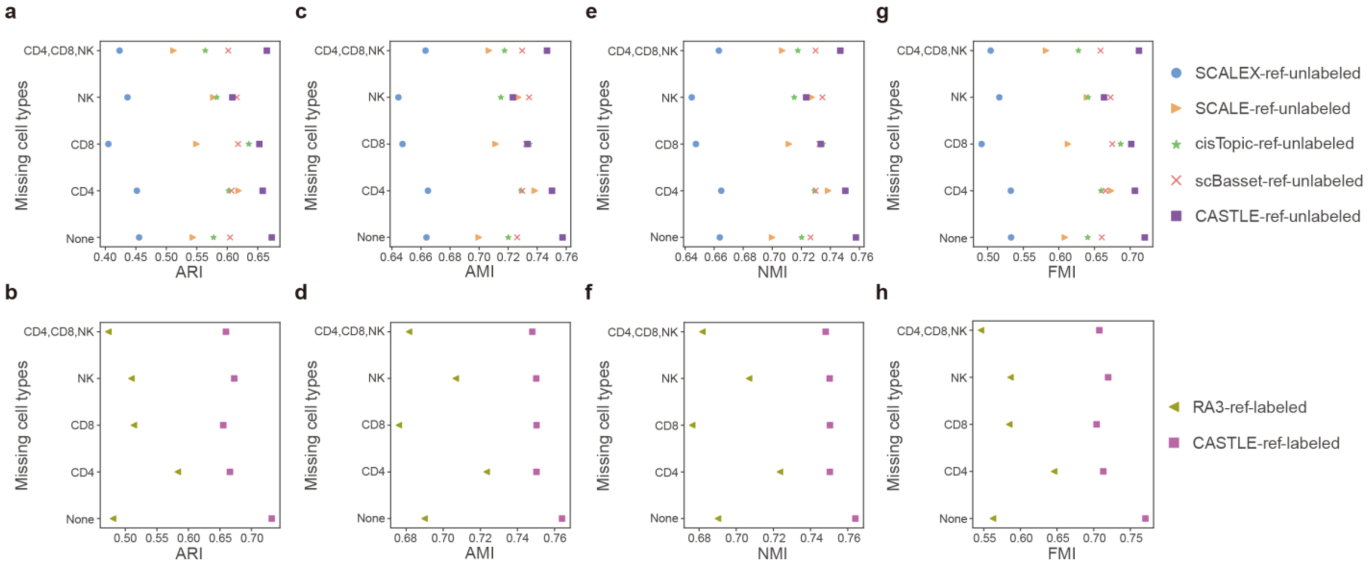
Supplementary Figure 26. Clustering performance under different dropout rates. a-f, AMI (**a**, **b**), NMI (**c**, **d**), FMI (**e**, **f**) of different methods on the Stimulated Droplet dataset when incorporating unlabeled (**a**, **c**, **e**) or labeled (**b**, **d**, **f**) reference dataset (Resting Droplet dataset) under different dropout rates.



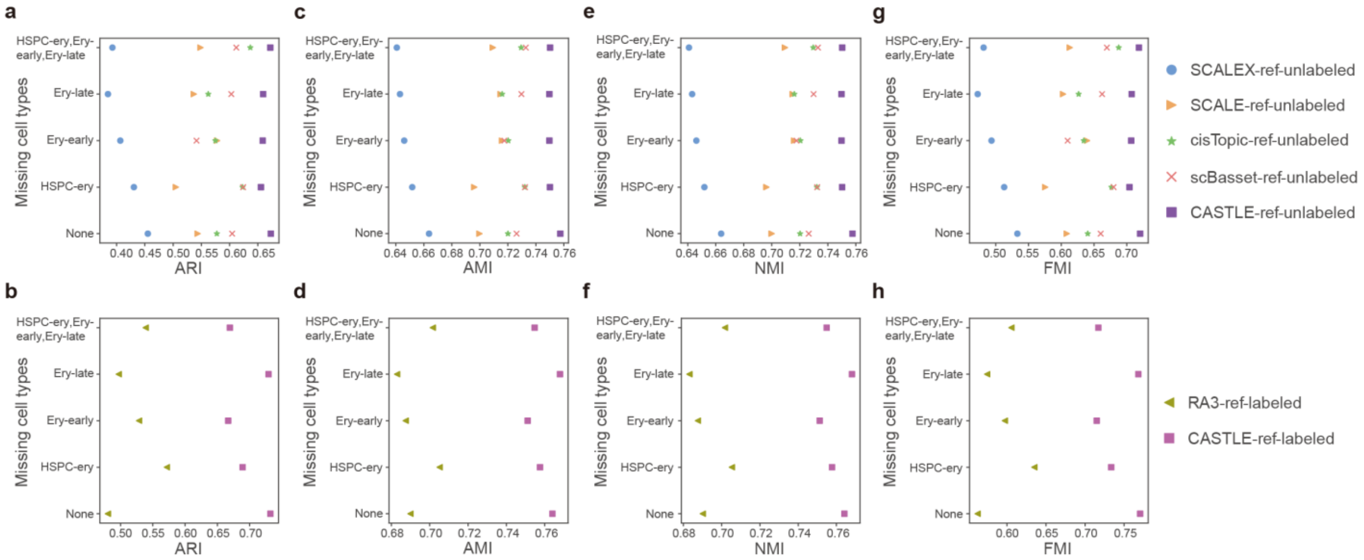
Supplementary Figure 27. Clustering performance with major cell types. a-h, ARI (**a**, **b**), AMI (**c**, **d**), NMI (**e**, **f**), FMI (**g**, **h**) of different methods on the Stimulated Droplet dataset when incorporating unlabeled (**a**, **c**, **e**, **g**) or labeled (**b**, **d**, **f**, **h**) reference dataset (Resting Droplet dataset) with different thresholds for reserving the major cell types.



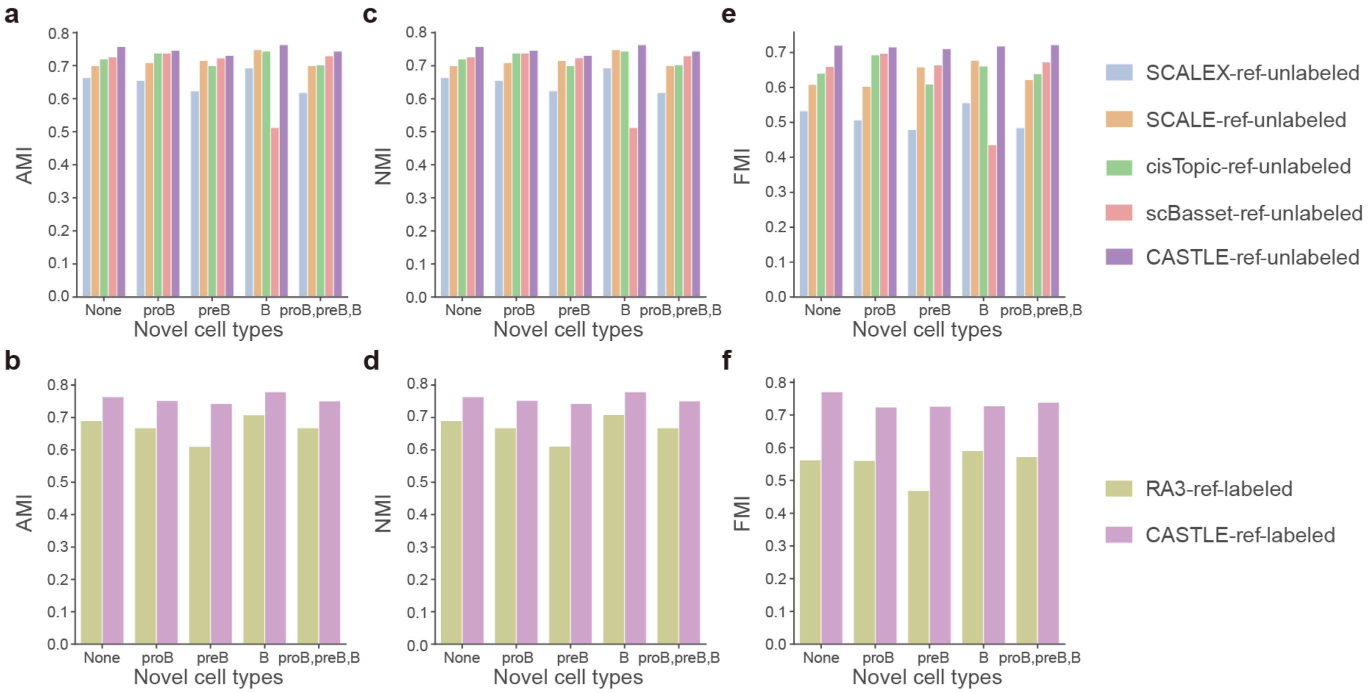
Supplementary Figure 28. Clustering performance with missing cell types. a-f, AMI (**a**, **b**), NMI (**c**, **d**), FMI (**e**, **f**) of different methods on the Stimulated Droplet dataset when incorporating unlabeled (**a**, **c**, **e**) or labeled (**b**, **d**, **f**) reference dataset (Resting Droplet dataset) with none, one or all missing cell types of proB, preB and B cells.



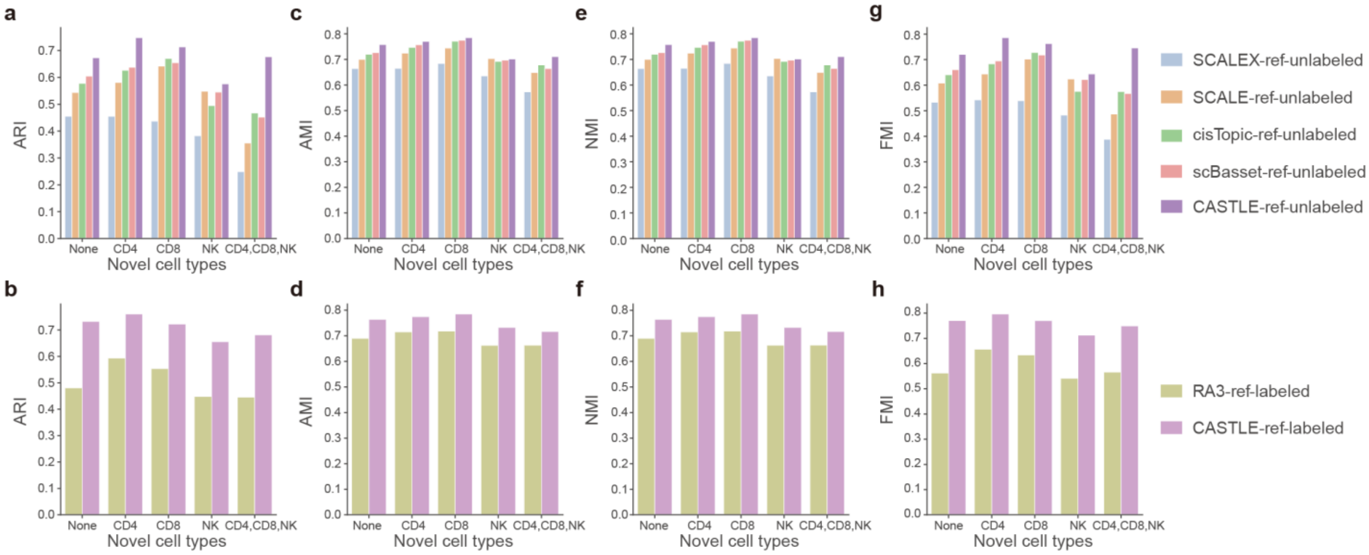
Supplementary Figure 29. Clustering performance with missing cell types. a-h, ARI (**a**, **b**), AMI (**c**, **d**), NMI (**e**, **f**), FMI (**g**, **h**) of different methods on the Stimulated Droplet dataset when incorporating unlabeled (**a**, **c**, **e**, **g**) or labeled (**b**, **d**, **f**, **h**) reference dataset (Resting Droplet dataset) with none, one or all missing cell types of CD4, CD8 and NK cells.



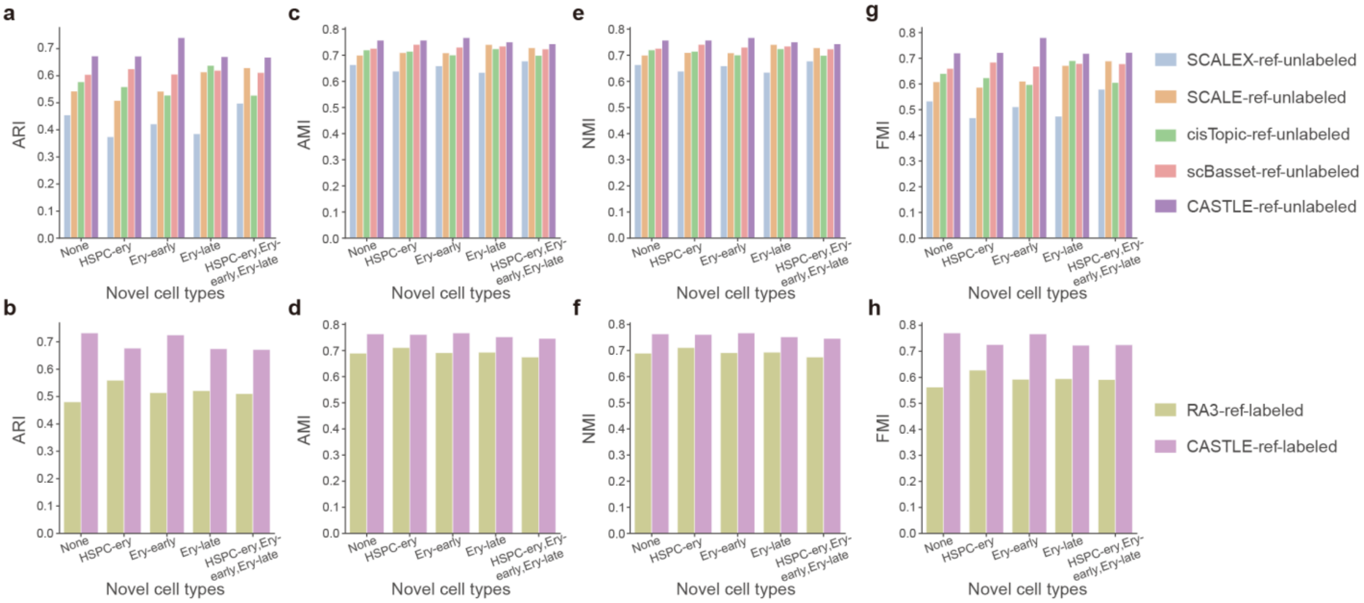
Supplementary Figure 30. Clustering performance with missing cell types. a-h, ARI (**a**, **b**), AMI (**c**, **d**), NMI (**e**, **f**), FMI (**g**, **h**) of different methods on the Stimulated Droplet dataset when incorporating unlabeled (**a**, **c**, **e**, **g**) or labeled (**b**, **d**, **f**, **h**) reference dataset (Resting Droplet dataset) with none, one or all missing cell types of HSPC-ery, Ery-early and Ery-late cells.



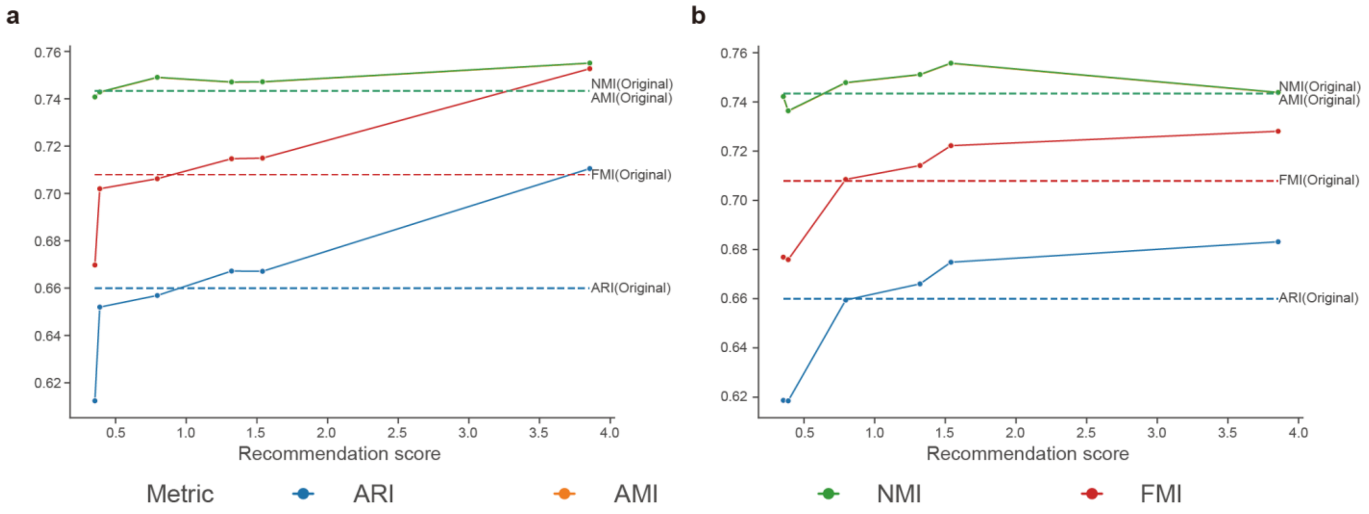
Supplementary Figure 31. Clustering performance with novel cell types. a-f, AMI (**a**, **b**), NMI (**c**, **d**), FMI (**e**, **f**) of different methods on the Stimulated Droplet dataset when incorporating unlabeled (**a**, **c**, **e**) or labeled (**b**, **d**, **f**) reference dataset (Resting Droplet dataset) with none, one or all novel cell types of proB, preB and B cells.



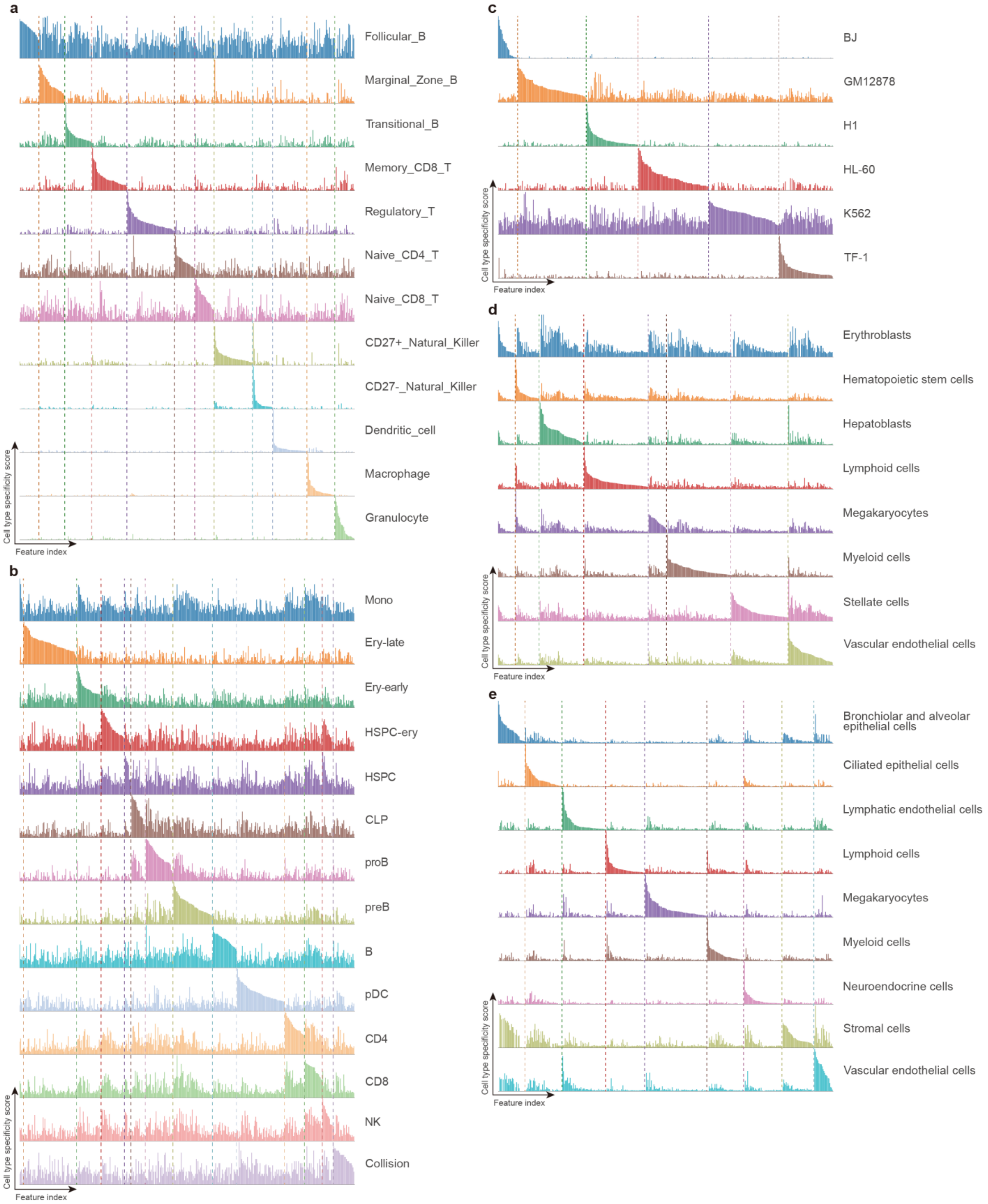
Supplementary Figure 32. Clustering performance with novel cell types. a-h, ARI (**a**, **b**), AMI (**c**, **d**), NMI (**e**, **f**), FMI (**g**, **h**) of different methods on the Stimulated Droplet dataset when incorporating unlabeled (**a**, **c**, **e**, **g**) or labeled (**b**, **d**, **f**, **h**) reference dataset (Resting Droplet dataset) with none, one or all novel cell types of CD4, CD8 and NK cells.



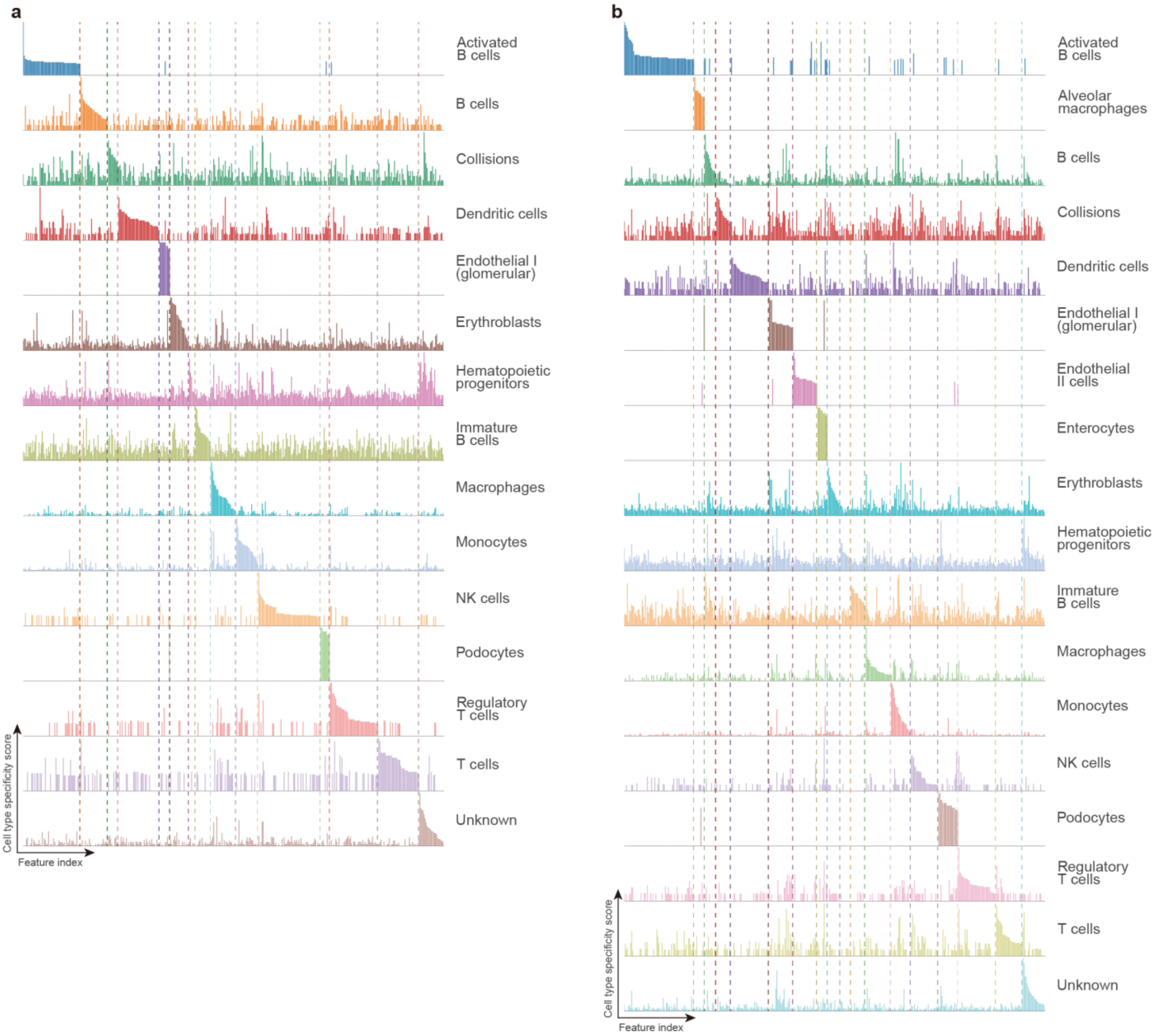
Supplementary Figure 33. Clustering performance with novel cell types. a-h, ARI (**a**, **b**), AMI (**c**, **d**), NMI (**e**, **f**), FMI (**g**, **h**) of different methods on the Stimulated Droplet dataset when incorporating unlabeled (**a**, **c**, **e**, **g**) or labeled (**b**, **d**, **f**, **h**) reference dataset (Resting Droplet dataset) with none, one or all novel cell types of HSPC-ery, Ery-early and Ery-late cells.



