

METHODS

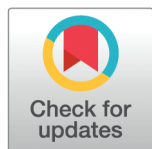
M-band wavelet-based multi-view clustering of cells

Tong Liu¹, Zihuan Liu², Wenke Sun³, Adeethya Shankar⁴, Yongzhong Zhao⁵, Xiaodi Wang⁶*

1 Department of Mathematical Sciences, Tsinghua University, Beijing, China, **2** Data and Statistical Science, AbbVie, Chicago, Illinois, United States of America, **3** School of Economics and Management, Dalian University of Technology, Dalian, China, **4** Brown University, Providence, Rhode Island, United States of America, **5** Frontage Labs, Exton, Pennsylvania, United States of America, **6** Department of Mathematics, Western Connecticut State University, Danbury, Connecticut, United States of America

* These authors contributed equally to this work are co-first authors.

* zhaoy@iscb.org (YZ); wangx@wcsu.edu (YW)



Abstract

Wavelet analysis has been recognized as a widely used and promising tool in the fields of signal processing and data analysis. However, the application of wavelet-based method in single-cell RNA sequencing (scRNA-seq) data is little known. Here, we present M-band wavelet-based scRNA-seq multi-view clustering of cells (WMC). We applied for integration of M-band wavelet analysis and uniform manifold approximation and projection (UMAP) to a panel of single cell sequencing datasets by breaking up the data matrix into an approximation or low resolution component and $M-1$ detail or high resolution components. Our method is armed with multi-view clustering of cell types, identity, and functional states, enabling missing cell types visualization and new cell types discovery. Distinct to standard scRNA-seq workflow, our wavelet-based approach is a new addition to uncover rare cell types with a fine resolution.

OPEN ACCESS

Citation: Liu T, Liu Z, Sun W, Shankar A, Zhao Y, Wang X (2025) M-band wavelet-based multi-view clustering of cells. PLoS Comput Biol 21(5): e1013060.
<https://doi.org/10.1371/journal.pcbi.1013060>

Editor: Wei Li, University of Maryland School of Medicine, UNITED STATES OF AMERICA

Received: September 9, 2024

Accepted: April 18, 2025

Published: May 23, 2025

Copyright: © 2025 Liu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The breast cancer datasets, including CID-3921, CID-4495, CID-4523, and CID-4463, are available from the Gene Expression Omnibus (GEO) with accession numbers GSM5354515, GSM5354530, GSM5354536, and GSM5354527, respectively. The colon cancer dataset is available from GEO with accession number GSM4143678. The Peripheral Blood Mononuclear Cells dataset is accessible at: https://satijalab.org/seurat/articles/pbm3k_tutorial.html. The FPKM dataset is embedded in

Author summary

We develop M-band wavelet-based multi-view clustering method of cells. Our new approach integrates M-band wavelet analysis and UMAP to a panel of single cell sequencing datasets via breaking up the data matrix into an approximation or low resolution component and $M-1$ detail or high resolution components. Our method enables us to examine multi-view clustering of cell types, identity, and functional states, potentializing missing cell types recovery, rare cell types discovery, as well as functional cell states exploration.

Introduction

Recent breakthroughs in methodology are enabling the study of the transcriptomes of individual cells, which paves the way for more objective investigations of cellular functions at

the SC3 R package. The ILC dataset is available at: <https://doi.org/10.6084/m9.figshare.27190692.v1>.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

single-cell level [1–3]. A primary objective of single-cell RNA sequencing (scRNA-seq) analysis is to identify and discover cell types [4–7], identities [8], states [6,9], as well as accompanied gene signatures [10,11], while conventional pipelines have been unsatisfying due to limited resolution even with missed critical cell types, identity, and functional states. Earlier studies based on microscopy [12], histology [13], and pathological criteria [14] have contributed to the resolution of this problem. Recent efforts have utilized both unsupervised and supervised clustering techniques to cluster cells alongside transcription signatures [5,7] yet with big room to be optimized, requesting novel approaches.

Those limitations in the existing methods have motivated us to develop a new method, aiming to recover missing cell clusters and uncover rare cell types. Wavelet frameworks have been successfully applied to numerous tasks in the biomedical domain, such as genomics [15], promoters [16,17] and GWAS [18]. However, wavelet-based method in scRNA-seq data analysis has been largely unexplored.

Our hybrid technique incorporated with M-band orthogonal wavelet enables us to illustrate multi-view clustering of cell types, identities, and functional states from a variety of perspectives simultaneously. Distinct from the regular RNA-Seq analysis pipelines, our strategy have the power to uncover missed and rare cell types, identities, states, via wavelet-based multi-view of clustering (WMC). By taking a weighted-average like process on the original RNA-Seq matrix, The trend component of the M-band wavelet enables us to take an approximation of the original data, while the detail components provides the detailed information.

Results

The overview of WMC

We aim to build a pipeline with multi-view of clustering on scRNA-seq data. Stemming from the original scRNA-seq workflow, our WMC integrates wavelet-based multi-view clustering into standard pipelines. In the light of wavelet analysis, we develop WMC to present multi-views of scRNA-seq in different resolution windows, potentializing recovery of missing cell types, cell states, cell identities, as well as rare cell types discovery. Our WMC consists of the following steps, including quality control, log-transformation, discrete wavelet transform (DWT), dimension reduction, multi-view clustering visualization, along with an assessment of intersecting components.

Given a raw scRNA-seq data matrix (Fig 1A), we remove the low-quality cells (Fig 1B) to ensure that technical effect does not distort downstream analysis results. Next, we apply a logarithmic transformation to the raw data matrix (Fig 1C), followed by M-band DWT, which decomposes the normalized data matrix into an approximation component and $M-1$ detail components (Fig 1D). This decomposition enables the extraction of multi-scale components, facilitating multi-view clustering of cell types, identities, and functional states from multiple perspectives. To further reduce dimensionality, we apply uniform manifold approximation and projection (UMAP) to both wavelet-transformed and non-transformed matrices for exploratory visualization (Fig 1F). Clustering is performed using the default method in Seurat [19]. In particular, on PCA-reduced matrix, we implement Louvain community detection algorithms [20] on single-cell K-Nearest Neighbor (KNN) graph to organize cells into clusters. To characterize the similarities and differences between the classic and our proposed method, we compare the clusters based on both original and wavelet-transformed multi-view matrices (Fig 1G). Of note, we use the single-cell Cluster-based automatic Annotation Toolkit for Cellular Heterogeneity (scCATCH) [21] for cluster annotation. In addition to the Seurat pipeline, we incorporate alternative clustering methods, including Single-Cell Consensus Clustering (SC3) and Hierarchical Graph-based

