Unraveling the stereoscopic gene transcriptional landscape of zebrafish using FishSED, a fish spatial expression database with multispecies scalability

Cheng Guo[1, 2†], Weidong Ye[1, 2†], Danying Cao[1, 2], Mijuan Shi[1, 3, *], Wanting Zhang[1,3], Yingyin Cheng[1,3], Yaping Wang[1, 2, 3], Xiao-Qin Xia[1, 2, 3, *]

[1]*Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China*

[2]*College of Advanced Agricultural Sciences, University of Chinese Academy of Sciences, Beijing, China*

[3]*The Innovative Academy of Seed Design, Chinese Academy of Sciences, Beijing 100101, China*

[†]These authors share first authorship.

[*]Corresponding authors: E-mails: shimijuan@ihb.ac.cn; xqxia@ihb.ac.cn.

Dear Editor,

The expression profile demonstrates how each gene was expressed in the target tissue or cell at a given time. Cell heterogeneity is a biological concept characterized by the presence of numerous cells with varying functions and expression profiles within organ/tissue samples. There is also cell-to-cell signaling, which implies that the extracellular environment and spatial location influence intracellular transcription, resulting in spatial heterogeneity of gene expression within the same cell type. In terms of methodology, traditional transcriptome studies frequently ignore the complexity within tissues and study them as a collection of homogeneous cells. Despite the fact that the single-cell transcriptome (scRNA-seq) includes heterogeneous cells, spatial information is lost during sample processing (Longo et al., 2021; Regev et al., 2017). Spatial transcriptome technology allows for the analysis of gene expression in a variety of cells while preserving cellular spatial information, thereby improving the presentation of the 3D transcriptional landscape of study targets via two variables: heterogeneous cells and cellular spatial location (Liao et al., 2021). Spatial transcriptome technologies have been widely used in biological research fields such as development (Wang et al., 2022), disease (Ou et al.), regeneration (Burkhard and Bakkers, 2018), and so on.

Several high-throughput sequencing technologies, such as spatial transcriptomics (ST), Geo-seq, and Tomo-seq, have been used in spatial transcriptomic research to interpret spatial patterns of gene expression in samples at high-resolution. Both 10x Genomics Visium and Stereo-seq can achieve single-cell spatial granularity. As spatial transcriptome data has accumulated, corresponding online databases such as SpatialDB (Fan et al., 2020), STOmicsDB, SODB (Yuan et al., 2023), and others have emerged to aid researchers in spatial transcriptome analysis.

Because of its short life cycle, ease of mass reproduction and rearing, and ease of genetic manipulation, the zebrafish (*Danio rerio*) is a significant model species that is frequently used in basic biology research. Furthermore, because zebrafish share 87% of the human protein sequence (Klee et al., 2012), they are becoming more important in medical research. Despite the fact that zebrafish spatial transcriptome data is rapidly expanding, the current spatial transcriptome database only contains insufficient and fragmented zebrafish data. Researchers could only obtain limited dataset information from existing databases. The various technologies used for zebrafish spatial transcriptome data from various laboratory sources have increased data heterogeneity and reuse complexity. To assist researchers working on fish spatial transcriptomes, we created the open access fish spatial expression database FishSED (http://bioinfo.ihb.ac.cn/fishsed).

In this study, we compiled existing articles and raw data on the zebrafish spatial transcriptome. We collected 3D gene expression profiles from 5 sequencing technologies and built FishSED, the only database dedicated to fish spatial transcriptome data and the largest and most comprehensive online resource for zebrafish spatial transcriptome data. FishSED provides a variety of visualization services based on various sequencing technologies, as well as searching multiple gene expression patterns and mapping across datasets, making it easy for researchers to conduct comparative analysis. Although zebrafish is the only fish with spatial transcriptome data to date, FishSED is designed to support multiple species, and once spatial transcriptomic techniques are applied to other fishes, a combobox will be available for users to select target fish species.

