haCCA: Multi-module Integrating of spatial transcriptomes and metabolomes.

abstract

Spatial multiomics techniques, such as spatial transcriptomics and MALDI-MSI, provide detailed molecular profiles across tissue sections, capturing mRNA expression and mass-to-charge (m/z) spectra, respectively. These methods generate matrices with spatial coordinates, offering insights into the biological states of specific tissue spots. However, integrating spatial transcriptomics with MALDI-MSI data is challenging due to the lack of shared features or spots, which limits conventional integration strategies.

We present haCCA, a workflow designed to integrate spatial transcriptomics and metabolomics data using high-correlated feature pairs and modified spatial morphological alignment. This approach ensures high-resolution and accurate spot-to-spot data integration. To evaluate haCCA, we generated benchmark datasets combining spatial transcriptomes with pseudo-spatial metabolomes. Our results show that haCCA outperforms existing morphological alignment methods, such as STUtility.

We applied haCCA to both publicly available 10X Visium and MALDI-MSI datasets from mouse brain tissue and a custom 10X VisiumHD and MALDI-MSI dataset from an intrahepatic cholangiocarcinoma (ICC) model. These applications demonstrated haCCA's effectiveness in enhancing multi-modular spatial data analysis. To facilitate its use, we developed and published an easy-to-use Python package, providing the research community with a robust tool for spatial multiomics integration.

Introduction

Spatial mutiomics techniques allows resolving the spatial molecular profiles at transcript, metabolite, protein or epi-genetic level(1). Spatial transcriptomes enables the measurement of mRNA expression, while MALDI-MSI allows the collection of mass-to-charge (m/z) spectra at defined raster spots across the tissue section. Both methodologies are most frequently applied spatial techniques. These 2 techniques generate gene-expression or peak-intensity matrix of sampling spots, with coordinates indicating their location. Numbers of spots can very from thousands to hundred of thousands determined by resolution.

An integration of spatial transcriptomes and MALDI-MSI to generate multi-module spatial data provides more comprehensive understanding and novel insight to biological state of section spots. However, such integration may be challenge. Till now 3 categories of integration strategies has been brought out for multi module single cell data(2), using mutual data features, using mutual cells, or using shared latent space if both mutal cells or features are not accessible. The first 2 strategies are able to keeps the origin resolution, which is single cell or spot, while the last strategy will lose it. Usually no mutual feature can be found between spatial transcriptomes and MALDI-MSI, neither mutual spots or cells, constrain the application of first 2 strategies. Luckily, spatial multi module techniques are often applied on neighbour sections, even the same section(3), ensuring the shape similarity. Hence morphological alignment of spatial shape followed by neighbour finding strategy has been applied on their integration most often(4-7).

Such integrate strategy brings certain flaws. First is that there is no benchmark data to access the performance of integration, second is that as no features are applied for assistance in present procedure, the efficiency of integration might be undermined. Using mutual features for alignment has been proved useful in integrate single cell data from difference batches, where by using linear combination of mutual features to establish latent space, allowing cells with the same labels to

appear together and be clustered into the same group (8-10). Although no mutual feature existed among MSI and spatial transcriptomes, a few studies revealed the high correlation between gene transcripts and metabolites(11, 12), giving us the inspiration to explore whether high correlated gene-metabolite pairs can be utilized in integrating.

In this study, we established haCCA, a workflow utilizing **h**igh Correlated fe**a**ture pairs **c**ombined with a modified spatial morphologi**ca**l alignment to ensure high resolution and accuracy of spot-to-spot data integration of spatial transcriptomes and metabolomes. We generated a series of benchmarks using spatial transcriptomes and paired pseudo spatial metabolomes, proving haCCA outperformed morphological-aligned method(STUtility). We applied haCCA in public available 10X Visium and MALDI-MSI data cohort of mouse brain and a home-brew 10X VisiumHD and MALDI-MSI data cohort of ICC(intrahepatic cholangiocarcinoma) model, to test the performance of haCCA in assisting multi-module spatial data analysis. An easy-to-use python package has been developed and published for use.

Method

1. Data source

For public data, 10X Visium and MSI-MALDI data from Parkinson's disease mouse brain was download from Macro et al's manuscript(3). 6 10X Visium,3 MSI-MALDI data of 6 sections from 3 mouses were included(Table).

2. For home-brew data, We generated Akt/Yap induced ICC(intrahepatic cholangiocarcinoma) mouse model in WT or PADI4-/- mouse. 10X VisiumHC and MALDI-MSI data of tumor was generated following standard procedure(see below).

Mouse model

10X VisiumHD

MALDI-MSI

Pseudo-MALDI-MSI data generating

We employed an iteration and sampling strategy on spatial transcription data to generate paired pseudo-MALDI-MSI data. These pairs served as benchmarks for evaluating the performance of our alignment strategy.

For the spatial transcription data, we have $Data(f,d,l) = \{f_m^n,\ d_m^2,\ l_m\}\ for\ \{m = 1, \dots, M\}$, where for each $data_m(f,d,l)$ f denotes the n-dimension feature-spot vector $f^1, f^2, \dots, f^n\}$, d represents the coordinate matrix (d_x, d_y), l represents the cluster label.

Given $Data(f,d,l)$, we use the following steps to mimic the morphological disagreement between neighboring sections and generate a pair of $(Data_{train}, Data'_{truth})$ from $Data(f,d,l)$ to evaluate $haCCA$ workflow.