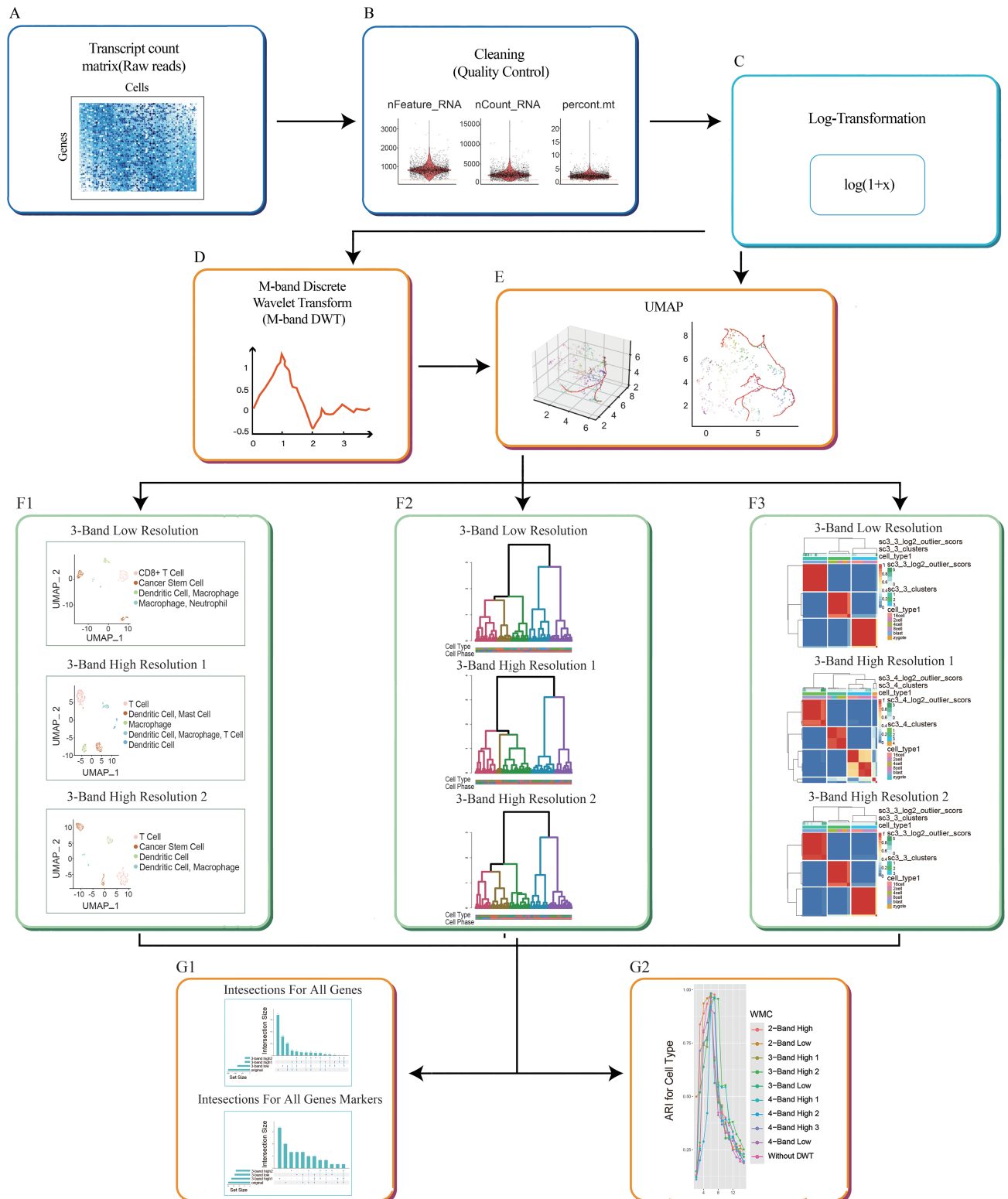


Fig 1. Illustration of the WMC workflow. (A) Raw scRNA-seq data matrix with row as transcripts and columns as individual cells. (B) Quality control diagrams, demonstrating the process of removing unqualified cells and transcripts involving mitochondrial genes. (C) Logarithm transformation of the raw matrix. (D) Applying M-band DWT on the logarithm transformed matrix. (E) PCA-based dimension reduction on matrices with and without DWT. (F) Visualization of clustering based on different methods, with (F1) for UMAP in Seurat, (F2) for SC3 and (F3) for HGC. (G) Assessing the performance of WMC via intersection analysis and ARI computation.

<https://doi.org/10.1371/journal.pcbi.1013060.g001>

Clustering (HGC), to wavelet-transformed matrices, demonstrating the compatibility of WMC with multiple clustering frameworks. To assess the performance of WMC multi-view clustering, we evaluate the overlap of marker genes identified by WMC and conventional methods without DWT. Additionally, the adjusted rand index (ARI) and normalized mutual information (NMI) are computed, resulting in similar cluster consistency quality across distinct approaches. Yet WMC enables more clusters revealed. Thus, WMC has a multi-view function without compromise of cluster quality.

Mathematical statement of WMC

Let $S = [s_1, \dots, s_p] \in \mathbb{R}^{n \times p}$ denotes the original data matrix and $W = [w_1, \dots, w_n]^T \in \mathbb{R}^{n \times n}$ be the corresponding M-Band DWT matrix [22,23]. Then, the M-Band DWT of s_i for $i = 1, \dots, p$ is given by

$$Ws_i = [a(i), d_1(i), d_2(i), \dots, d_{M-1}(i)]^T \quad (1)$$

where $n = Mk$, $a(i) = [a_{i1}, a_{i2}, \dots, a_{ik}]^T$ and $d_j(i) = [d_{j,i,1}, d_{j,i,2}, \dots, d_{j,i,k}]^T$ for $j = 1, 2, \dots, M-1$. For brevity, we denote Ws_i by $\tilde{s}_i$. Let $\tilde{S} = [\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_p]$, then

$$WS = \tilde{S} = [a, d_1, \dots, d_{M-1}]^T \quad (2)$$

where $a = [a(1), \dots, a(p)]$ being the low resolution component (or trend) of S , and $d_1 = [d_1(1), \dots, d_1(p)]$, $\dots$, $d_{M-1} = [d_{M-1}(1), \dots, d_{M-1}(p)]$ being the high resolution components (or fluctuations) of S in wavelet domain. Since W is an orthogonal matrix, $\{w_1, w_2, \dots, w_n\}$ forms an orthonormal basis of $\mathbb{R}^n$. The components of $\tilde{s}_i$ are also called the wavelet coefficients of s_i . Therefore, the components of $\tilde{s}_i$ are coordinates of s_i under this wavelet basis. Due to the orthonormality of W , for $i = 1, 2, \dots, p$, we have $\|\tilde{s}_i\| = \|s_i\|$ and $W^T = W^{-1}$. Hence the M-Band DWT preserve the length or energy of vectors it transforms. If we multiply both sides of (2) by W^T , we obtain

$$S = W^T \tilde{S} = A + D_1 + \dots + D_{M-1} \quad (3)$$

and

$$\begin{aligned} \|s_i\|^2 &= \|A_i\|^2 + \|D_{1,i}\|^2 + \dots + \|D_{M-1,i}\|^2 \\ &= \|a(i)\|^2 + \|d_1(i)\|^2 + \dots + \|d_{M-1}(i)\|^2, \end{aligned} \quad (4)$$

where

$$\begin{aligned} A &= [A_1, A_2, \dots, A_p] = [w_1, w_2, \dots, w_k]a, \\ D_j &= [D_{j1}, D_{j2}, \dots, D_{jp}] = [w_{kj+1}, w_{kj+2}, \dots, w_{kj+k}]d_j, j = 1, 2, \dots, M-1. \end{aligned} \quad (5)$$

In other words, we use M-band DWT to decompose S into the sum of M orthogonal components (multi-view windows), where one low-resolution component represents the approximation part, while the remaining $M-1$ components, which are of high resolution, correspond to the detail parts, as shown in (3)

We next aim to find A_i and D_j for $i = 1, \dots, p$ and $j = 1, 2, \dots, M-1$. Define $\mathcal{V} = \text{span}\{w_1, w_2, \dots, w_k\}$, $\mathcal{W}_i = \text{span}\{w_{ki+1}, w_{ki+2}, \dots, w_{ki+k}\}$, $i = 1, 2, \dots, M-1$, then $\mathcal{V}, \mathcal{W}_1, \dots$, and $\mathcal{W}_{M-1}$ are orthogonal subspaces of $\mathbb{R}^n$. It follows that,

$$\mathcal{V} \oplus \mathcal{W}_1 \oplus \dots \oplus \mathcal{W}_{M-1} = \mathbb{R}^n \quad (6)$$

Let $\mathcal{P}_Y \mathbf{x}$ be a projection of $\mathbf{x}$ on Y , we have

$$\begin{aligned} A_i &= \mathcal{P}_V \mathbf{s}_i \\ D_j &= \mathcal{P}_{W_j} \mathbf{s}_i \end{aligned} \quad (7)$$

We then apply different clustering methods on S and its corresponding components $A, D_1, \dots, D_{M-1}$ to obtain their multiview images as $\mathcal{I}(S), \mathcal{I}(A_1), \mathcal{I}(D_1), \dots, \mathcal{I}(D_{M-1})$, resulting in multiview clusters with principal cell names, encompassing both approximation and detail parts of S . Thus, the ensemble of annotated gene clusters enables us a multi-views cell types, identities, and states with sc-RNA seq data.

