Batch alignment of mass spectrometry imaging (MSI) metabolome through data integration

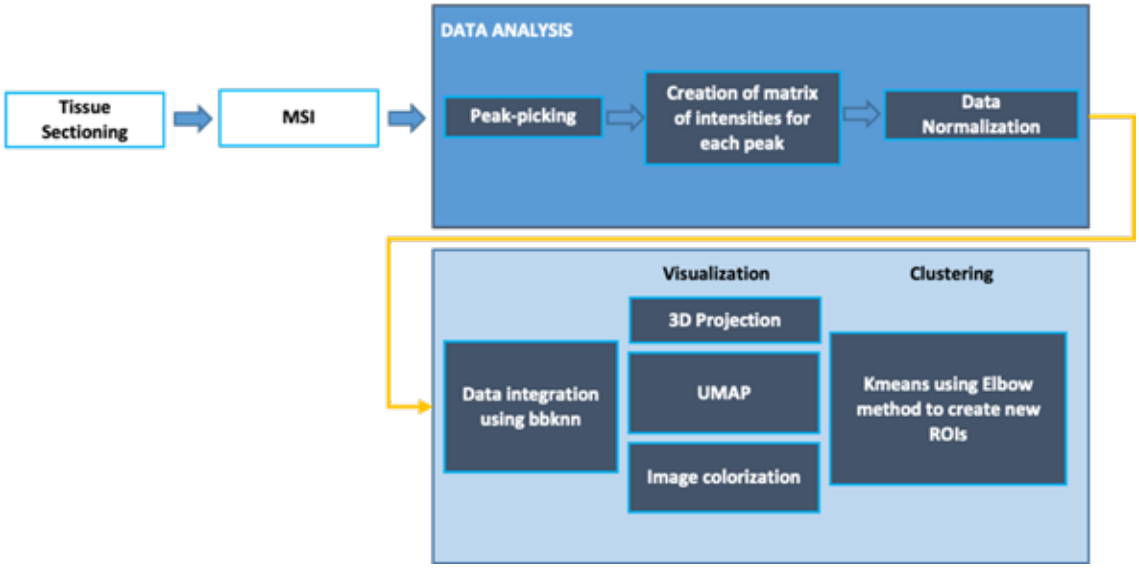
SUMMARY

Mass spectrometry imaging (MSI) has become an indispensable tool for spatially resolved molecular investigation of tissues in biomedical research. While many of the initial technical challenges have now been resolved, so-called batch effects appear to impede reliable comparison of data from large-scale studies performed in translational clinical research for biomarker discovery or multivariate classification. In this paper, we discuss the development of a batch correction method to minimize this detrimental effect, allowing the reliable identification of biological clusters and their comparison.

APPROACH

To enable the reliable interpretation of subtle biological differences between different tissue sections, and to avoid false discoveries in MALDI-MSI studies, here we present the development and application of a new procedure (Figure 1) which can correct for batch effect. An extremely fast graph-based data integration algorithm, BBKNN (batch balanced k nearest neighbors), is applied. Its output can be immediately used for dimensionality reduction (UMAP) and clustering.

FIGURE 1. Presentation of the data analysis workflow with data integration algorithm.



CASE ILLUSTRATION STUDY

To comprehensively evaluate the efficiency of BBKNN data integration, we used four colorectal cancer (CRC) tissue samples to perform an unbiased spatial metabolic characterization using MALDI-Mass Spectrometry Imaging (MSI) to identify *in-situ* biomarkers related to disease progression.

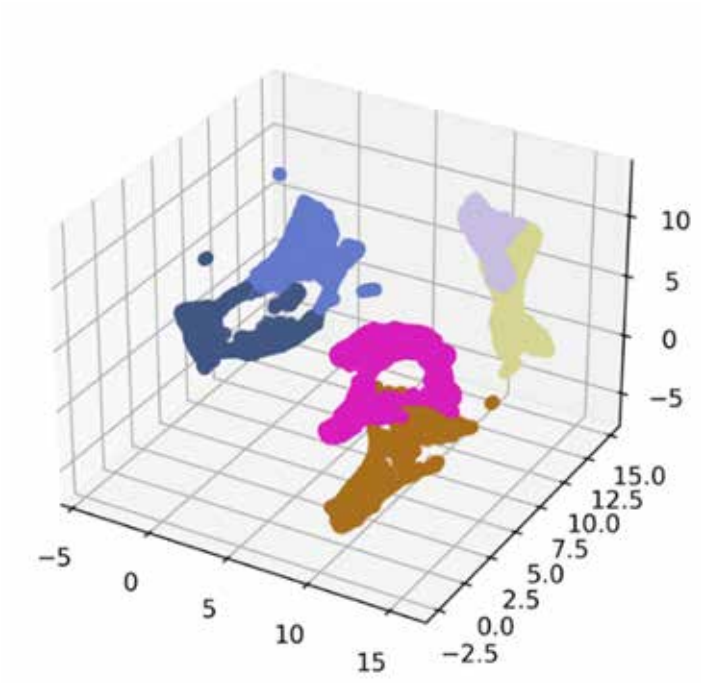
Automatic regions of interest (ROIs) detection based on biological similarities