Step 1: Split feature set f by randomly sample $j\ where\ j \in (0,n)$ features from f and assign them to $(Data_{train}, Data'_{truth})$

$$Data_{train} = \{f_m^j, d_m^2, l_m\}$$
$$Data'_{truth} = \{f_m^{\bar{j}}, d_m^2, l_m\}$$

Step 2: apply Gaussian translation to both d_{train} and d_{truth} to mimick the morphological

disagreement between neighborsections.

$$Data_{train} = \{f_m^j, \tilde{d}_m^2, l_m\}$$

$$Data'_{truth} = \left\{f_m^{\tilde{j}}, \tilde{\tilde{d}}_m^2, l_m\right\}$$

haCCA workflow

Given two spatial data $\boldsymbol{Data_A}$ and $\boldsymbol{Data_B}$ where

$$\boldsymbol{Data_A} = \{\boldsymbol{f_{m_a}^{n_a}}, \boldsymbol{d_{m_a}^2}, \boldsymbol{l_{m_a}}\}, \boldsymbol{m_a} \sim \{\boldsymbol{1}, \dots, \boldsymbol{A}\}$$

$$\boldsymbol{Data_B} = \{\boldsymbol{f_{m_b}^{n_b}}, \boldsymbol{d_{m_b}^2}, \boldsymbol{l_{m_b}}\}, \boldsymbol{m_b} \sim \{\boldsymbol{1}, \dots, \boldsymbol{B}\}$$

$\boldsymbol{haCCA_{A\to B}}$ aligns the points in $\boldsymbol{Data_A}$ to $\boldsymbol{Data_B}$ by finding a point-to-point map from A to B which minimizing the following loss object

$$loss_{hacca_{a\to b}} \sim \sum_{(m_b, m_a)} \left\|d_{m_b} - d_{m_a}\right\| + ||f_{m_b}^{cca} - f_{m_a}^{cca}||$$

Where $\{(m_b, m_a), m_b \in Data_b, m_a \in Data_a\}$ is the aligned points pairs.

Once $\{(m_b, m_a)\}$ is determined, the result of $haCCA_{B\to A}$ and aggregate result for A can be represented as

$$Data_{A\to B} = \left\{\boldsymbol{f_{m_a}^{n_b}}, \boldsymbol{d_{m_a}^2}, \boldsymbol{l_{m_a}}\right\}, (\boldsymbol{m_b}, \boldsymbol{m_a}) \in \{(\boldsymbol{m_a}, \boldsymbol{m_b})\}$$

$$aggregated\ Data_B = Data_B \cup Data_{A\to B}$$

To minimize the $loss_{hacca_{a\to b}}$ we performed a heuristic two-step method: In the first step, we apply the gross alignment plus further alignment to $Data_B$, which minimize the first item of $loss_{hacca_{a\to b}}$. In the second step, we make use the alignment result from first step and identify a set of anchor points, which is used to determine the high correlation feature pairs and minimize the second item of the loss function. We will introduce each step in detail in the below sections.

Step 1: minimize the first item by aligning A to B using location information $\{d\}$

In this step, we apply two alignment strategies here: gross alignment and further alignment.

In gross alignment, manual alignment or classical cloud mesh alignment algorithm like ICP or FGW is used to determine an affine translation matrix M to minimize the L2 distance between points in $Data_A$ and $Data_B$

$$M \sim gross_alignment(d_A, d_B)$$

$$d_B^{\square} = M * d_B$$

In further alignment, we find a rotation and moving matrix M' to minimize the distance between outlier spots in $\{d_B\}$ and their nearest neighbors in $\{d_A\}$. the outlier spots in $\{d_B\}$ is defined as the spots where it does not have neighbor within a given distance in $\{d_A\}$. After the outlier spots

in $\{d_B\}$ is found, we calculate the affine translation matrix M' using outlier spots in $\{d_B\}$ and its nearest neighbor in $\{d_A\}$.

$$M' \sim further_alignment(d_A, d_B)$$
$$d_B^{\square} = M' * d_B$$

Step 2: minimize the second item by reducing the distance of high-correlation feature pairs identified by anchor spot pairs $\{(m_a^{anchor}, m_b^{anchor})\}, m_a \in M_A, m_b \in M_B$

The anchor spot pair are paired points from (M_A, M_B) where each pair (m_a, m_b) satisfies with the following conditions: 1. The distance between m_a and m_b are smaller than a given threshold, and they are the closest point to each other. 2. The neighborhood of both m_a and m_b is of low variety. We use Simpson index to evaluate the variety.

The anchor spot pairs are identified using the following steps: 1. For each points $\{m_a\} \in M_A$, found a set of points $\{m_b\}, m_b \in M_B$ where each m_b is within a given distance. 2. For each item in $\{m_a : \{m_b\}\}, m_a \in M_A, m_b \in M_B$, calculate the Simpson index for each item and remove the items where its Simpson index is higher than a given threshold. 3. For the remaining items in $\{m_a : \{m_b\}\}, m_a \in M_A, m_b \in M_B$, the final anchor spot pairs $\{(m_a^{anchor}, m_b^{anchor})\}, m_a \in M_A, m_b \in M_B$ are identified as the closest point pair in each item.

After finding anchor spot pairs $\{(m_a^{anchor}, m_b^{anchor})\}, m_a \in M_A, m_b \in M_B$, the high correlated feature pairs $\left\{(f_A^i, f_B^j)_k\right\}, f_A \in F_{Data_A}, f_B \in F_{Data_B}$ where i, j identifies the ith and j^{th} feature in each dataset are further identified by the top k score of Pearson correlation matrix on aggregated feature sets $\left\{f_{m_a}^{n_a} \cup f_{m_b}^{n_b}\right\}, (m_a, m_b) \in \{(m_a^{anchor}, m_b^{anchor})\}$

After the high correlated feature pairs $\left\{(f_A^i, f_B^j)_k\right\}, f_A \in F_{Data_A}, f_B \in F_{Data_B}$ are identified, we calculate the cca feature for both $\{m_a\}$ and $\{m_b\}$ using canonical variables found by Canonical Correlation Analysis (CCA). The idea behind is to further maximize the correlation and normalize the value between the high correlated feature pairs. The new features are noted as f^{cca}. f^{cca} will be added to $Data_A$ and $Data_B$ as an individual feature after they are calculated for all the points $\{m_a\}$ in $Data_A$ and $\{m_b\}$ in $Data_B$.