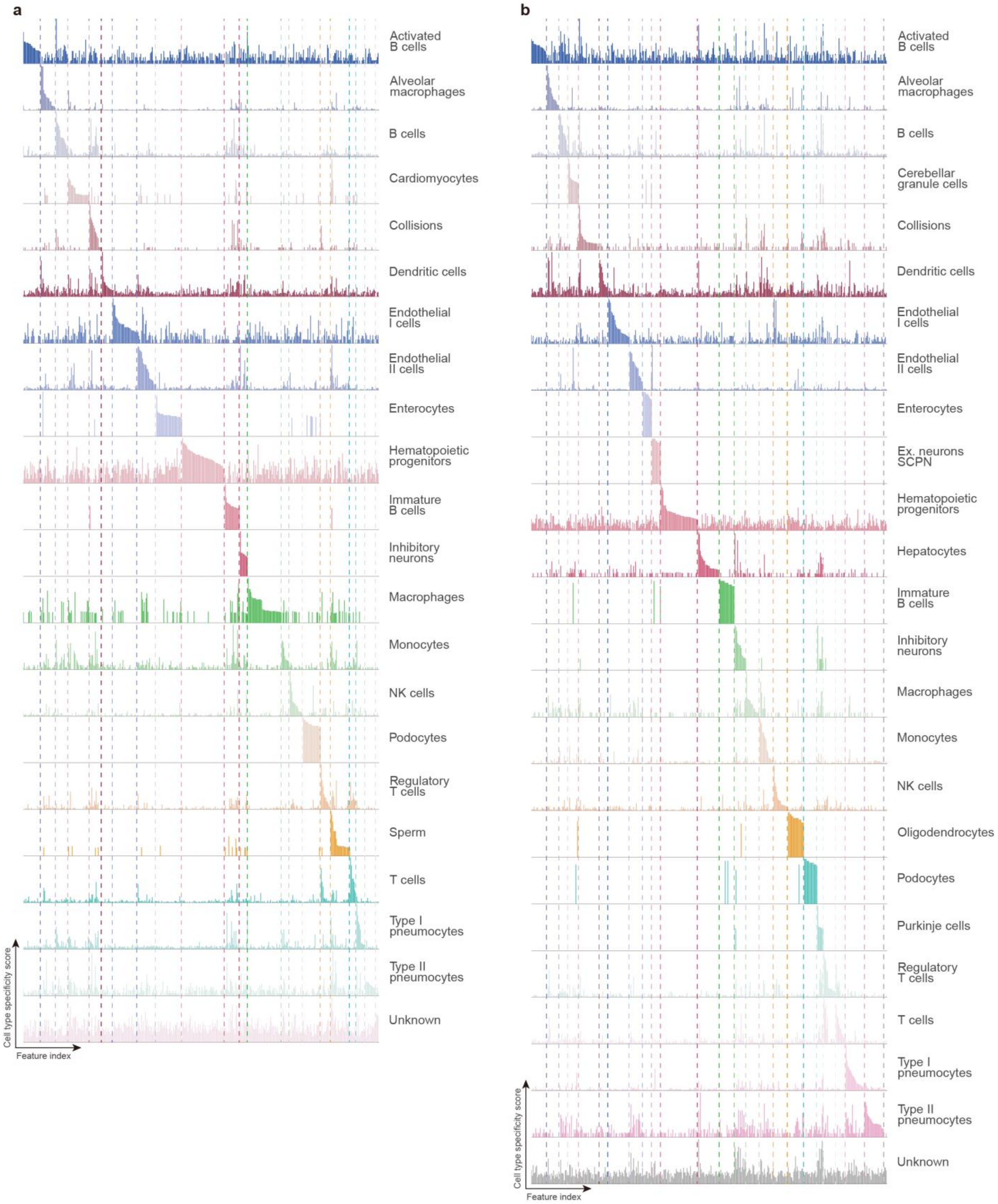
Supplementary Figure 34. Recommendation scores of reference datasets. a-b, Correlation between clustering performance and Recommendation scores for different unlabeled (**a**) or labeled (**b**) reference datasets on the Stimulated Droplet dataset.



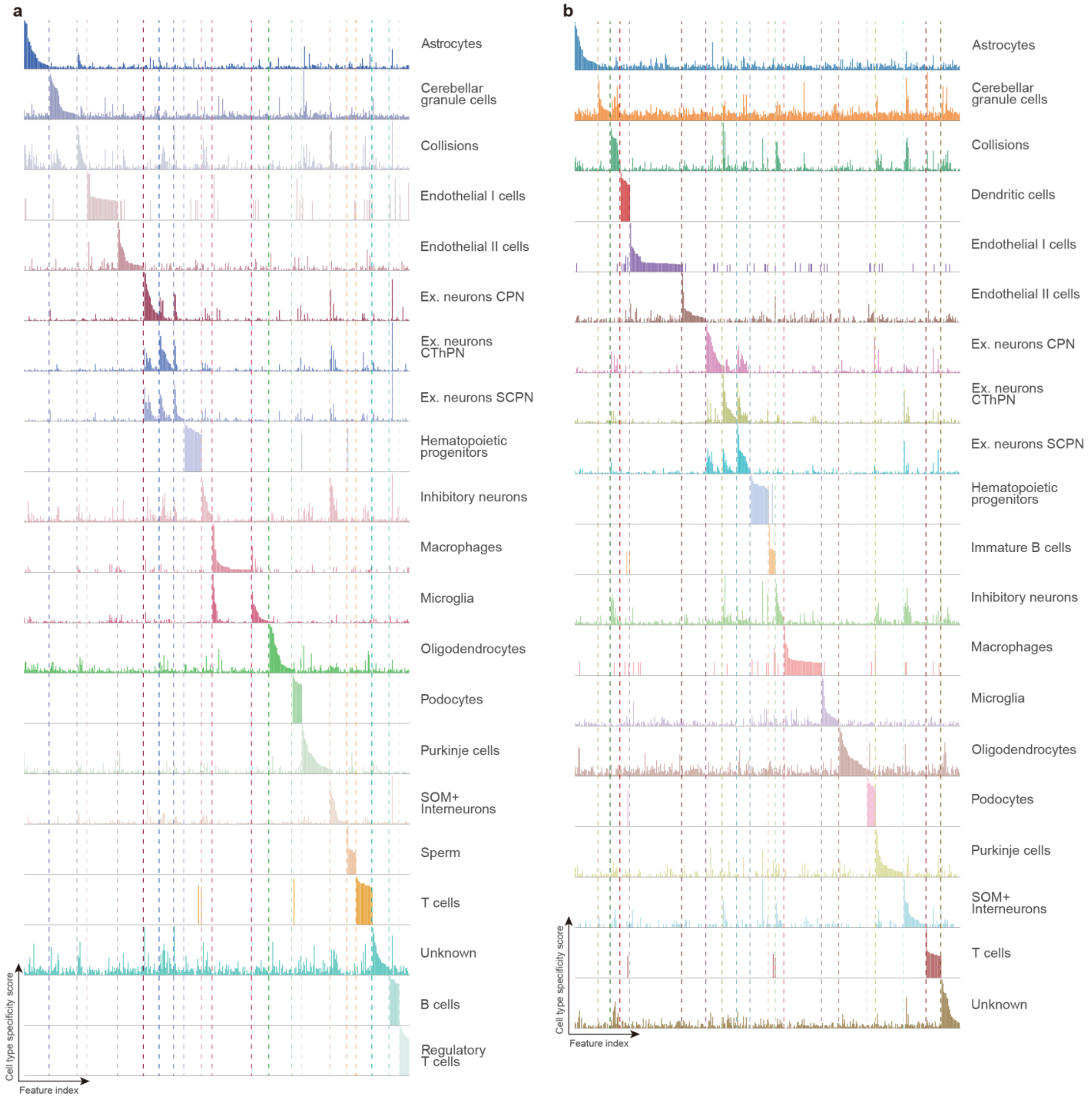
Supplementary Figure 35. Feature spectrum. a-e, Feature spectrum for each cell type of the Splenocyte (**a**), Resting Droplet (**b**), InSilico (**c**), Fetal Liver (**d**) and Fetal Lung (**e**) datasets.



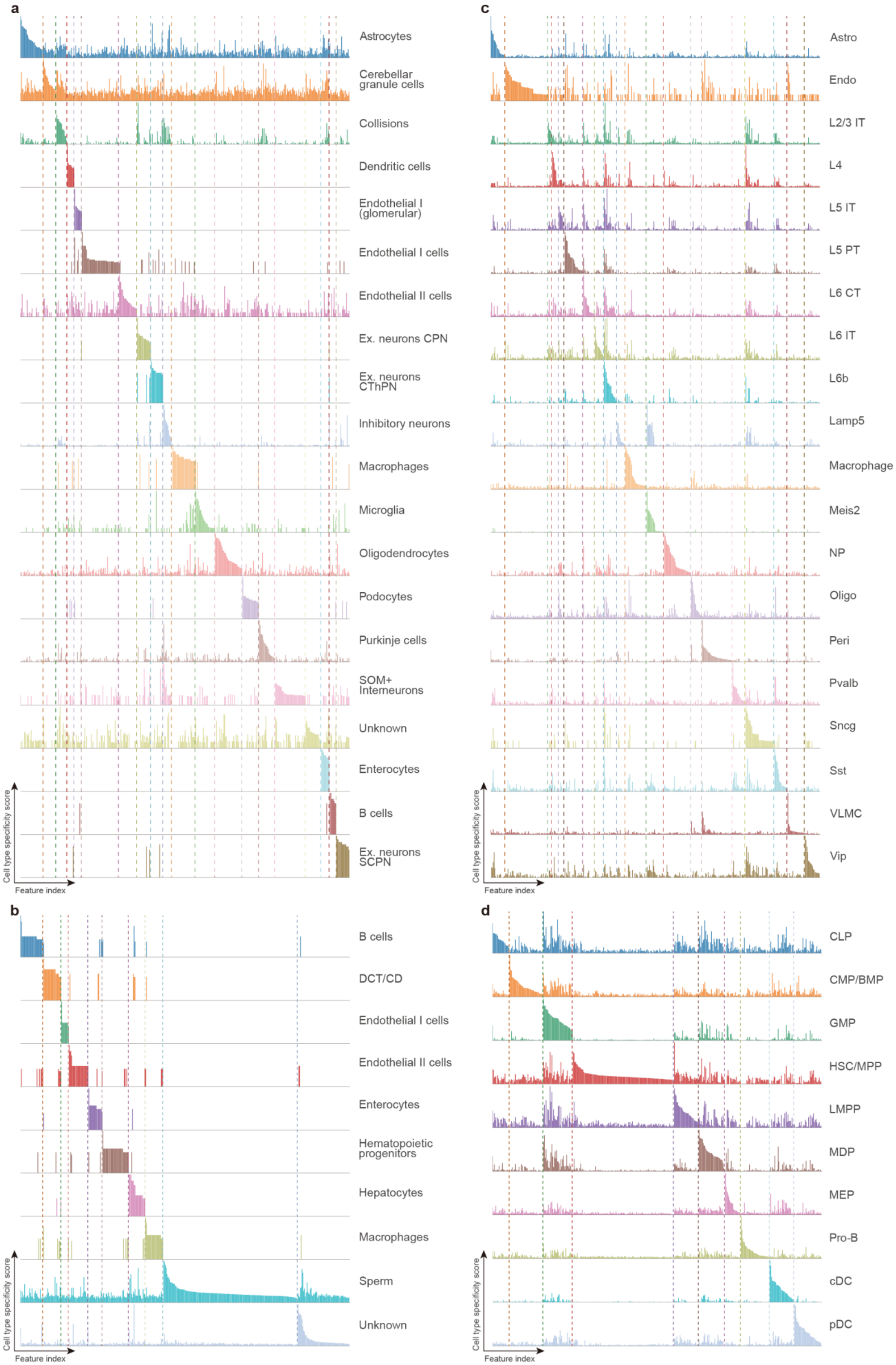
Supplementary Figure 36. Feature spectrum. a-b, Feature spectrum for each cell type of the Bone Marrow A (**a**) and Bone Marrow B (**b**) datasets.



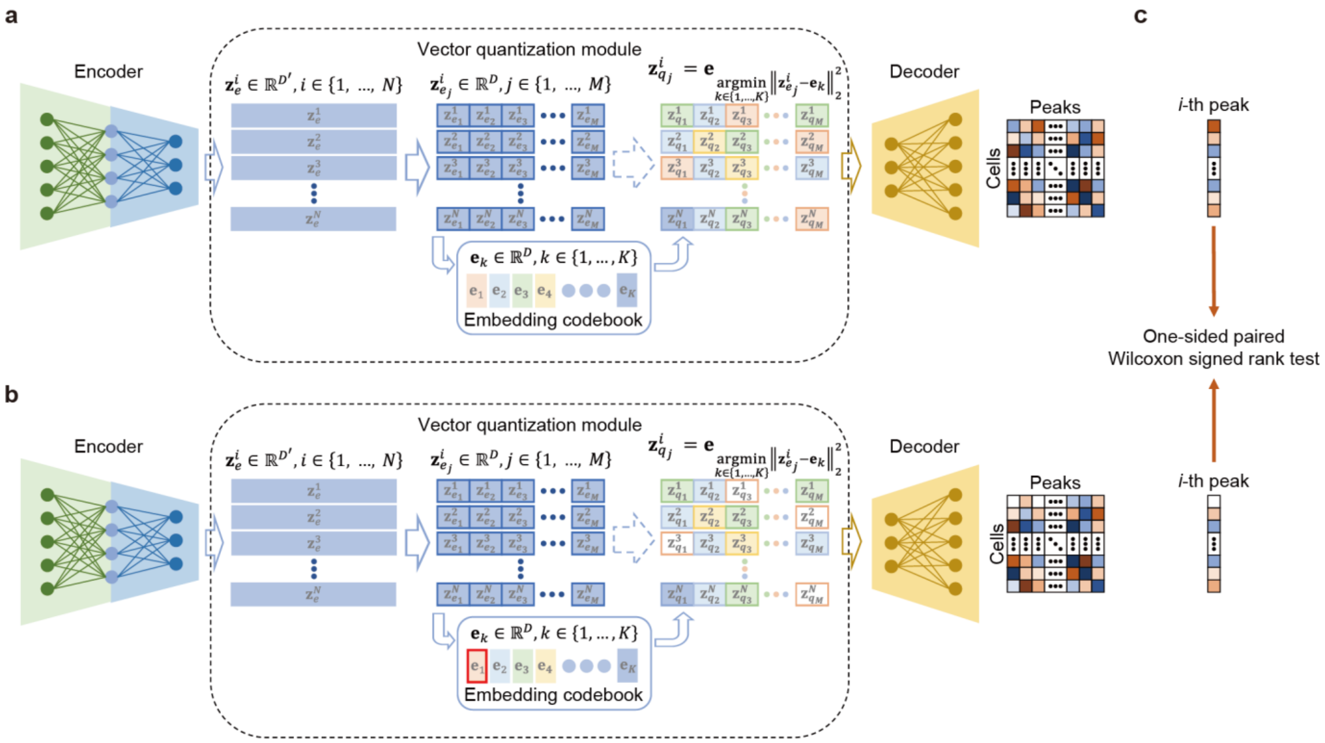
Supplementary Figure 37. Feature spectrum. a-b, Feature spectrum for each cell type of the Lung A (**a**) and Lung B (**b**) datasets.



Supplementary Figure 38. Feature spectrum. a-b, Feature spectrum for each cell type of the Whole Brain A (**a**) and Whole Brain B (**b**) datasets.

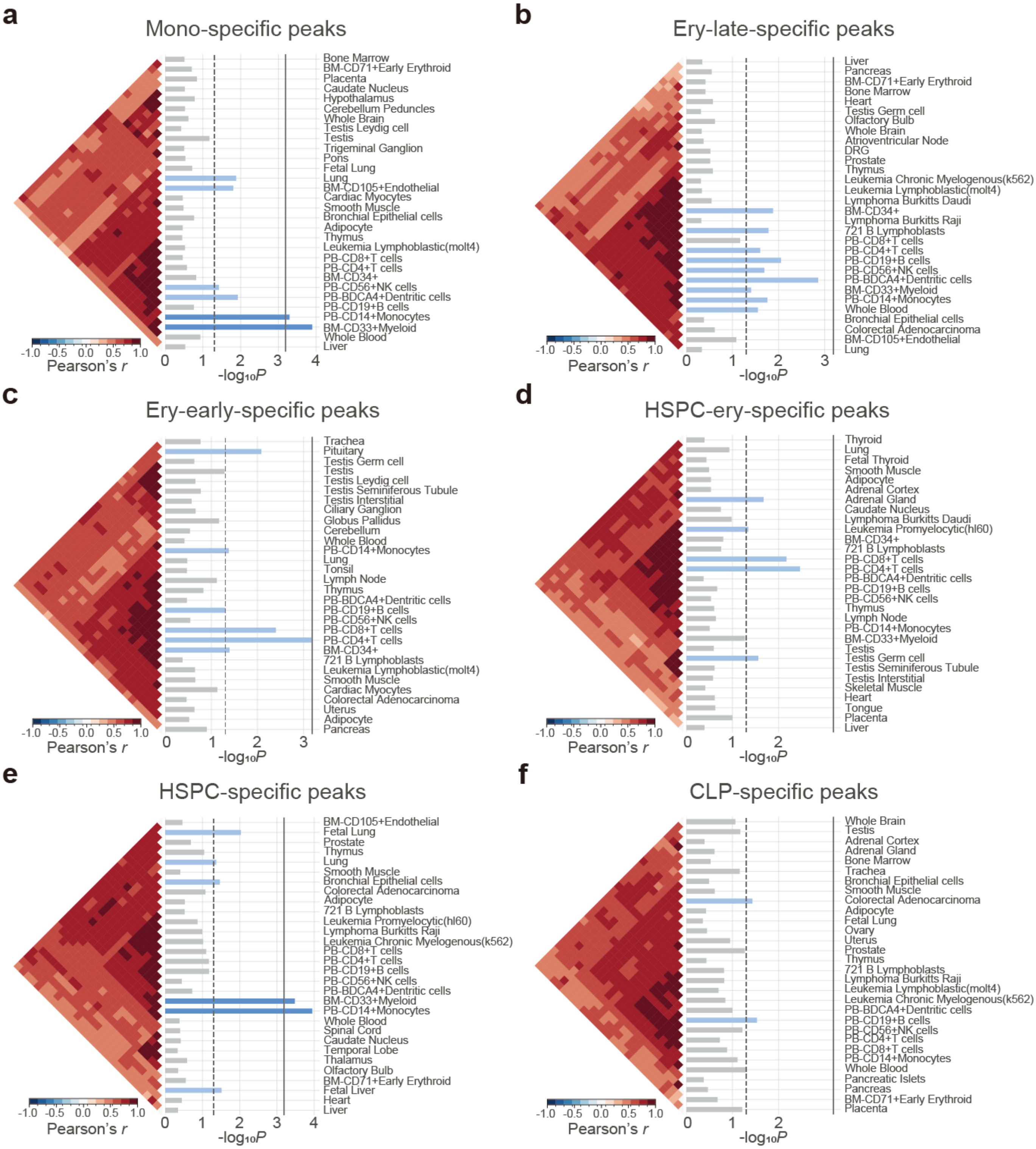


Supplementary Figure 39. Feature spectrum. a-d, Feature spectrum for each cell type of the Cerebellum (**a**), Testes (**b**), Brain (**c**) and Immune (**d**) datasets.

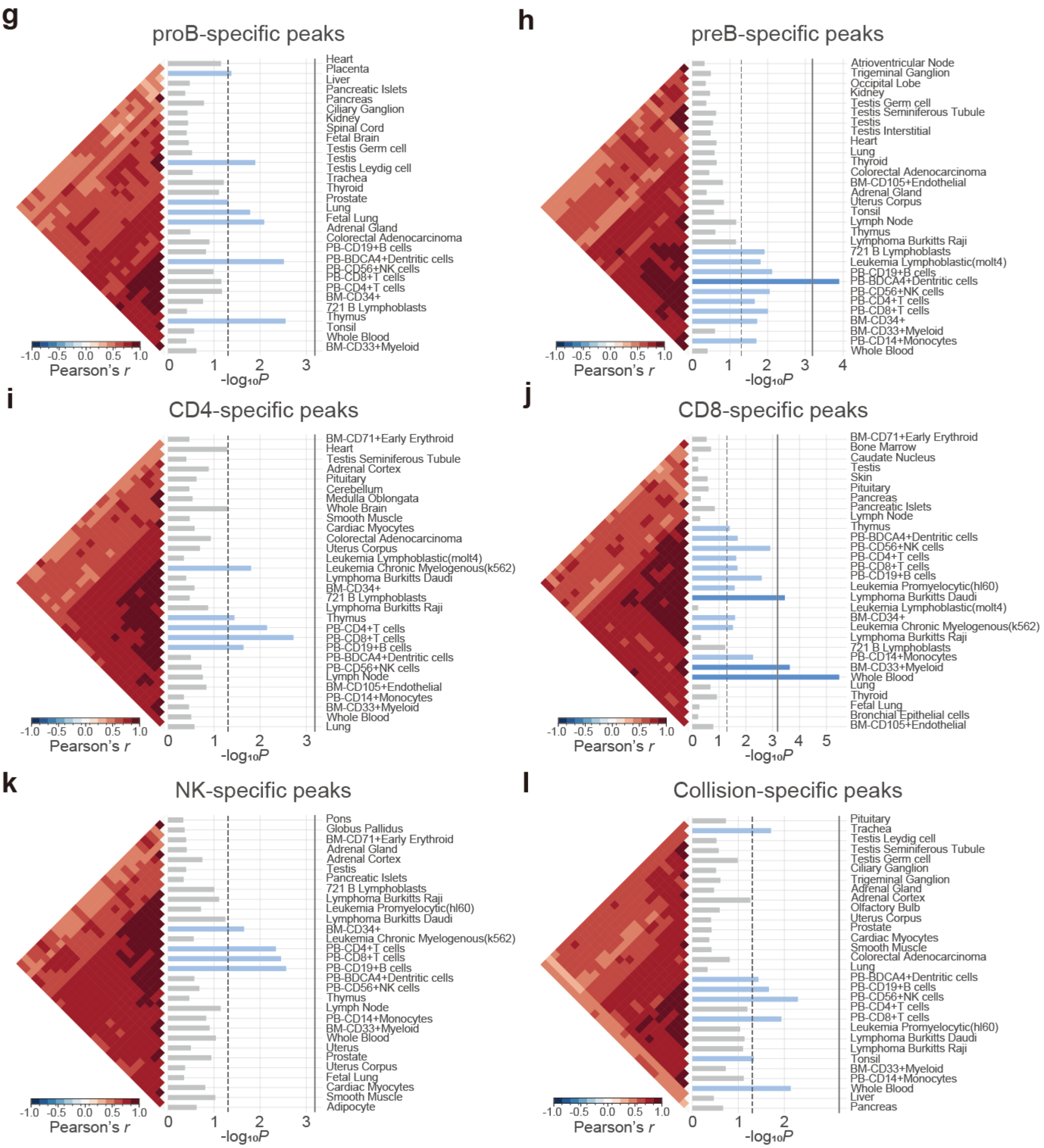


a
Encoder
Vector quantization module
$\mathbf{z}_e^i \in \mathbb{R}^{D'}, i \in \{1, \ldots, N\}$
$\mathbf{z}_{e_j}^i \in \mathbb{R}^{D}, j \in \{1, \ldots, M\}$
$\mathbf{z}_{q_j}^i = \mathbf{e}_{\operatorname{argmin}_{k\in\{1,\ldots,K\}} \|\mathbf{z}_{e_j}^i - \mathbf{e}_k\|_2^2}$
$\mathbf{e}_k \in \mathbb{R}^D, k \in \{1, \ldots, K\}$
Embedding codebook
Decoder
Peaks
Cells
b
Encoder
Vector quantization module
Decoder
Embedding codebook
Peaks
Cells
c
i-th peak
One-sided paired
Wilcoxon signed rank test
i-th peak

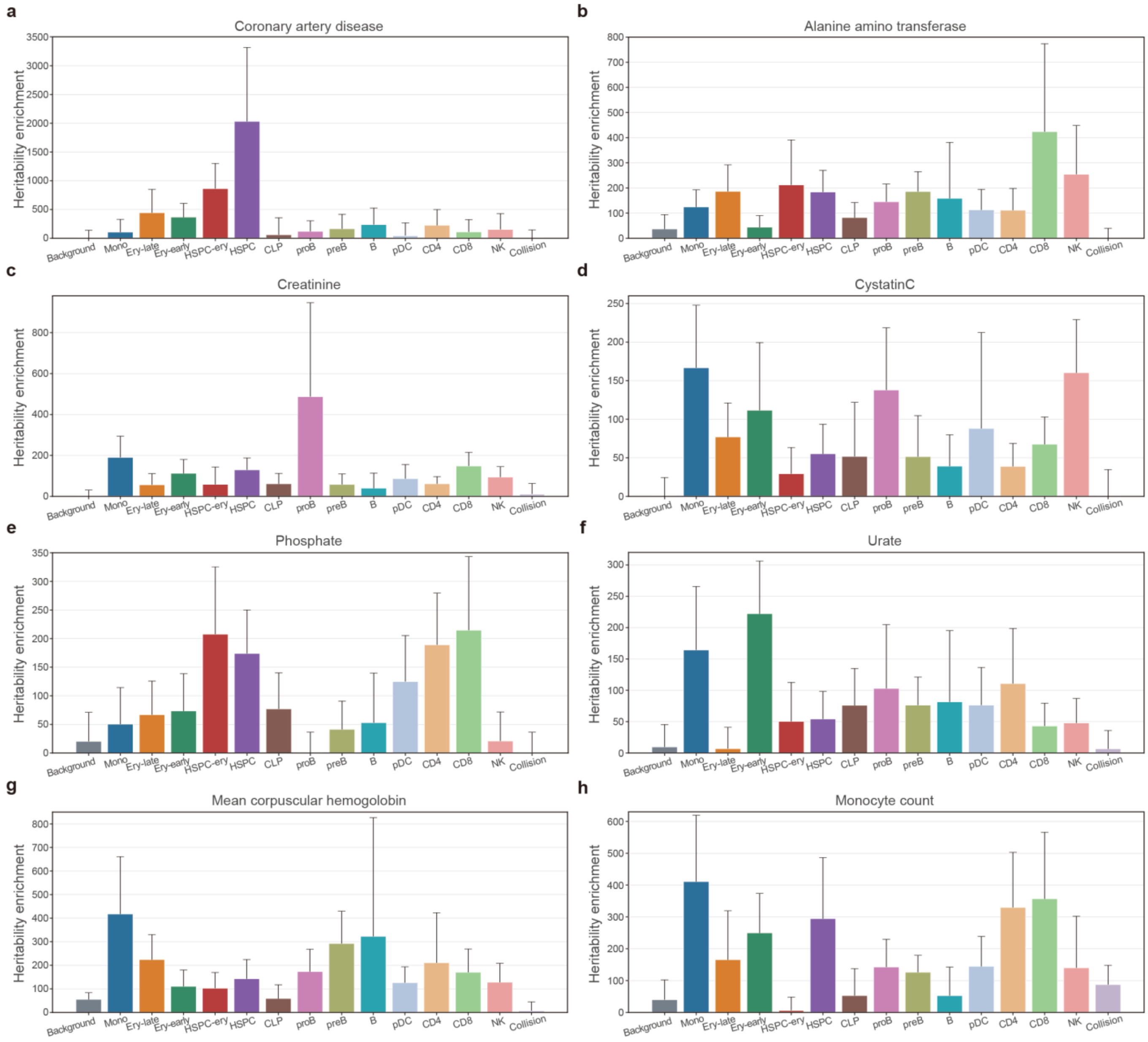
Supplementary Figure 40. Process of identifying the cell type-specific peaks linked with cell type-specific features. **a**, The standard process of CASTLE. **b**, Process of CASTLE when information derived from a certain cell type-specific feature in codebook is removed by zeroing it out before passing the cell embeddings again through the decoder. **c**, By performing the one-sided paired Wilcoxon signed-rank test between the reconstructed cell-by-region matrices with or without zeroing, we identified the cell type-specific peaks that are significantly impacted, and thus causally linked, to that feature.