The database currently contains 56 published datasets, of which 9 are based on the highest resolution 10x Genomics Visium and Stereo-seq, and 44 are based on Tomo-seq. These datasets cover 4 different zebrafish tissue types, including 41 embryonic datasets that span different stages of embryo development (Figure S1A in Supporting Information). The web interface of FishSED has five main pages: "Home", "Search", "Expression", "BLAST", and "Browse" (Figure S2D in Supporting Information). Users can browse all datasets in the main menu's "Browse" page for sample information, sequencing technology type, data source, and project information, among other things. All datasets are free to download (Figure S3 in Supporting Information).

The three basic types of search functionality are comprehensive search, advanced search, and homology search (BLAST). Entering a gene symbol or a gene/project/dataset ID in the search box in the middle of the "Home" page (Figure S1B in Supporting Information) allows users to conduct a quick comprehensive search. The "Search" menu item in the main menu implements advanced search, providing more detailed gene/project/dataset information options for users to fill out, and the page will return a list of results that satisfy all inputs (Figure S4 in Supporting Information).

The BLAST tool has been integrated into the main menu "BLAST" page (Figure S5 in Supporting Information). Users can set parameters, paste or upload sequence files, and align with all transcript sequences in FishSED to generate a descending list of similarity. BLAST output is tabular with comment lines, and FishSED supports one-click download of aligned sequences.

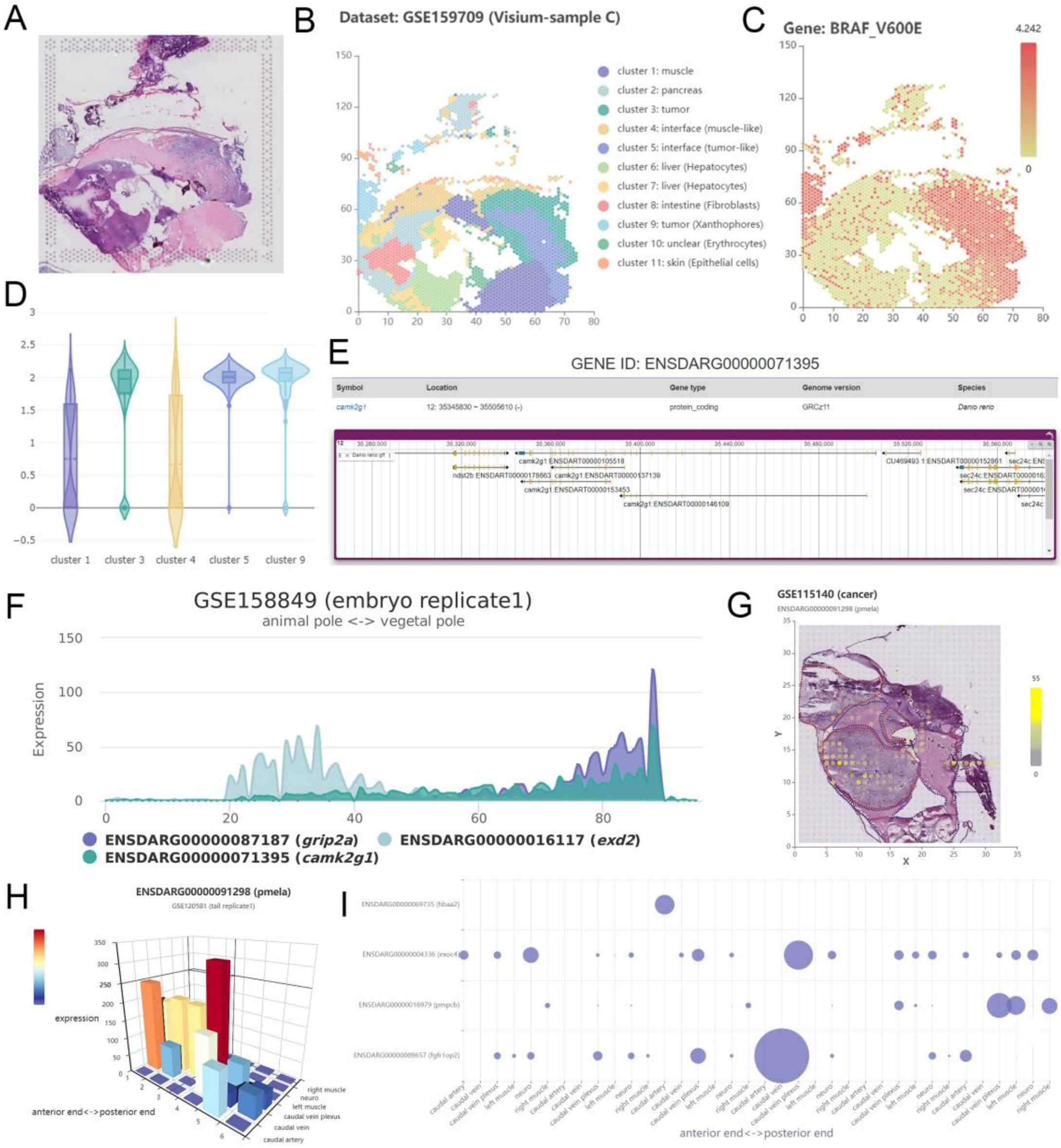
The comprehensive search and advanced search can be used to access the detailed page of a project, dataset, or gene. Cells in the "interface" state were discovered in the microenvironment where cancer cells met non-cancer cells during the study of tumor progression (Hunter et al., 2021). This "interface" region is histomorphologically indistinguishable from muscle tissue, but its transcriptional status is highly correlated with tumors. When users search for projects on the advanced search page using the key word "interface", a project (GSE159709) is returned, and its detailed page (Figure S6 in Supporting Information) provides information about the articles that are associated with the project. The abstract of the article is used to generate a word cloud, and the distribution of samples is shown in a bubble chart. Users can also view the information for each dataset covered by the project in a list on this page.

