时空组学
STOmics

STOMICS FILE FORMAT MANUAL

Version: V4.1.0

Manual Version: A0.1

Revision History

Manual Version:	A0
Software Version:	V4.1.0
Date:	Apr. 2022
Description:	Initial release

Manual Version:	A0.1
Software Version:	V4.1.0
Date:	Jun. 2022
Description:	Revised the data type of some data elements in the GEF file.

Note: Please download the latest version of the manual and use it with the software specific for this manual.

TABLE OF CONTENTS

CHAPTER 1
OVERVIEW

1.1. About Software

STOmics Analysis Workflow[1] (SAW) software suite is a set of pipelines that are bundled to position sequenced reads to their spatial location on the tissue section, and quantify spatial gene expression.

SAW download (Docker Hub): https://hub.docker.com/r/stomics/saw

SAW Github: https://github.com/BGIResearch/SAW

1.2. About Manual

This manual includes descriptions of key files format generated from SAW, which help users better understand and make use of information from analysis results.

1.3. Terminologies and Concepts

Table 1-1 Terminologies and Concepts

Abbreviation	Full Name	Description
SN	**Serial Number**	Unique ID for STOmics Gene Expression Chip.
RIN	**RNA Integrity Value**	RNA integrity value measures the RNA degradation degree to indicate the integrity of RNA and evaluate the quality of the RNA sample. RIN values range from 1 (totally degraded) to 10 (intact). In STOmics analysis, only tissue sample with a pre-measured RIN value greater than 7 should be used for further sequencing and bioinformatics analysis.
CID	**Coordinate ID**	Spatial position identifier, the artificially synthesized barcode sequence unique to each spot on the STOmics Gene Expression Chip.
MID	**Molecular ID**	Molecular identifier (same as UMI), the artificially synthesized sequence unique to each mRNA molecule captured from the sample which helps to differentiate the number of reads contributed by mRNA expression level due to amplification. Two copies of native transcripts from the same molecule captured on one DNB will result in two independent reads with the same CID but different MID. In contrast, two reads with identical CID and MID were originated from the same transcript but got amplified.
DNB	**DNA Nanoball**	DNA nanoball is the product of rolling-circle amplification (RCA) that is linearly amplified from the original circular single-stranded DNA template. DNB is the smallest capture unit on the STOmics Gene Expression Chip.
Bin	**Bin**	Bin (or Square Bin) is the analysis unit on the gene expression heat map. A bin is a fixed-sized square in which the expression value in this square is accumulated. Bins are not overlapped. The value followed by “Bin” represents the side length of the square. For bin 1, each DNB on the STOmics Gene Expression Chip is shown as a spot, which means one spot only contains the data from one DNB. Bin N means one spot on the heat map is an aggregation of data from N×N neighbor DNBs. For example, a spot of bin 100 covers data from 10,000 DNBs.

CHAPTER 2
FILE FORMAT

2.1 BAM

The BAM[2] file format is a binary format for saving sequence alignment and gene annotation data. SAW **mapping** BAM adds custom tags in the BAM optional field to record reads coordinates, CID and MID information. **count** BAM adds annotation information in the tag field. Custom tags are described in Table 2-1.

Table 2-1 BAM custom tags

Tag	Description
Cx:i	x coordinate of CID.
Cy:i	y coordinate of CID.
UR:Z	The hexadecimal representation of uncorrected binary-encoded MID.
XF:Z	Mapping region on the reference genome. Valid value: 0=EXONIC, 1=INTRONIC, 2=INTERGENIC.
GE:Z	Annotated gene name.
GS:Z	'+' or '-', indicating forward/reverse strand respectively.
UB:Z	The hexadecimal representation of count corrected binary-encoded MID.

Example of `mapping` BAM:

```
E100026571L1C009R00301275185    16      1       3000095 255     26M121066N74M
*       0       0       GGCTTTTTTTTTTTTTTTTTTTTTTTTTTCTAAATATTGGGTTTTATTAGC
ACCATGATAACTGTATATTAATTTGCACTGACTGTCATAACAAAATAC      G+:GFFGGFGFFGFFGFGGFFGF
FFFFCFGFCFGGGFGGFGFFFFGGFGGFGFFFGGFFGFFFGFGFGFFGFFGFGFFFFGFFFFFFFFGGFFGGFFGEF
NH:i:1  HI:i:1    AS:i:88 nM:i:0  Cx:i:4826       Cy:i:11598      UR:Z:6FA29
```