We assess the performance of WMC via computing the intersection of $\mathcal{I}(S), \mathcal{I}(A_1), \mathcal{I}(D_1), \dots, \mathcal{I}(D_{M-1})$. Since the space of different components are orthogonal to each other, few genes could be found in the intersection of different components theoretically. Therefore, the smaller number of genes appearing in the intersection of these components yields a better resolution of clustering.

Cluster analysis of the scRNA-Seq data set by Wavelet-Seurat routine

We apply Wavelet-Seurat to a benchmark scRNA-seq dataset profiling 2,613 peripheral blood mononuclear cells (PBMCs) from $10 \times$ Genomics. Specifically, WMC identifies 8, 10, and 13 cell types under 2-band, 3-band, and 4-band UMAP settings, respectively, whereas only six cell types are identified using the standard method without DWT (Fig A in [S1 Text](#)). In addition to the six cell types which have been identified by the regular method, WMC discovers novel clusters, including natural killer (NK) cells with 2-band and 3-band DWT, non-switched memory B cells with 3-band and 4-band DWT, as well as regulatory T cells, memory T cells, and SLC16A7+ cells with 4-band DWT. Furthermore, additional unknown cell types are identified from the detailed components in 3-band and 4-band DWT.

To further evaluate our Wavelet-Seurat approach, we have analyzed a published scRNA-seq dataset from a colorectal cancer patient. A similar trend is observed, as Fig B in [S1 Text](#) illustrates that the standard UMAP setting identifies six cell types, while 2-band, 3-band, and 4-band UMAP settings reveal 9, 12, and 12 cell types, respectively. Beyond the cell clusters observed in both the conventional and WMC methods, our approach identifies additional clusters, including mitotic fetal germ cells, astrocytes, monocytes, fibroblasts, plasma cells, and unknown cell types.

Moreover, we visualize the cell-type representations learned by Wavelet-Seurat in a three-dimensional UMAP space using the innate lymphoid cell (ILC) differentiation dataset (Figs 2 and G in [S1 Text](#)) [24]. After quality control, this dataset contains 2,544 cells spanning three developmental phases. Using the standard Seurat framework, we identify clusters corresponding to ILC2P, ILCP, NKP, cEILP, and sEILP families, along with six types of alpha-LP cells. However, applying wavelet transform further resolves alpha-LP cells into 10 distinct subtypes, revealing additional clusters such as alphaLP1.7, alphaLP1.8, alphaLP1.9, and alphaLP1.10, resulting in a total of 21 distinct cell clusters. Overall, wavelet transformation provides a multi-view perspective by projecting the dataset into different component spaces, enhancing the resolution of cell-type identification and enabling finer-grained clustering.

Cluster analysis using Wavelet-SC3 routine

In addition to the Wavelet-Seurat analysis, WMC is also performed on ILC differentiation cells dataset using the SC3 clustering method [25] (Figs 3 and H in [S1 Text](#)). As shown in Fig 3, different wavelet components yield distinct hierarchical structures, with greater

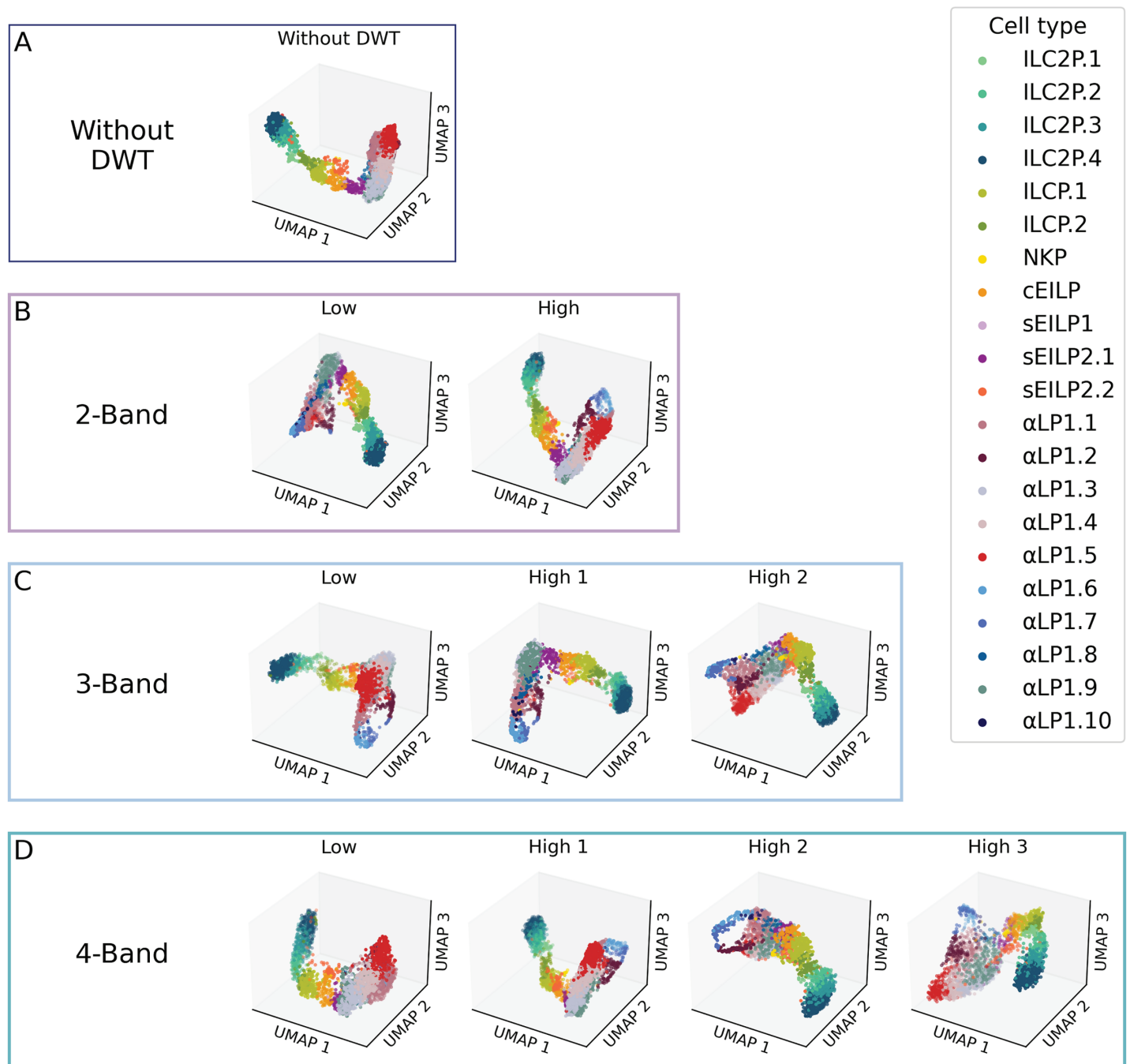


Fig 2. Multi-view of clusters of the innate lymphoid cell differentiation dataset. (A) UMAP visualization of cell types based on data matrix without DWT. (B)-(D) are clusters under wavelet analysis, with (B) for 2-band DWT, (C) for 3-band DWT, and (D) for 4-band DWT.