To see the expression patterns of a target gene in the interface, tumor, and other regions, users can click on the ID of a dataset, for example, GSE159709 (Visium-sample C), to go to its detailed page (Figure S7 in Supporting Information), which displays more information such as sample information, the type of sequencing technique used, the dataset's source, and so on. This page allows users to enter multiple gene IDs or gene symbols at the same time to view their expression patterns in this dataset. Users can also specify parameters in the form, such as exact or fuzzy matching, to achieve the desired result. A sample slice plot, a scatter plot of cell fractionation data, and a scatter and violin plot of expression patterns (Figure 1A-D) will be returned. We looked for the expression pattern of the previously identified tumor marker *BRAF*^{V600E} (Figure 1A-D). In comparison to the sample slice plot (Figure 1A), the scatterplot of cell fractionation data can show the spatial distribution of various tissues (Figure 1B), including tumor (cluster 3, 9) and interface (cluster 5) regions. A scatterplot for the expression pattern (Figure 1C) shows a high expression (red area) in both the tumor and the interface region, illustrating the changes in the expression level of *BRAF*^{V600E} at different spatial sites. The violin plot directly shows *BRAF*^{V600E} expression levels in different groups (Figure 1D), and its expression levels tend to be consistent in tumor (cluster 3, 9) and interface (cluster 5) groups.

In addition to the methods mentioned above, users can visit the gene detail page to learn more about the target gene's functions. Zebrafish have been widely used in developmental biology research, and the large Tomo-seq datasets can assist researchers in exploring gene transcription patterns. For example, the gene *camk2g1* promotes the activity of calmodulin-dependent protein kinase, which is involved in pronephros structural organization and acts upstream of or within animal organ development and protein autophosphorylation. A simple search for *camk2g1* will bring up its detail page (Figure S8 in Supporting Information), where users can look up its location, gene type, genome version, and other basic information, as well as visualize the structural components of the gene using a JBrowse2-embedded window (Figure 1E). Users can browse the PPI network of *camk2g1* and enriched GO terms and KEGG pathways based on the genes in the network in the "Function" section. Users can also browse the expression of this gene by selecting sequencing technologies and datasets in the "Expression" section. Tomo-seq is the sequencing technology with the most datasets in FishSED, and relevant datasets covering various stages of zebrafish embryo development were organized using a folded tree. The gene expression curves in these datasets will be displayed at the bottom of this page by clicking on the mirror buttons to the left of the dataset ID of interest.