$$\boldsymbol{Data_A = \left\{f_{m_a}^{n_a}, d_{m_a}^2, l_{m_a}\right\} \cup \{f_{m_a}^{cca}\}, m_a \sim \{1, \dots, A\}}$$

$$\boldsymbol{Data_B = \left\{f_{m_b}^{n_b}, d_{m_b}^2, l_{m_b}\right\} \cup \{f_{m_b}^{cca}\}, m_b \sim \{1, \dots, B\}}$$

Finally, we minimize the second item on both $\{l\}$ and $\{f^{cca}\}$ using classic cloud mesh alignment algorithm like ICP or FGW.

$$M \sim ICP(\{\boldsymbol{d_{m_a}^2} \cup \boldsymbol{f_{m_a}^{cca}}\}, \{\boldsymbol{d_{m_b}^2} \cup \boldsymbol{f_{m_b}^{cca}}\})$$

$$\{\boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_{b'}}} \cup \boldsymbol{f}^{\boldsymbol{cca}}_{\boldsymbol{m_{b'}}}\} = M * \{\boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_b}} \cup \boldsymbol{f}^{\boldsymbol{cca}}_{\boldsymbol{m_b}}\}$$

$$J^* = arg\ min_j\ d\left(\{\boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_{a'i}}} \cup \boldsymbol{f}^{\boldsymbol{cca}}_{\boldsymbol{m_{a'j}}}\}, \{\boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_b}} \cup \boldsymbol{f}^{\boldsymbol{cca}}_{\boldsymbol{m_b}}\}\right)$$

$$\boldsymbol{Data_{B'}} = \left\{\boldsymbol{f}^{\boldsymbol{n_a}}_{\boldsymbol{m_{aJ*}}}, \boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_{aJ*}}}, \boldsymbol{l}_{\boldsymbol{m_{aJ*}}}\right\}$$

FGW algorithm is described below:

Finding a transport plan π, where

$$\mathrm{C}^{\mathrm{P}}_{\mathrm{ij}} = ||\boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_b i}} - \boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_b j}}||^{\,2}$$

$$\mathrm{C}^{\mathrm{Q}}_{\mathrm{kl}} = ||\boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_{ak}}} - \boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_{al}}}||^{\,2}$$

$$F_{ik} = ||\boldsymbol{f}^{\boldsymbol{cca}}_{\boldsymbol{m_{bl}}} - \boldsymbol{f}^{\boldsymbol{cca}}_{\boldsymbol{m_{ak}}}||$$

$$\pi = \arg\min_{\pi} \sum_{\mathrm{i,j,k,l}} \left[\mathrm{a}(\mathrm{C}^{\mathrm{P}}_{\mathrm{ij}} - \mathrm{C}^{\mathrm{Q}}_{\mathrm{kl}})^2 + (1-\mathrm{a})\mathrm{F}_{\mathrm{ik}}\right]\pi_{\mathrm{i,k}}\pi_{\mathrm{j,l}}$$

$$\boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_{b'}}} = \pi * \boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_a}}$$

$$J^* = arg\ min_j\ d\left(\boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_{bi}}}, \boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_{aj}}}\right)$$

$$\boldsymbol{Data_{B'}} = \left\{\boldsymbol{f}^{\boldsymbol{n_a}}_{\boldsymbol{m_{aJ*}}}, \boldsymbol{d}^{\boldsymbol{2}}_{\boldsymbol{m_{aJ*}}}, \boldsymbol{l}_{\boldsymbol{m_{aJ*}}}\right\}$$

B' is the haCCA translated B.

Evaluation

Label Transfer Accuracy

Label transfer accuracy measures how well labels from one dataset can be transferred to and correctly applied in another dataset.

$$Accuracy\ =\ \frac{Ncorrect}{Ntotal}$$

Where $N_{correct}$ is the number of correctly transferred labels in the target dataset, N_{total} is the total number of samples in the target dataset.

Label Transfer ARI

Label Transfer ARI (Adjusted Rand Index) is a measure of the similarity between two data clusterings or label assignments. It adjusts the Rand Index (RI) to account for chance grouping.

$$\mathrm{ARI}\ =\ \frac{\mathrm{RI} - \mathrm{Expected\ RI}}{\mathrm{Max\ RI} - \mathrm{Expected\ RI}}$$

$$\mathrm{Expected\ RI} = \frac{\sum_{\mathrm{i}}\binom{\mathrm{ai}}{2} * \sum_{\mathrm{j}}\binom{\mathrm{bj}}{2}}{\binom{\mathrm{n}}{2}}$$

$$\mathrm{Max\ RI} = 0.5 * (\sum_{\mathrm{i}}\binom{\mathrm{ai}}{2} + \sum_{\mathrm{j}}\binom{\mathrm{bj}}{2})$$

$$RI = \frac{a+b}{a+b+c+d}$$

n: Total number of elements in the dataset.

a: Number of pairs of elements that are in the same cluster in both clusterings.

b: Number of pairs of elements that are in different clusters in both clusterings.

c: Number of pairs of elements that are in the same cluster in the first clustering but in different clusters in the second clustering.

d: Number of pairs of elements that are in different clusters in the first clustering but in the same cluster in the second clustering.

$\binom{n}{2}$: Total number of possible pairs of elements, calculated as n(n-1)/2.

Loss

For Data(X(m,n), D(m,2), Label(m,1)) and groundtruth Data'(X'(m,n), D'(m,2), Label'(m,1)), L2 difference between X and X' was defined as Loss.

Data analysis

Statistical analyses were performed using the R 4.0.4 software. Paired Student t test were used for comparisons between groups. For all tests, significance was determined with a 95% confidence interval (ns, $P > 0.05$; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$).