<https://doi.org/10.1371/journal.pcbi.1013060.g002>

intra-cluster similarity observed in wavelet-transformed data compared to clustering without DWT.

To investigate cell-type-level resolution, we have employed Wavelet-SC3 to partition the dataset into 21 clusters per component, as determined by SC3 without DWT. As illustrated

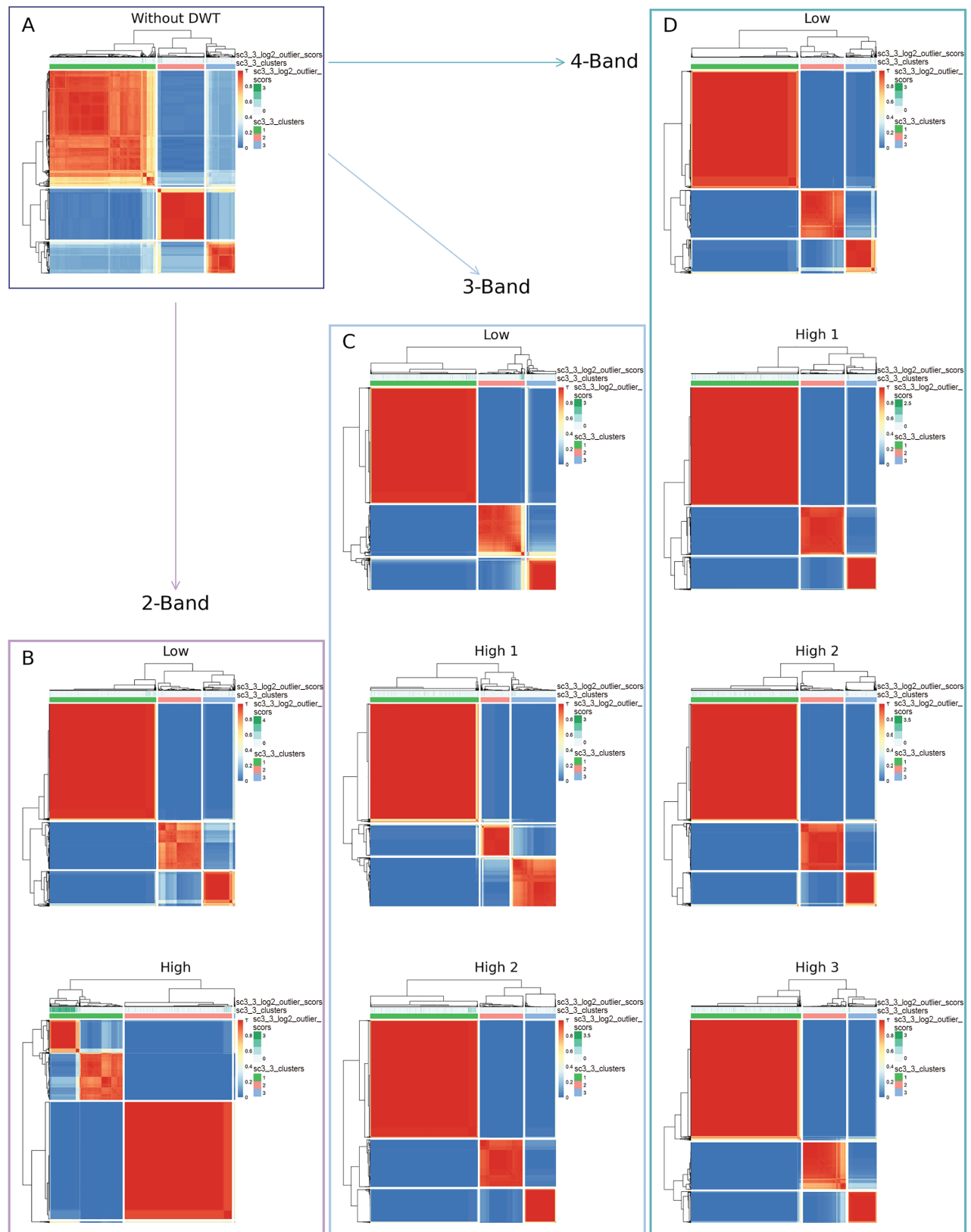


Fig 3. Consensus matrices among different cell phases of ILC differentiation dataset by Wavelet-SC3 method. (A) Consensus matrix using SC3 without DWT. (B)-(D) are consensus matrices under Wavelet-SC3, with (B) for 2-band DWT, (C) for 3-band DWT, and (D) for 4-band DWT.

<https://doi.org/10.1371/journal.pcbi.1013060.g003>

in Fig H in [S1 Text](#), the approximation component under multi-band DWT closely mirrors the clustering pattern of the original dataset, whereas the detail components reveal additional, distinct cluster structures.

We further apply Wavelet-SC3 to the PBMC-3k dataset (Fig I in [S1 Text](#)). This dataset exhibits a complex hierarchical structure, with SC3 without DWT identifying 14 clusters. To evaluate whether Wavelet-SC3 improves clustering resolution compared to the original SC3 method, we perform clustering while fixing the number of clusters at 14, applying SC3 with and without wavelet transformation. Indeed, as shown in Fig I in [S1 Text](#), it appears that the clustering pattern under 2-band DWT closely resembles the original SC3 clustering. However, with multi-band DWT, the approximation components enable more compact clusters captured with higher correlation, while the detail components are of additional, distinct cluster structures. Thus, Wavelet-SC3 has the capacity of enhancing better clustering resolution and uncovering finer-scale biological heterogeneity.

Hierarchical analysis using Wavelet-HGC routine

To explore the hierarchical structure of clusters in scRNA-seq data, we have applied WMC in combination with the HGC method to the ILC differentiation dataset ([Fig 4](#)) [24]. For cell cycle phases, we observe that 3-band DWT provides the most effective partitioning for G1-phase cells (marked in red), while detail components from 2-band and 4-band DWT successfully distinguish G2M-phase cells. In cell type clustering, the approximation component from 2-band wavelet-transformed data closely resembles the clustering structure obtained without DWT, whereas the detail components introduce alternative hierarchical structures, revealing additional biological insights.

Moreover, a similar pattern can be observed in the PBMC dataset using Wavelet-HGC ([Fig J](#) in [S1 Text](#)). Without DWT, HGC method yields nine cluster from the original data; under wavelet transformation, each component identifies eight or nine clusters. The low-resolution components provide an overview of major cell types, while the high-resolution components reveal finer-grained cellular structures, highlighting the multi-scale clustering capabilities of Wavelet-HGC.

Distinct gene signatures of cell clusters from different wavelet-based multi-view windows

To quantify the improvement in multi-view clustering of cells from our algorithms, we assess the performance of clustering at two levels, overlapping's of cells between clusters and intersections of gene markers.

