Published online 12 November 2019 *Nucleic Acids Research, 2020, Vol. 48, Database issue* ***D233–D237***
doi: 10.1093/nar/gkz934

SpatialDB: a database for spatially resolved transcriptomes

Zhen Fan[1], Runsheng Chen[1,2,*] and Xiaowei Chen[1,*]

[1]Center for High Throughput Sequencing, Core Facility for Protein Research, Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China and [2]University of Chinese Academy of Sciences, Beijing, 100049, China

Received August 06, 2019; Revised September 30, 2019; Editorial Decision October 04, 2019; Accepted October 08, 2019

ABSTRACT

Spatially resolved transcriptomic techniques allow the characterization of spatial organization of cells in tissues, which revolutionize the studies of tissue function and disease pathology. New strategies for detecting spatial gene expression patterns are emerging, and spatially resolved transcriptomic data are accumulating rapidly. However, it is not convenient for biologists to exploit these data due to the diversity of strategies and complexity in data analysis. Here, we present SpatialDB, the first manually curated database for spatially resolved transcriptomic techniques and datasets. The current version of SpatialDB contains 24 datasets (305 subdatasets) from 5 species generated by 8 spatially resolved transcriptomic techniques. SpatialDB provides a user-friendly web interface for visualization and comparison of spatially resolved transcriptomic data. To further explore these data, SpatialDB also provides spatially variable genes and their functional enrichment annotation. SpatialDB offers a repository for research community to investigate the spatial cellular structure of tissues, and may bring new insights into understanding the cellular microenvironment in disease. SpatialDB is freely available at https://www.spatialomics.org/SpatialDB.

INTRODUCTION

Cells are recognized as the fundamental unit of multicellular organisms. The establishment of single-cell RNA sequencing (scRNA-seq) (1,2) has boosted the development of modern cellular and molecular biology. scRNA-seq unmasks cell subsets within the bulk RNA-seq data and provides information of the transcriptome of each individual cell to group subpopulations of cells with similar transcription patterns. Cells are then characterized into various types accordingly. However, when applied to solid tissues, a dissociation step must be performed to obtain cell suspension for the subsequent scRNA-seq analysis. This process loses spatial information, which is critical to cellular fate and property.

To fully understand cell type identity in a multicellular organism (such as human and mouse), one must integrate individual cells' transcriptome profiles with their spatial position in a certain tissue. Several methods have been developed to preserve spatial information. Single molecule RNA fluorescence *in situ* hybridization (smFISH) (3) has been applied to quantitate RNA transcripts at single-cell resolution within a particular tissue context. But only a small number of genes can be measured. To improve the throughput, other imaging-based approaches, such as multiplexed error-robust FISH (MERFISH) (4) and sequential FISH (seqFISH) (5), were emerged. Meanwhile, sequencing-based methods, such as laser capture microdissection sequencing (LCM-seq) (6), Tomo-seq (7), spatial transcriptomics (ST) (8) and Slide-seq (9), take advantages of high-throughput sequencing technology to obtain spatially resolved gene expression even at the single-cell or subcellular resolution.

The development of spatially resolved transcriptomic techniques have profoundly impact many fields, including neuroscience (5,10–11), developmental biology (7,12,13) and immunology (14). Besides, these techniques have also been applied to cancer tissues. In 2016, a study of breast cancer using ST uncovered unexpected heterogeneity within a biopsy (which would be impossible to detect by regular transcriptome analysis), and provided more detailed prognostic information (8). Profiling >6000 tissue regions in a single prostate by ST, Emelie *et al*. measured spatial gene expression in prostate cancer tissue sections and identified gene expression gradients for re-stratifying the tumor micro-environment (15).