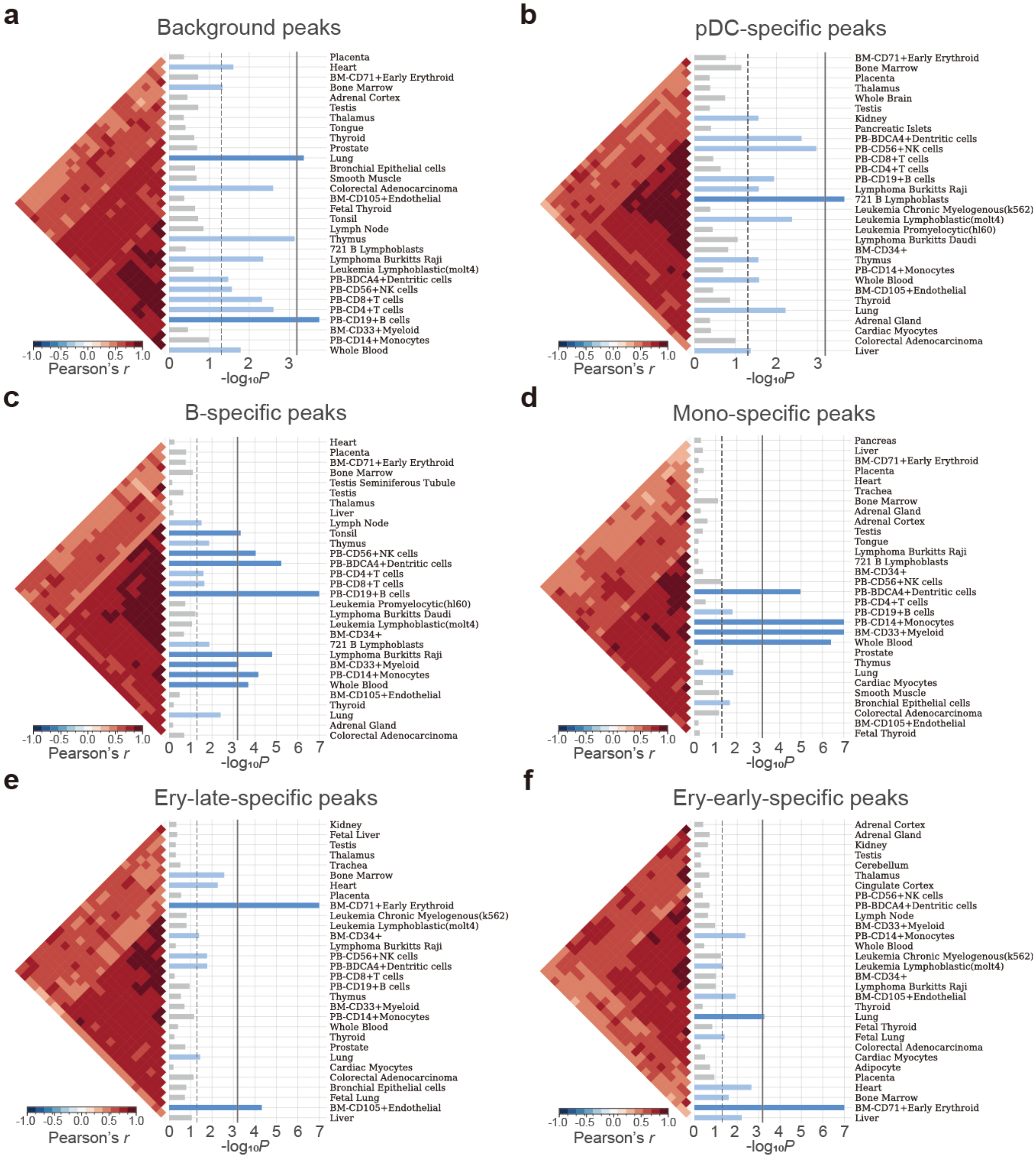
Supplementary Figure 41. SNP enrichment analysis for cell type-specific peaks identified by CASTLE on the Stimulated Droplet dataset. a-l, Top 30 significantly enriched tissues in SNPsea analysis on Mono-specific peaks (**a**), Ery-late-specific peaks (**b**), Ery-early-specific peaks (**c**), HSPC-ery-specific peaks (**d**), HSPC-specific peaks (**e**), CLP-specific peaks (**f**), proB-specific peaks (**g**), preB-specific peaks (**h**), CD4-specific peaks (**i**), CD8-specific peaks (**j**), NK-specific peaks (**k**) and Collision-specific peaks (**l**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



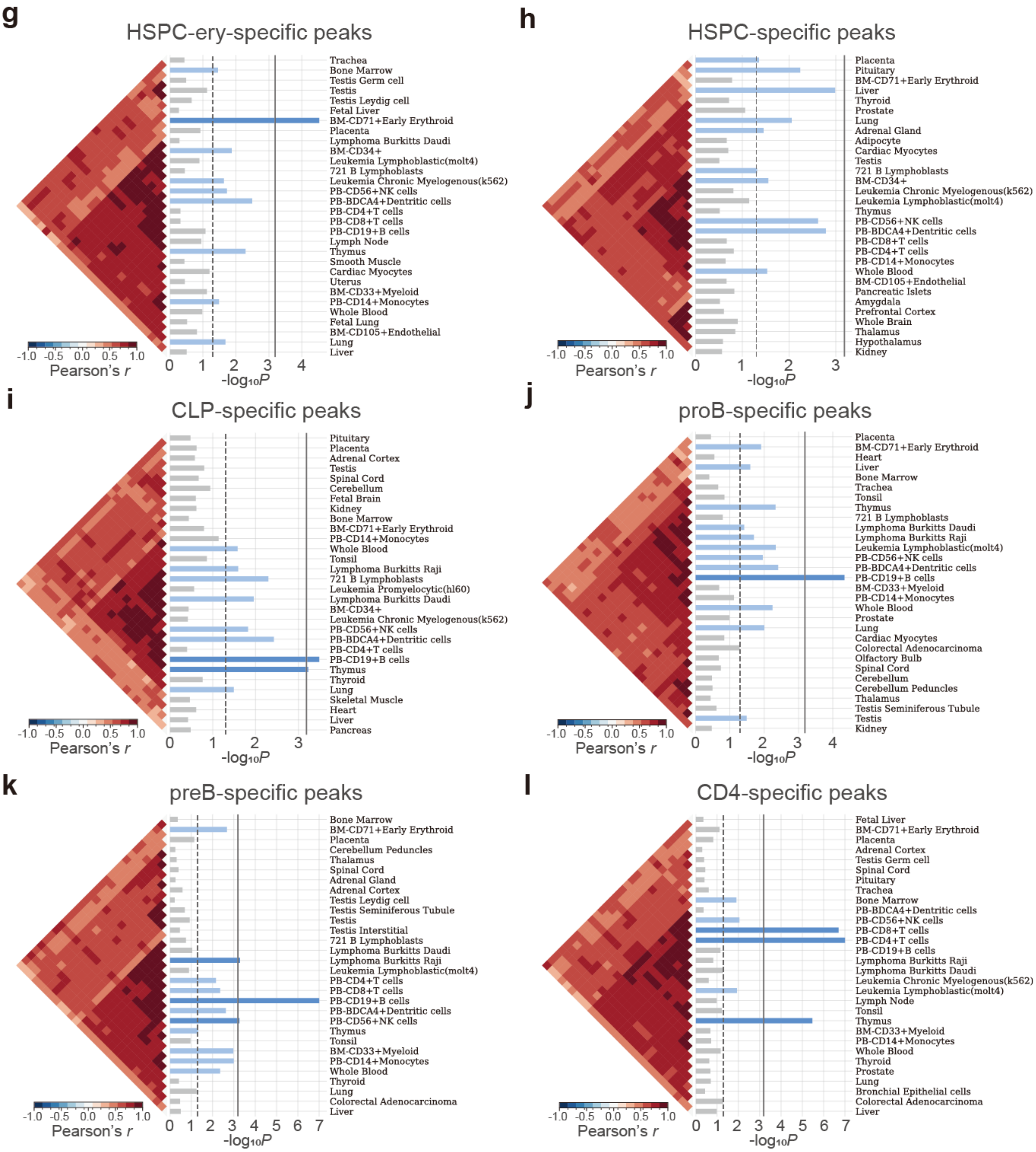
Supplementary Figure 41 (continue). SNP enrichment analysis for cell type-specific peaks identified by CASTLE on the Stimulated Droplet dataset. a-l, Top 30 significantly enriched tissues in SNPsea analysis on Mono-specific peaks (**a**), Ery-late-specific peaks (**b**), Ery-early-specific peaks (**c**), HSPC-ery-specific peaks (**d**), HSPC-specific peaks (**e**), CLP-specific peaks (**f**), proB-specific peaks (**g**), preB-specific peaks (**h**), CD4-specific peaks (**i**), CD8-specific peaks (**j**), NK-specific peaks (**k**) and Collision-specific peaks (**l**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



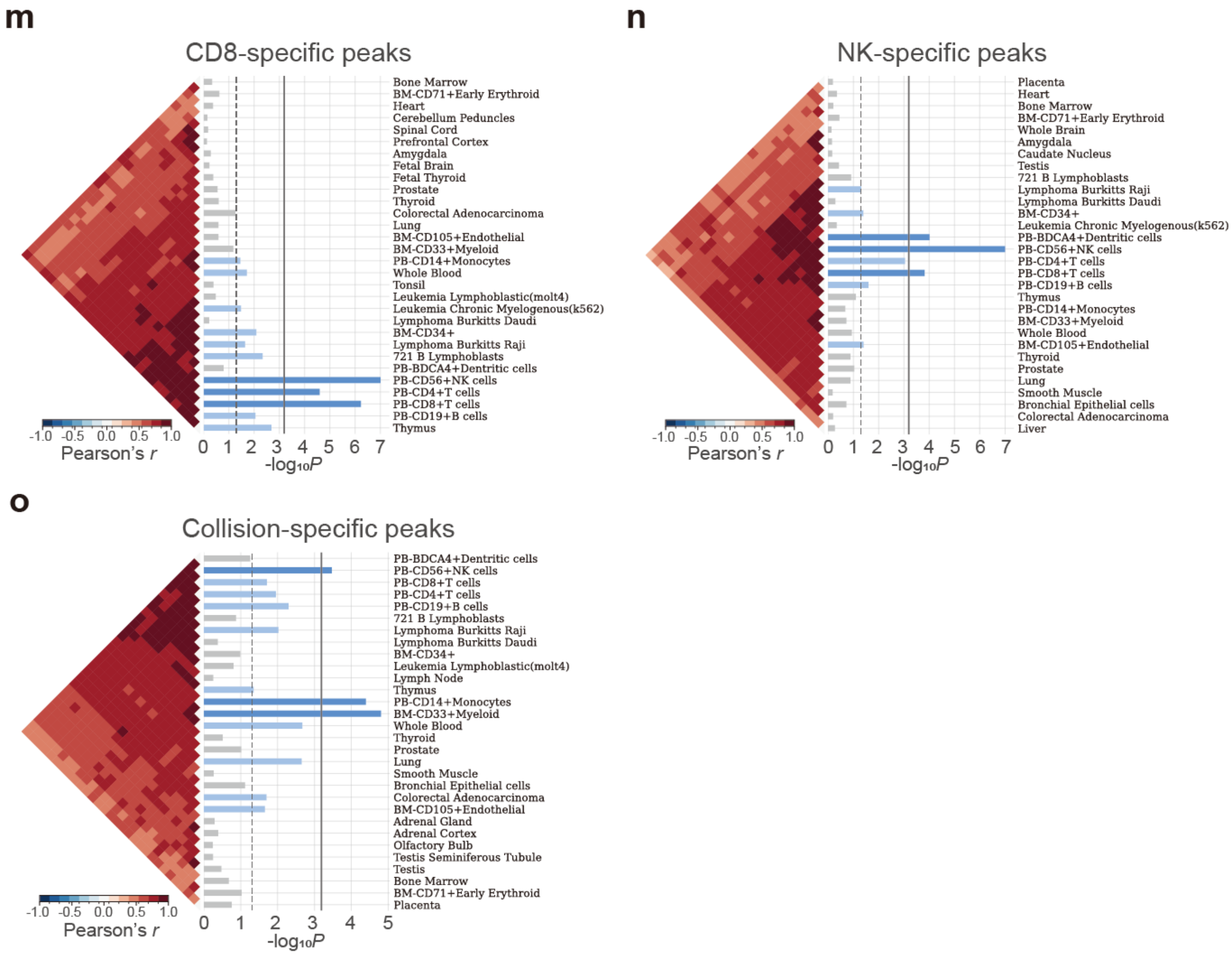
Supplementary Figure 42. Heritability enrichment analysis for cell type-specific peaks identified by CASTLE on the Stimulated Droplet dataset. a-h, Heritability enrichments estimated by LDSC within cell type-specific peaks identified by CASTLE and the background peaks for blood-related traits including coronary artery disease (**a**), alanine amino transferase (**b**), creatinine (**c**), cystatinC (**d**), phosphate (**e**), urate (**f**), mean corpuscular hemogolobin (**g**) and monocyte count (**h**). The error bars denote jackknife standard errors over 200 equally sized blocks of adjacent SNPs about the estimates of enrichment, and the centers of error bars represent the average value.



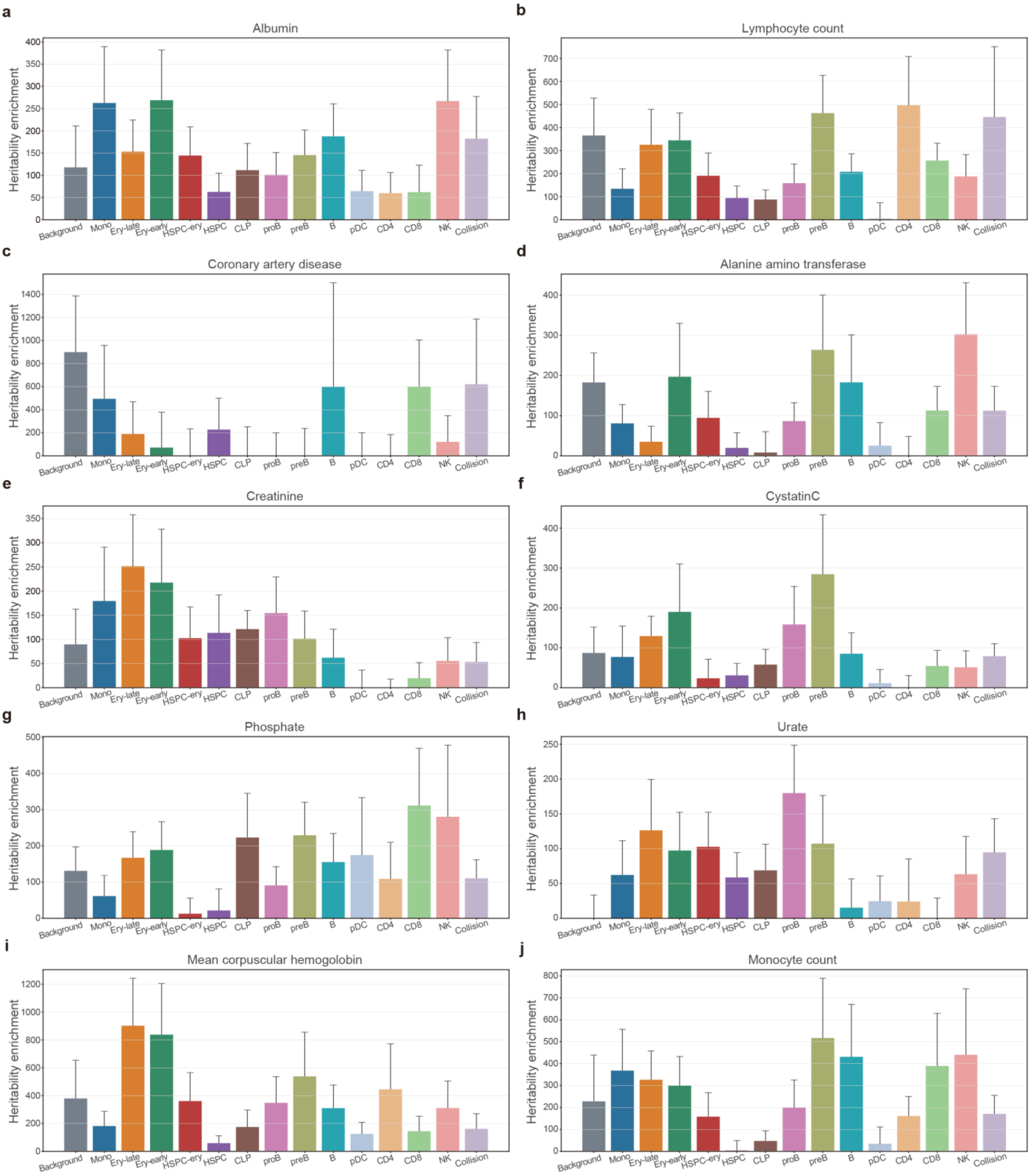
Supplementary Figure 43. SNP enrichment analysis for cell type-specific peaks identified by epiScanpy on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by epiScanpy. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



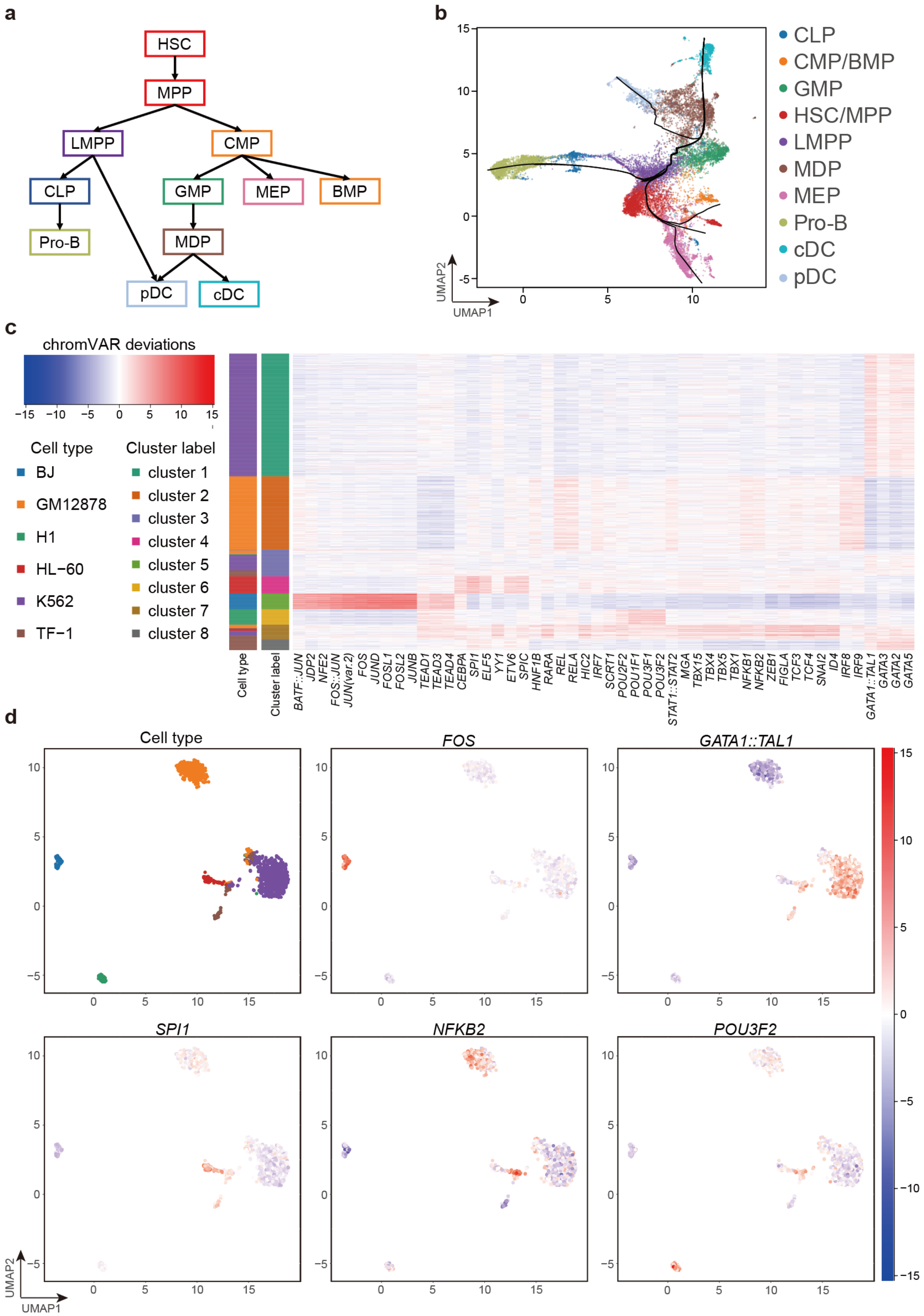
Supplementary Figure 43 (continue). SNP enrichment analysis for cell type-specific peaks identified by epiScanpy on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by epiScanpy. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



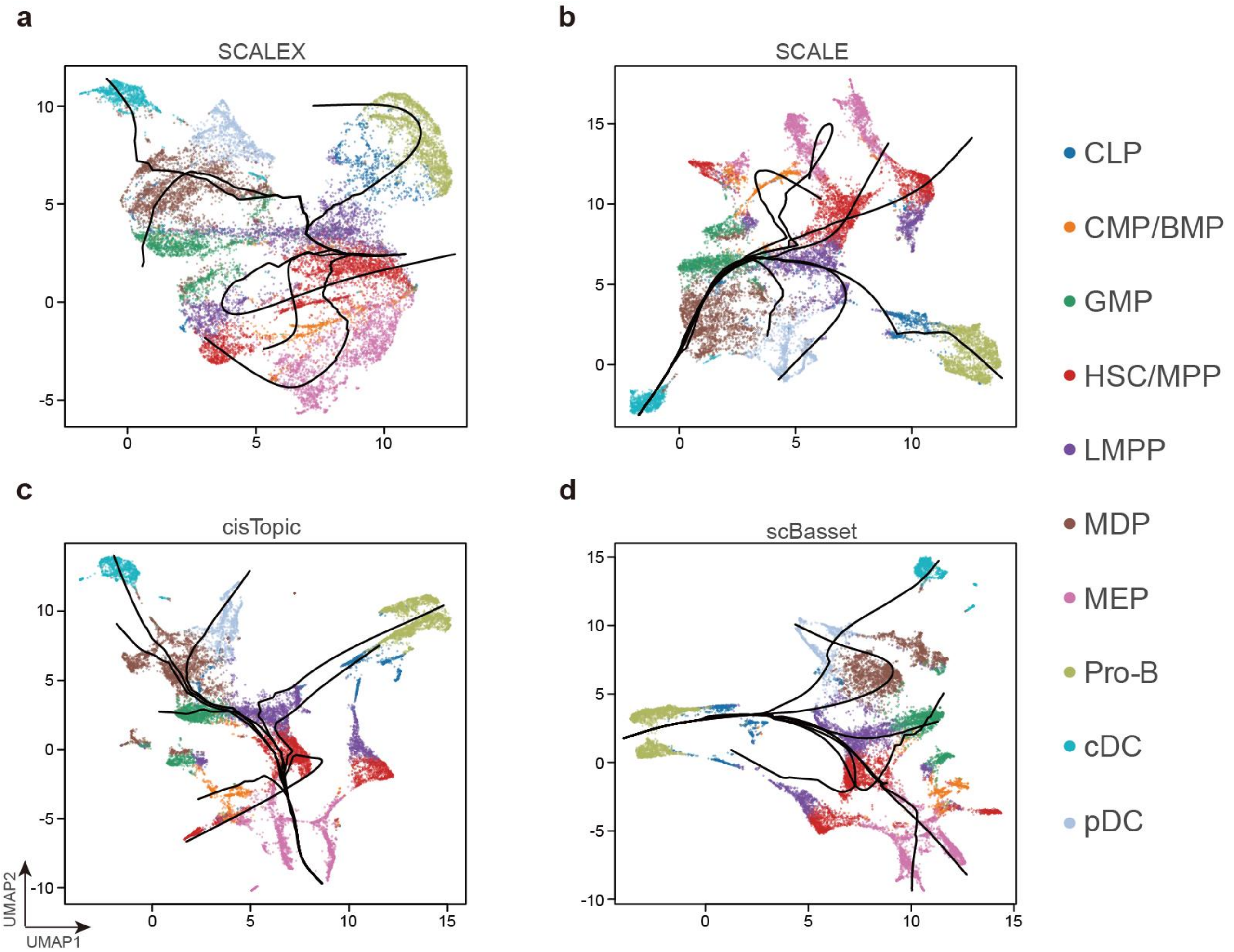
Supplementary Figure 43 (continue). SNP enrichment analysis for cell type-specific peaks identified by epiScanpy on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by epiScanpy. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