We find 10 significant clusters using 2-band DWT for the breast cancer data, and one more cluster from detail component can be found when compared with data without DWT. However, the effects are marginal when utilizing multi-band DWT. Of note, increasing the number of intersections between clusters in different resolution components may result in some subsets containing a small number of cells with a relatively high p-value. We discover 22 significant clusters using 3-band DWT, where 10, 8 and 11 clusters are found in approximation, detail-1 and detail-2 component-based data, respectively. Similarly, a total of 28 significant clusters are revealed for 4-band DWT, with 10, 8, 9 and 9 clusters in approximation, detail-1, detail-2 and detail-3 component-based data, respectively. Consequently, one can use WMC to divide some clusters in original data into a few subgroups, resulting in the clusters with greater significance.

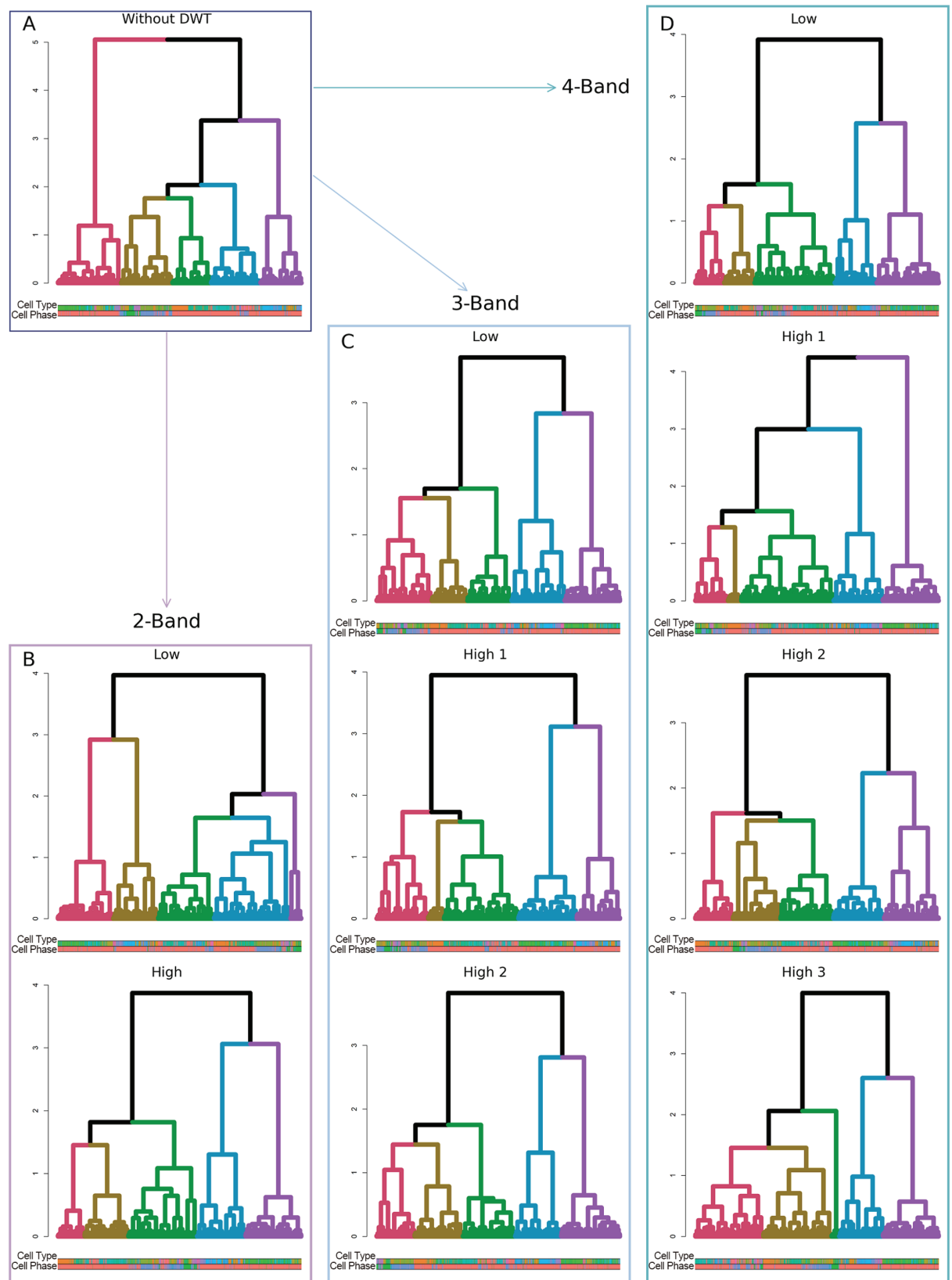


Fig 4. Multi-view hierarchical clusters on ILC dataset using Wavelet-HGC method. (A) Dendrogram for ILC dataset without DWT. (B)-(D) are dendrograms under wavelet analysis, with (B) for 2-band DWT, (C) for 3-band DWT, and (D) for 4-band DWT.

<https://doi.org/10.1371/journal.pcbi.1013060.g004>

Apparently, the cell types in different resolution components can overlap or be distinct. Indeed, comprehensive intersection analysis of gene signatures further support the power of multi-view of WMC (Fig 5).

Daubechies wavelet family, belonging 2-band wavelet families, produced one low-resolution component and one high-resolution component. Low-resolution parts are traditionally considered as an approximation of original data and share a number of common genes with original data. We find 528 more genes in the 2-band approximation component-based data than in original data (Fig 5A). Similarly, we discover 510 additional genes in 3-band approximation component-based data compared to original data (Fig 5B), and 321 more genes in 4-band approximation component (Fig 5C) than in original data.

DWT has the ability to decompose original data into orthogonal components, which is another important property. This means that different resolution parts are not overlapped