Example of `count` BAM:

```
E100026571L1C002R00703943265    1040    1       3082766 255     11M132671N89M
*       0       0       CTGCTGCAGCTTTTTTTTCTTTGAGATTTATTTTTATGCTATGTGTATGGGT
ATTTTGCCTGCATATATGTCTATGCACCATGTGTGTGCAGTGCTTGAG      FFFFFECGFDCFGDGDFEE@EEG
IBFGGCGFFGACGFCGFFDGDGFFFFFFEGCDFCGFFGG@FFF=EFFDGGGGGFDGFFFGGGFGFFGGGFFGGGDFG
NH:i:1  HI:i:1    AS:i:88 nM:i:0  Cx:i:7767       Cy:i:18052      UR:Z:7AE49
XF:i:0  GE:Z:Xkr4       GS:Z:-  UB:Z:79E49
```

2.2 Mapped CID List with Reads Count File

mapping pipeline outputs mapped CID list file with reads counts for each CID. This file stores CID coordinates and reads count for each coordinate. The list does not have header. The three columns are x coordinate, y coordinate and MID count.

Example of mapped CID list with reads count file:

```
14195   16619   1
19945   14450   2
14548   9438    1
```

2.3 Gene Expression File

Gene expression file (GEF) is a type of gene expression distribution visualization file particularly designed for Stereo-seq. GEF is a hierarchically structured data model that stores multiple gene expression matrices in different bin sizes.

Each GEF container organizes a collection of spatial gene expression matrices. It includes two primary data objects, Group and Dataset. A Dataset is a multidimensional array of data elements. Group object is analogous to file system directory which organizes Datasets and other Groups in hierarchies.

The first level of GEF includes three Group object: “geneExp”(required), “wholeExp”(optional), and “stat” (optional). Group “geneExp” contains Groups of gene spatial expression data in one or multiple bin size. Group “wholeExp” contains Datasets that record expression level and gene type count of each coordinate in one or multiple bin sizes. Group “stat” saves gene names, total MID count and spatial pattern enrichment score of each gene. “Attributes” in each Dataset records the key metrics of that Dataset. Check https://www.processon.com/view/link/610cc49c7d9c087bbd1ab7ab#map to get the schematic diagram of GEF. The field names and field data types are described in Table 2-2. SAW outputs three GEF files in the whole process, please check the Table 2-3 to find the description.

Table 2-2 Gene Expression FileText Fields Description

Attributes			
File Attributes	**Data Type**	**Example**	**Description**
version	**uint32**	**1**	Gene expression file format version.
geftool_ver	**uint32**	**0,6,2**	geftool version. It can be used as an individual tool to manipulate GEF files.

geneExp/binN/expression: Dataset “expression” is a 1D array which stores coordinates and MID counts of each gene in the bin size of N, aggregated by gene name.			
Dataset Attributes	**DataType**	**Example (bin1)**	**Description**
minX	**uint32**	**59820**	Minimum x coordinate in bin N.
minY	**uint32**	**102086**	Minimum y coordinate in bin N.
maxX	**uint32**	**73040**	Maximum x coordinate in bin N.
maxY	**uint32**	**120539**	Maximum y coordinate in bin N.
maxExp	**uint32**	**28**	Maximum MID count in a spot when the bin size is N.
resolution	**uint32**	**500**	Physical pitch (nm) between neighbor spots.
Dataset DataType: compound	**DataType**	**Example (bin1)**	**Description**
x	**uint32**	**71032**	x coordinate in bin N.
y	**uint32**	**103180**	y coordinate in bin N.
count	**uint8/uint16/ uint32**	**1**	MID count at (x, y) when bin size is N. Data type for “count” is consistent with “maxExp” in the “Attributes.”

geneExp/binN/gene:
Dataset "gene" is a 1D array which stores the gene names, the starting row indexes in dataset "expression", and row counts.

Dataset DataType: compound	Data Type	Example (bin1)	Description
gene	S32	b'Gm16045'	Gene name.
offset	uint32	21	The starting row index in dataset "expression" for the gene. In this example, the gene expression data for gene "Gm16045" starts from row 21 in the dataset "expression."
count	uint32	2	Row count. In this example, expression data for gene "Gm16045" is recorded in row 21 and 22 (2 rows) in the dataset "expression."

wholeExp/binN:
Dataset "binN" is a 2D array (matrix) which stores the MID count and gene type count at each spot.