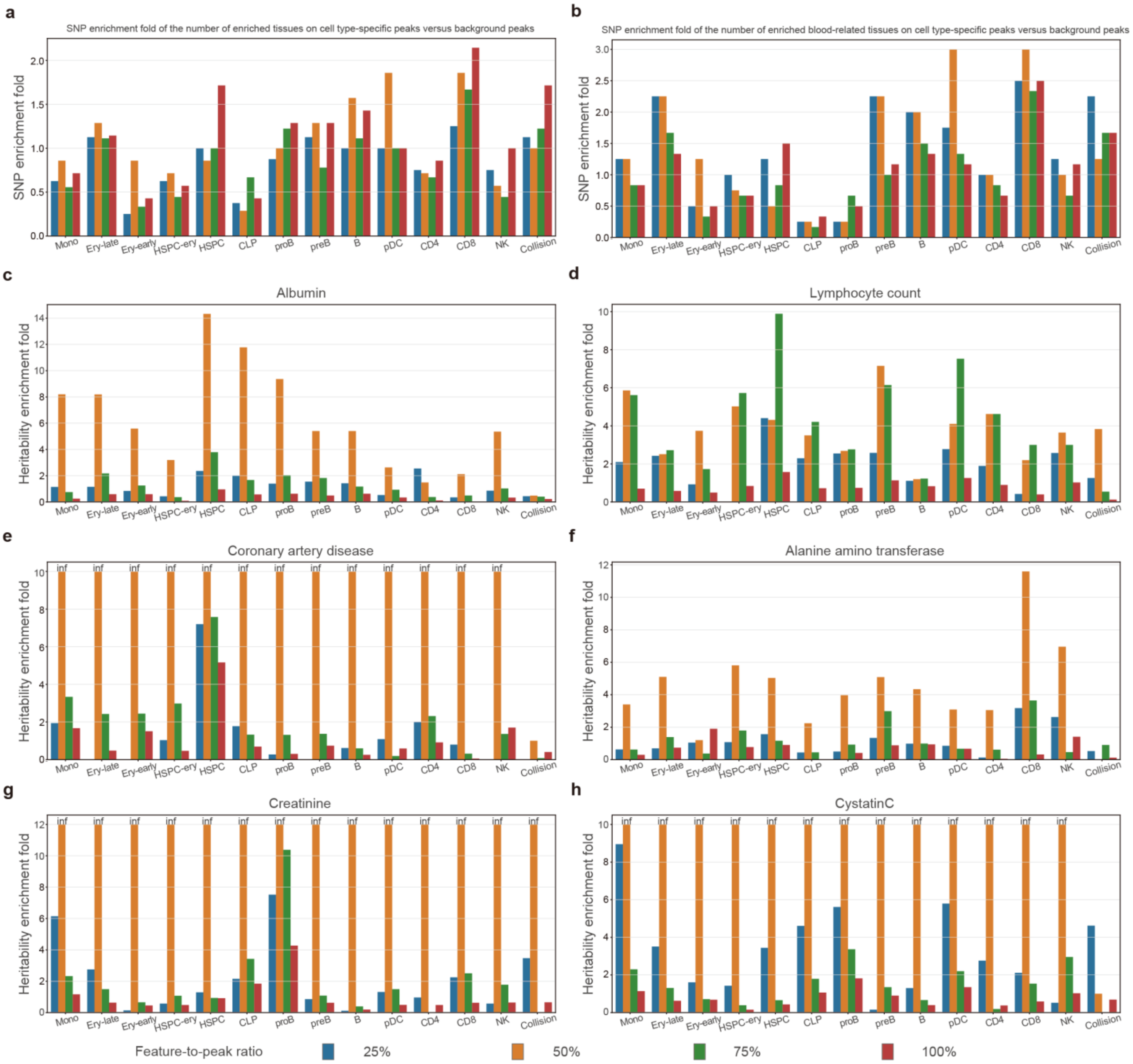
Supplementary Figure 44. Heritability enrichment analysis for cell type-specific peaks identified by epiScanpy on the Stimulated Droplet dataset. a-j, Heritability enrichments estimated by LDSC within cell type-specific peaks and the background peaks for blood-related traits including albumin (**a**), lymphocyte count (**b**), coronary artery disease (**c**), alanine amino transferase (**d**), creatinine (**e**), cystatinC (**f**), phosphate (**g**), urate (**h**), mean corpuscular hemogolobin (**i**) and monocyte count (**j**). The error bars denote jackknife standard errors over 200 equally sized blocks of adjacent SNPs about the estimates of enrichment, and the centers of error bars represent the average value.



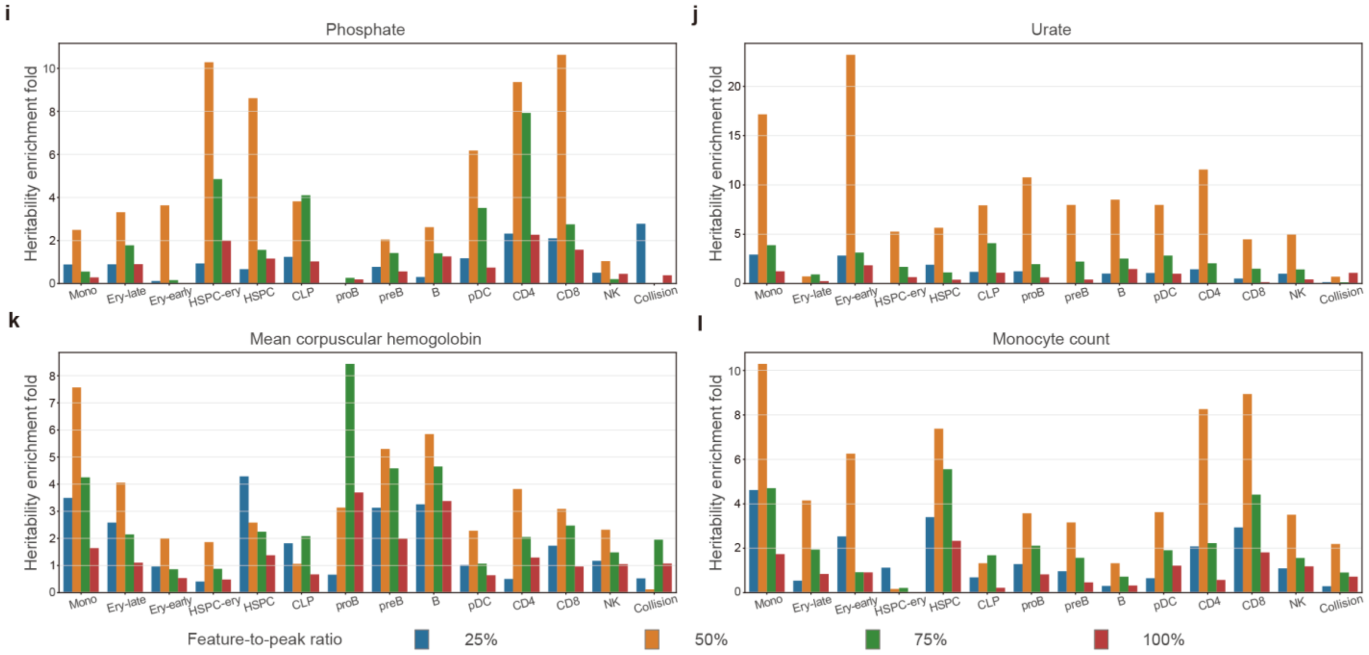
Supplementary Figure 45. Trajectory inference and motif enrichment analysis. **a**, A schematic of the authentic hematopoietic differentiation tree for the Immune dataset, with each color representing a different cell type. **b**, UMAP visualization of cell embeddings from the Immune dataset and the inferred trajectory with Slingshot. **c**, Heatmap of the top 50 most variable TF binding motifs within the peaks specific to each cluster by CASTLE and Louvain clustering for the InSilico dataset. The deviations calculated by chromVAR are shown. **d**, UMAP visualization of cell embeddings for the InSilico dataset, and chromVAR-derived deviation scores of five literature-validated motifs.



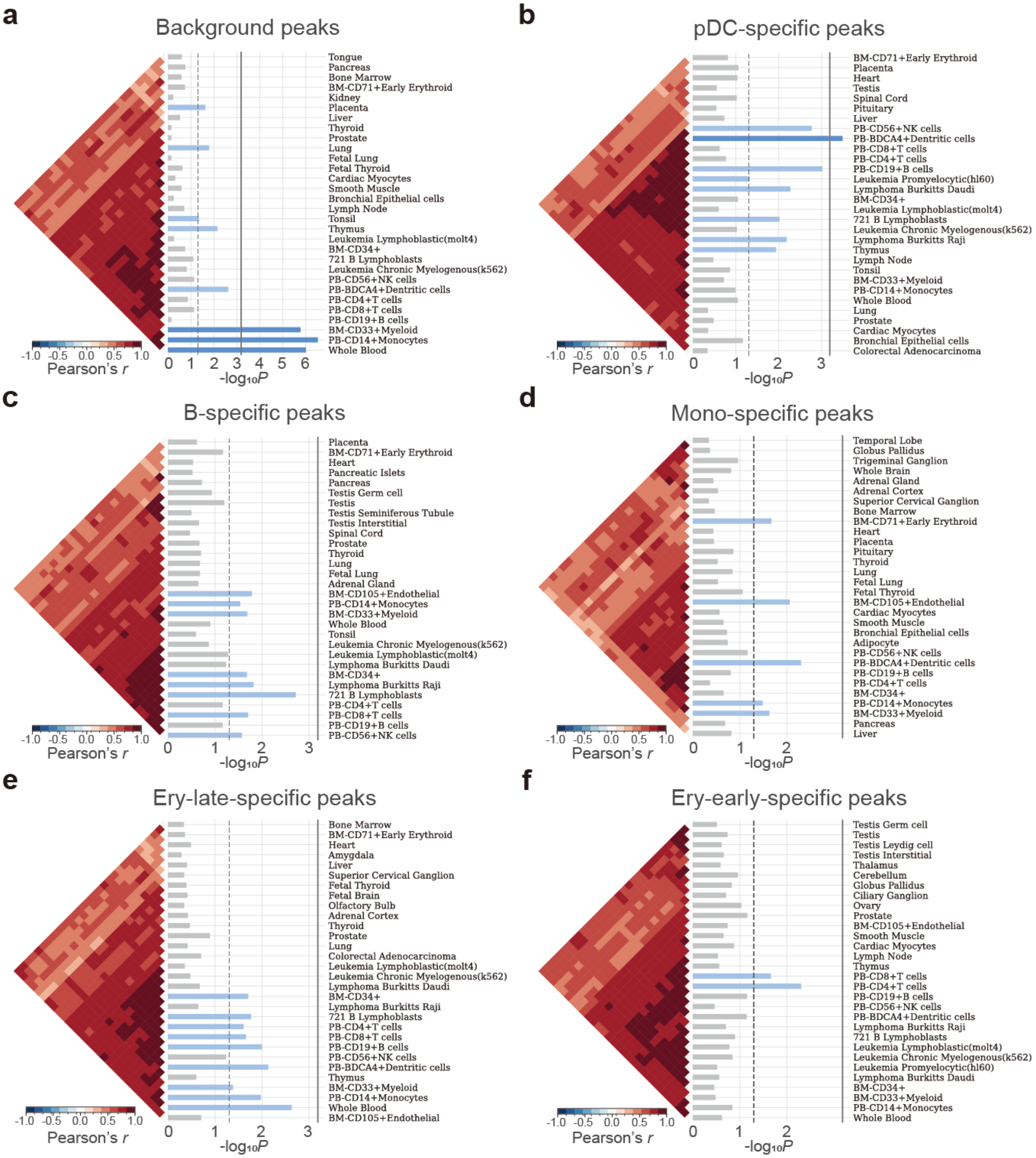
Supplementary Figure 46. Trajectory inference based on cell embeddings from baseline methods. a-d, UMAP visualization of cell embeddings on the Immune dataset with SCALEX (**a**), SCALE (**b**), cisTopic (**c**), scBasset (**d**), and the inferred trajectories with Slingshot.



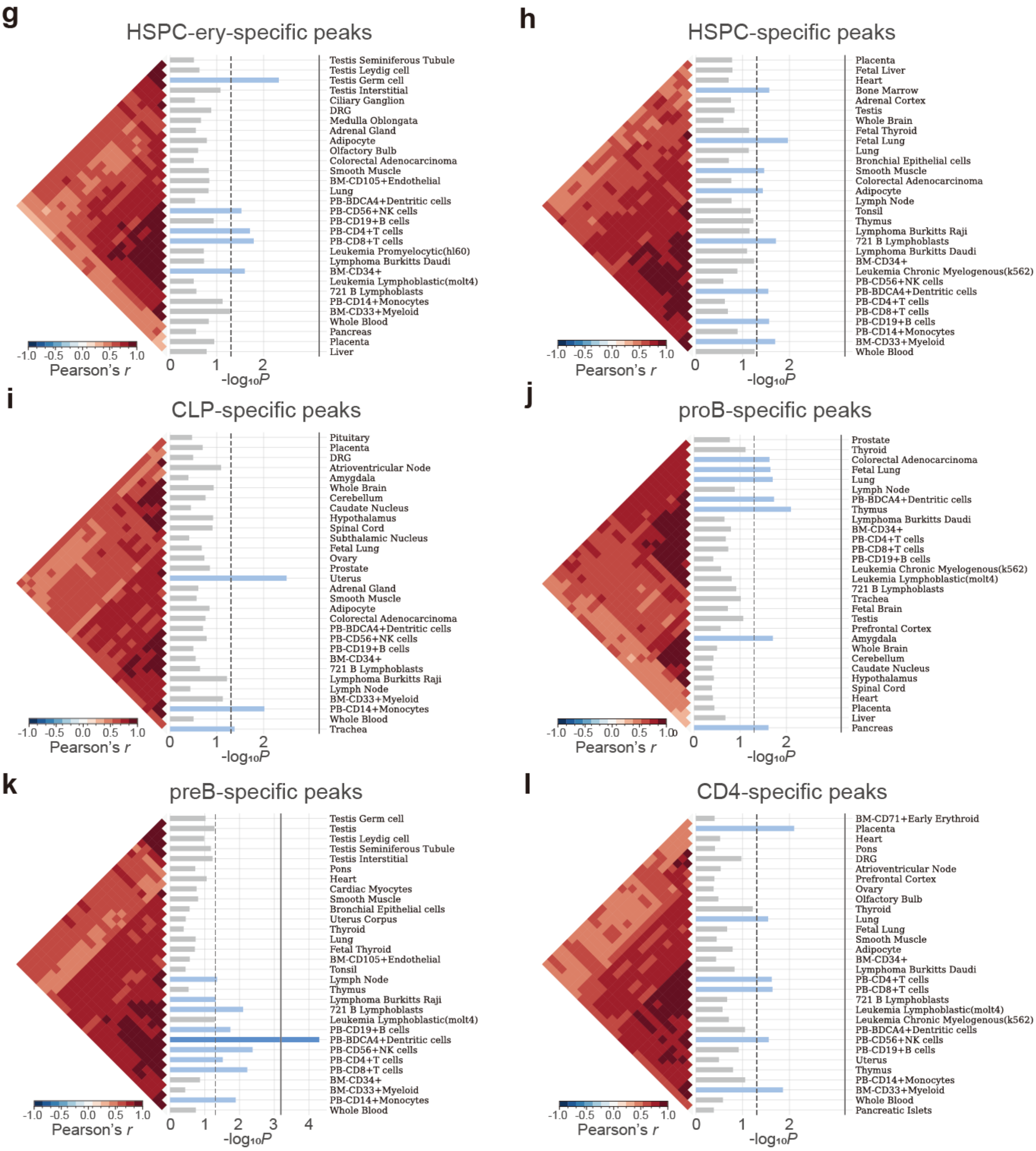
Supplementary Figure 47. SNP enrichment and heritability enrichment fold changes between the cell type-specific peaks and background peaks with different feature-to-peak ratios on the Stimulated Droplet dataset. a-b, SNP enrichment fold between the number of significantly enriched tissues (**a**) and significantly enriched blood-related tissues (**b**) on the cell type-specific peaks versus background peaks with the feature-to-peak ratio of 25%, 50%, 75% and 100%. **c-l**, Fold changes between the heritability enrichments of cell type-specific peaks and background peaks for blood-related traits including albumin (**c**), lymphocyte count (**d**), coronary artery disease (**e**), alanine amino transferase (**f**), creatinine (**g**), cystatinC (**h**), phosphate (**i**), urate (**j**), mean corpuscular hemogolobin (**k**) and monocyte count (**l**) with the feature-to-peak ratio of 25%, 50%, 75% and 100%. “inf” indicates infinity.



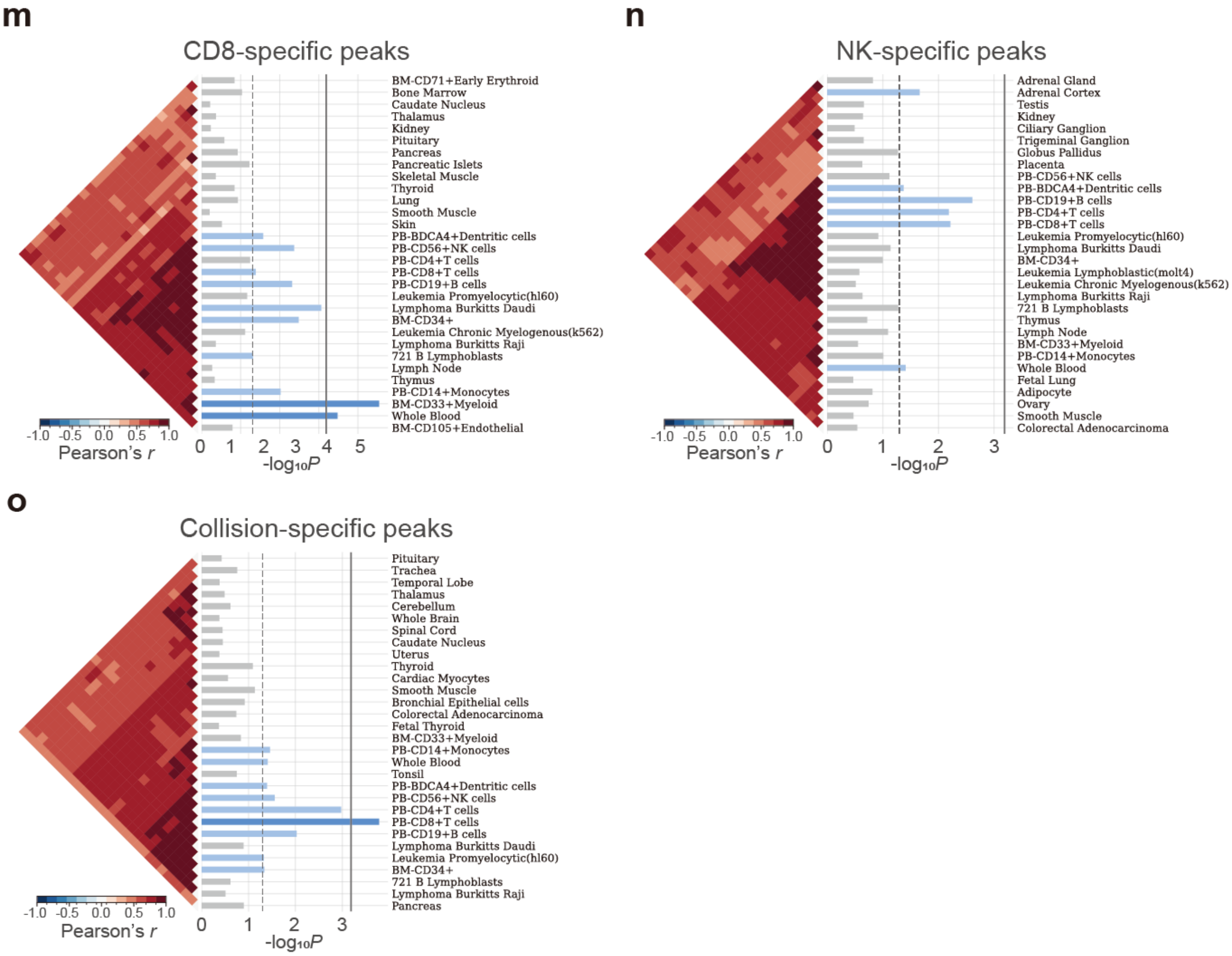
Supplementary Figure 47 (continue). SNP enrichment and heritability enrichment fold changes between the cell type-specific peaks and background peaks with different feature-to-peak ratios on the Stimulated Droplet dataset. **a-b**, SNP enrichment fold between the number of significantly enriched tissues (**a**) and significantly enriched blood-related tissues (**b**) on the cell type-specific peaks versus background peaks with the feature-to-peak ratio of 25%, 50%, 75% and 100%. **c-l**, Fold changes between the heritability enrichments of cell type-specific peaks and background peaks for blood-related traits including albumin (**c**), lymphocyte count (**d**), coronary artery disease (**e**), alanine amino transferase (**f**), creatinine (**g**), cystatinC (**h**), phosphate (**i**), urate (**j**), mean corpuscular hemogolobin (**k**) and monocyte count (**l**) with the feature-to-peak ratio of 25%, 50%, 75% and 100%. "inf" indicates infinity.



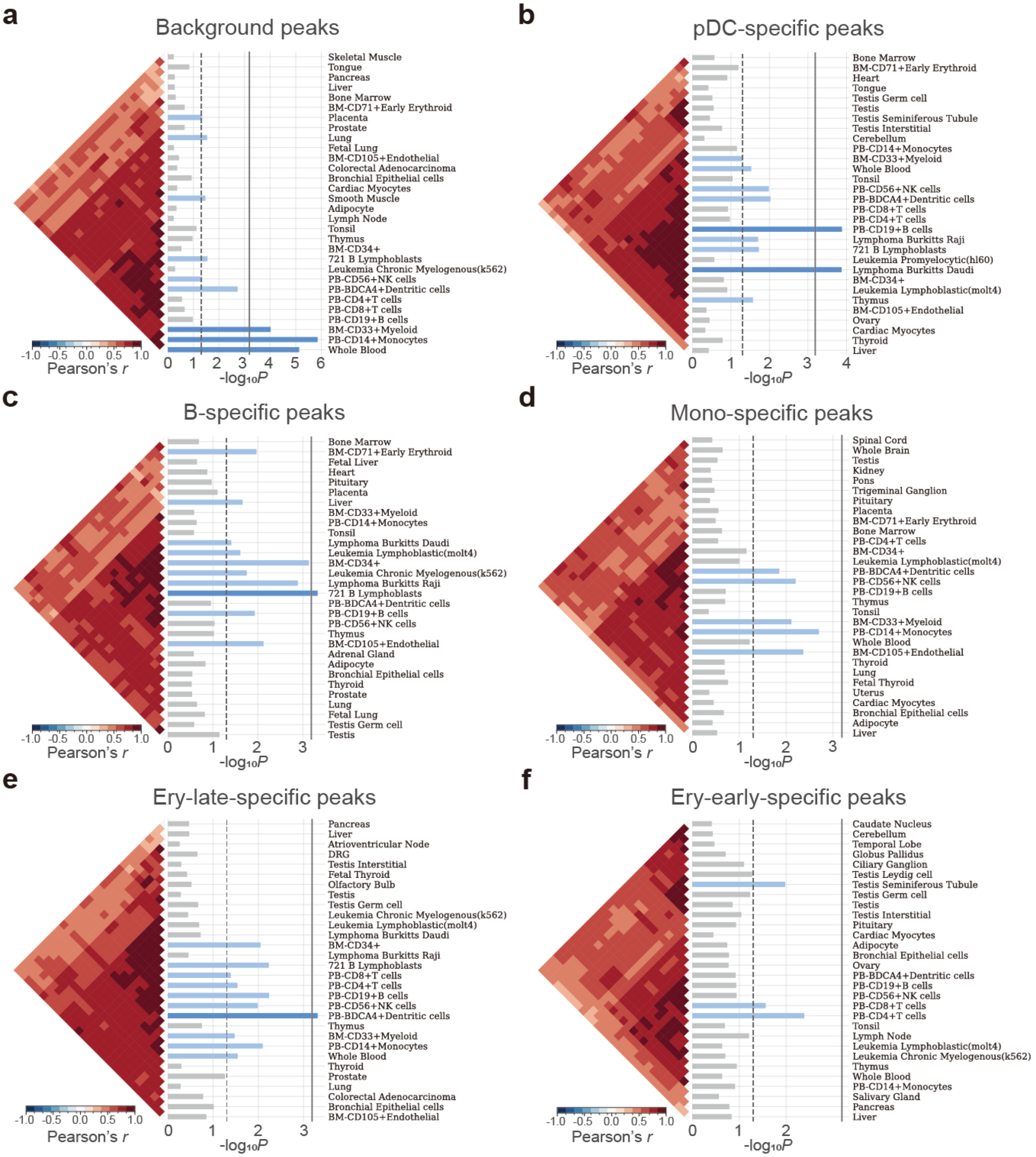
Supplementary Figure 48. SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 25% on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