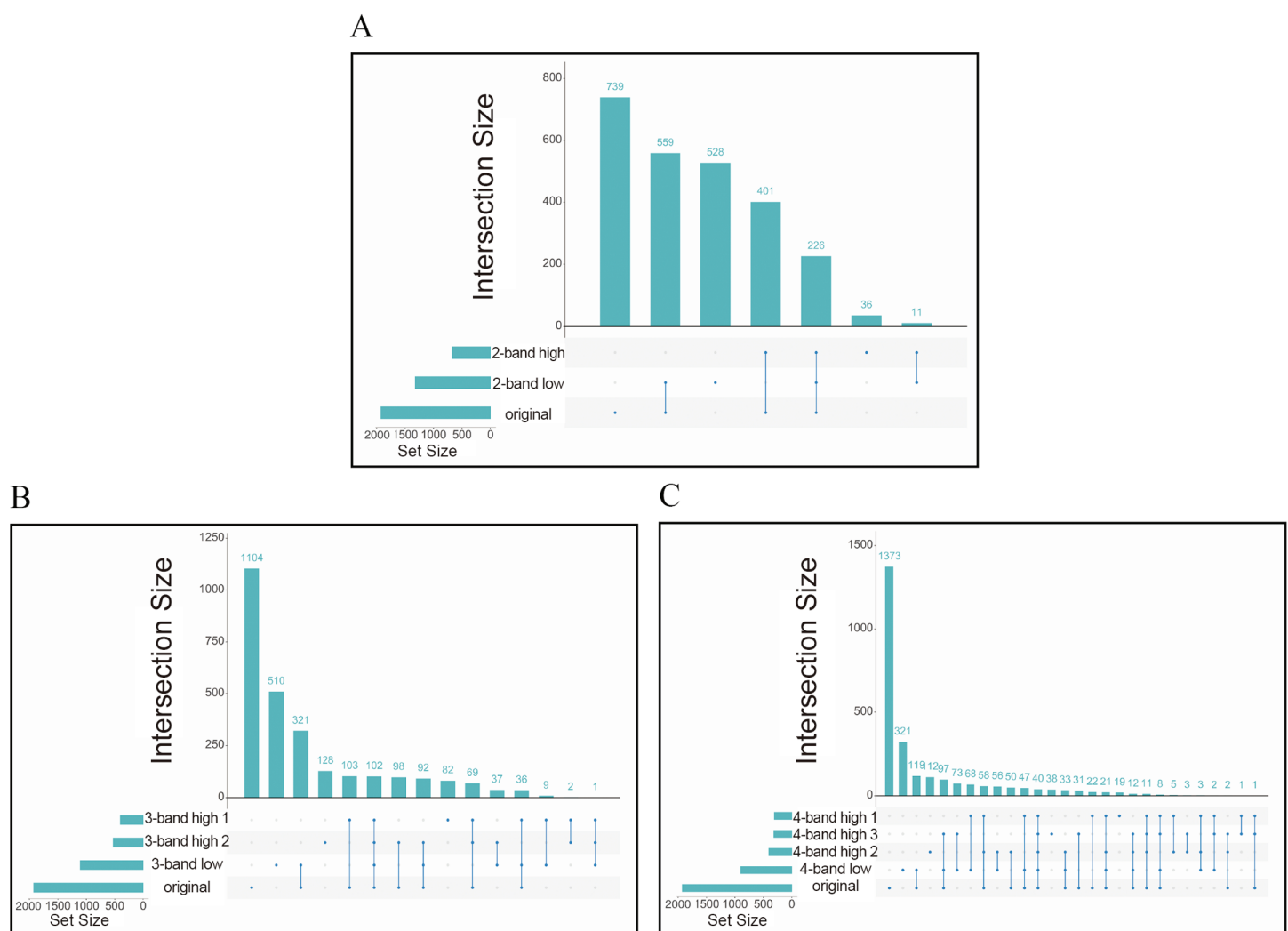


Fig 5. Assessing the performance of WMC on the breast cancer dataset. The x-axes of panels (A) through (C) represents the total number of genes in each component and the y-axes shows the number of genes in each intersection set. The cyan dots represent the number of genes belonging to the corresponding components, for gray dots vice versa. (A) The intersection of genes between original data and its different wavelet transformed components using 2-band DWT, while (B) for 3-band and (C) for 4-band DWT, respectively.

<https://doi.org/10.1371/journal.pcbi.1013060.g005>

theoretically. Using 2-band DWT, there are 237 genes that overlap in both resolution components, but only 11 (4.6%) of them are new, and the remaining 226 genes appeared in the original data (Fig 5A).

High-resolution components are particularly valuable for identifying cell types, as they capture fine-grained characteristics that may be overlooked in traditional analyses. Compared to the original dataset, the detail components extract through multi-band DWT reveal a greater number of cell-type-specific gene markers (Fig 5A–5C), underscoring the potential of wavelet transformation for enhancing marker gene detection.

Similar phenomena can be found in colorectal and PBMC dataset. In PBMC dataset, there are 133 genes that are identified at the intersection of two resolution components using 2-band DWT, whereas only 2 genes are absent from original data (Fig N in S1 Text). For the colorectal dataset, we only discover 80 at the intersection between two components using 2-band DWT, which is significantly less than the number of new genes only appearing in one single component. For example, we discover 301 and 201 genes in 2-band approximation and detail component-based data respectively (Fig O in S1 Text).

To further assess the consistency of clustering across wavelet-transformed components, we compute ARI and NMI as summarized in Table 1. The ARI values range from 0.6 to 0.8, while NMI values range from 0.57 to 0.73, indicating a moderate level of agreement. Notably, both ARI and NMI decrease as the number of wavelet bands increases from 2-band to 4-band DWT. It appears that the diversity of clustering outcomes introduced by multi-band wavelet transformation increases.

In Fig 6, we present the compared average ARI across breast cancer datasets on performance of clustering methods with and without DWT. The detail components extracted via DWT consistently improve clustering accuracy across all breast cancer subtypes. The highest ARI values are observed for the 4-band detail component in ER+ (0.96) and TNBC (0.98) subtypes, while the 2-band detail component achieves the highest ARI for HER2+ (0.97). These results demonstrate the strong compatibility of WMC with breast cancer classification. Thus, WMC is robust in multi-view clustering.

Discussion

In this study, we develop WMC, an M-band DWT-based approach for presenting scRNA-seq matrices via multi-view clustering of cells. WMC not only offers a unique capability that overcomes key limitations of existing single-cell analysis techniques, but also helps to identify missing cell types and rare cell types alongside activity. WMC can help discover rare cell types missed by conventional method with a fine resolution. Finally, WMC can potentially

Table 1. Average ARI and NMI for clusters of M-band DWT components

The type of DWT	components	ARI	NMI
2-band	low vs high	0.80	0.73
3-band	low vs h1	0.72	0.65
3-band	low vs h2	0.76	0.70
3-band	h1 vs h2	0.73	0.67
4-band	low vs h1	0.66	0.59
4-band	low vs h2	0.61	0.57
4-band	low vs h3	0.64	0.58
4-band	h1 vs h2	0.69	0.63
4-band	h1 vs h3	0.73	0.65
4-band	h2 vs h3	0.65	0.61