Individual tissues were imaged, and as a first step batch integration was applied on all tissue MSI data. Its output was compared to the output without batch integration for improving the identification of the biological clusters. UMAP is applied on the integrated data (Figure 2). The UMAP 3D-projection of the metabolic data for all samples is then used for KMEANS clustering and transposition on tissue.

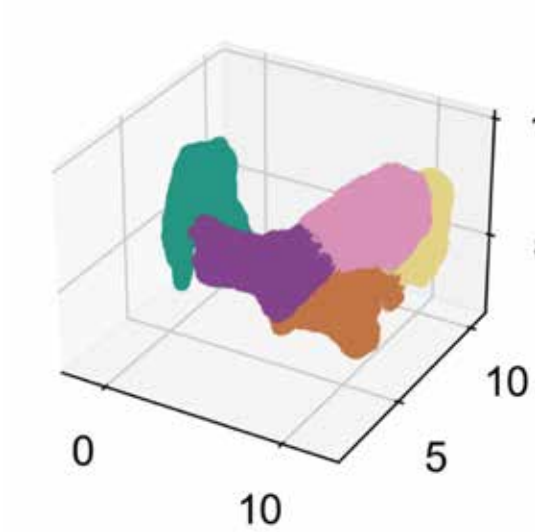
Once the clusters are transposed and colorized in each tissue section (Figure 3), the better differentiation of the biological clusters appears to be easily recognizable when batch integration is applied. Indeed, one can see that the clusters are shared between samples with batch data integration. In contrast, clusters without the batch integrations appear to be remated to the sample itself and therefore do not represent a biological variation in metabolic content.

FIGURE 2. UMAP of metabolites clusters present in the colorectal tissue.