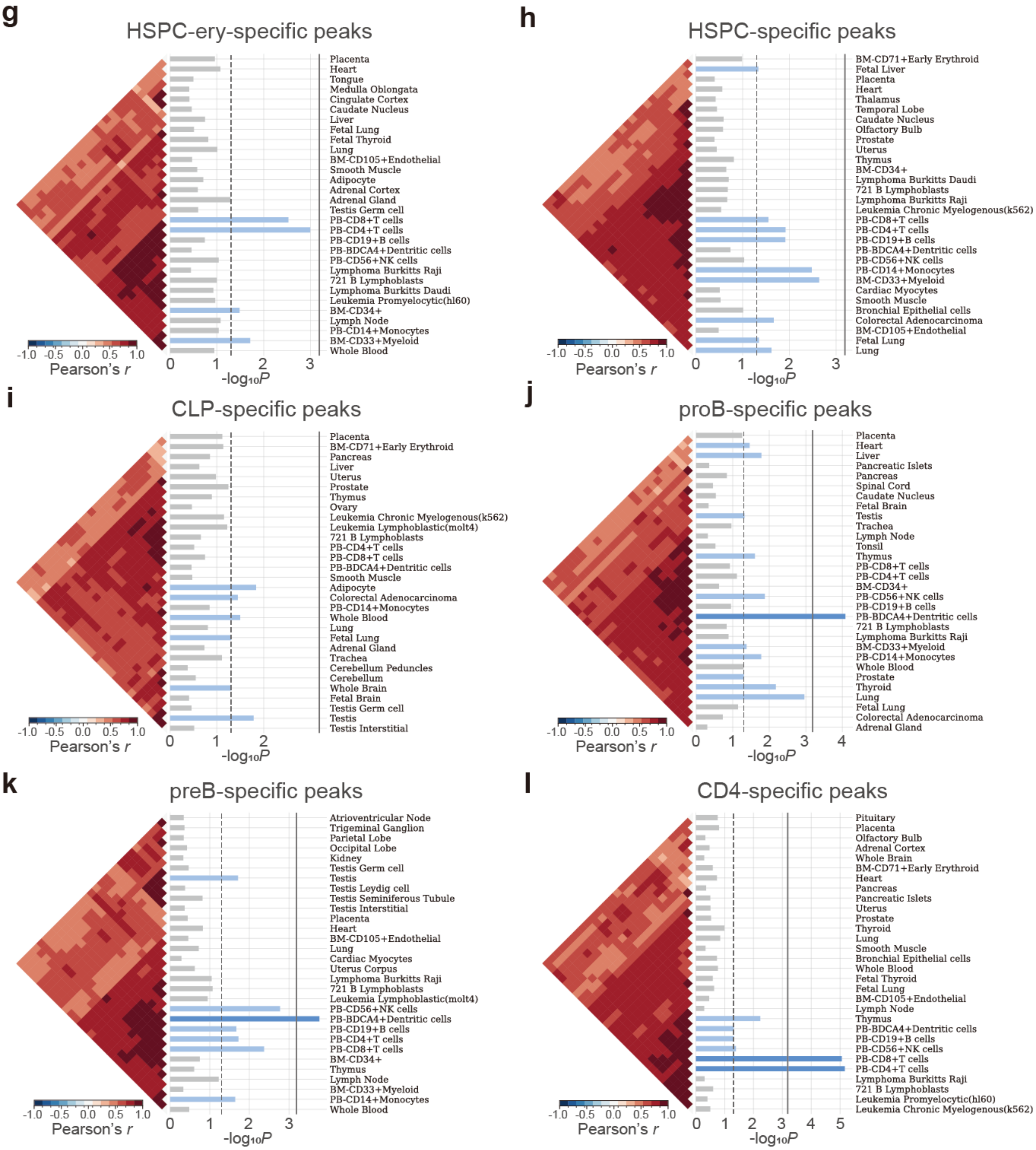
Supplementary Figure 48 (continue). SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 25% on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



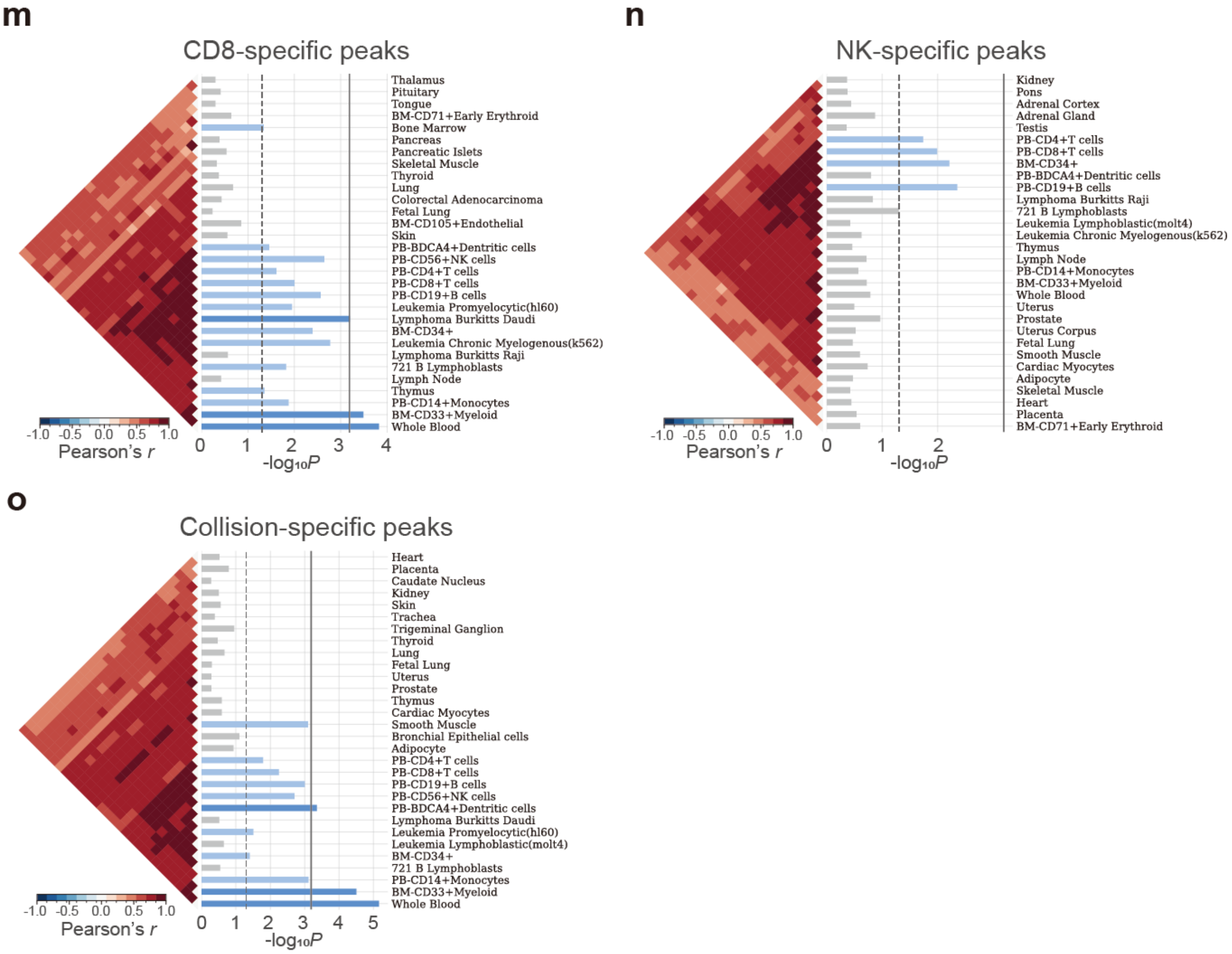
Supplementary Figure 48 (continue). SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 25% on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



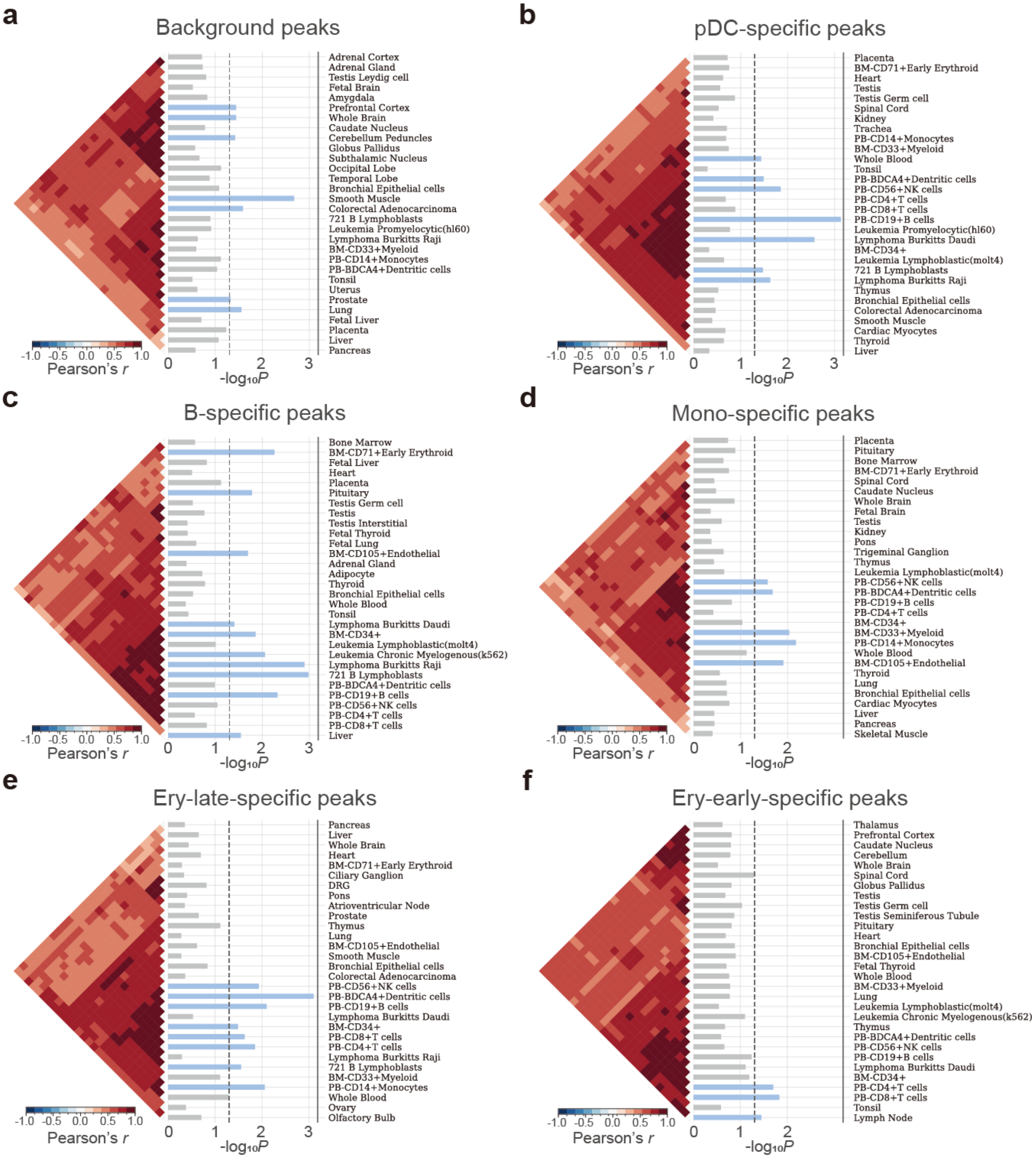
Supplementary Figure 49. SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 75% on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



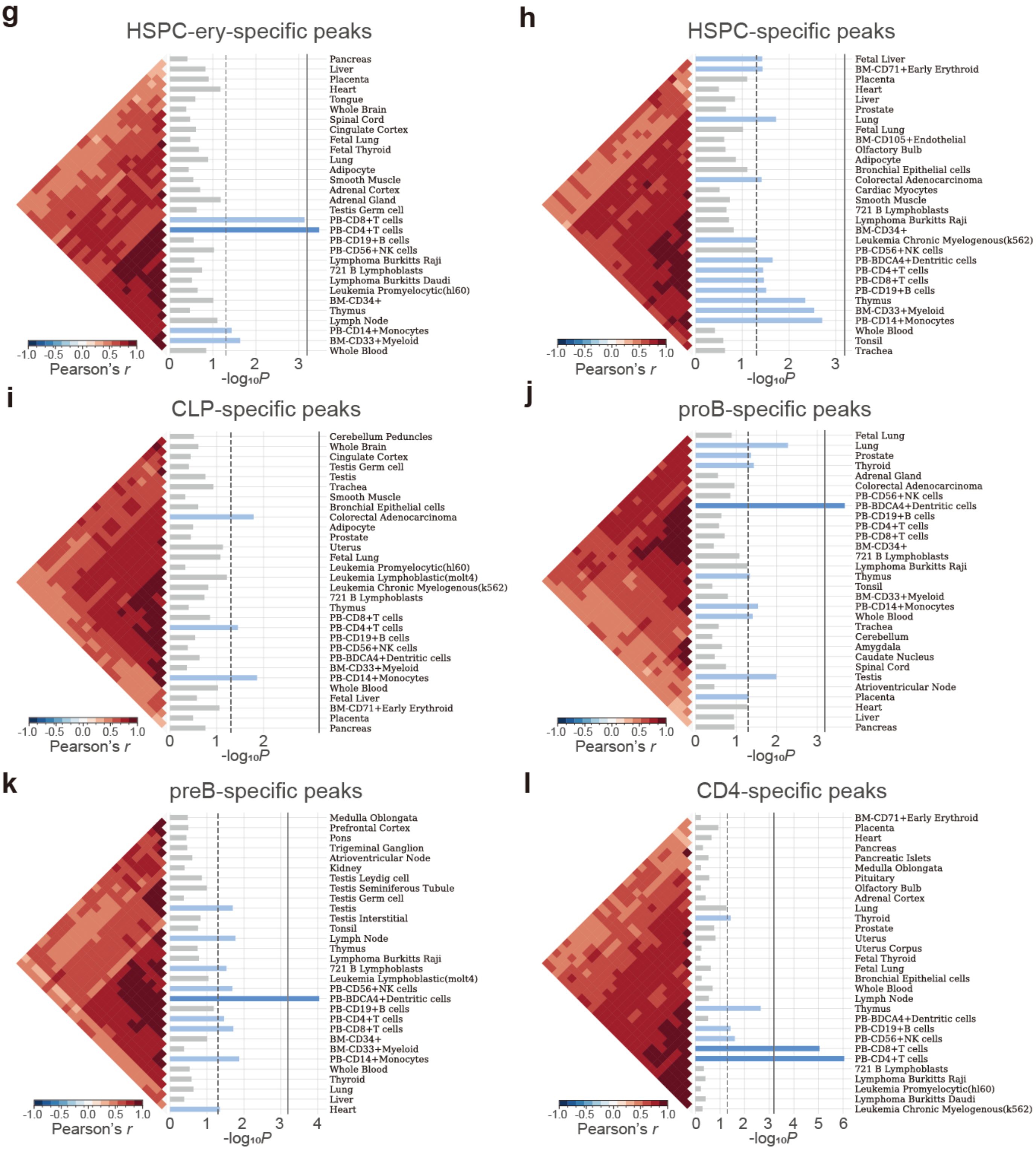
Supplementary Figure 49 (continue). SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 75% on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



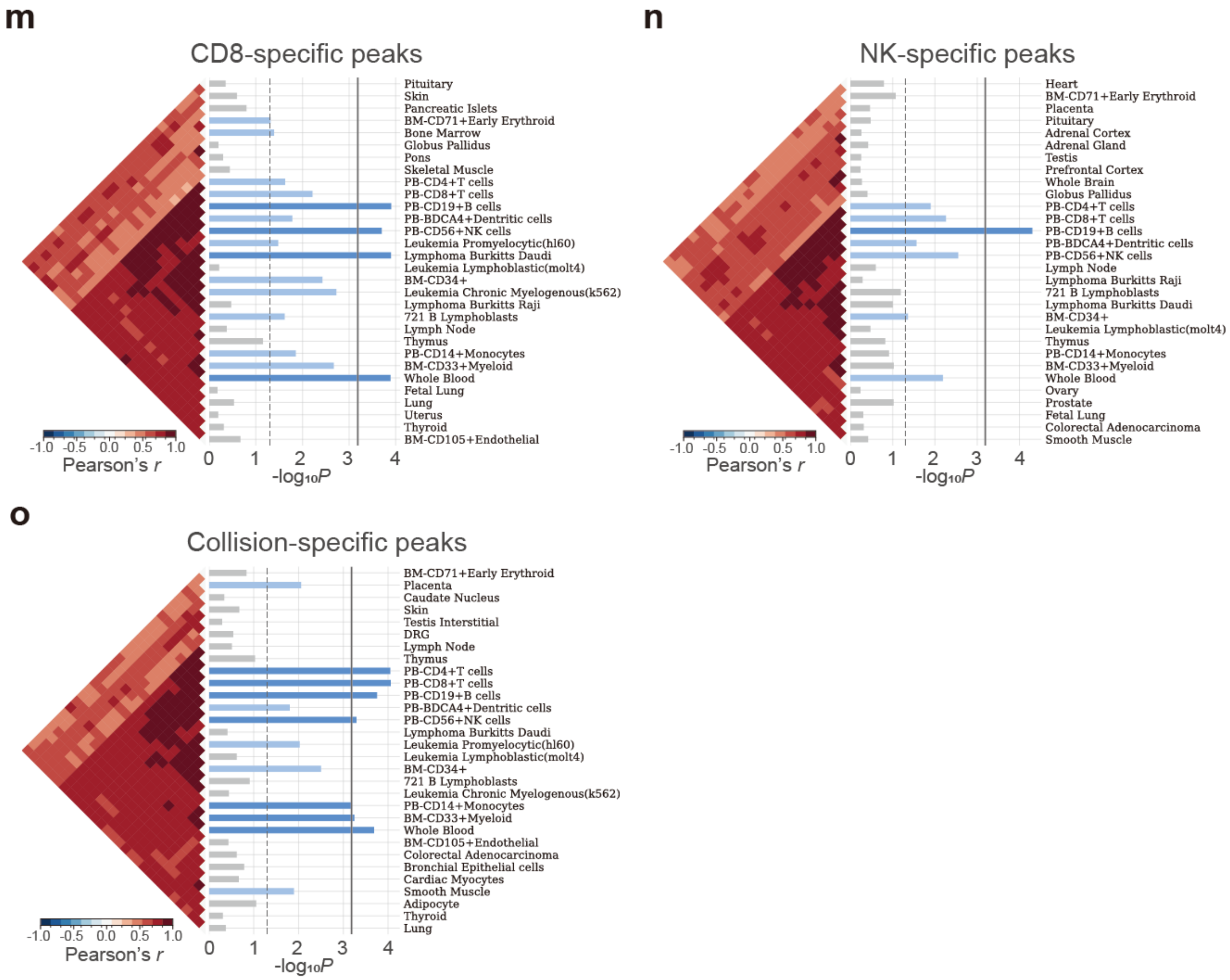
Supplementary Figure 49 (continue). SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 75% on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



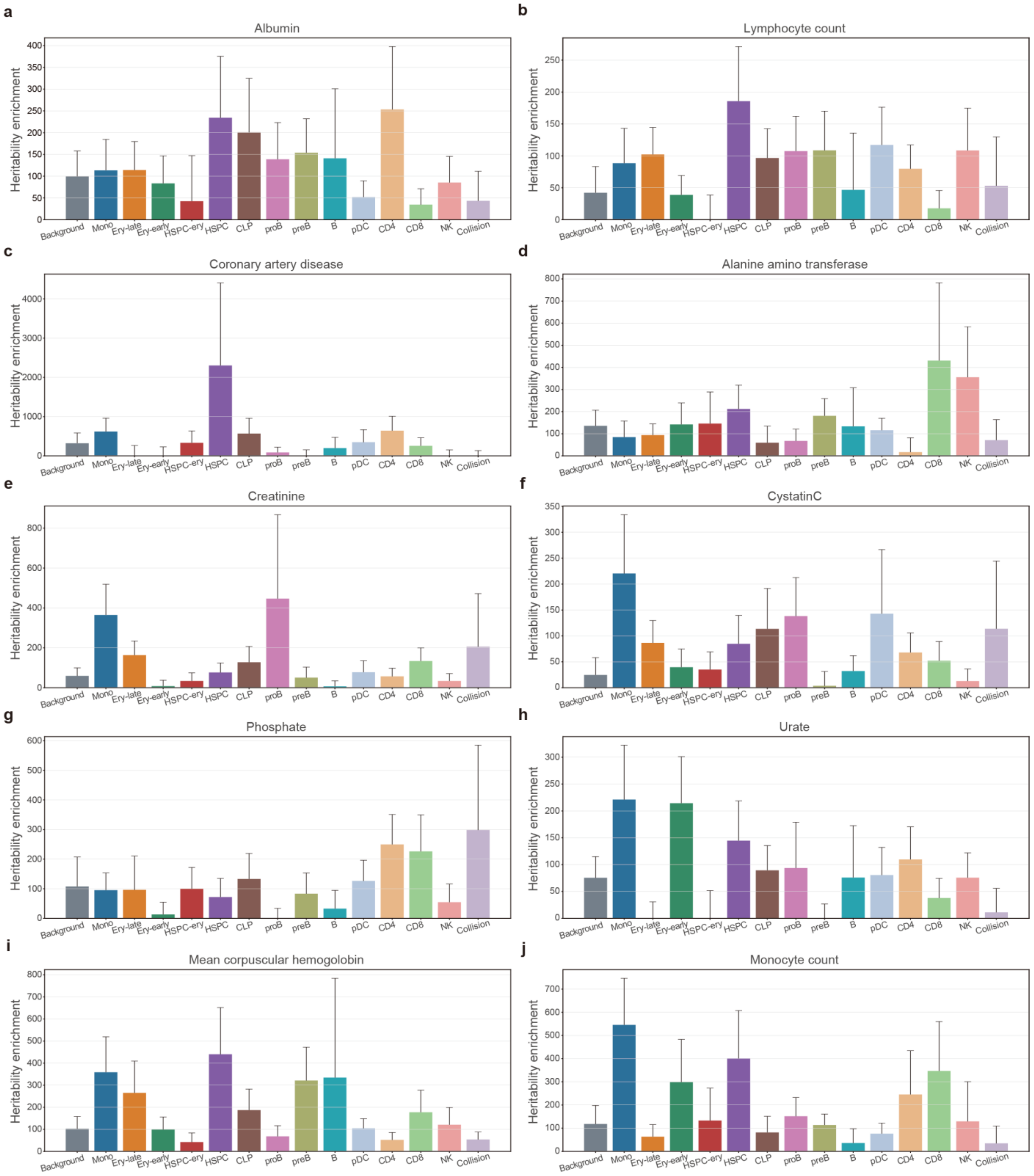
Supplementary Figure 50. SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 100% on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



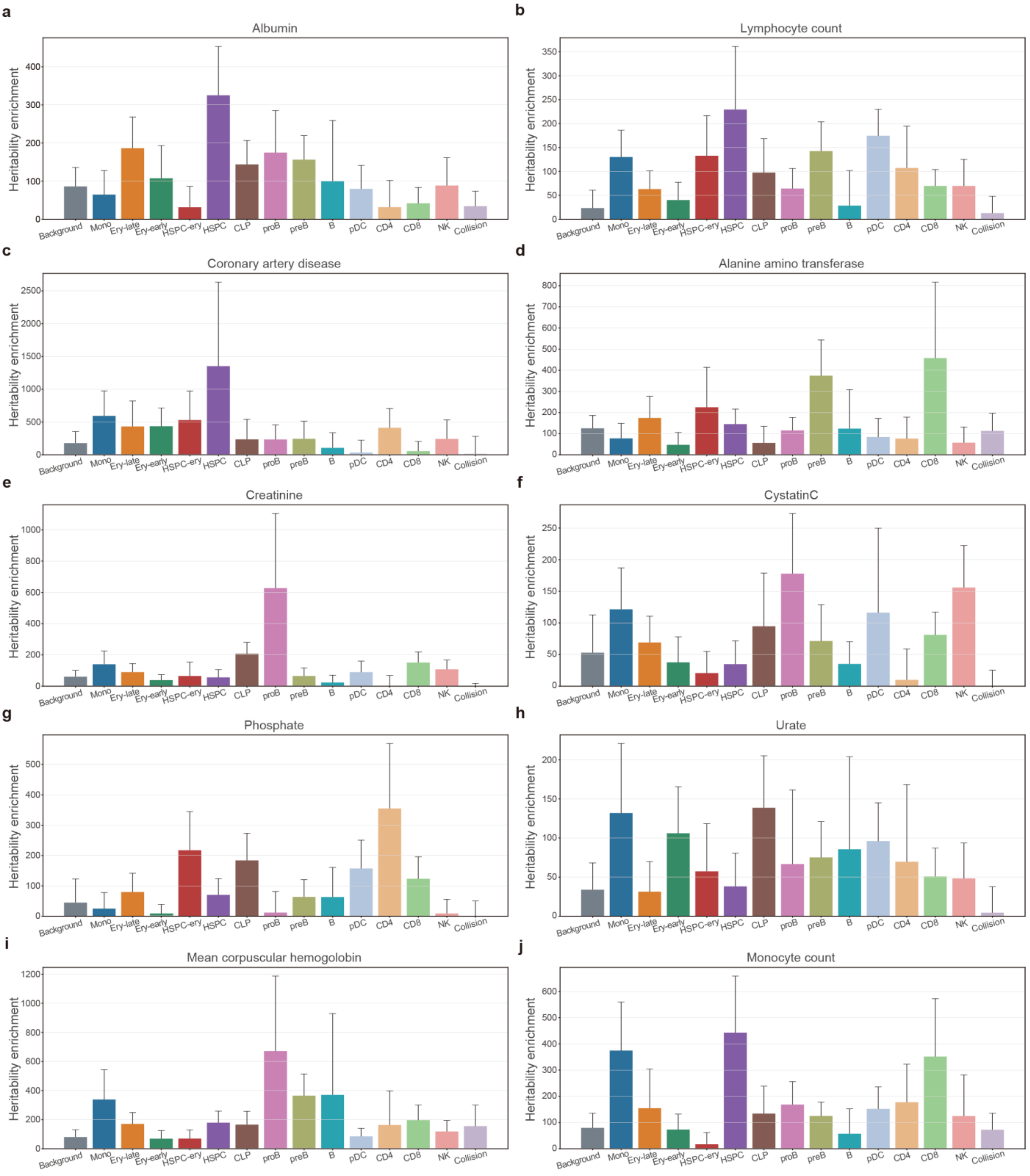
Supplementary Figure 50 (continue). SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 100% on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