Improvements of the spatially resolved transcriptomic techniques led to the rapid accumulation of complex datasets with positional information. Due to the dramatic differences in the available approaches, a database to achieve handy comparison, integration and visualiza-

*To whom correspondence should be addressed. Tel: +86 10 6488 8543; Fax: +86 10 6487 1293; Email: chenrs@sun5.ibp.ac.cn
Correspondence may also be addressed to Xiaowei Chen. Tel: +86 10 64888561; Fax: +86 10 6487 1293; Email: chenxiaowei@ibp.ac.cn

tion of spatially resolved transcriptomic data is lacking. Therefore, we developed SpatialDB, a manually curated resource of spatially resolved transcriptomes for researchers to efficiently investigate and reuse these published data. The current version of SpatialDB includes 24 spatially resolved transcriptomic datasets in 5 species (human, mouse, drosophila, *Caenorhabditis elegans* and zebrafish) generated by 8 spatially resolved transcriptomic techniques, including ST (8), Slide-seq (9), LCM-seq (6), seqFISH (5), MERFISH (4), Liver single cell zonation (16), Geo-seq (12) and Tomo-seq (7). SpatialDB provides an online tool for visualization of spatially resolved transcriptomic data and quick retrieval of spatial gene expression in a certain tissue of interest. Moreover, spatially variable (SV) genes were identified in 10 datasets, and functional enrichment analysis were performed. We expect that SpatialDB may serve as a helpful resource to facilitate the exploring of spatial organization of cells in tissues.

DATA COLLECTION AND PROCESSING

Data collection

We collected published spatially resolved transcriptomic datasets by searching PubMed with the following keywords: 'spatial' AND ('transcriptome' OR 'transcriptomics' OR 'RNA-seq' OR 'RNA sequencing'). We obtained 8 spatially resolved transcriptomic techniques and 24 datasets from the search results (Figure 1 and Table 1). For each technique, we extracted a brief description and a schematic from papers. For each dataset, we read the original paper, and extracted the corresponding metadata, including publication information, data description, experimental design, samples, data availability, etc. We downloaded the gene expression matrix data and spatial position information from supplementary materials of the papers, GEO database and custom data hubs. If a dataset contains biological/technical replicates or samples from different tissues or treatments, we manually divided it into multiple sub-datasets. A total of 305 sub-datasets were obtained (Table 1). Datasets generated by four techniques (Table 1) were at single-cell resolution. All the collected datasets and studies were published before May 2019.

Data processing

We performed median ratio normalization on ST datasets using DESeq2 (version 1.22.2) (17). Two LCM-seq datasets (6,18) did not provide precise coordinates of samples. We obtained the 2D positions of samples by using t-SNE algorithm (19).

Five pucks of coronal hippocampus, sagittal cerebellum, kidney, liver and sagittal cortex from Slide-seq dataset (9) were processed for visualization. Cells that had less than 100 read counts were removed. Genes that had less than 300 read counts in all cells were removed.

In order to facilitate online visualization, expression matrix and spatial position information of Slide-seq and MERFISH datasets were represented in JSON (JavaScript Object Notation) format for each gene. Datasets from the other six techniques were stored to MySQL (version 5.5.60), and one data table was created for each sub-dataset (Table 1).

We used two methods, SpatialDE (20) and trendsceek (21), to identify SV genes in 10 datasets (Table 1). The source codes of the two methods were obtained from the GitHub website. The q-value threshold of significant SV genes identified by SpatialDE was 0.05. Trendsceek performed four statistical tests (Emark, Vmark, MarkCorr and MarkVario) for each gene. Genes were considered to be significantly SV if q-values were <0.05 for at least one of the four tests. GO and KEGG enrichment analysis of SV genes were performed using clusterProfiler package (version 3.12.0) (22). The package was obtained from Bioconductor (release 3.9). The parameters of enrichment analysis were as follows. GO: ont = 'ALL', pAdjustMethod = 'BH', pvalueCutoff = 0.05, qvalueCutoff = 0.2, keyType = 'ENTREZID'. KEGG: pvalueCutoff = 0.05, pAdjustMethod = 'BH', minGSSize = 10, maxGSSize = 500, qvalueCutoff = 0.2, use_internal_data = FALSE.