Result

The workflow of high Correlated feature pairs combined with spatial morphological alignment(haCCA) for Multi-module Spatial assay Integrating

the haCCA integrate Multi-module Spatial assay Integrating by finding and integrating data between paired spot data from 2 spatial assay. The worklflow include 6 steps: **1: data preparing.** Following the standard processing with Scanpy, the spatial data along with the peak/intensity or feature matrix are extracted, normalized, and scaled. **2: gross alignment by feature point aided affine translation:** Three feature points in both datasets, sharing high similarity in image characteristics, are manually selected to generate an affine matrix. This matrix enables the rotation, scaling, and translation of the spatial data in transcript spatial data to align with the MALDI data. **3: a further alignment by enclosing un-neighboured spot:** After gross alignment, a small subset of spots in transcript spatial data are not overlay with MALID. Moreover, the spot-spot distance between MALID and transcript spatial data is not minimized. We extract outlier spot in transcript spatial data, which is identified as having no neighbour spot in MALID within certain distance.(dist_min). The nearest spot in MALDI for each outlier spot in transcript spatial data is identified as the potential neighbor spot. A translation matrix was found by minimizing the summary of distance between outliers and their potential neighbor spot. After this step, we found a satisfied alignment between MALDI spatial data and transcript spatial data. **4: identification of high correlated feature-peak pairs.** After further alignment, we assume the spatial data of MALID and transcript data are aligned. Anchor spot pairs between MALID and transcript data were admitted if they meet 2 criteria: 1. very close distance(default 0.5) 2. neighbour spots (distance less than certain value, default 2) are with low diversity(Simpson index <0.2). Anchor spots pairs are considered as spot with the same coordinates and their gene feature data and metabolic peak data are applied for pearson correlation test, in order to figure out high correlated feature-peak pairs. **5: combined feature and morphological alignment.** paired peak/intensity-features are

underwent CCA transfer to generate a high-correlated variate as a third dimension. Alignment using spatial data and high-correlated variate between MALD and transcript data is carried out by ICP(Iterative Closest Point) algorithm. **6: integration:** after ICP algorithm, MALDI and transcript data is projected to a 3-dimention, spatial-variable space. For each spot in transcript data, nearest MALD spot data is integrated to and generate multi-module data.

haCCA perform effective integration on pseudo muti-module spatial data

To quantitatively evaluate the performance on integrating spatial transcript data and spatial metabolic data of haCCA, we generate pseudo spatial metabolic data using transcript data. The pseudo spatial metabolic data has a few characters: 1. Small iteration was added to the spatial coordinates of pseudo spatial metabolic data to ensure few overlap with parent spatial transcript data. 2. no mutual features with parent spatial transcript data. We assessed accuracy(AC) of the cluster between true and integrating method identified paired-spots to evaluate performance. In all, pseudo spatial metabolic data was generated form 3 mouse brain, 1 rat brain and 1 human gastric cancer. First we ask how each steps in haCCA contribute to integrating. We found that compared with directalign, which means do-nothing-and-merge strategy, after gross and further alignent the AC increased significantly. Moreover, the feature-aid combined coordinate alignment strategy also increased the AC from 6%-20%, depending on the AC in previous steps. This indicated the efficiency of combing coordinate and feature alignment. Till now there is only very few public method to integrate mutimodule spatial data. Most of them relied on similarity of spatial plot. We found that compared with directalignment, SUTility and ICP, hCCA showed the best performance, with the highest CA and ARI. This results indicating the efficiency of the combing spatial and feature strategy in integrating spatial transcription and metabolic data.

haCCA performs consistent integration of 10X Visium and MALID-MSI spatial data

Next we asked how haCCA performed in real-world generated data. We applied haCCA on a dataset which is consisted of 10X Visium and MALID-MSI from 3 mouse brain. For each mouse brain we included 3 spatial data from 2 section. To be noted, 1 run of 10X Visium and 1 run of MALID-MSI data was performed on the same section, with a novel procedure. The third run of 10X Visium was performed on the neighbour section. STUtility was applied to integrated the 10X Visium and MALID-MSI data carried on the same section by spatial plot alignment, which was used as the ground truth. We used haCCA to perform integration between 10X Visium data and MALID-MSI data from the same section or the neighbour section. It turns out that haCCA is able to effectively integrated 10X Visium and MALID-MSI spatial data from the same section, with the highly accuracy of 87.5%, 79.5% and 87.6%(A,C,E). as we applied STUtility outcomes as ground truth, and STUtility is of inferior performance in simulated data compared with haCCA, it is highly likely that haCCA outcomes is more accurate. In the mission of integrating 10X Visium and MALID-MSI spatial data from neighbour section, where exists a larger difference between spatial shape. haCCA effectively adjust the spatial shape of 10X Visium data, making it more assemble to the MALID-MSI data, and performed highly consistent integration, shown by the similar pattern of the metabolic cluster distribution between MALID-MSI data and haCCA integrated data.(B,D,F) Using dopamine and 3-MT as example, it turns out haCCA is able to reconstruct the metabolic data in integrated data. It leverage the connivance to identify the genes associated with metabolite. Not surprisingly, genes associated with dopamine metabolism, such as Th and Slc6a3, shown an overlapped distribution pattern with dopamine.

These results indicated haCCA is able to perform intergration between 10X Visium and MALID-MSI

spatial data effectively.

haCCA generated muti-module data allows integrative analysis of transcript and metabolite

After performing haCCA workflow, We performed WNN(weighted nearest neighbors) on the muti-module data contain spatial, transcript and metabolite data. Compared with using transcript or metabolite data alone, integrated data enable a more dedicate exploration to the heterogeneity of mouse brain by finding more clusters. Different umap plots were shown when using transcript, metabolite or muti-module data for embedding, and a sanky plot shows the relationship between them. Spatially projection was shown to visualize the distribution of integrated data derived clusters(A,B). We further found the marker gene or metabolites for each cluster(C,D). generally, integrated data has the highest heterogeneity, while metabolic data has the lowest. Transcript data derived cluster is highly agreed with the integrated data. These results indicated that haCCA allows integrative analysis of transcript and metabolite data. The heterogeniety of spatial data is mainly from transcription level.