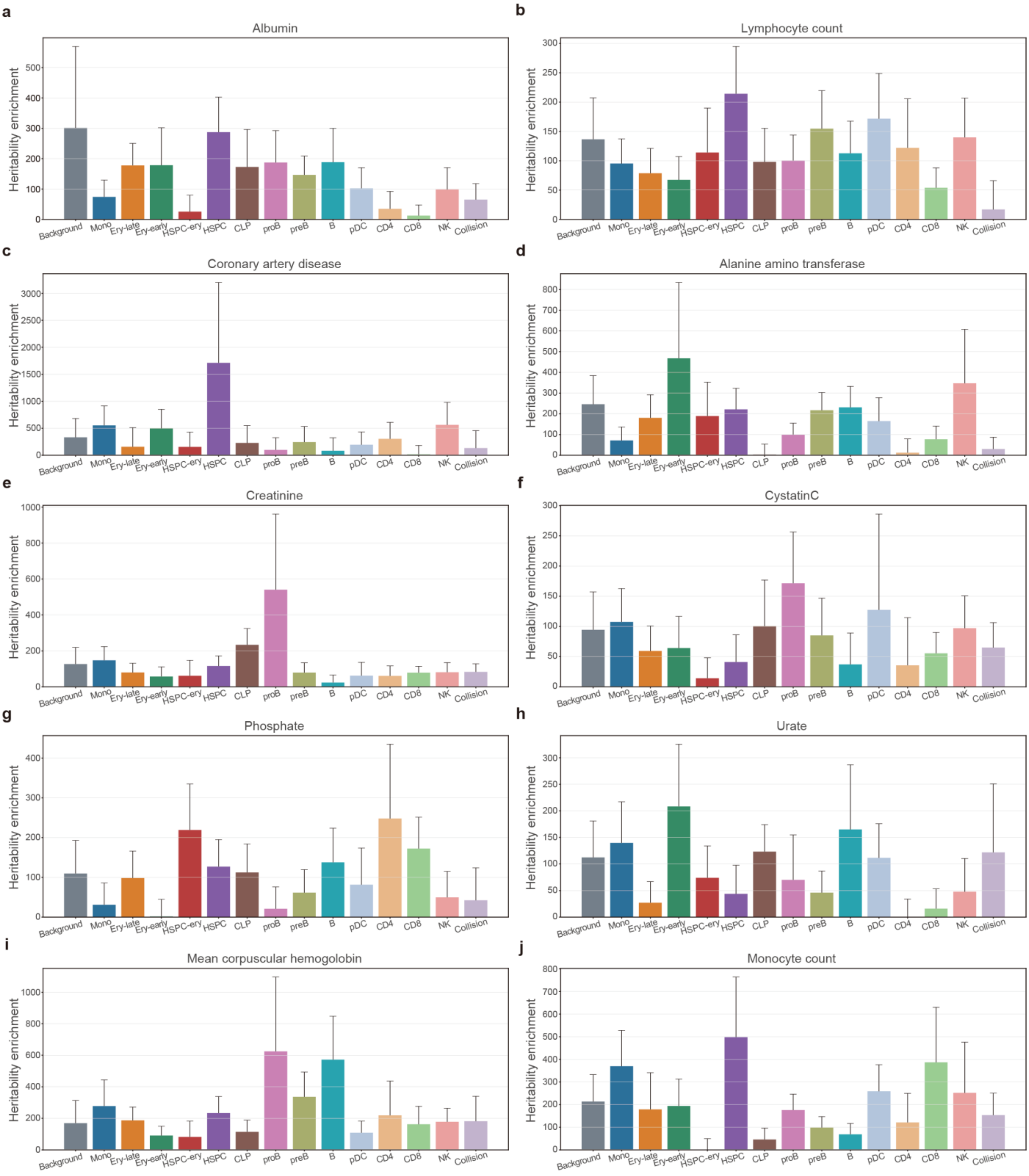
Supplementary Figure 50 (continue). SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 100% on the Stimulated Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



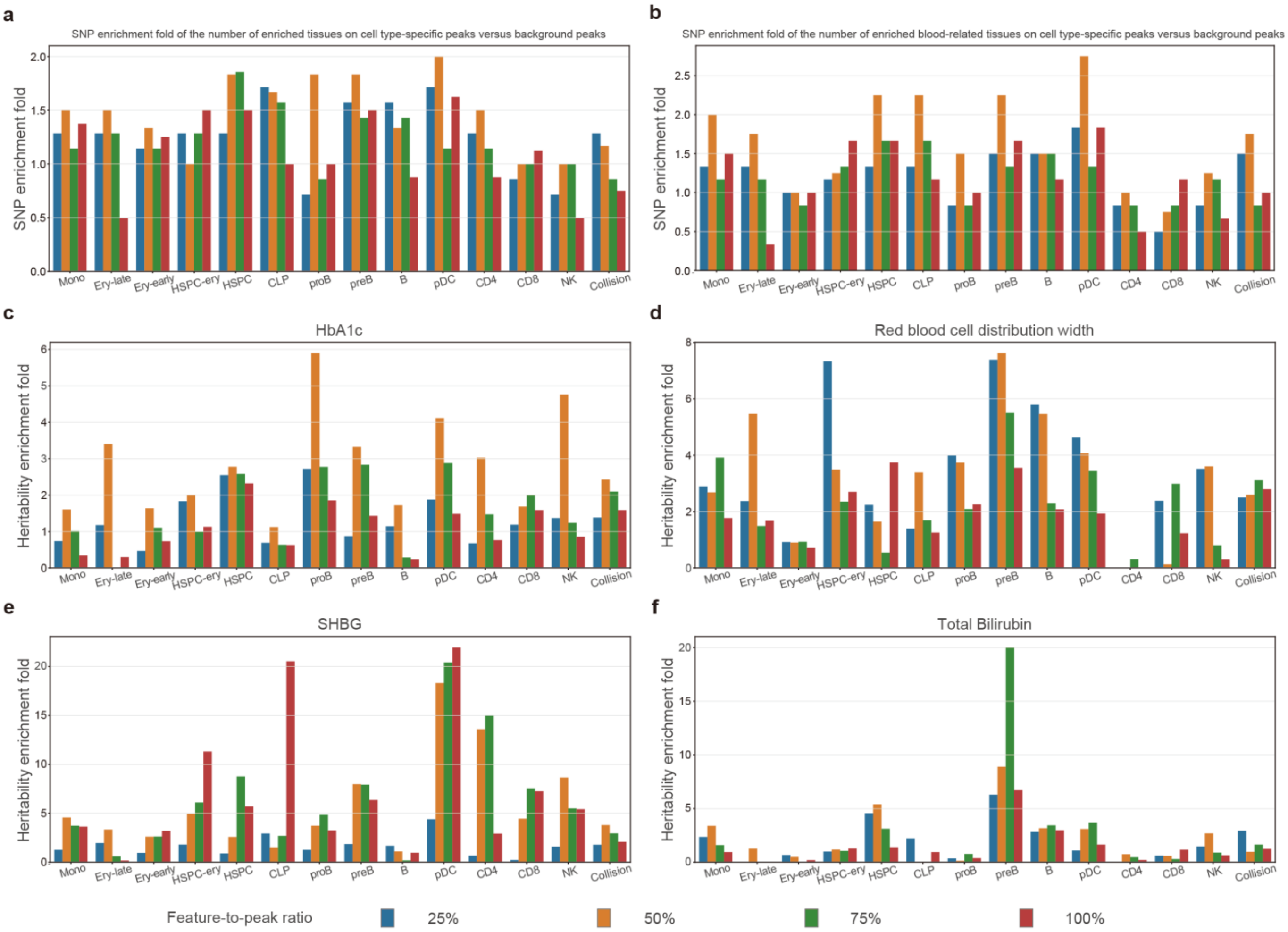
Supplementary Figure 51. Heritability enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 25% on the Stimulated Droplet dataset. a-j, Heritability enrichments estimated by LDSC within cell type-specific peaks identified by CASTLE and the background peaks for blood-related traits including albumin (**a**), lymphocyte count (**b**), coronary artery disease (**c**), alanine amino transferase (**d**), creatinine (**e**), cystatinC (**f**), phosphate (**g**), urate (**h**), mean corpuscular hemogolobin (**i**) and monocyte count (**j**). The error bars denote jackknife standard errors over 200 equally sized blocks of adjacent SNPs about the estimates of enrichment, and the centers of error bars represent the average value.



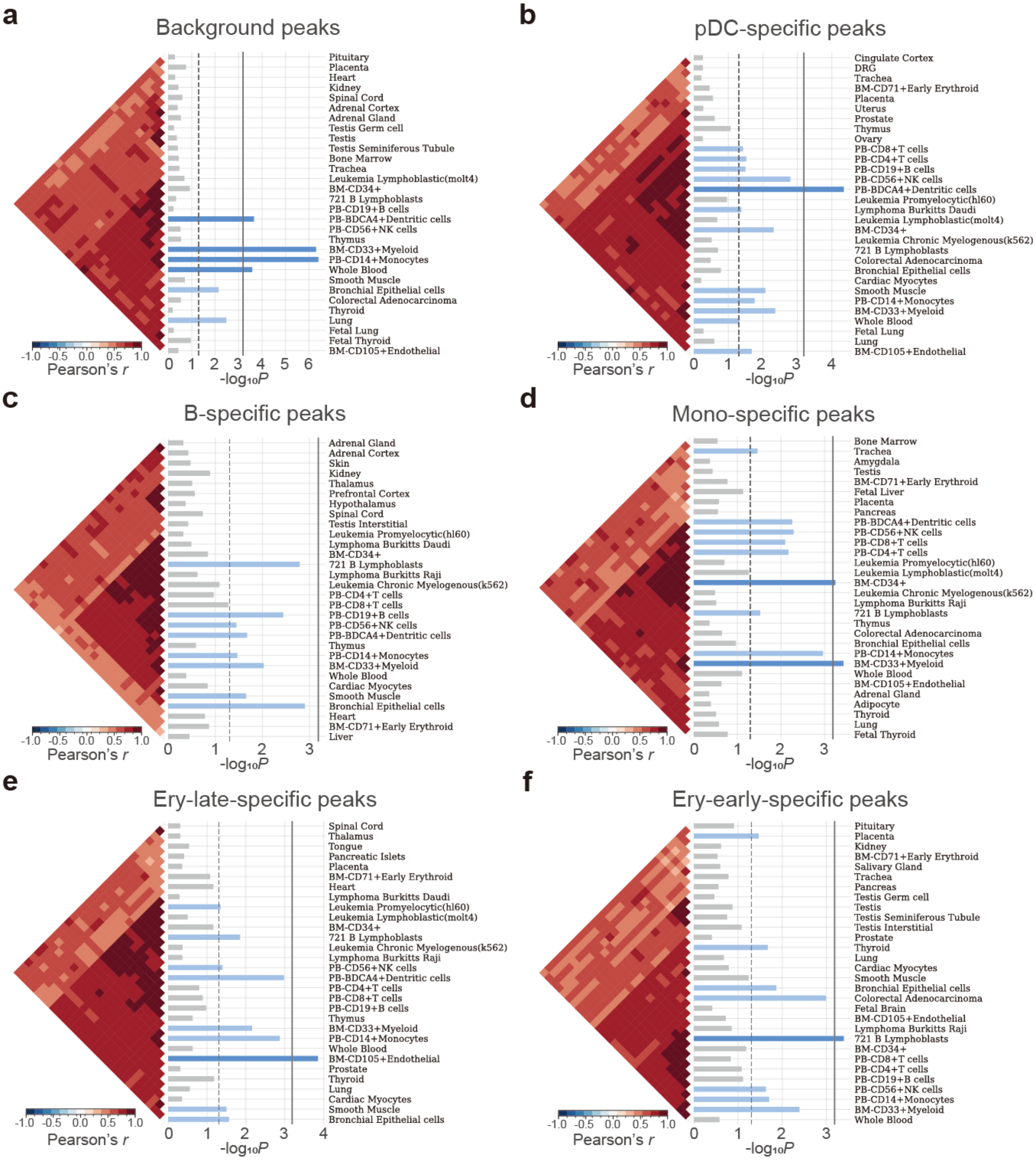
Supplementary Figure 52. Heritability enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 75% on the Stimulated Droplet dataset. a-j, Heritability enrichments estimated by LDSC within cell type-specific peaks identified by CASTLE and the background peaks for blood-related traits including albumin (**a**), lymphocyte count (**b**), coronary artery disease (**c**), alanine amino transferase (**d**), creatinine (**e**), cystatinC (**f**), phosphate (**g**), urate (**h**), mean corpuscular hemogolobin (**i**) and monocyte count (**j**). The error bars denote jackknife standard errors over 200 equally sized blocks of adjacent SNPs about the estimates of enrichment, and the centers of error bars represent the average value.



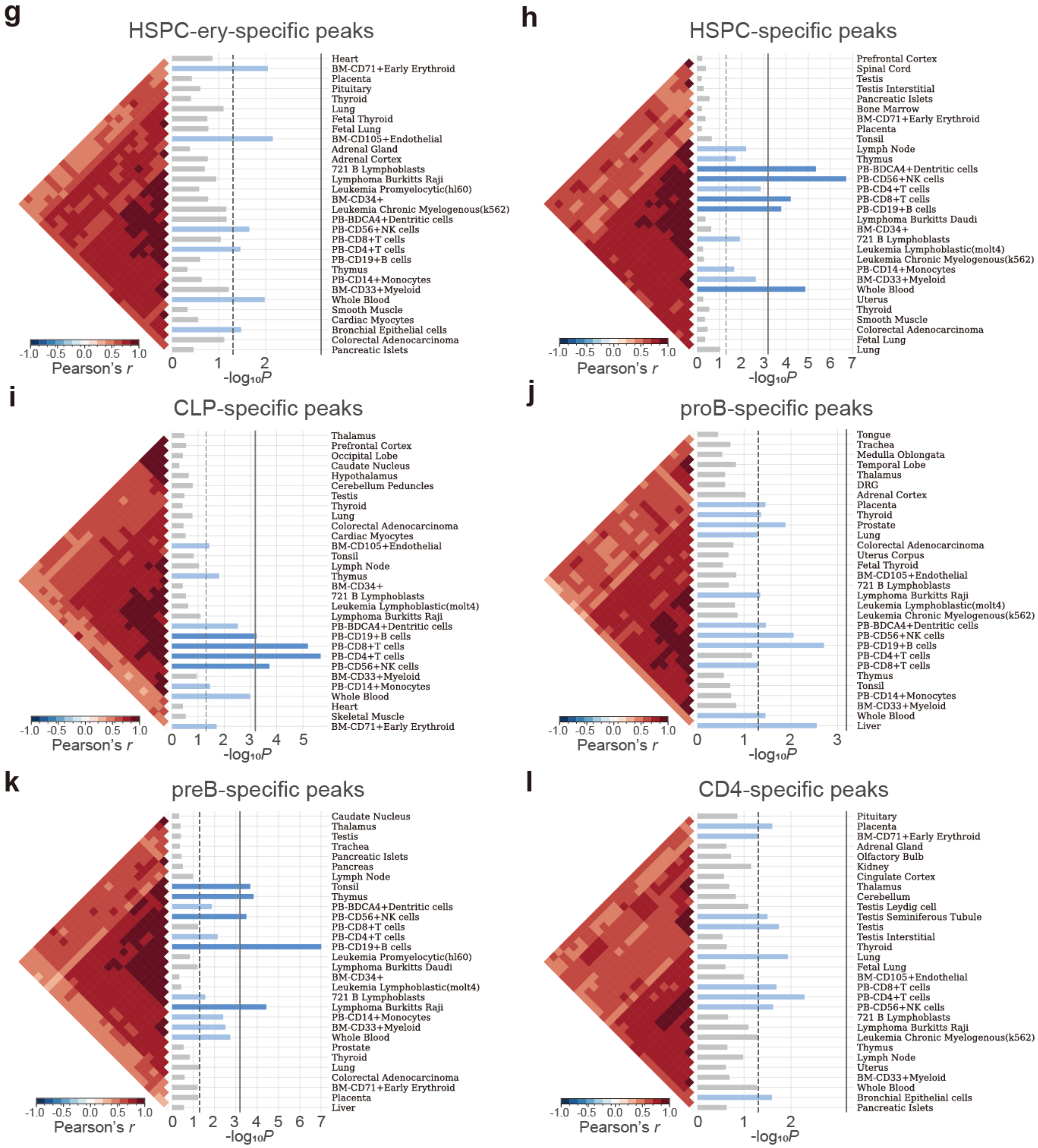
Supplementary Figure 53. Heritability enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 100% on the Stimulated Droplet dataset. a-j, Heritability enrichments estimated by LDSC within cell type-specific peaks identified by CASTLE and the background peaks for blood-related traits including albumin (**a**), lymphocyte count (**b**), coronary artery disease (**c**), alanine amino transferase (**d**), creatinine (**e**), cystatinC (**f**), phosphate (**g**), urate (**h**), mean corpuscular hemogolobin (**i**) and monocyte count (**j**). The error bars denote jackknife standard errors over 200 equally sized blocks of adjacent SNPs about the estimates of enrichment, and the centers of error bars represent the average value.



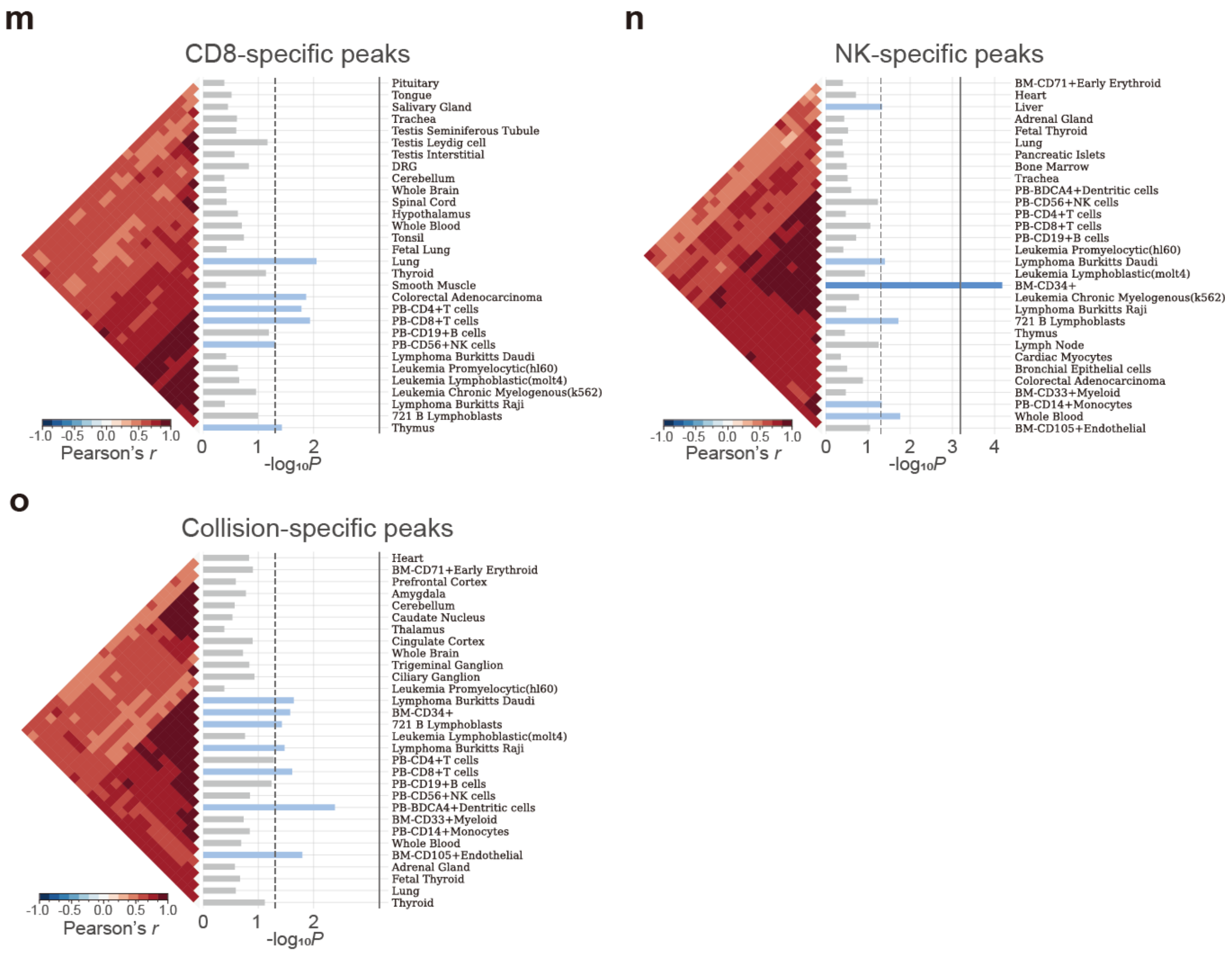
Supplementary Figure 54. SNP enrichment and heritability enrichment fold changes between the cell type-specific peaks and background peaks with different feature-to-peak ratios on the Resting Droplet dataset. a-b, SNP enrichment fold between the number of significantly enriched tissues (**a**) and significantly enriched dataset-related tissues (**b**) on the cell type-specific peaks versus background peaks with the feature-to-peak ratio of 25%, 50%, 75% and 100%. **c-f**, Fold changes between the heritability enrichments of cell type-specific peaks and background peaks for dataset-related traits including HbA1c (**c**), Red blood cell distribution width (**d**), SHBG (**e**) and Total Bilirubin (**f**) with the feature-to-peak ratio of 25%, 50%, 75% and 100%.



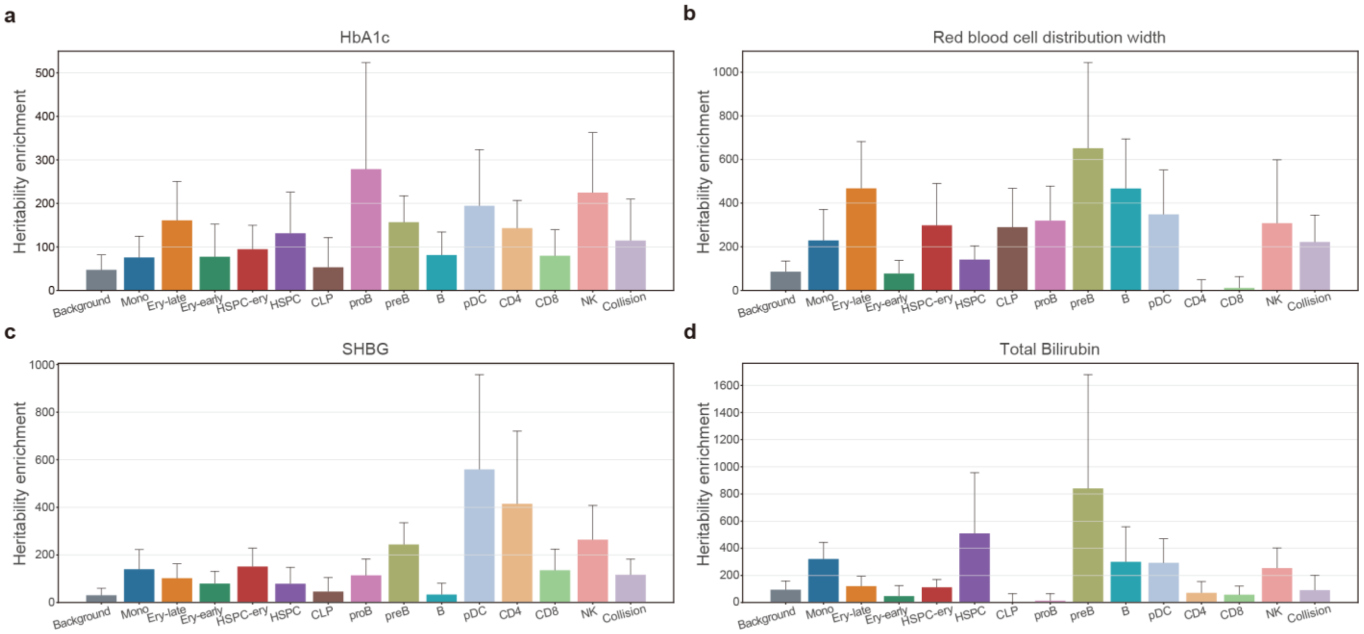
Supplementary Figure 55. SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Resting Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



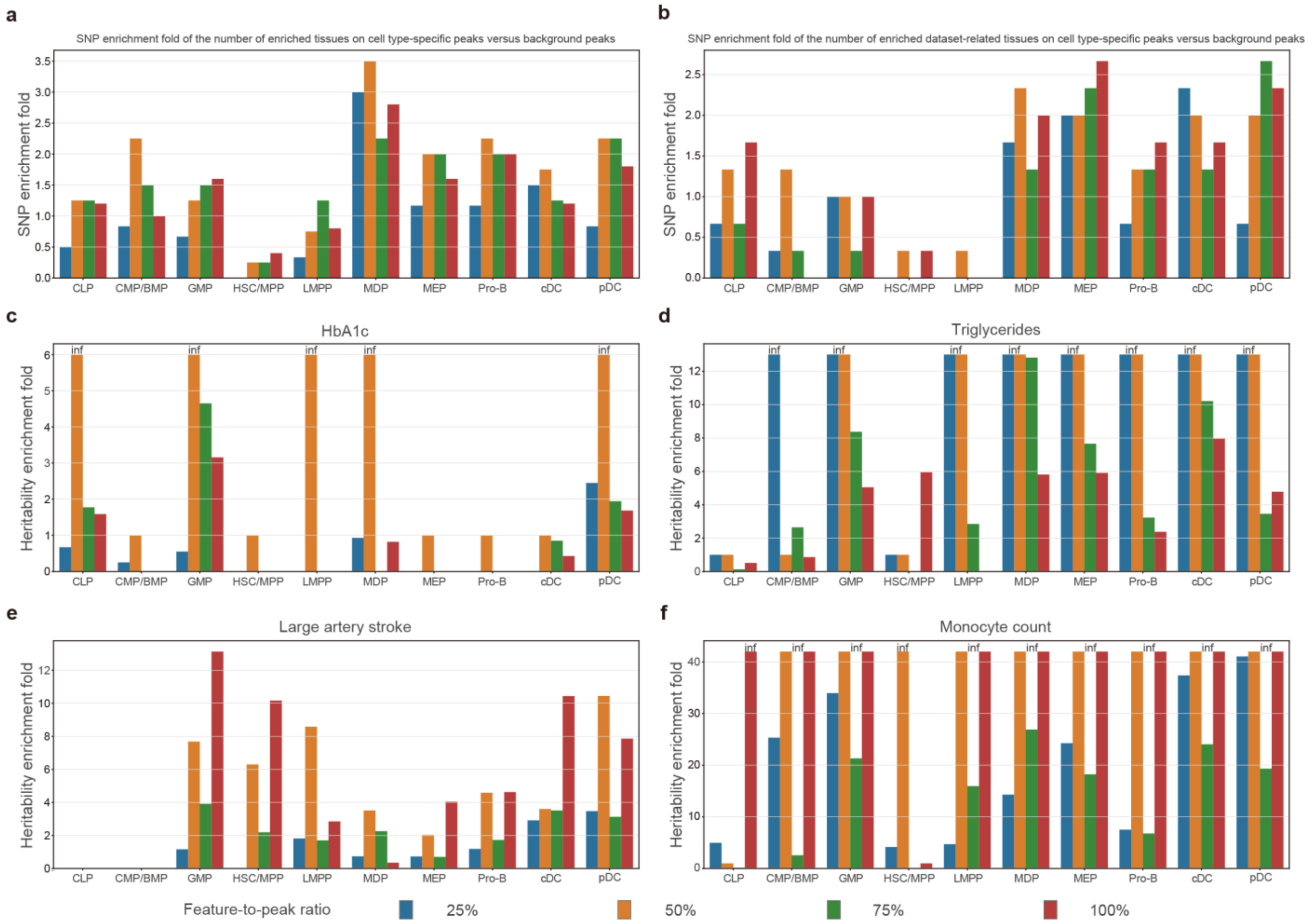
Supplementary Figure 55 (continue). SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Resting Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