DATABASE CONSTRUCTION AND CONTENT

Database construction

SpatialDB was constructed on a CentOS Linux server (version 7.6). The web services were built using Apache (version 2.4.6). The website was developed using PHP (version 7.0.33). The front-end of the website was developed using Bootstrap framework (version 3.3.7). The DataTables framework (version 1.10.19) was used to display data in tables. The online visualization of spatially resolved transcriptomes was implemented using Highcharts (version 7.1.1), jQuery (version 3.4.0) and d3 (version 3.3.10) JavaScript libraries. SpatialDB is freely available to the research community at https://www.spatialomics.org/SpatialDB and requires no registration or login.

Visualization of spatially resolved transcriptomes

We combined scatter module with heatmap module of Highcharts framework to implement the visualization of spatial gene expression profiles generated by seven techniques (Figure 1 and Table 1). Users can browse the spatial expression profile of the gene of interest in the selected sample. One point in the chart represents a group of cells or one single cell. Users can set the radius and symbol of the point for each chart manually. The color of the point represents the gene expression level. A color bar is shown at the bottom of the chart. A popup box will open to show detailed information (read counts, coordinates, etc.) of the point by hovering the mouse over it. Users can view in full screen, print chart, view data table and download the image or data in various formats by clicking the chart context menu on the top right corner of the chart. Users can click and hold down the left mouse button, then drag a rectangle in the chart to zoom in. Hold down the 'Shift' key and the left mouse button to drag in the chart to move the field of view. The histological images of tissue sections are shown as the background of charts for two ST datasets (8,15). Diagrams of murine small intestinal villus and liver lobule are displayed as the background of charts for the LCM-seq dataset (23) and the liver single cell zonation datasets (16,24), respectively. Users can clearly browse the spatial positions of