The "Expression" menu is used to compare the expression patterns of multiple genes. For example, users can select one dataset (or more) from the embryonic stage, such as GSE158849 (embryo replicate1) at the one-cell stage, and enter *camk2g1*, *grip2a* and *exd2*. As a vegetally localized gene, *grip2a* is thought to enable signaling receptor complex adaptor activity, whereas *exd2*, which localizes to the animal pole, is thought to enable 3'-5' exonuclease activity. As a result, *camk2g1* expression resembled *grip2a* but did not overlap with *exd2*, indicating that *camk2g1* plays a more important role in the vegetal pole (Figure 1F). When compared to the other three sequencing technologies, ST and Geo-seq datasets have distinct characteristics. FishSED illustrates the expression patterns using sample slice plots and scatter plots for the former (Figure 1G), and 3D bar/bubble charts for the latter (Figure 1H-I) based on the different spatial information resolutions.

Finally, we created FishSED, an easy-to-use online platform to help researchers conduct more effective studies of zebrafish gene spatial expression patterns. FishSED covers a broader range of sequencing technologies and a greater number of datasets for zebrafish data than the four previously published spatial transcriptome databases (Figure S1C in Supporting Information). With the rapid growth of spatial transcriptome data, we will track and integrate spatial transcriptome-related data in fish, as well as maintain and update FishSED on a regular basis, to provide the most recent and comprehensive public resources for spatial transcriptome research in fish.



Symbol	Location	Gene type	Genome version	Species
camk2g1	12: 35345830 ~ 35505610 (-)	protein_coding	GRCz11	*Danio rerio*

Figure 1. A, The sample slice plot of dataset GSE159709 (Visium-sample C). The slice was obtained by transversely cutting the whole fish to a thickness of 10 μm, and the slice plot from left to right represents the direction from the ventral side to the dorsal side of the fish body, and from top to bottom represent the direction from the right side to the left side of the fish body. The chip size of 10x Genomics Visium is 6.5mm × 6.5mm. B, The scatterplot of cell fractionation data of dataset GSE159709 (Visium-sample C). The x-axis and y-axis are in the same direction as those of the slice plot (Figure 1A). C, The scatterplot for the expression pattern of *BRAF*^{V600E} in dataset

GSE159709 (Visium-sample C). The direction of the x-axis and y-axis is the same as those of Figure 1A. D, The violin-box plot of expression levels of the gene *BRAF*^{V600E} in muscle, tumor and interface groups in dataset GSE159709 (Visium-sample C). The x-axis represents the cell clusters of muscle, tumor and interface, and the y-axis represents expression levels. E, A part of the detailed page of gene *camk2g1* (gene id: ENSDARG00000071395). F, Visualization of gene expression patterns in the Tomo-seq dataset. The x-axis represents the orientation of the embryo from the animal pole to the vegetal pole, as stated in the subtitle. Note that the magnitude of the x-axis represents the position of each section and the actual thickness of each section varies from dataset to dataset. G, Visualization of a ST-based dataset. This is the spatial expression pattern of the gene *pmela* (gene id: ENSDARG00000091298) in a section of zebrafish transplanted tumor from dataset GSE115140 (cancer). Gene *pmela* enables ATP binding activity and metal ion binding activity and acts upstream of or within several processes, including amyloid fibril formation, melanocyte differentiation and melanosome organization. Each spot printed on the array is 100μm in diameter and covers an area of 6.2 mm × 6.6 mm. H-I, Visualization of expression levels respectively for a single gene and multiple genes in the Geo-seq dataset.