Discussion

Spatial muti-module technologies requires efficient and accurate integration of spatial data. Although recently some integration workflow has been developed for spatial transcription data(13-15), their application to muti-module data is limited due to lack of mutual features. Here we proposed haCCA, a method to perform high correlated feature pairs assisted morphological alignment and integration between spatial transcription data and MSI-MALDI data. haCCA is composed of 2 main steps: a modified morphological alignment of spatial coordinates, following by identifying high correlated feature pairs and turning them into new high correlated variables. Using aligned spatial coordinates and high correlated variable to perform integration with more accuracy than only using morphological alignment.

By using spatial transcription data and pseudo paired MSI-MALDI data, we shown that haCCA outperformed morphological alignment, represented by STUtility by 6-20%, evaluating by label transfer accuracy, label transfer ARI and loss. Besides, the modified morphological alignment of spatial coordinates of haCCA itself is able to outperform STUtility, and the combing alignment is able to further increase the performance. It indicated the efficiency of combing alignment strategy of haCCA. When using real world generated data of normal tissue and tumor, haCCA successfully integrated MSI-MALDI data into spatial transcription data, and generated simultaneous gene expression as well as bio molecule profiling. Such multi-module spatial data allows more sophisticated deciphering to heterogeneity, for example, we identified different metabolic states in tumor region from ICC model. It also facilitates both gene and bio molecule marker detection for certain clusters.

haCCA may be improved in the following aspects. First, only spatial transcription and spatial metabolome data has been used. haCCA can expend its capability to integrate more spatial techniques, such as image-based sequencing, spatial proteinomes or histological images. Second, how haCCA behaves on higher resolution data should be further explored.

Figure legend

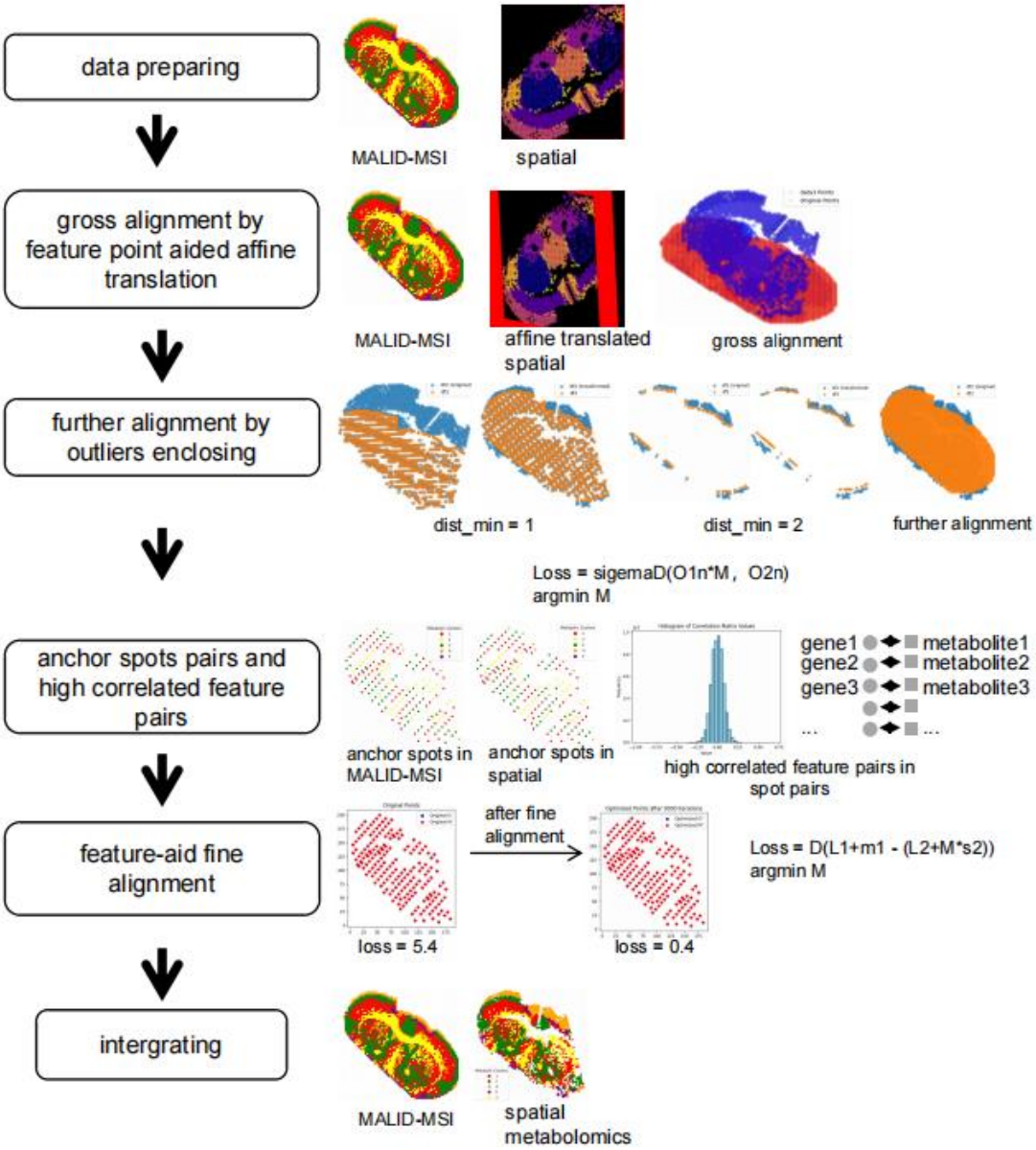
Figure 1: an overview illustrator of **h**igh Correlated fe**a**ture pairs-aid combined **CCA**-ICP **M**ulti-module **S**patial assay **I**ntegrating. The procedure consist of 6 main steps: **A:** Data preparing: Peak-intensity in MALID-MSI and gene count in 10X Visium spatial transcriptome were processed and clustering with scanpy. Spatial matrix(2*N) and feature matrix(m*N) was generated for aligning **B:**