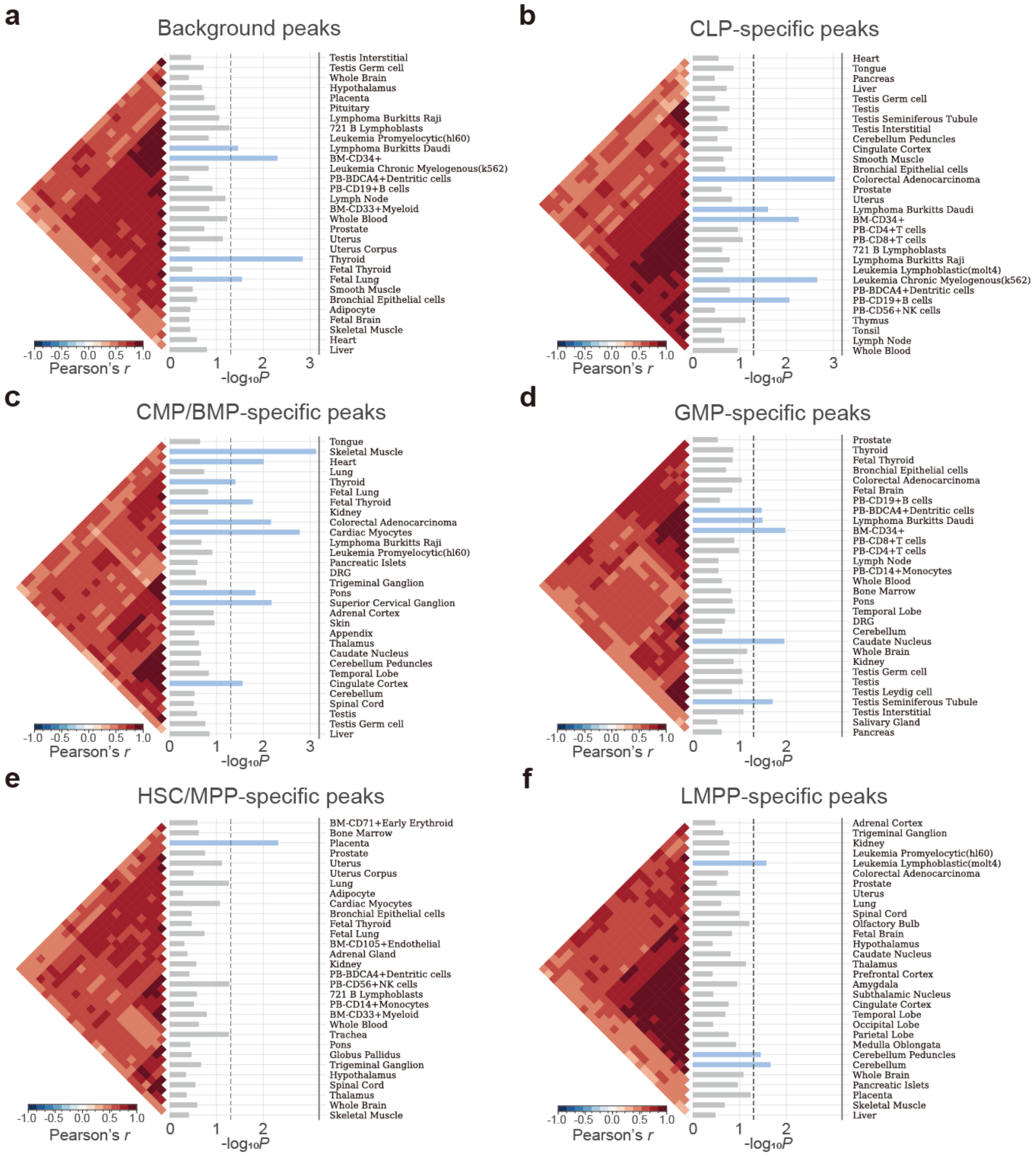
Supplementary Figure 55 (continue). SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Resting Droplet dataset. a-o, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), pDC-specific peaks (**b**), B-specific peaks (**c**), Mono-specific peaks (**d**), Ery-late-specific peaks (**e**), Ery-early-specific peaks (**f**), HSPC-ery-specific peaks (**g**), HSPC-specific peaks (**h**), CLP-specific peaks (**i**), proB-specific peaks (**j**), preB-specific peaks (**k**), CD4-specific peaks (**l**), CD8-specific peaks (**m**), NK-specific peaks (**n**) and Collision-specific peaks (**o**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



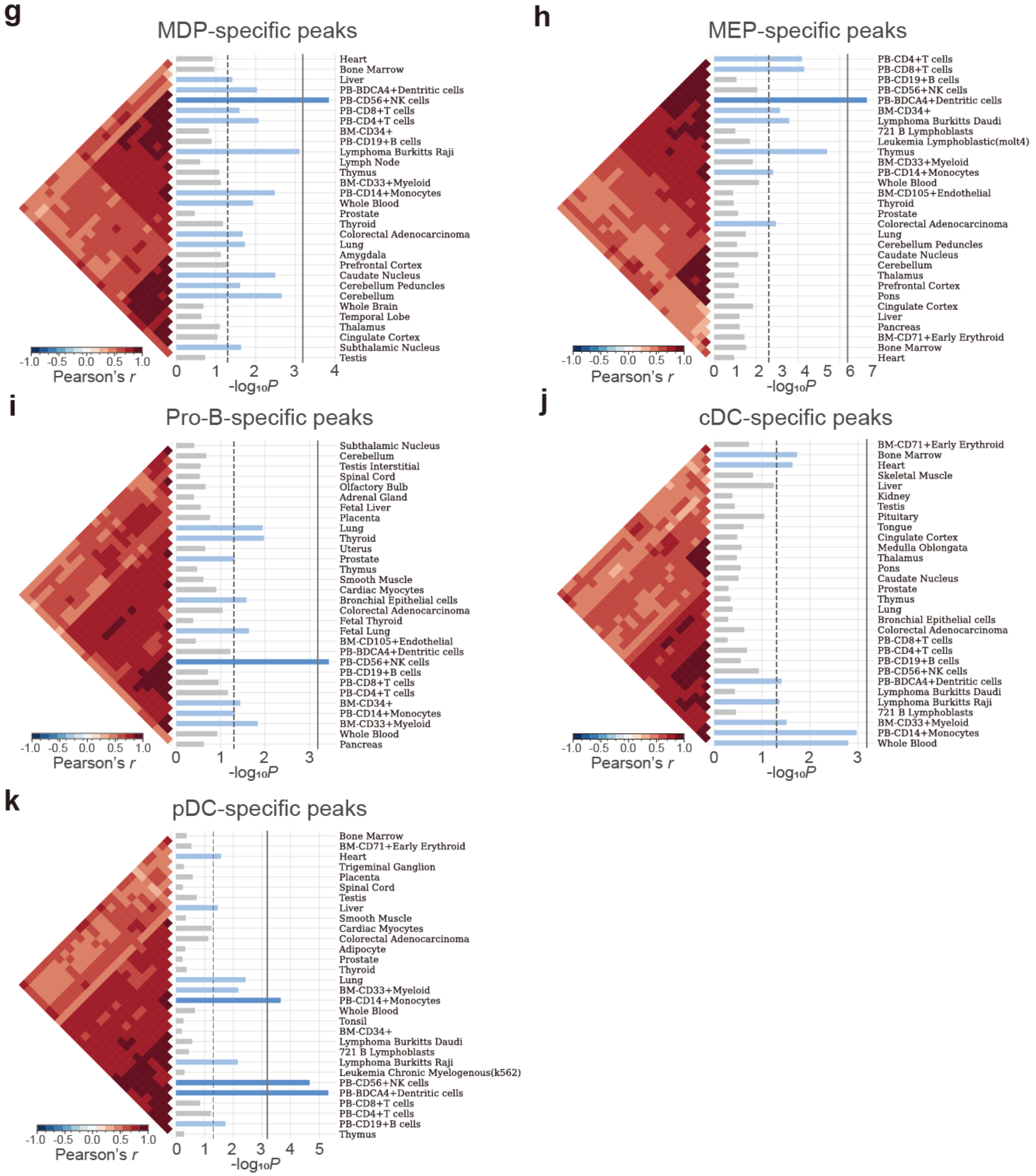
Supplementary Figure 56. Heritability enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Resting Droplet dataset. a-d, Heritability enrichments estimated by LDSC within cell type-specific peaks identified by CASTLE and the background peaks for dataset-related traits including HbA1c (**a**), Red blood cell distribution width (**b**), SHBG (**c**) and Total Bilirubin (**d**). The error bars denote jackknife standard errors over 200 equally sized blocks of adjacent SNPs about the estimates of enrichment, and the centers of error bars represent the average value.



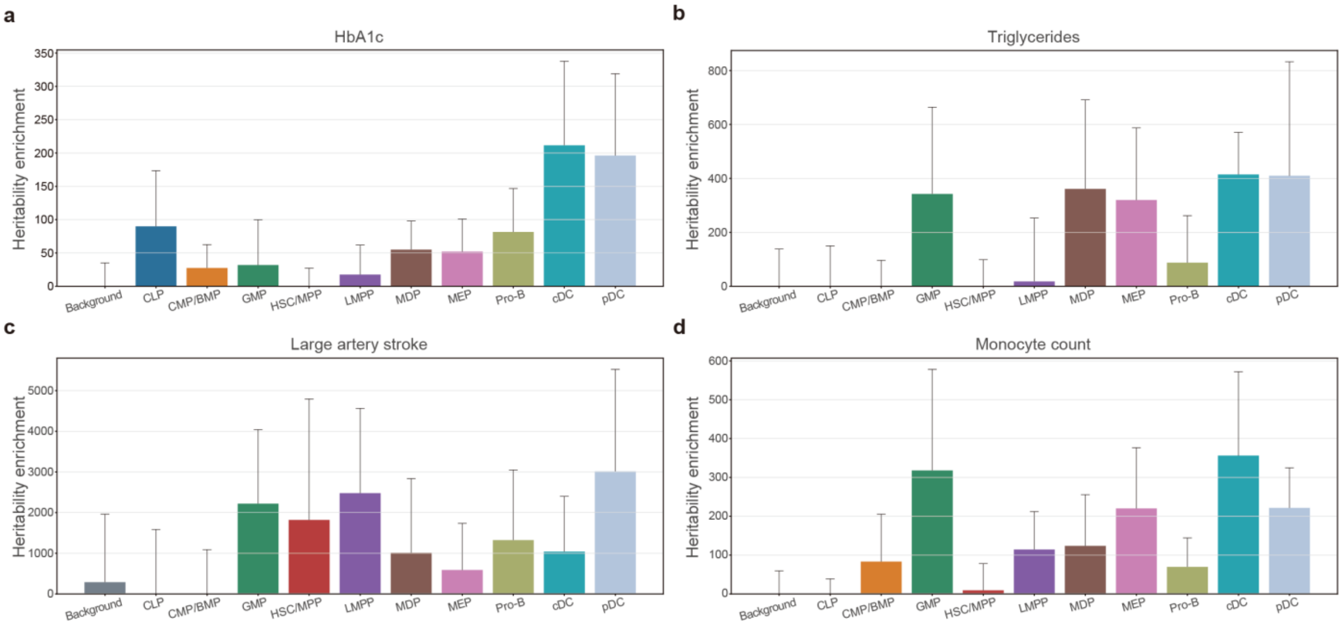
Supplementary Figure 57. SNP enrichment and heritability enrichment fold changes between the cell type-specific peaks and background peaks with different feature-to-peak ratios on the Immune dataset. a-b, SNP enrichment fold between the number of significantly enriched tissues (**a**) and significantly enriched dataset-related tissues (**b**) on the cell type-specific peaks versus background peaks with the feature-to-peak ratio of 25%, 50%, 75% and 100%. **c-f**, Fold changes between the heritability enrichments of cell type-specific peaks and background peaks for dataset-related traits including HbA1c (**c**), Triglycerides (**d**), Large artery stroke (**e**) and Monocyte count (**f**) with the feature-to-peak ratio of 25%, 50%, 75% and 100%. “inf” indicates infinity.



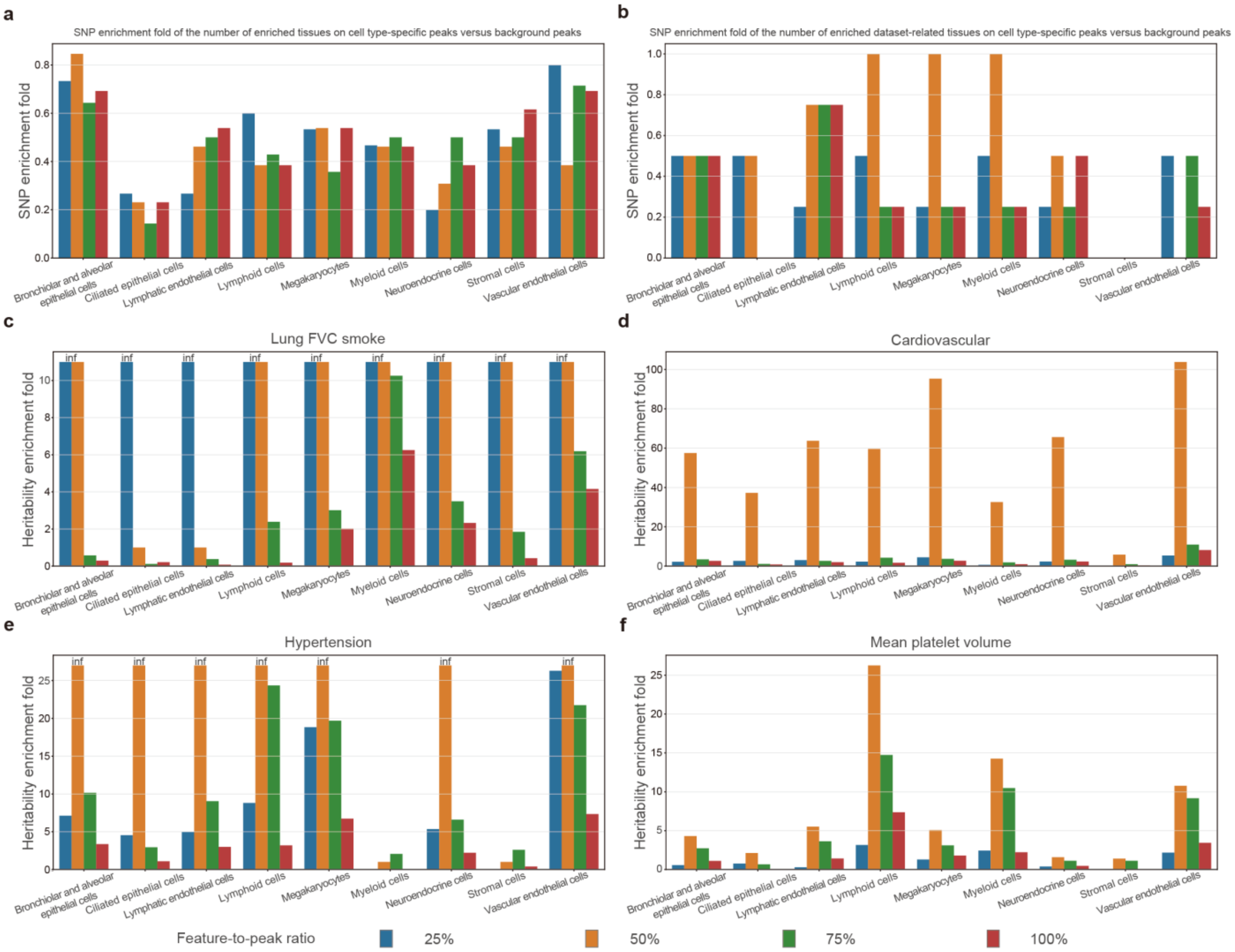
Supplementary Figure 58. SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Immune dataset. a-k, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), CLP-specific peaks (**b**), CMP/BMP-specific peaks (**c**), GMP-specific peaks (**d**), HSC/MPP-specific peaks (**e**), LMPP-specific peaks (**f**), MDP-specific peaks (**g**), MEP-specific peaks (**h**), Pro-B-specific peaks (**i**), cDC-specific peaks (**j**) and pDC-specific peaks (**k**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



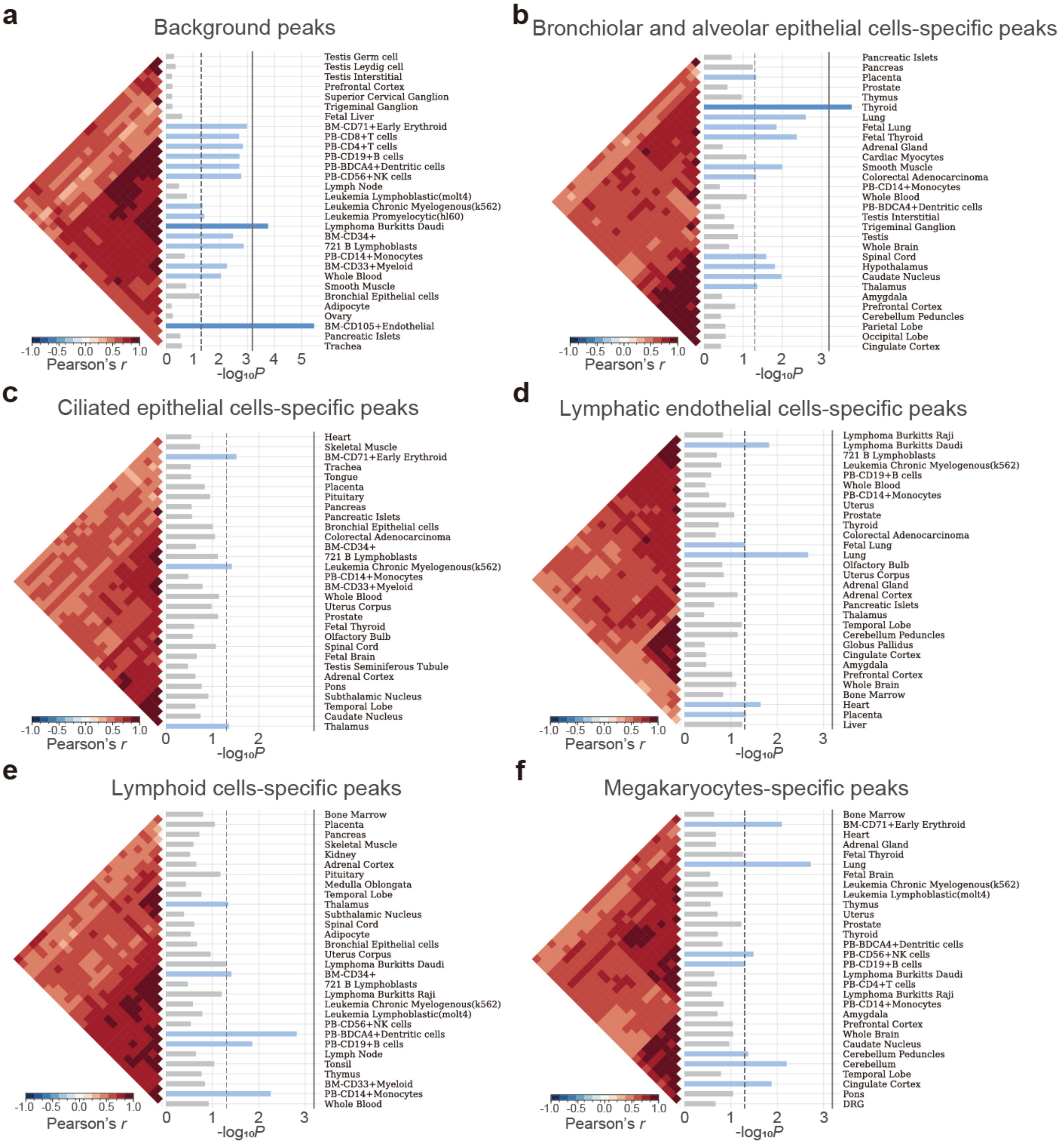
Supplementary Figure 58 (continue). SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Immune dataset. a-k, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), CLP-specific peaks (**b**), CMP/BMP-specific peaks (**c**), GMP-specific peaks (**d**), HSC/MPP-specific peaks (**e**), LMPP-specific peaks (**f**), MDP-specific peaks (**g**), MEP-specific peaks (**h**), Pro-B-specific peaks (**i**), cDC-specific peaks (**j**) and pDC-specific peaks (**k**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



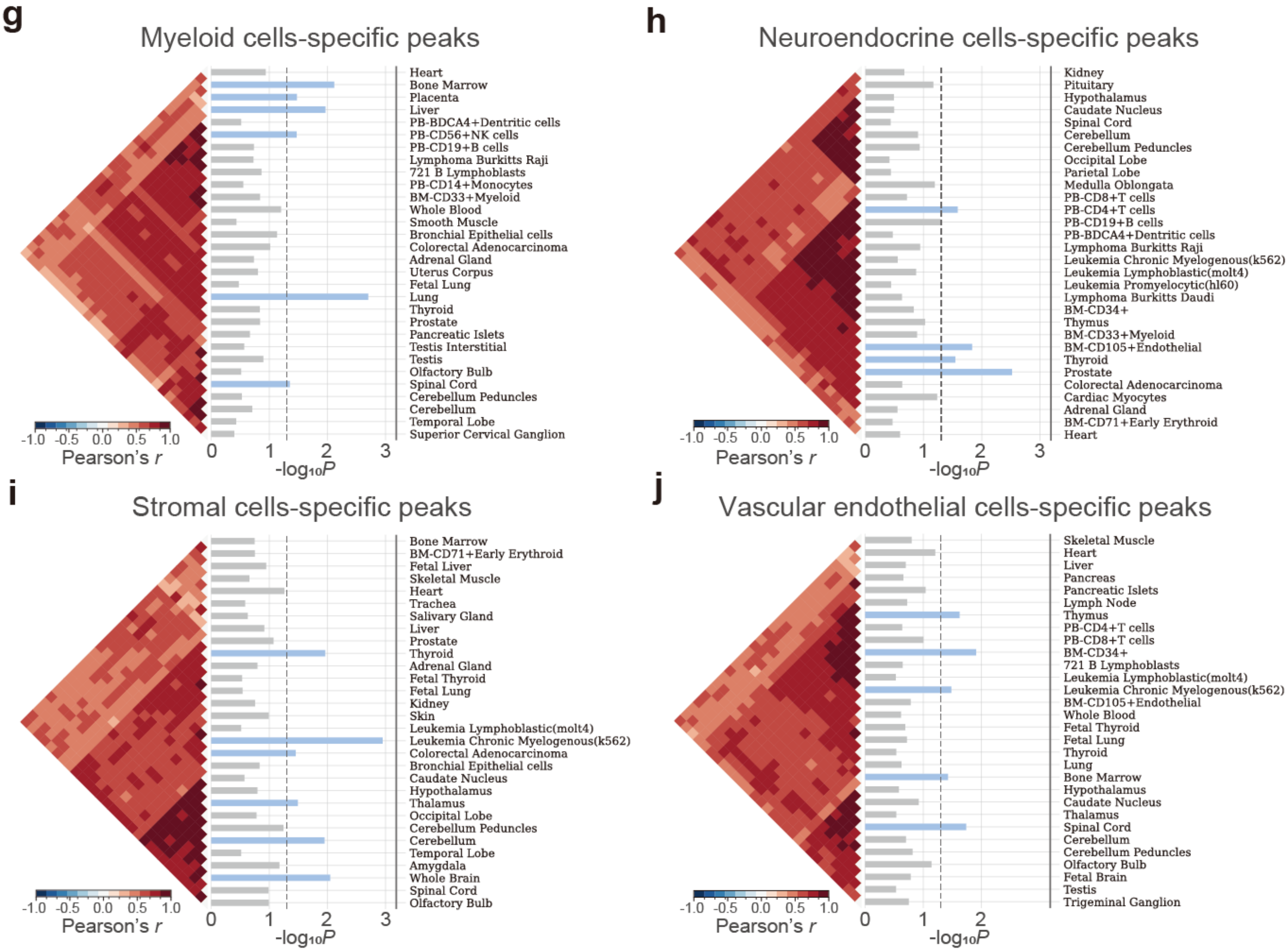
Supplementary Figure 59. Heritability enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Immune dataset. a-d, Heritability enrichments estimated by LDSC within cell type-specific peaks identified by CASTLE and the background peaks for dataset-related traits including HbA1c (**a**), Triglycerides (**b**), Large artery stroke (**c**) and Monocyte count (**d**). The error bars denote jackknife standard errors over 200 equally sized blocks of adjacent SNPs about the estimates of enrichment, and the centers of error bars represent the average value.