<https://doi.org/10.1371/journal.pcbi.1013060.t001>

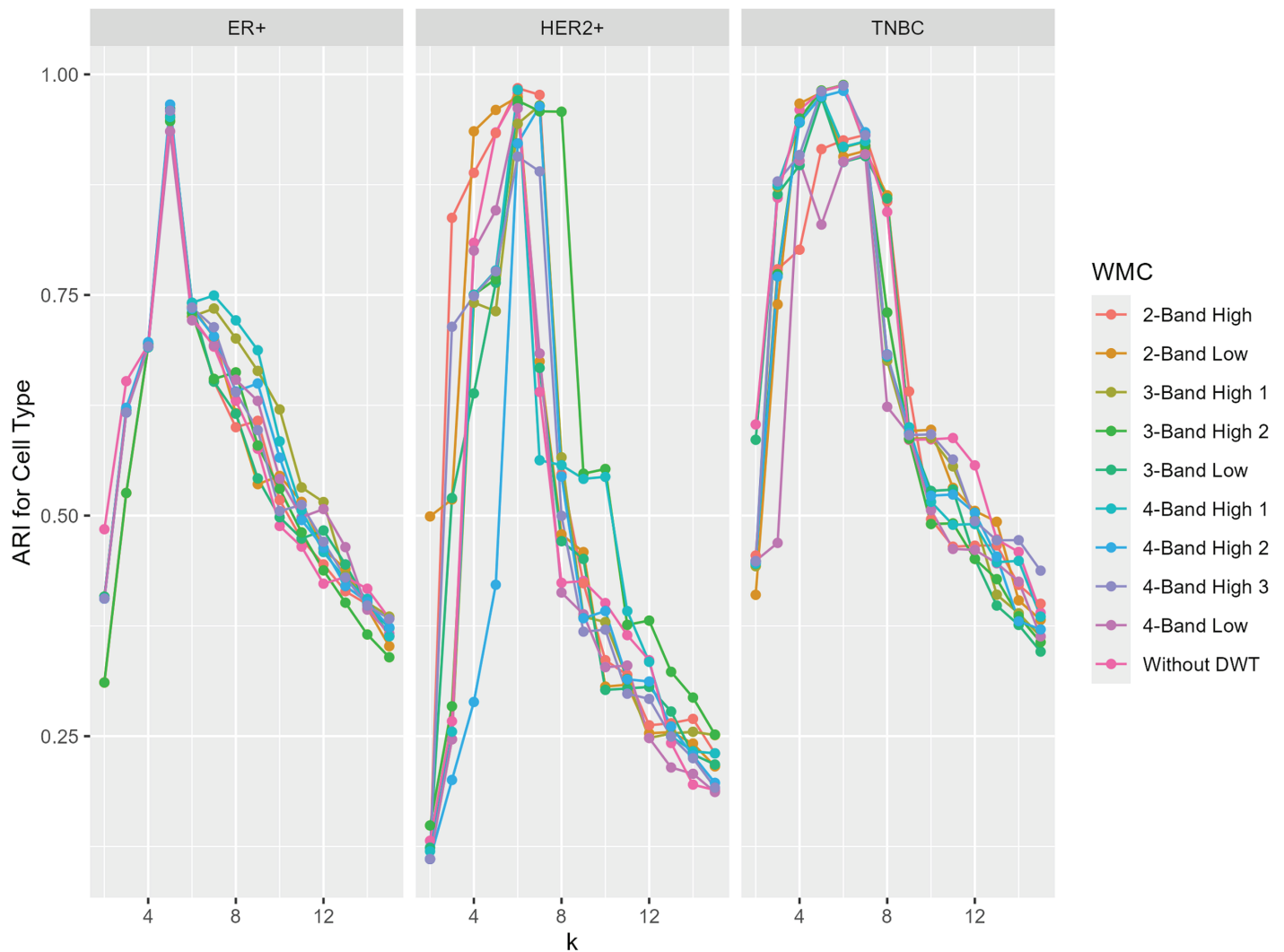


Fig 6. Average ARI results for breast cancer datasets on different subtypes of cancer cells.

<https://doi.org/10.1371/journal.pcbi.1013060.g006>

prioritize rare cell types for experimental validation by incorporating intersection analysis on multi-view of clusters.

Our study has certain limitations, as the current WMC framework is implemented with a maximum of 4-band DWT, balancing computational efficiency and multi-view clustering performance. It is plausible that using more than four bands could further enhance multi-scale clustering resolution. A key open question is determining the optimal choice of DWT type and band number for different biological datasets. We aim to systematically address this challenge through clustering optimization strategies in future work. For general WMC applications, the Daubechies (2-band) wavelet basis is recommended as the preferred option due to its computational efficiency and ease of implementation.

Moreover, even though multi-view clusters can detect unidentified cell types that may represent novel cell types in the biological field, these cell types are yet to functionally validate in a biological experiment. For future research, we envision that DWT-based framework

integrated with other omics information is promising. Meanwhile, WMC has a great potential to analysis spatial single-cell gene expression data. Thus, our WMC paves the way to deploy wavelet tools in scRNA-seq data analysis.

Materials and methods

Data preprocessing

Any type of high-dimensional single-cell data can be transformed using WMC. However, before analysing the single-cell gene expression data, the raw scRNA-seq data often necessitates particular preprocessing and normalization to ensure that technical effect does not distort downstream analysis results. Given a raw data generated by sequencing machine, the quality control preformed based on following three criteria: 1) the number of counts per barcode (count depth), 2) the number of genes per barcode, and 3) the proportion of counts from mitochondrial genes per barcode. Since low-quality cells or empty droplets frequently have very few genes and a large number of detected genes may represent doublets, we begin by filtering cells with unique genes detected in excess of 3,000 or fewer than 200. Next, we filter out genes that are not expressed in fewer than 20 cells. In addition, cells with a mitochondrial count greater than 5 percent are removed, as cells with a relatively high mitochondrial count may be implicated in respiratory processes. We next perform a log normalization on the quality-controlled data matrix to reduce the variability of data before applying WMC.

Discrete wavelet analysis

In our method to deal with the scRNA-seq data matrix, we use M-band DWT to decompose the count matrix (denoted by S) into M different resolution components, as shown in (3). In this paper, we select $M = 2, 3$, and 4, respectively, where the case of $M = 2$ corresponds to Daubechies wavelets family. The decomposition can be performed using the DWT matrix (see S1 Text for more details). The DWT projects the data into orthogonal subspaces with different resolutions, allowing us to extract the information partially hidden by the original data matrix. The wavelet-based filters satisfy a set of orthogonal condition. Thus, the DWT preserves the energy of the data since the l_2 -norm remains unchanged under orthonormal transform.

Assessment using intersection analysis

We assess the performance of WMC via computing the intersection of $\mathcal{I}(S), \mathcal{I}(A), \dots, \mathcal{I}(D_{M-1})$, where $\mathcal{I}(X)$ denotes that the data X is transformed by PCA and UMAP. Without wavelet transform, $\mathcal{I}(S)$ gives a single-view clustering, while $\mathcal{I}(A), \mathcal{I}(D_1), \dots, \mathcal{I}(D_{M-1})$ provide a multi-view clustering of RNA-seq data. To examine the number of total clusters in multi-view windows, we calculate the distribution of significant genes among different resolutions components. Specifically, let G_0 be the set of significant genes in the raw data, G_1 be that in the approximation component, and $G_2, \dots, G_M$ be that in the detail components, respectively. For each non-empty subset, $\Lambda \subset \{0, 1, \dots, M\}$, we calculate $N_\Lambda = (\cap_{i \in \Lambda} G_i) \cap (\cap_{j \notin \Lambda} \bar{G}_j)$ i.e., the set of genes in the selected components but not in the other components. If $|\Lambda| = 1$, $|N_\Lambda|$ denotes the number of exclusive genes in the corresponding component. Note that the components of different resolutions under the DWT are orthogonal to each other, if $|\Lambda| \geq 2$ and $0 \notin \Lambda$, only a few genes will contain in such N_Λ theoretically. The smaller number of $|N_\Lambda|$ for such intersection yields a better resolution of clustering.

Adjusted rand index and normalized mutual information

The ARI is a widely used metric for quantifying the similarity between two clustering schemes. Let n_{ij} denote the number of cells in cluster i of scheme A and cluster j in scheme B . The ARI between clustering schemes A and B is computed as follows:

$$\text{ARI}(A, B) = \frac{\sum_{i,j} \binom{n_{ij}}{2} - (\sum_i \binom{n_{i+}}{2}) \sum_j \binom{n_{+j}}{2}) / \binom{N}{2}}{(\sum_i \binom{n_{i+}}{2} + \sum_j \binom{n_{+j}}{2}) / 2 - (\sum_i \binom{n_{i+}}{2}) \sum_j \binom{n_{+j}}{2}) / \binom{N}{2}} \quad (8)$$

where $n_{i+} = \sum_j n_{ij}$ represents the number of cells in cluster i of scheme A , $n_{+j} = \sum_i n_{ij}$ represents the number of cells in cluster j of scheme B , and N is the total number of cells.

In this study, ARI is utilized for two primary purposes: (1) evaluating the performance of WMC by comparing ARI values between WMC and the corresponding conventional framework without wavelet transformation; (2) assessing clustering consistency across different wavelet-transformed components, by computing ARI values for clustering results obtained from distinct resolution components.

ARI values range from -1 to 1, where a value closer to 1 indicates a high degree of similarity between clustering schemes, while a negative value suggests poor agreement and mismatching.

Another widely used metric for quantifying clustering similarity is the NMI, defined as:

$$\text{NMI}(A, B) = \frac{\sum_{i,j} \frac{n_{ij}}{N} \log \frac{n_{ij}N}{n_{i+}n_{+j}}}{\sqrt{\sum_i n_{i+} \log \frac{n_{i+}}{N} \sum_j n_{+j} \log \frac{n_{+j}}{N}}}. \quad (9)$$

The NMI score ranges from 0 to 1, where 1 indicates a perfect correspondence between the two clustering schemes. In this study, NMI is also employed to evaluate the consistency of multi-view clustering results across different wavelet components, providing insights into the robustness and diversity of the clustering outcomes.