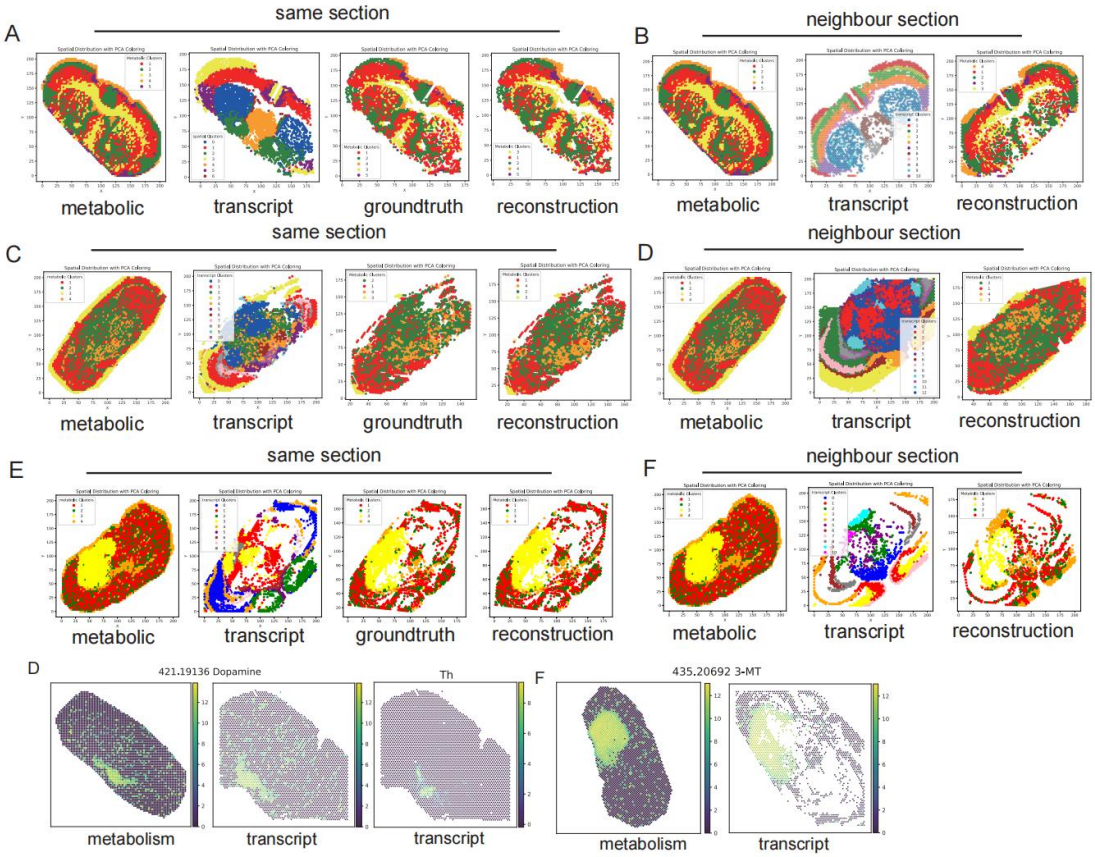
Gross spatial alignment: Feature points were selected based on mutual spatial region with high similarity between MALID-MSI and 10X Visium spatial transcriptome. Affine translation matrix was applied on Spatial matrix of 10X Visium spatial transcriptome to achieve maximum similarity, shown in gross alignment. Red refers to MALID-MSI and blue refers to affine translated 10X Visium spatial transcriptome. **C:** Further alignment by outliers enclosing: after gross alignment, spots in 10X Visium spatial transcriptome within certain distance(dist_min) to MALID-MSI were one-to-one aligned. If no MALID-MSI spot was found, then such spots in 10X Visium spatial transcriptome was defined as outliers. Distance of outliers to nearest MALID-MSI spots was calculated and a transfer matrix was found to minimize the sum. Then the transfer matrix was applied to the addine transferred 10X Visium spatial transcriptome. As shown in C, after further alignment a maxim alignment between MALID-MSI and 10X Visium spatial was achieved. **D**: Anchor spot pairs and high correlated feature pairs identification: After further alignment, anchor spots, defined as spots pairs with the same location(distance < 0.1) were separately identified in Spatial matrix of MALID-MSI and 10X Visium spatial. Then high correlated features between anchor spots pairs was found by pearson correlation test. **E**: feather-aid fine alignment: spatial matrix and feature matrix of anchor spot pairs of MALID-MSI and 10X Visium spatial were contact to one spatial-feature matrix(Msf), a transfer matrix M was found to minimize the sum of distance define as L1+Msf-M*Msf. **F**: Integrating: after fine alignment, feature matrix in MALID was allocated to the nearest spot in 10X Visium spatial, where distance was defined as Ms(spatial matrix)+ M*Msf.

Figure 2: using APC-MSI to integrate MALID-MSI and 10X Visium data. A-F: spatial plot of MALID-MSI(metabolic), 10X Visium(transcript), groundtruth and metabolic features of integrated data(reconstruction) of a same coronal section of mouse brain(A,C,E), or neighbour section(B,D,F).