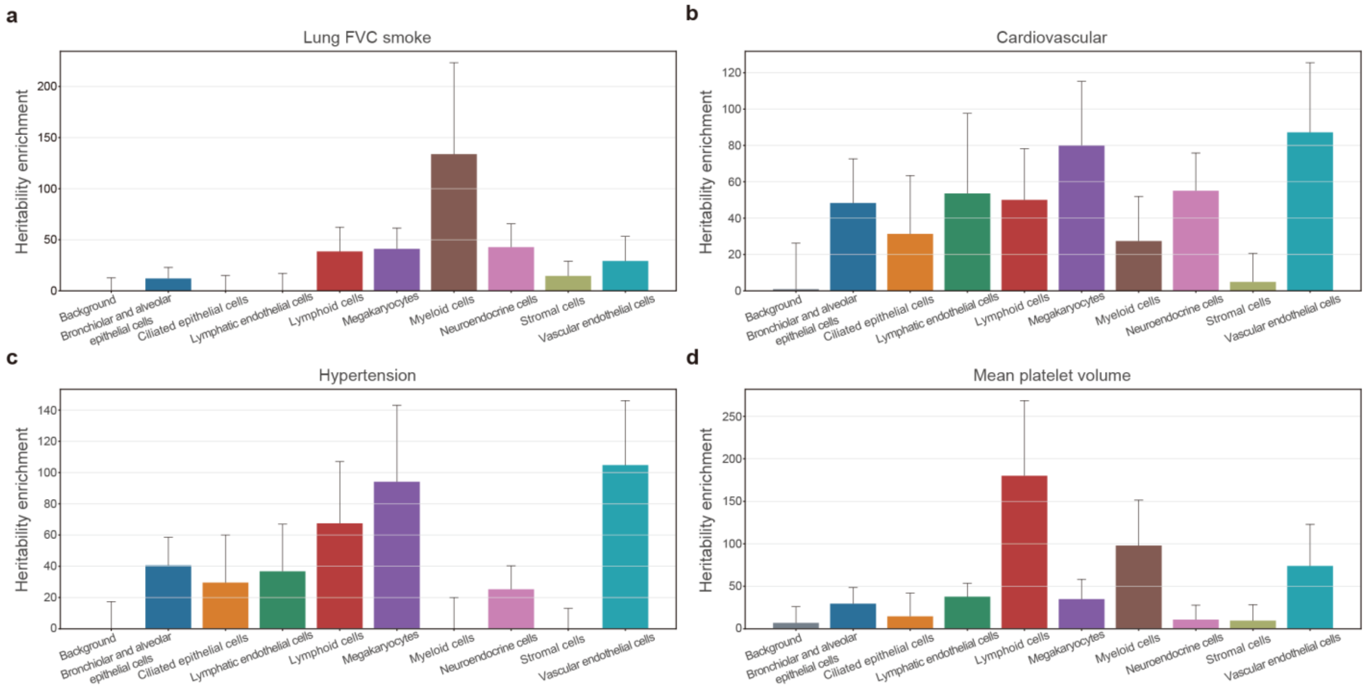
Supplementary Figure 60. SNP enrichment and heritability enrichment fold changes between the cell type-specific peaks and background peaks with different feature-to-peak ratios on the Fetal Lung dataset. a-b, SNP enrichment fold between the number of significantly enriched tissues (**a**) and significantly enriched dataset-related tissues (**b**) on the cell type-specific peaks versus background peaks with the feature-to-peak ratio of 25%, 50%, 75% and 100%. **c-f**, Fold changes between the heritability enrichments of cell type-specific peaks and background peaks for dataset-related traits including Lung FVC smoke (**c**), Cardiovascular (**d**), Hypertension (**e**) and Mean platelet volume (**f**) with the feature-to-peak ratio of 25%, 50%, 75% and 100%. “inf” indicates infinity.



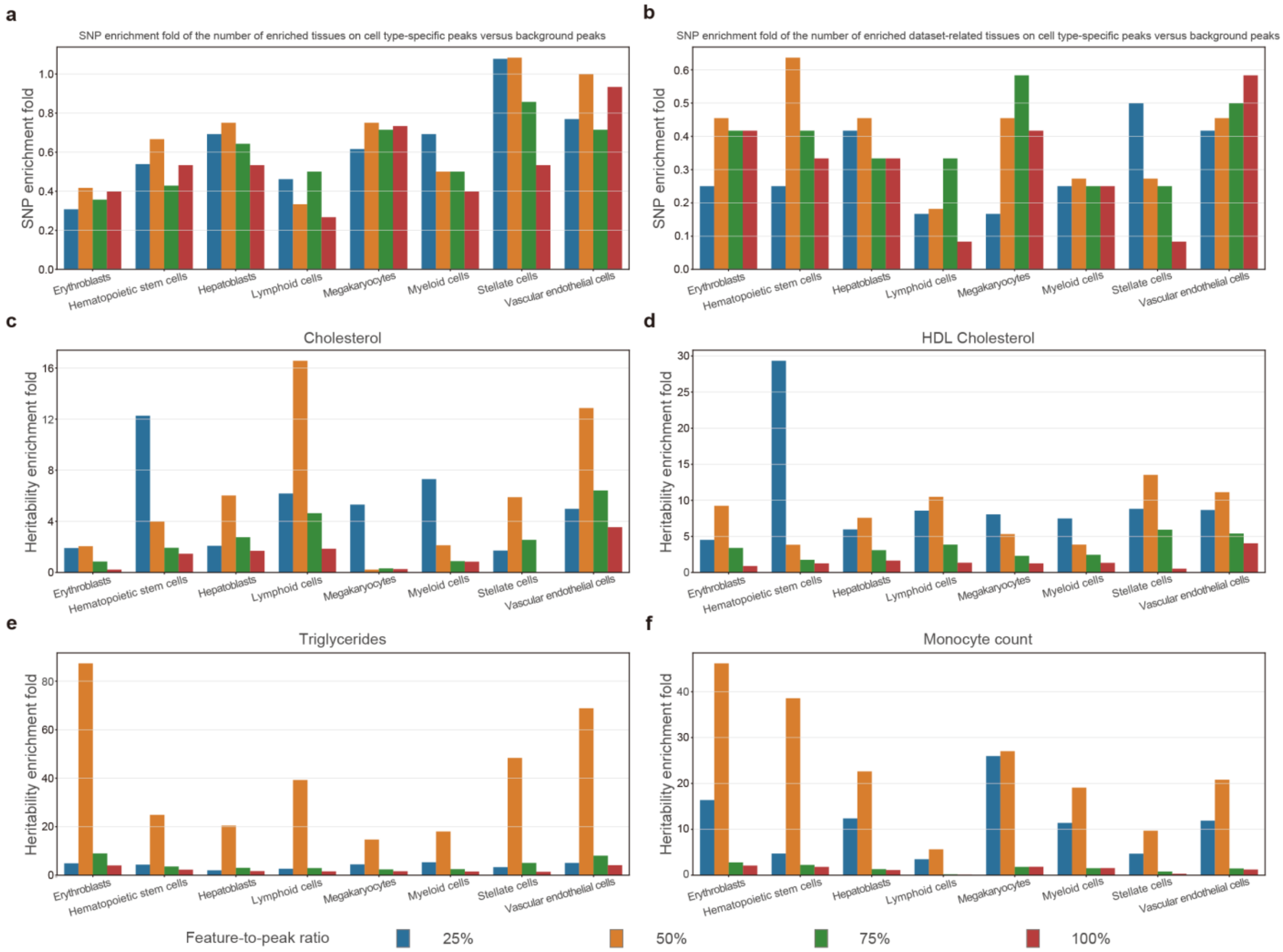
Supplementary Figure 61. SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Fetal Lung dataset. a-j, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), Bronchiolar and alveolar epithelial cells-specific peaks (**b**), Ciliated epithelial cells-specific peaks (**c**), Lymphatic endothelial cells-specific peaks (**d**), Lymphoid cells-specific peaks (**e**), Megakaryocytes-specific peaks (**f**), Myeloid cells-specific peaks (**g**), Neuroendocrine cells-specific peaks (**h**), Stromal cells-specific peaks (**i**) and Vascular endothelial cells-specific peaks (**j**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



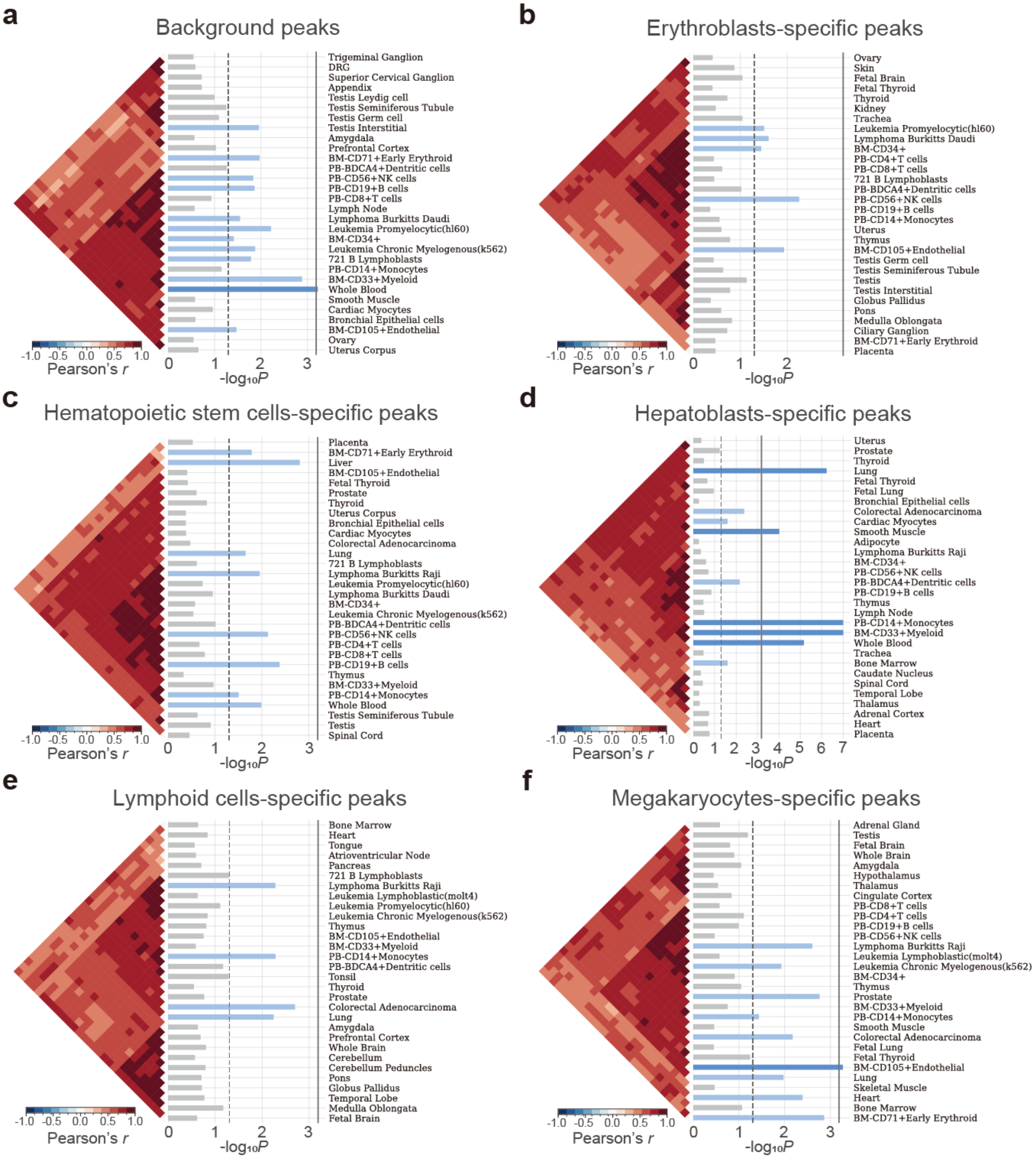
Supplementary Figure 61 (continue). SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Fetal Lung dataset. a-j, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), Bronchiolar and alveolar epithelial cells-specific peaks (**b**), Ciliated epithelial cells-specific peaks (**c**), Lymphatic endothelial cells-specific peaks (**d**), Lymphoid cells-specific peaks (**e**), Megakaryocytes-specific peaks (**f**), Myeloid cells-specific peaks (**g**), Neuroendocrine cells-specific peaks (**h**), Stromal cells-specific peaks (**i**) and Vascular endothelial cells-specific peaks (**j**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



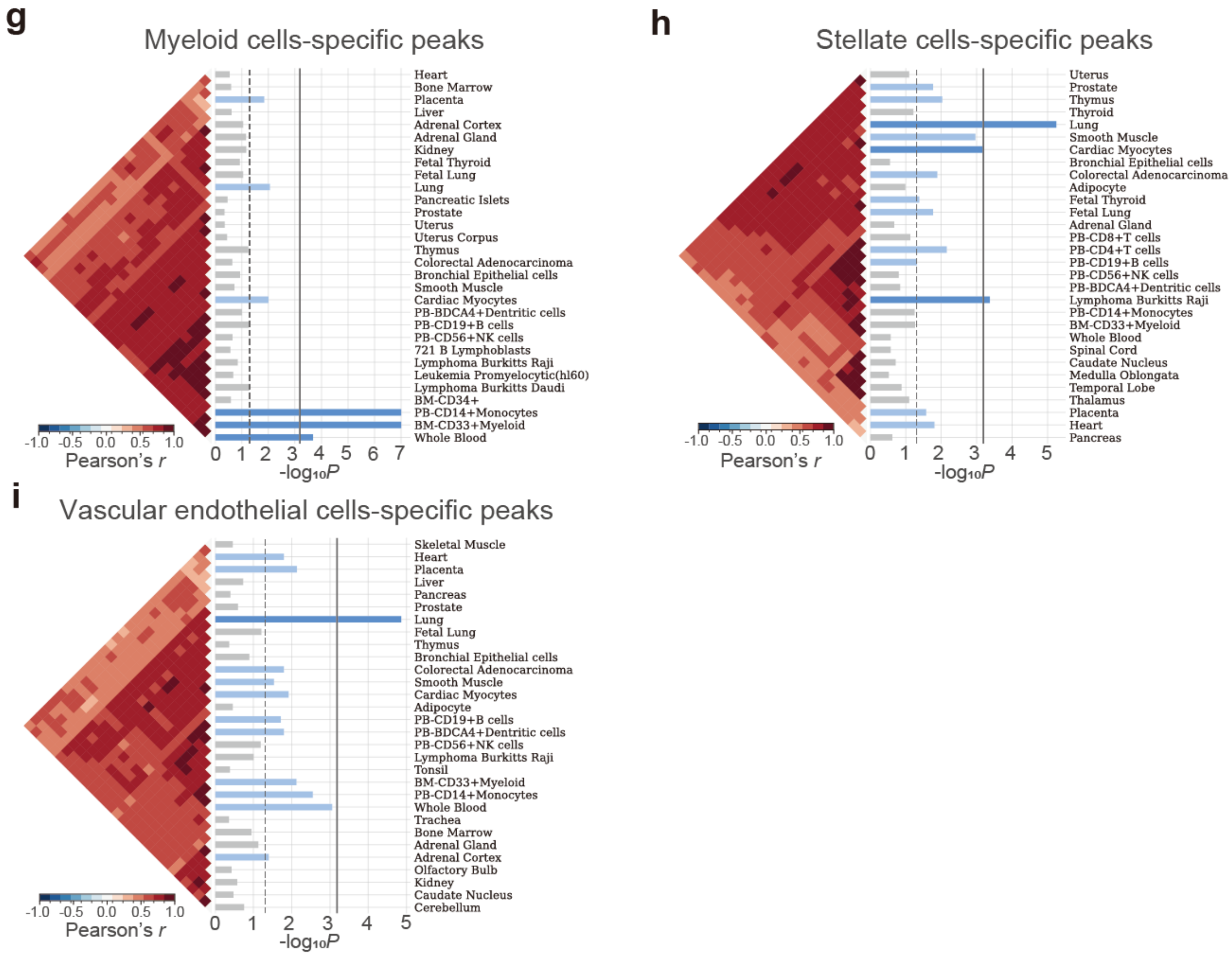
Supplementary Figure 62. Heritability enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Fetal Lung dataset. a-d, Heritability enrichments estimated by LDSC within cell type-specific peaks identified by CASTLE and the background peaks for dataset-related traits including Lung FVC smoke (**a**), Cardiovascular (**b**), Hypertension (**c**) and Mean platelet volume (**d**). The error bars denote jackknife standard errors over 200 equally sized blocks of adjacent SNPs about the estimates of enrichment, and the centers of error bars represent the average value.



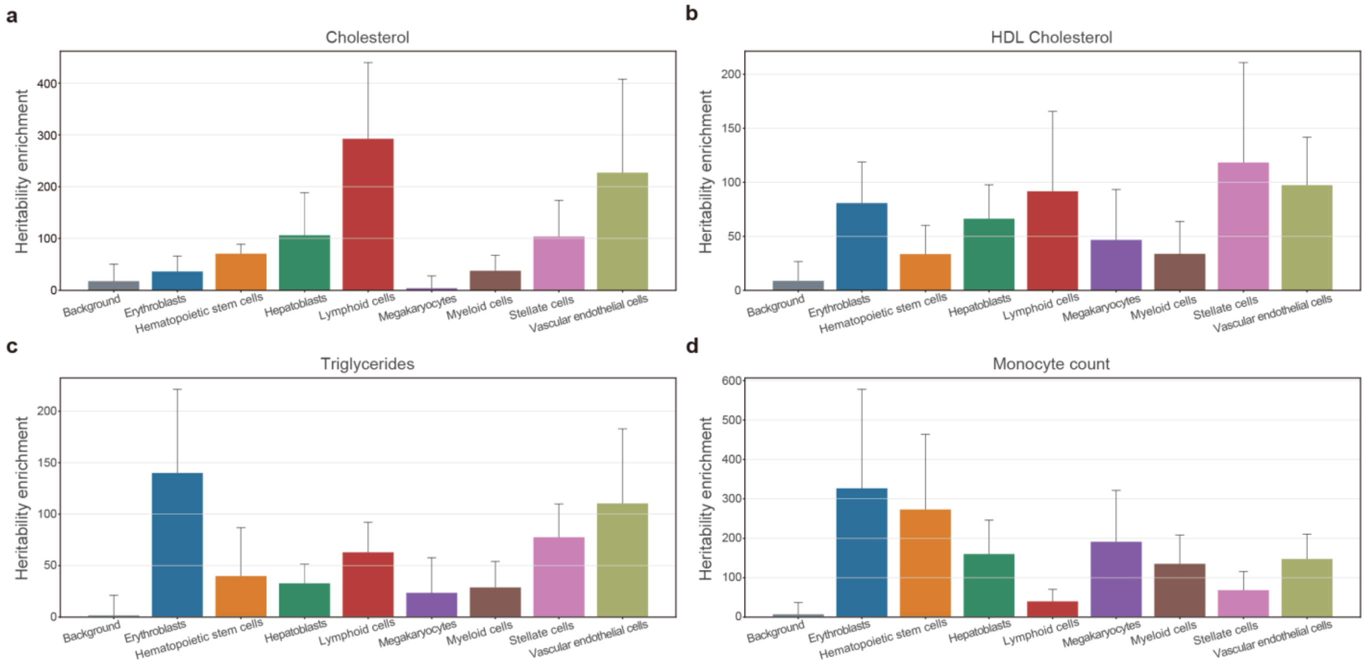
Supplementary Figure 63. SNP enrichment and heritability enrichment fold changes between the cell type-specific peaks and background peaks with different feature-to-peak ratios on the Fetal Liver dataset. a-b, SNP enrichment fold between the number of significantly enriched tissues (**a**) and significantly enriched dataset-related tissues (**b**) on the cell type-specific peaks versus background peaks with the feature-to-peak ratio of 25%, 50%, 75% and 100%. **c-f**, Fold changes between the heritability enrichments of cell type-specific peaks and background peaks for dataset-related traits including Cholesterol (**c**), HDL Cholesterol (**d**), Triglycerides (**e**) and Monocyte count (**f**) with the feature-to-peak ratio of 25%, 50%, 75% and 100%.



Supplementary Figure 64. SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Fetal Liver dataset. a-i, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), Erythroblasts-specific peaks (**b**), Hematopoietic stem cells-specific peaks (**c**), Hepatoblasts-specific peaks (**d**), Lymphoid cells-specific peaks (**e**), Megakaryocytes-specific peaks (**f**), Myeloid cells-specific peaks (**g**), Stellate cells-specific peaks (**h**) and Vascular endothelial cells-specific peaks (**i**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



Supplementary Figure 64 (continue). SNP enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Fetal Liver dataset. a-i, Top 30 significantly enriched tissues in SNPsea analysis on the background peaks (**a**), Erythroblasts-specific peaks (**b**), Hematopoietic stem cells-specific peaks (**c**), Hepatoblasts-specific peaks (**d**), Lymphoid cells-specific peaks (**e**), Megakaryocytes-specific peaks (**f**), Myeloid cells-specific peaks (**g**), Stellate cells-specific peaks (**h**) and Vascular endothelial cells-specific peaks (**i**) identified by CASTLE. The vertical dashed line represents the one-sided p-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided p-value with Bonferroni correction. Using hierarchical clustering with unweighted pair-group method with arithmetic means (UPGMA), the expression profiles were ordered. The Pearson correlation coefficients, denoting the correlation between expression profiles, were displayed in the heatmaps.



Supplementary Figure 65. Heritability enrichment analysis for cell type-specific peaks with the feature-to-peak ratio of 50% on the Fetal Liver dataset. a-d, Heritability enrichments estimated by LDSC within cell type-specific peaks identified by CASTLE and the background peaks for dataset-related traits including Cholesterol (**a**), HDL Cholesterol (**b**), Triglycerides (**c**) and Monocyte count (**d**). The error bars denote jackknife standard errors over 200 equally sized blocks of adjacent SNPs about the estimates of enrichment, and the centers of error bars represent the average value.

Supplementary Tables

Supplementary Table 1. Summary of datasets used in this study.

Dataset	Species	No. of cells	No. of peaks	No. of cell types	Sparsity (%)
InSilico[34]	Human	1377	68,069	6	96.849%
Stimulated Droplet[21]	Human	75,968	156,311	14	99.330%
Resting Droplet[21]	Human	60,495	156,311	14	99.458%
Fetal Lung[41]	Human	72,662	1,050,819	9	99.658%
Fetal Liver[41]	Human	183,175	1,050,819	8	99.823%
Immune[28]	Human	18,489	571,400	10	98.453%
Splenocyte[1]	Mouse	3166	77,453	12	83.386%
Bone Marrow A[42]	Mouse	4033	436,206	15	99.095%
Bone Marrow B[42]	Mouse	4370	436,206	18	99.207%
Lung A[42]	Mouse	5122	436,206	22	99.174%
Lung B[42]	Mouse	4874	436,206	25	99.174%
Whole Brain A[42]	Mouse	5494	436,206	21	98.537%
Whole Brain B[42]	Mouse	3272	436,206	20	98.325%
Cerebellum[42]	Mouse	2278	436,206	20	99.178%
Testes[42]	Mouse	2723	436,206	10	98.936%
Brain[31, 43]	Mouse	13,671	479,127	20	98.923%

Reference

1. Chen, X., Miragaia, R.J., Natarajan, K.N. & Teichmann, S.A. A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.* **9**, 5345 (2018).
2. Xiong, L. et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.* **10**, 4576 (2019).
3. Groeneveld, R.A. & Meeden, G. Measuring skewness and kurtosis. *J. Roy. Stat. Soc. Ser. D. (Statistician)* **33**, 391-399 (1984).
4. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193-218 (1985).
5. Romano, S., Vinh, N.X., Bailey, J. & Verspoor, K. Adjusting for chance clustering comparison measures. *J. Mach. Learn. Res.* **17**, 4635-4666 (2016).
6. Strehl, A. & Ghosh, J. Cluster ensembles---a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583-617 (2002).
7. Fowlkes, E.B. & Mallows, C.L. A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.* **78**, 553-569 (1983).
8. Ding, J., Condon, A. & Shah, S.P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **9**, 2002 (2018).
9. Chen, X. et al. Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. *Nat. Mach. Intell.* **4**, 116-126 (2022).
10. Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 1-19 (2019).
11. Bravo González-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397-400 (2019).
12. Yuan, H. & Kelley, D.R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat. Methods* **19**, 1088-1096 (2022).
13. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 1-13 (2020).
14. Bekkar, M., Djemaa, H.K. & Alitouche, T.A. Evaluation measures for models assessment over imbalanced data sets. *J. Inf. Eng. Appl.* **3** (2013).
15. Johnson, J.M. & Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **6**, 1-54 (2019).
16. Luecken, M.D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41-50 (2022).
17. Chen, S., Wang, R., Long, W. & Jiang, R. ASTER: accurately estimating the number of cell types in single-cell chromatin accessibility data. *Bioinformatics* **39**, btac842 (2023).
18. Kobayashi, H., Cheveralls, K.C., Leonetti, M.D. & Royer, L.A. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nat. Methods* **19**, 995-1003 (2022).
19. Razavi, A., Van den Oord, A. & Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. In *Proc. Advances in Neural Information Processing Systems* **32** (2019).
20. Van Den Oord, A. & Vinyals, O. Neural discrete representation learning. In *Proc. Advances in Neural Information Processing Systems* **30** (2017).
21. Lareau, C.A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916-924 (2019).
22. Li, Z., Chen, X., Zhang, X., Jiang, R. & Chen, S. Latent feature extraction with a prior-based self-attention framework for spatial transcriptomics. *Genome Res.* **33**, 1757-1773 (2023).
23. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometrics Intellig. Lab. Syst.* **2**, 37-52 (1987).
24. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math.* **20**, 53-65 (1987).
25. Danese, A. et al. EpiScanpy: integrated single-cell epigenomic analysis. *Nat. Commun.* **12**, 5228 (2021).
26. Finucane, H.K. et al. Partitioning heritability by functional annotation using genome-wide association summary

statistics. *Nat. Genet.* **47**, 1228-1235 (2015).

27. Slowikowski, K., Hu, X. & Raychaudhuri, S. SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* **30**, 2496-2497 (2014).

28. Satpathy, A.T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925-936 (2019).

29. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547-554 (2019).

30. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 1-16 (2018).

31. Xiong, L. et al. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nat. Commun.* **13**, 6118 (2022).

32. Zamanighomi, M. et al. Unsupervised clustering and epigenetic classification of single cells. *Nat. Commun.* **9**, 2410 (2018).

33. Schep, A.N., Wu, B., Buenrostro, J.D. & Greenleaf, W.J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975-978 (2017).

34. Buenrostro, J.D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).

35. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44**, D110-D115 (2016).

36. Saretzki, G., Feng, J., Zglinicki, T.v. & Villeponteau, B. Similar gene expression pattern in senescent and hyperoxic-treated fibroblasts. *J. Gerontol. A Biol. Sci. Med. Sci.* **53**, B438-B442 (1998).

37. Gerstein, M.B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).

38. Smith, L.T., Hohaus, S., Gonzalez, D.A., Dziennis, S.E. & Tenen, D.G. PU. 1 (Spi-1) and C/EBP alpha regulate the granulocyte colony-stimulating factor receptor promoter in myeloid cells. (1996).

39. Cusanovich, D.A. et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).

40. Wang, J. et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798-1812 (2012).

41. Domcke, S. et al. A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).

42. Cusanovich, D.A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309-1324. e1318 (2018).

43. Fang, R. et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).