Dataset Attributes	DataType	Example (bin1)	Description
number	uint64	22879557	Number of non-zero spots in the dense matrix.
minX	uint32	59820	Minimum x coordinate in bin N.
lenX	uint32	13221	Length of x.
minY	uint32	102086	Minimum Y coordinate in bin N.
lenY	uint32	18454	Length of y.
maxMID	uint32	2155	Maximum MID count in a spot.
maxGene	uint32	846	Maximum gene type count in a spot.
resolution	uint32	500	Pitch (nm) between neighbor spots.
Dataset DataType: 2D array (XxY), compound	**DataType**	**Example (bin1)**	**Description**
MIDcount	uint32	1	MID count in the spot. The spot coordinate can be identified from the row and column index of the 2D matrix plus the "minX" and "minY" specified in the attributes.
genecount	uint16	1	Gene count in the spot. The spot coordinate can be identified from "Attributes" and the indexes of the 2D array.

stat/gene: Dataset "gene" is a 1D array which stores the MID count and spatial pattern enrichment score (E10) of each gene. The array is order by the MID count in descending order.			
Dataset Attributes	**DataType**	**Example**	**Description**
maxE10	**float32**	**65.53**	Maximum E10 score.
minE10	**float32**	**0.**	Minimum E10 score.
cutoff	**float32**	**0.1**	Threshold for filtering spots that will be used for computing E10. In this example, 0.1 means that the spots whose MID count is in the top 10% are used for calculating the spatial enrichment score.
Dataset DataType: compound	**DataType**	**Example**	**Description**
gene	**S32**	**b'Ptgds'**	Gene name.
MIDcount	**uint32**	**229502**	MID counts for the gene.
E10	**float32**	**65.53**	The spatial pattern enrichment score (E10) for the gene.

The distinctions of each SAW output GEF files are explained in Table 2-3.

Table 2-3 SAW Output GEF Files Description

GEF Name	SAW Pipeline	Example	Description
SN.raw.gef	**count**	**SS200000135TL_D1.raw.gef**	`count` output raw GEF, it only includes geneExp Group for the bin size of 1. The origin of expression matrix has been calibrated to (0,0).
SN.gef	**tissueCut**	**SS200000135TL_D1.gef**	`tissueCut` output full GEF file. It contains geneExp Group and wholeExp Group for the bin size of 1, 10, 20, 50, 100, 200, and 500. SN.gef is also the only one that includes stat Group. The origin of expression matrix has been calibrated to (0,0), and its offsets are the same with SN.raw.gef. SN.gef is the input file for visualization.
SN.tissue.gef	**tissueCut**	**SS200000135TL_D1.tissue.gef**	`tissueCut` output GEF file for the tissue-covered region. It only includes geneExp Group for the bin size of 1. The coordinates in the matrix and the offsets are all same with SN.raw.gef.

2.4 Gene Expression Matrix

Gene expression matrix stores genes spatial expression data. SAW generates multiple gene expression matrix files in the workflow, the basic format requires four columns with a header row that show the column names. The four columns are gene name, x coordinate, y coordinate, and MID count. The origin of `tissueCut` generated gene expression matrices have been calibrated to (0, 0). The header of expression matrix for maximum area enclosing rectangle region has six annotation rows start with "#" before the column rows. The header field names and field types are described in Table 2-4.