Supporting information

S1 Text. Implementation details and supplementary results. This file provides detailed mathematical background on the DWT, implementation notes for the WMC framework, and additional results that support and extend the main findings presented in the manuscript. (PDF)

Author contributions

Conceptualization: Tong Liu, Zihuan Liu, Wenke Sun, Yongzhong Zhao, Xiaodi Wang.

Data curation: Tong Liu, Zihuan Liu, Wenke Sun, Adeethya Shankar.

Formal analysis: Tong Liu, Zihuan Liu, Wenke Sun, Adeethya Shankar.

Funding acquisition: Yongzhong Zhao, Xiaodi Wang.

Investigation: Tong Liu, Zihuan Liu, Wenke Sun, Adeethya Shankar.

Methodology: Tong Liu, Zihuan Liu, Wenke Sun, Adeethya Shankar, Yongzhong Zhao, Xiaodi Wang.

Project administration: Yongzhong Zhao, Xiaodi Wang.

Resources: Yongzhong Zhao, Xiaodi Wang.

Software: Tong Liu, Zihuan Liu, Wenke Sun, Adeethya Shankar.

Supervision: Yongzhong Zhao, Xiaodi Wang.

Validation: Yongzhong Zhao, Xiaodi Wang.

Visualization: Tong Liu, Zihuan Liu, Wenke Sun, Adeethya Shankar.

Writing – original draft: Tong Liu, Zihuan Liu, Wenke Sun, Adeethya Shankar, Yongzhong Zhao, Xiaodi Wang.

Writing – review & editing: Tong Liu, Zihuan Liu, Wenke Sun, Adeethya Shankar, Yongzhong Zhao, Xiaodi Wang.

References

1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*. 2009;6(5):377–82. <https://doi.org/10.1038/nmeth.1315> PMID: 19349980
2. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep*. 2012;2(3):666–73. <https://doi.org/10.1016/j.celrep.2012.08.003> PMID: 22939981
3. Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol*. 2011;29(12):1120–7. <https://doi.org/10.1038/nbt.2038> PMID: 22081019
4. Jindal A, Gupta P, Sengupta D. Discovery of rare cells from voluminous single cell expression data. *Nat Commun*. 2018;9(1):4719. <https://doi.org/10.1038/s41467-018-07234-6> PMID: 30413715
5. Brbić M, Zitnik M, Wang S, Pisco AO, Altman RB, Darmanis S, et al. MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nat Methods*. 2020;17(12):1200–6. <https://doi.org/10.1038/s41592-020-00979-3> PMID: 33077966
6. Trapnell C. Defining cell types and states with single-cell genomics. *Genome Res*. 2015;25(10):1491–8. <https://doi.org/10.1101/gr.190595.115> PMID: 26430159
7. Miao Z, Moreno P, Huang N, Papatheodorou I, Brazma A, Teichmann SA. Putative cell type discovery from single-cell gene expression data. *Nat Methods*. 2020;17(6):621–8. <https://doi.org/10.1038/s41592-020-0825-9> PMID: 32424270
8. Michielsen L, Reinders MJT, Mahfouz A. Hierarchical progressive learning of cell identities in single-cell data. *Nat Commun*. 2021;12(1):2799. <https://doi.org/10.1038/s41467-021-23196-8> PMID: 33990598
9. Saviano A, Henderson NC, Baumert TF. Single-cell genomics and spatial transcriptomics: discovery of novel cell states and cellular interactions in liver physiology and disease biology. *J Hepatol*. 2020;73(5):1219–30. <https://doi.org/10.1016/j.jhep.2020.06.004>
10. Levitin HM, Yuan J, Cheng YL, Ruiz FJ, Bush EC, Bruce JN, et al. De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. *Mol Syst Biol*. 2019;15(2):e8557. <https://doi.org/10.15252/msb.20188557> PMID: 30796088
11. Lu Y-C, Jia L, Zheng Z, Tran E, Robbins PF, Rosenberg SA. Single-cell transcriptome analysis reveals gene signatures associated with T-cell persistence following adoptive cell therapy. *Cancer Immunol Res*. 2019;7(11):1824–36. <https://doi.org/10.1158/2326-6066.CIR-19-0299> PMID: 31484655
12. Plass M, Solana J, Wolf FA, Ayoub S, Misios A, Glažar P, et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*. 2018;360(6391):eaq1723. <https://doi.org/10.1126/science.aq1723> PMID: 29674432
13. Wang W, Vilella F, Alama P, Moreno I, Mignardi M, Isakova A, et al. Single-cell transcriptomic atlas of the human endometrium during the menstrual cycle. *Nat Med*. 2020;26(10):1644–53. <https://doi.org/10.1038/s41591-020-1040-z> PMID: 32929266
14. Fernandez DM, Rahman AH, Fernandez NF, Chudnovskiy A, Amir E-AD, Amadori L, et al. Single-cell immune landscape of human atherosclerotic plaques. *Nat Med*. 2019;25(10):1576–88. <https://doi.org/10.1038/s41591-019-0590-4> PMID: 31591603

15. Song J, Ware T, Liu SL, Surette M. Comparative genomics via wavelet analysis for closely related bacteria. *EURASIP J Adv Signal Process*. 2004;2004:497292. <https://doi.org/10.1155/S111086570430921X>
16. Sawaya S, Bagshaw A, Buschiazio E, Kumar P, Chowdhury S, Black MA, et al. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS One*. 2013;8(2):e54710. <https://doi.org/10.1371/journal.pone.0054710> PMID: 23405090
17. Zhou X, Li Z, Dai Z, Zou X. Predicting promoters by pseudo-trinucleotide compositions based on discrete wavelet transform. *J Theor Biol*. 2013;319:1–7. doi:10.1016/j.jtbi.2012.10.008
18. Denault WRP, Gjessing HK, Juodakis J, Jacobsson B, Jugessur A. Wavelet screening: a novel approach to analyzing GWAS data. *BMC Bioinformatics*. 2021;22(1):464. <https://doi.org/10.1186/s12859-021-04356-5> PMID: 34620077
19. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36(5):411–20. <https://doi.org/10.1038/nbt.4096> PMID: 29608179
20. Shao X, Liao J, Lu X, Xue R, Ai N, Fan X. scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data. *iScience*. 2020;23(3):100882. <https://doi.org/10.1016/j.isci.2020.100882> PMID: 32062421
21. De Meo P, Ferrara E, Fiumara G, Proveti A. Generalized Louvain method for community detection in large networks. In: 2011 11th International Conference on Intelligent Systems Design and Applications. 2011, pp. 88–93. <https://doi.org/10.1109/isda.2011.6121636>
22. Steffen P, Heller PN, Gopinath RA, Burrus CS. Theory of regular m-band wavelet bases. *IEEE Trans Signal Process*. 1993;41(12):3497–511. doi:10.1109/78.258083
23. Lin T, Xu S, Shi Q, Hao P. An algebraic construction of orthonormal M-band wavelets with perfect reconstruction. *Appl Math Comput*. 2006;172(2):717–30. <https://doi.org/10.1016/j.amc.2004.11.025>
24. Tao W, Li M, Zhu X, Zhou Q, Zong J, Zhang L, et al. Dynamic regulation of innate lymphoid cell development during ontogeny. *Mucosal Immunol*. 2024;17(6):1285–300. doi:10.1016/j.mucimm.2024.03.004
25. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods*. 2017;14(5):483–6. <https://doi.org/10.1038/nmeth.4236> PMID: 28346451