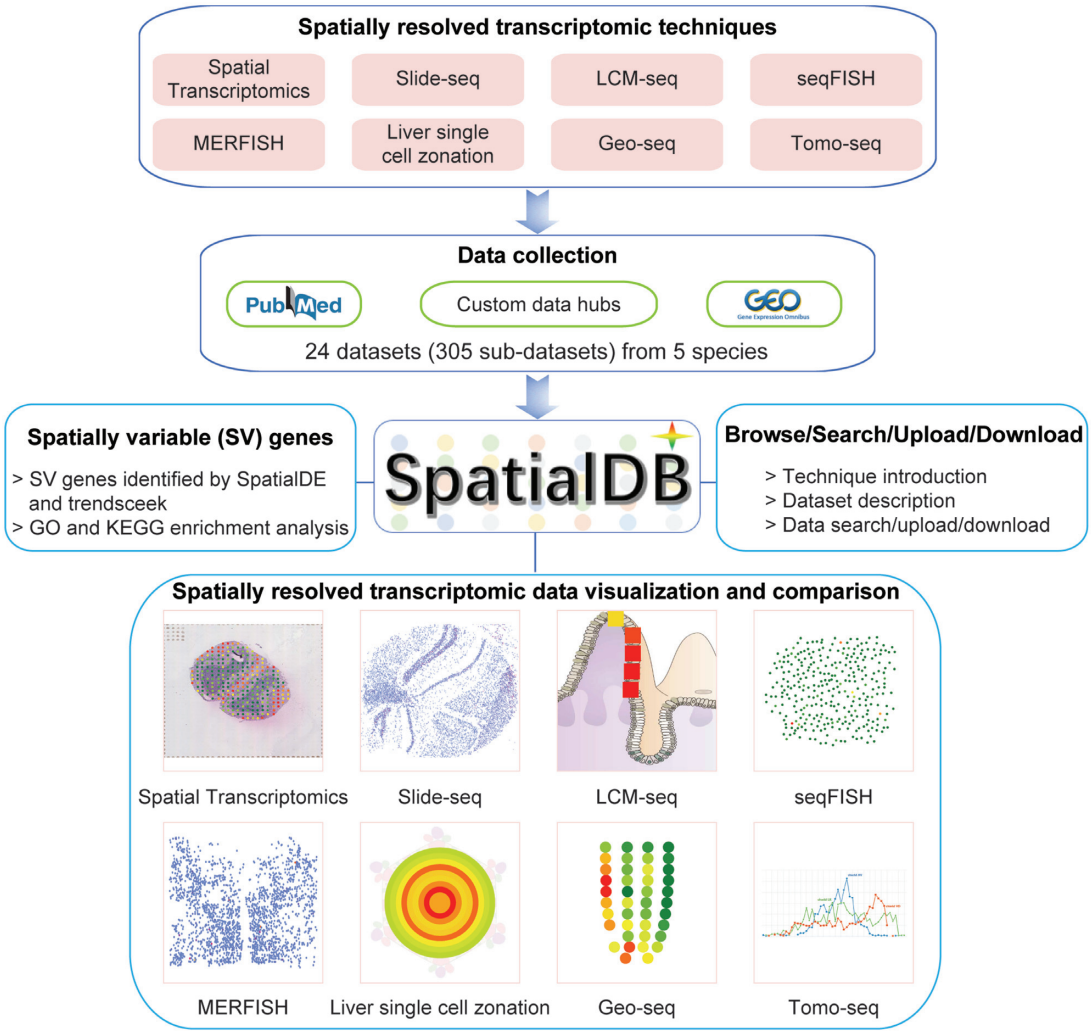

Figure 1. Overview of SpatialDB database. Spatially resolved transcriptomic data generated by eight techniques were collected from public resources. SpatialDB provided a web interface for online visualization and comparison of these data. Users can browse, search and download the datasets, SV genes and their functional annotations.

points in the original tissues. A data transformation drop-down list is provided for each chart. Users can manually apply data transformation, including log2, ln and log10. We used line module of Highcharts to implement the visualization of datasets generated by Tomo-seq (Figure 1 and Table 1). In the charts of Tomo-seq datasets, the coordinates on *X*-axis represent the serial numbers of tissue slices. The *Y*-axis represents the gene expression level.

Side-by-side comparison of spatial gene expression

We implemented the comparison of spatial gene expression profiles in the SpatialDB web interface. We provided two web pages to compare the heatmap charts and the line charts, respectively. Users can compare the spatial gene expression of two datasets generated by the same or different techniques at the same time side by side. Taking the heatmap charts comparison web page as an example, users can select 'Dataset1' and 'Dataset2' from the drop-down list, then click 'Compare' button to show the two charts side by side. For each chart, users can input a gene of interest, select a certain sample and click 'Submit' to show its spatial expression profile. The charts in the comparison web page contain all the options and properties mentioned in the above section.

Spatially variable (SV) genes

For each of the 10 datasets, SV genes identified in all sub-datasets by the same method were collected together and displayed in one data table. Taking ST as an example, a description of ST technique and a table of ST datasets will display by clicking 'Dataset' in the navigation bar of the web interface. Then, users can select a dataset (for example: dataset 27365449) and click 'Details' in the fifth column of the first row. A table containing dataset details will display. Users can find a tab named 'SV genes' under the dataset details table. Users can click the column title to sort the SV gene table in ascending or descending order. A fuzzy search box is provided on the top right of the table. Users

Table 1. Statistics and description of spatially resolved transcriptomic techniques in SpatialDB

Techniques	Datasets No.[a]	SV genes[b]	Single-cell resolution	Data storage	Highcharts module
Spatial Transcriptomics	5 (46)	4		MySQL	Scatter, Heatmap
Slide-seq	1 (5)	1	Single cell	JSON	Scatter, Heatmap
LCM-seq	4 (9)	1		MySQL	Scatter, Heatmap
seqFISH	3 (35)	2	Single cell	MySQL	Scatter, Heatmap
MERFISH	1 (181)	1	Single cell	JSON	Scatter, Heatmap
Liver single cell zonation	2 (2)		Single cell	MySQL	Scatter, Heatmap
Geo-seq	1 (3)	1		MySQL	Scatter, Heatmap
Tomo-seq	7 (24)			MySQL	Line
Total	24 (305)	10	4		

[a]The number of datasets (sub-datasets) for each technique.
[b]The number of datasets in which the SV genes were identified.

can quickly search the table by keywords of interest. Functional enrichment analysis results were listed below the SV gene table. The 50 most enriched GO or KEGG items were shown in the dot plot.