UMAP without batch correction



UMAP with batch correction



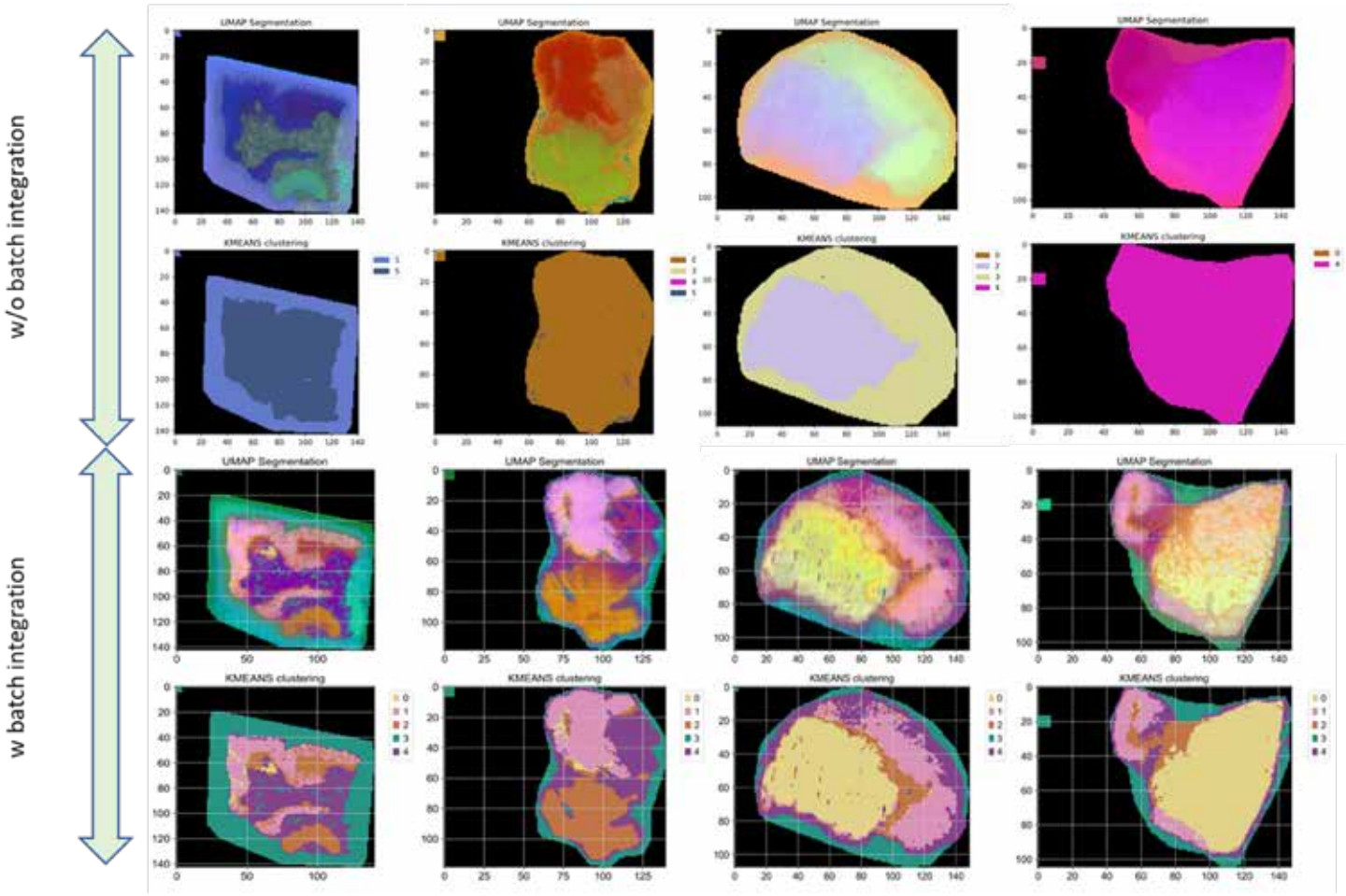
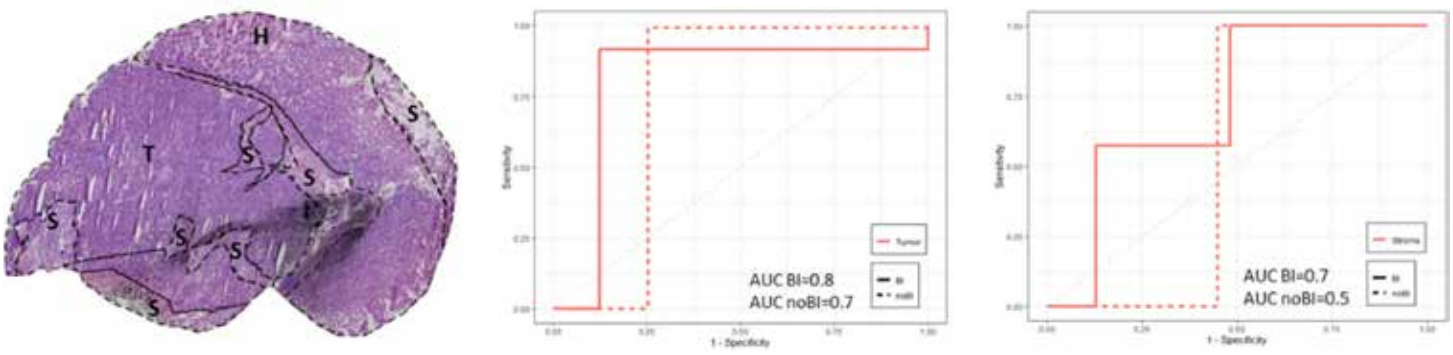


FIGURE 3. Color-based clustering comparison of MSI metabolic data.

Comparison to pathological annotations

The clusters defined above were compared to the pathological tissue annotations considered as ground truth (Figure 4). Each cluster was designated either as Stroma, Tumor or Healthy to match the ground truth. The multi-class ROC curves show that the clusters defined with batch integration better represent the biological heterogeneity of the tissue than the clusters defined without. For the Tumor class, the AUC score of the batch integration model exceeded 0.8, which indicated a very good performance and was also above the AUC score of the non-batch integration model. The same conclusion is observed for the Stroma class. Including the batch integration in the workflow produces more accurate and biologically relevant results.



H: Healthy
S: Stroma
T: Tumor

FIGURE 4. Comparison of Tumor and Stroma histological regions ROC curves with and without batch integration for one grade III CRC sample.

PERSPECTIVE

The meaningful analysis of data generated by large-scale studies is critically dependent on the statistical power required for systems biology and translational medicine studies. However, great power comes with batch effect baggage and requires specialized tools to handle this problem. Across the batches, even if the data are generated from the same machine, we often observe different data characteristics, and while normalization makes the samples more comparable, it only aligns their global patterns. Therefore, batch effects affecting specific metabolites or metabolite groups might still represent a major source of variance even after normalization. Here, we discuss the application of established approaches for batch effect adjustment and provide guidelines and tools to make the extraction of true biological signal from large studies more robust and transparent, ultimately facilitating reliable and reproducible research.

For more information visit www.aliribio.com.