Table 2-4 Gene Expression Matrix Header Fields Description

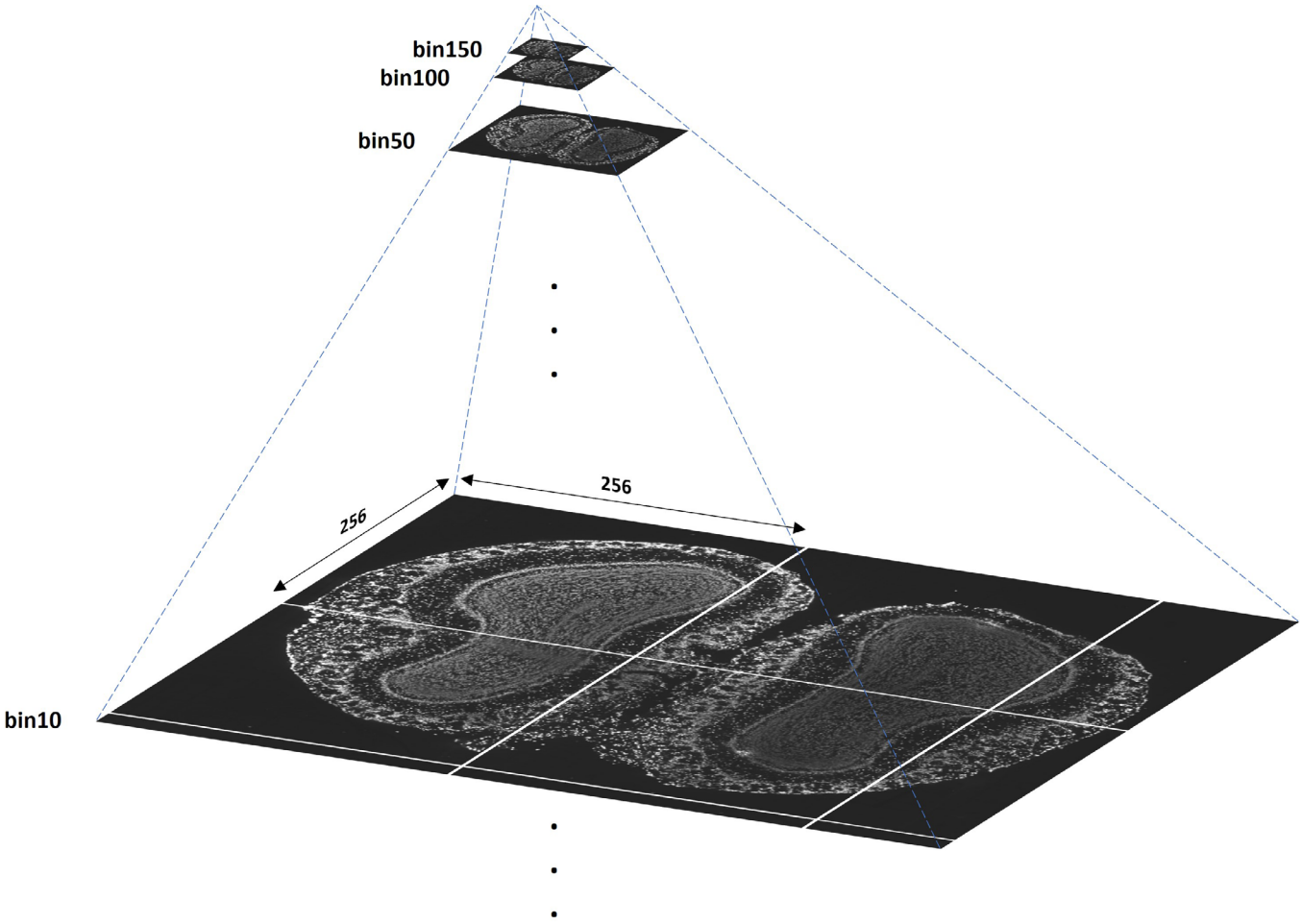
Fields	Data Type	Example	Description
#FileFormat	**string**	**GEMv0.1**	Gene expression matrix file format version.
#SortedBy	**string**	**None**	Gene expression matrix sorting strategy. Valid values: "geneID", "x", "y", "MIDCount", "None".
#BinSize	**uint16**	**1**	(Please check 1.3 Terminologies and Concepts Bin)
#StereoChip	**string**	**SS200000135TL_D1**	STOmics Gene Expression Chip SN.
#OffsetX	**uint32**	**1**	X coordinate of the origin before calibration.
#OffsetY	**uint32**	**1**	Y coordinate of the origin before calibration.
geneID	**string**	**Cr2**	Gene name.
x	**uint32**	**16809**	X coordinate of the spot.
y	**uint32**	**8546**	Y coordinate of the spot.
MIDCount	**uint32**	**1**	Number of MIDs at (x, y) for the gene in the corresponding row.

Example of `GEM`:

```
#FileFormat=GEMv0.1
#SortedBy=None
#BinSize=1
#StereoChip=SS200000135TL_D1.gem
#OffsetX=0
#OffsetY=0
geneID  x       y       MIDCount
Ccl27   4665    19736   1
CR974586.5      10207   15257   1
CR974586.5      4707    18336   1
```

2.5 Image Pyramid

The image pyramid model is a multi-resolution hierarchical model that is used to store and display images in different resolutions. For the same field of view, the layer of the image pyramid that is closest to the bottom includes the most detailed information and has the largest scale. **`register`** pipeline performs the down-sampling step on the registered image, and the resulted images are layered to construct a pyramid. For each resolution layer, the intact registered image is split into 256 pixels x 256 pixels tiles. If the size of a layer is smaller than 256 x 256, the image will then remain intact. The suffix of the file can be ".ssDNA.rpi" or ".conA.rpi" which indicates a cell nucleus or membrane stained image were used/registered respectively.
Schematic diagram of image pyramid:



References

1. BGIResearch/SAW. Accessed October 13, 2021. https://github.com/BGIResearch/SAW
2. Sequence Alignment/Map Format Specification.; 2021. Accessed May 21, 2021. https://github.com/samtools/hts-specs.

Contact Us

BGI-Research, Shenzhen

stereomap.cngb.org/SAP

Email: support@stereomics.com