Search, download and upload

Users can easily obtain the spatial expression of a gene of interest across all the datasets from different techniques by searching SpatialDB. A quick search box has been embedded in the homepage of the web interface. A fuzzy search box is also provided on the top right of the search results table. Users can narrow the search scope by selecting species from the drop-down list. In addition, users can download all data via the 'Download' web page. If users would like to share their data, they can send necessary information to us through the 'Upload' web page. We will process the data and add to SpatialDB. A detailed tutorial for the usage of the database was also provided on the 'Help' page.

DISCUSSION

One of the main goals of the Human Cell Atlas Project (25) is to characterize the spatial relationship of all cell types. Spatially mapping multiple cell types based on expression signatures simultaneously may help to study the interaction between different cell types. Some strategies have been developed to detect the spatial gene expression profiles at single-cell resolution, such as Slide-seq (9), seqFISH (5) or MERFISH (10). In order to explore the spatial gene expression data more efficiently for the research community, we constructed SpatialDB—a database for spatially resolved transcriptomes. To our best knowledge, SpatialDB is the first dedicated database to curate spatially resolved transcriptomes. The spatial gene expression online visualization, comparison tools and SV gene annotations will be helpful for biologists to explore the spatial organization of cells.

The Human Cell Atlas Project have greatly propelled life sciences researches at single-cell level. It is expected that more spatial transcriptomic techniques will be developed, and spatial gene expression data will accumulate rapidly. We will keep collecting new techniques and datasets to update SpatialDB. Furthermore, we will integrate more tools and data sources to analyze the data.

ACKNOWLEDGEMENTS

We thank the staff from Core Facility for Protein Research, Center of Big Data Research in Health and Prof. Chen's lab, Institute of Biophysics, Chinese Academy of Sciences for the technique support.

FUNDING

National Natural Science Foundation of China [31871307, 31801072, 31701122, 31520103905]; CAS Key Technology Talent Program; National Key Research and Development Project [2018YFA0106901]. Funding for open access charge: National Natural Science Foundation of China [31871307].
Conflict of interest statement. None declared.

REFERENCES

1. Islam,S., Kjallquist,U., Moliner,A., Zajac,P., Fan,J.B., Lonnerberg,P. and Linnarsson,S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, **21**, 1160–1167.
2. Ramskold,D., Luo,S.J., Wang,Y.C., Li,R., Deng,Q.L., Faridani,O.R., Daniels,G.A., Khrebtukova,I., Loring,J.F., Laurent,L.C. *et al.* (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**, 777–782.
3. Battich,N., Stoeger,T. and Pelkmans,L. (2013) Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nat. Methods*, **10**, 1127–1133.
4. Chen,K.H., Boettiger,A.N., Moffitt,J.R., Wang,S.Y. and Zhuang,X.W. (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, **348**, aaa6090.
5. Shah,S., Lubeck,E., Zhou,W. and Cai,L. (2016) In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, **92**, 342–357.
6. Nichterwitz,S., Chen,G., Benitez,J.A., Yilmaz,M., Storvall,H., Cao,M., Sandberg,R., Deng,Q. and Hedlund,E. (2016) Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. *Nat. Commun.*, **7**, 12139.
7. Junker,J.P., Noel,E.S., Guryev,V., Peterson,K.A., Shah,G., Huisken,J., McMahon,A.P., Berezikov,E., Bakkers,J. and van Oudenaarden,A. (2014) Genome-wide RNA tomography in the zebrafish embryo. *Cell*, **159**, 662–675.
8. Ståhl,P.L., Salmén,F., Vickovic,S., Lundmark,A., Navarro,J.F., Magnusson,J., Giacomello,S., Asp,M., Westholm,J.O., Huss,M. *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, **353**, 78–82.
9. Rodriques,S.G., Stickels,R.R., Goeva,A., Martin,C.A., Murray,E., Vanderburg,C.R., Welch,J., Chen,L.M., Chen,F. and Macosko,E.Z. (2019) Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, **363**, 1463–1467.
10. Moffitt,J.R., Bambah-Mukku,D., Eichhorn,S.W., Vaughn,E., Shekhar,K., Perez,J.D., Rubinstein,N.D., Hao,J., Regev,A., Dulac,C.