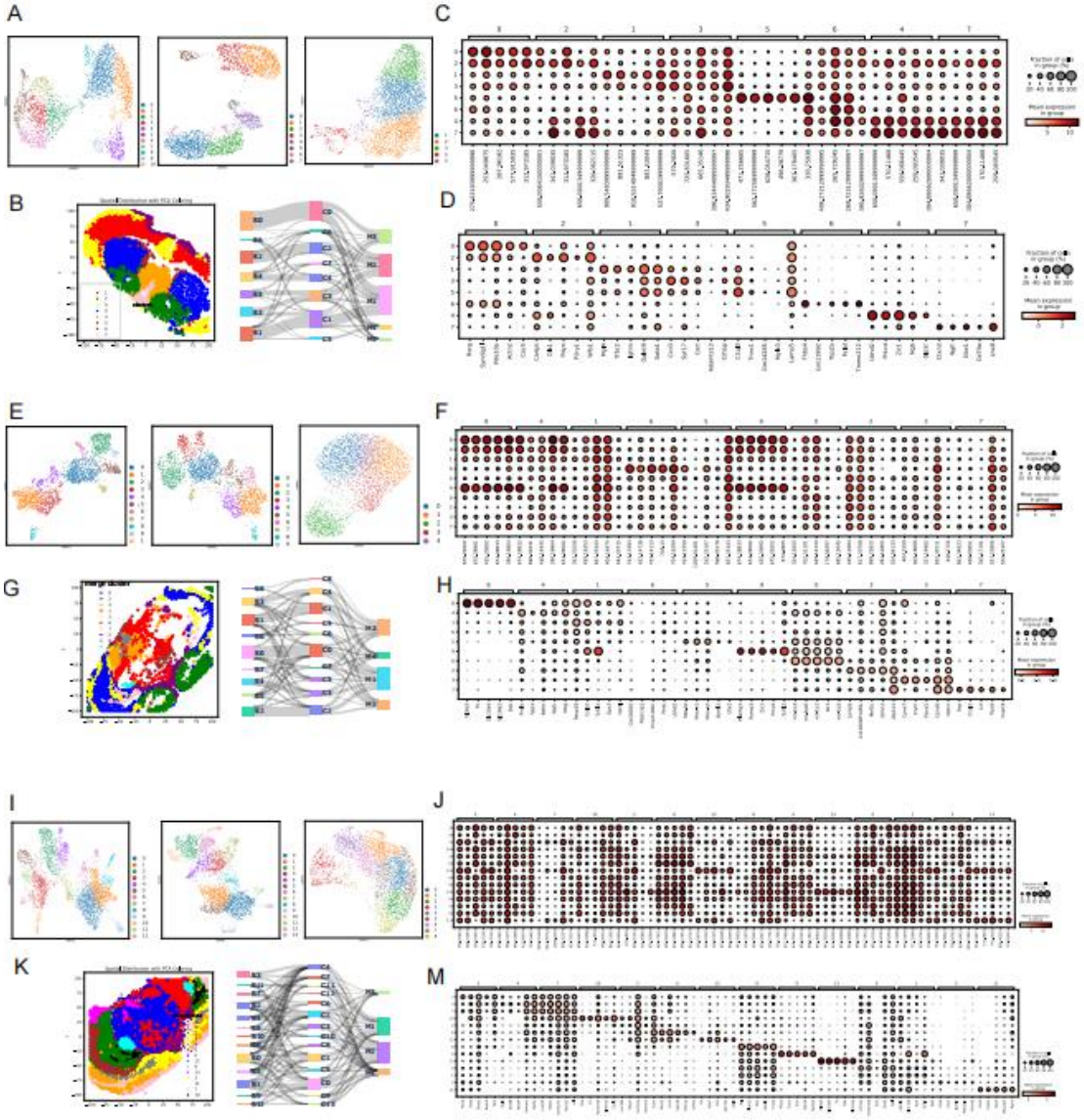
Figure 3: APC-MSI deciphering spatial heterogeneity with higher resolution. Umap plot showing cluster of mouse brain using 10X Visium(left), MALID-MSI(right) and APC-MSI generated muti-module data(middle)(A,E,I). cluster of APC-MSI generated muti-module data was projected to 10X Visium spatial plot(B,G,K left). A sanky plot showed the relations of 3 clusters(B,G,K right). marker gene (D,H,M)or metabolites(C,F,J) of clusters in APC-MSI generated muti-module data of mouse brain.

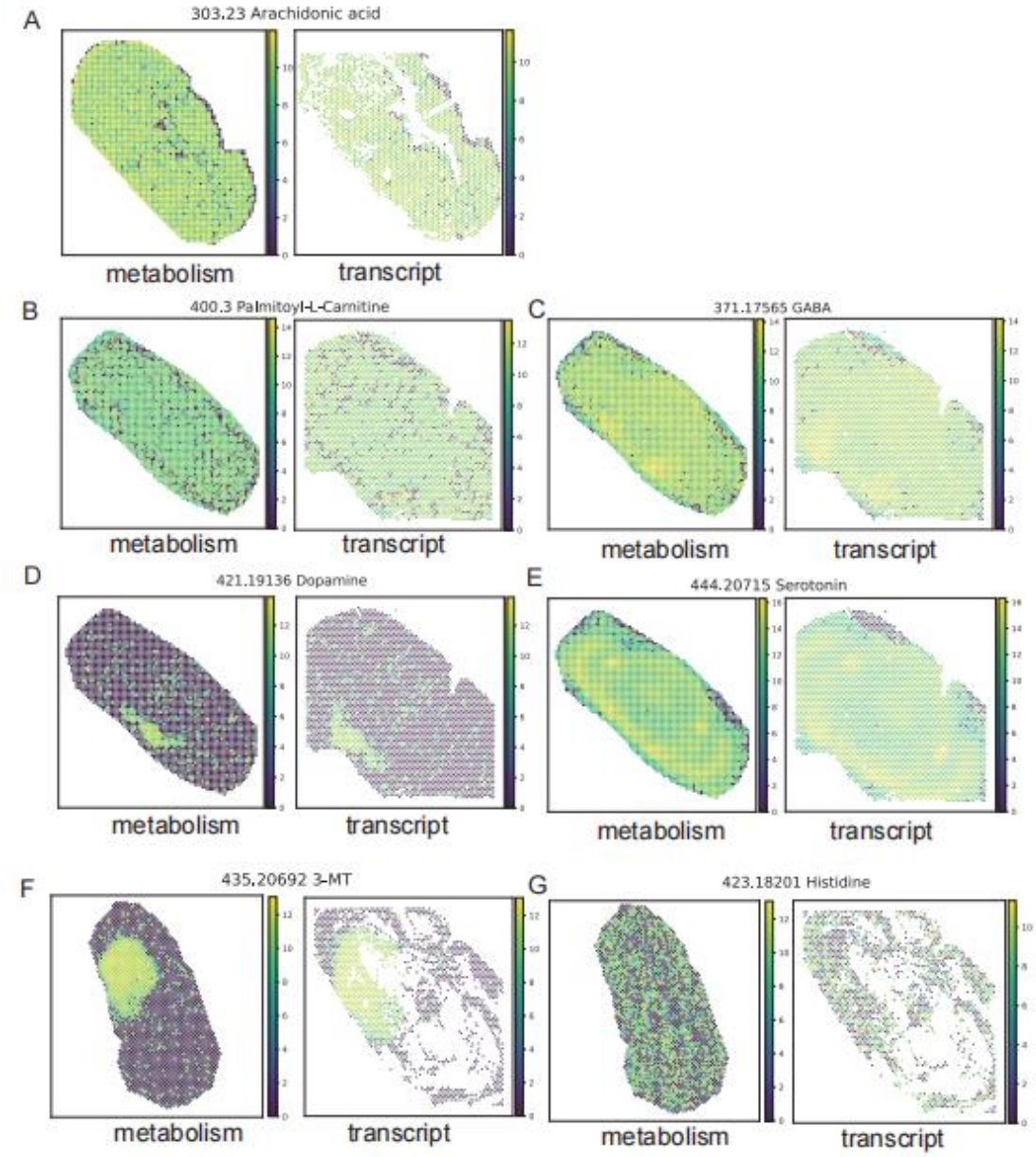
Figure 4: APC-MSI allows metabolite display in 10X Visium. A-G: Similar distribute pattern of representative metabolite in MALID-MSI(left, metabolism) and APC-MSI generated muti-module data(right).