Acknowledgments

This research was funded by the National Key R&D Program of China (Grant No. 2021YFD1200804 and 2018YFD0901201) and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA24010206). We thank the Analysis and Testing Center at Institute of Hydrobiology for technical supports. This computational work in this study was supported by the Wuhan Branch, Supercomputing Center, Chinese Academy of Sciences, China.

Author Contributions

Cheng Guo: Conceptualization, methodology, formal analysis, investigation, writing—original draft preparation. Weidong Ye: Conceptualization, methodology, formal analysis, investigation. Danying Cao: Writing—original draft preparation. Mijuan Shi: Conceptualization, investigation, writing—review and editing. Wanting Zhang: data curation. Yingyin Cheng: validation, resources, supervision. Yaping Wang: project administration, funding acquisition. Xiao-Qin Xia: Conceptualization, investigation, writing—review and editing, project administration, funding acquisition.

Reference

Burkhard, S.B., and Bakkers, J. (2018). Spatially resolved RNA-sequencing of the embryonic heart identifies a role for Wnt/beta-catenin signaling in autonomic control of heart rate. Elife 7.

Fan, Z., Chen, R., and Chen, X. (2020). SpatialDB: a database for spatially resolved transcriptomes. Nucleic Acids Res 48, D233-D237.

Hunter, M.V., Moncada, R., Weiss, J.M., Yanai, I., and White, R.M. (2021). Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. Nat Commun 12, 6278.

Klee, E.W., Schneider, H., Clark, K.J., Cousin, M.A., Ebbert, J.O., Hooten, W.M., Karpyak, V.M., Warner, D.O., and Ekker, S.C. (2012). Zebrafish: a model for the study of addiction genetics. Hum Genet 131, 977-1008.

Liao, J., Lu, X., Shao, X., Zhu, L., and Fan, X. (2021). Uncovering an organ's molecular architecture at single-cell resolution by spatially resolved transcriptomics. Trends Biotechnol 39, 43-58.

Longo, S.K., Guo, M.G., Ji, A.L., and Khavari, P.A. (2021). Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat Rev Genet 22, 627-644.

Ou, Z.H., Lin, S.T., Qiu, J.Y., Ding, W.C., Ren, P.D., Chen, D.S., Wang, J.X., Tong, Y.H., Wu, D., Chen, A., Deng, Y., Cheng, M.N., Peng, T., Lu, H.R., Yang, H.M., Wang, J., Jin, X., Ma, D., Xu, X., Wang, Y.Z., Li, J.H., and Wu, P. Single-nucleus RNA sequencing and spatial transcriptomics reveal the immunological microenvironment of cervical squamous cell carcinoma. Advanced Science 9, e2203040.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Gottgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J.C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C.P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T.N., Shalek, A., Shapiro, E., Sharma, P., Shin, J.W., Stegle, O., Stratton, M., Stubbington, M.J.T., Theis, F.J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N., and Human Cell Atlas Meeting, P. (2017). The human cell atlas. Elife 6.

Wang, M., Hu, Q., Lv, T., Wang, Y., Lan, Q., Xiang, R., Tu, Z., Wei, Y., Han, K., Shi, C., Guo, J., Liu, C., Yang, T., Du, W., An, Y., Cheng, M., Xu, J., Lu, H., Li, W., Zhang, S., Chen, A., Chen, W., Li, Y., Wang, X., Xu, X., Hu, Y., and Liu, L. (2022). High-resolution 3D spatiotemporal transcriptomic maps of developing *Drosophila* embryos and larvae. Dev Cell 57, 1271-1283 e1274.

Yuan, Z., Pan, W., Zhao, X., Zhao, F., Xu, Z., Li, X., Zhao, Y., Zhang, M.Q., and Yao, J. (2023). SODB facilitates comprehensive exploration of spatial omics data. Nature Methods 20, 387-399.