et al. (2018) Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, **362**, eaau5324.
11. Eng,C.-H.L., Lawson,M., Zhu,Q., Dries,R., Koulena,N., Takei,Y., Yun,J., Cronin,C., Karp,C., Yuan,G.-C. *et al.* (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature*, **568**, 235–239.
12. Peng,G., Suo,S., Chen,J., Chen,W., Liu,C., Yu,F., Wang,R., Chen,S., Sun,N., Cui,G. *et al.* (2016) Spatial transcriptome for the molecular annotation of lineage fates and cell identity in Mid-gastrula mouse embryo. *Dev. Cell*, **36**, 681–697.
13. Ebbing,A., Vertesy,A., Betist,M.C., Spanjaard,B., Junker,J.P., Berezikov,E., van Oudenaarden,A. and Korswagen,H.C. (2018) Spatial transcriptomics of C. elegans males and hermaphrodites identifies Sex-Specific differences in gene expression patterns. *Dev. Cell*, **47**, 801–813.
14. Medaglia,C., Giladi,A., Stoler-Barak,L., De Giovanni,M., Salame,T.M., Biram,A., David,E., Li,H., Iannacone,M., Shulman,Z. *et al.* (2017) Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. *Science*, **358**, 1622–1626.
15. Berglund,E., Maaskola,J., Schultz,N., Friedrich,S., Marklund,M., Bergenstrahle,J., Tarish,F., Tanoglidi,A., Vickovic,S., Larsson,L. *et al.* (2018) Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.*, **9**, 2419.
16. Halpern,K.B., Shenhav,R., Matcovitch-Natan,O., Toth,B., Lemze,D., Golan,M., Massasa,E.E., Baydatch,S., Landen,S., Moor,A.E. *et al.* (2017) Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*, **542**, 352–356.
17. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
18. Brunskill,E.W., Potter,A.S., Distasio,A., Dexheimer,P., Plassard,A., Aronow,B.J. and Potter,S.S. (2014) A gene expression atlas of early craniofacial development. *Dev. Biol.*, **391**, 133–146.
19. Van der Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.
20. Svensson,V., Teichmann,S.A. and Stegle,O. (2018) SpatialDE: identification of spatially variable genes. *Nat. Methods*, **15**, 343–346.
21. Edsgard,D., Johnsson,P. and Sandberg,R. (2018) Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods*, **15**, 339–342.
22. Yu,G., Wang,L.G., Han,Y. and He,Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.
23. Moor,A.E., Harnik,Y., Ben-Moshe,S., Massasa,E.E., Rozenberg,M., Eilam,R., Bahar Halpern,K. and Itzkovitz,S. (2018) Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis. *Cell*, **175**, 1156–1167.
24. Halpern,K.B., Shenhav,R., Massalha,H., Toth,B., Egozi,A., Massasa,E.E., Medgalia,C., David,E., Giladi,A., Moor,A.E. *et al.* (2018) Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat. Biotechnol.*, **36**, 962–970.
25. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M. *et al.* (2017) The Human Cell Atlas. *Elife*, **6**, e27041.