data preparing
MALID-MSI
spatial
gross alignment by feature point aided affine translation
MALID-MSI
affine translated spatial
gross alignment
further alignment by outliers enclosing
dist_min = 1
dist_min = 2
further alignment
Loss = sigemaD(O1n*M , O2n)
argmin M
anchor spots pairs and high correlated feature pairs
anchor spots in MALID-MSI
anchor spots in spatial
high correlated feature pairs in spot pairs
gene1
gene2
gene3
metabolite1
metabolite2
metabolite3
feature-aid fine alignment
after fine alignment
loss = 5.4
loss = 0.4
Loss = D(L1+m1 - (L2+M*s2))
argmin M
intergrating
MALID-MSI
spatial metabolomics


A
same section
metabolic
transcript
groundtruth
reconstruction
B
neighbour section
metabolic
transcript
reconstruction
C
same section
metabolic
transcript
groundtruth
reconstruction
D
neighbour section
metabolic
transcript
reconstruction
E
same section
metabolic
transcript
groundtruth
reconstruction
F
neighbour section
metabolic
transcript
reconstruction

D
421.19136 Dopamine
Th
metabolism
transcript
transcript
F
435.20692 3-MT
metabolism
transcript


A
B
C
D
E
F
G
H
I
J
K
M



1. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet. 2023;24(8):494-515.

2. Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. Nat Biotechnol. 2021;39(10):1202-15.

3. Vicari M, Mirzazadeh R, Nilsson A, Shariatgorji R, Bjarterot P, Larsson L, et al. Spatial multimodal analysis of transcriptomes and metabolomes in tissues. Nat Biotechnol. 2023.

4. Ravi VM, Will P, Kueckelhaus J, Sun N, Joseph K, Salie H, et al. Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. Cancer Cell. 2022;40(6):639-55 e13.

5. Sun C, Wang A, Zhou Y, Chen P, Wang X, Huang J, et al. Spatially resolved multi-omics highlights cell-specific metabolic remodeling and interactions in gastric cancer. Nat Commun. 2023;14(1):2692.

6. Zhi Y, Wang Q, Zi M, Zhang S, Ge J, Liu K, et al. Spatial Transcriptomic and Metabolomic Landscapes of Oral Submucous Fibrosis-Derived Oral Squamous Cell Carcinoma and its Tumor Microenvironment. Adv Sci (Weinh). 2024;11(12):e2306515.
7. Zheng P, Zhang N, Ren D, Yu C, Zhao B, Zhang Y. Integrated spatial transcriptome and metabolism study reveals metabolic heterogeneity in human injured brain. Cell Rep Med. 2023;4(6):101057.
8. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-902 e21.
9. Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. Jointly defining cell types from multiple single-cell datasets using LIGER. Nat Protoc. 2020;15(11):3632-62.
10. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):111.
11. Singh S, Singh PK, Jha A, Naik P, Joseph J, Giri S, et al. Integrative metabolomics and transcriptomics identifies itaconate as an adjunct therapy to treat ocular bacterial infection. Cell Rep Med. 2021;2(5):100277.
12. Iturria-Medina Y, Adewale Q, Khan AF, Ducharme S, Rosa-Neto P, O'Donnell K, et al. Unified epigenomic, transcriptomic, proteomic, and metabolomic taxonomy of Alzheimer's disease progression and heterogeneity. Sci Adv. 2022;8(46):eabo6764.
13. Guo T, Yuan Z, Pan Y, Wang J, Chen F, Zhang MQ, et al. SPIRAL: integrating and aligning spatially resolved transcriptomics data across different experiments, conditions, and technologies. Genome Biol. 2023;24(1):241.
14. Liu X, Zeira R, Raphael BJ. Partial alignment of multislice spatially resolved transcriptomics data. Genome Res. 2023;33(7):1124-32.
15. Zeira R, Land M, Strzalkowski A, Raphael BJ. Alignment and integration of spatial transcriptomics data. Nat Methods. 2022;19(5):567-75.