

STOmics 分析流程结果文件 格式说明

版本历史

说明书版本：	A0
软件版本：	V 4.1.0
修订日期：	2022 年 4 月
修订内容：	首次发布

说明书版本：	A0.1
软件版本：	V 4.1.0
修订日期：	2022 年 6 月
修订内容：	修改 GEF 文件中一些数据元素的数据类型

提示:请下载最新版说明书, 与相应版本的软件使用。

©2022 深圳华大生命科学研究院保留所有权利。

1. 本产品仅用于研究, 不用于诊断。
- 2.本手册上的内容可能全部或部分受到适用的知识产权法的保护。深圳华大生命科学研究院和/或相应权利主体依法拥有其知识产权, 包括但不限于商标权、版权等。
- 3.深圳华大生命科学研究院不授予或暗示使用我们或任何第三方的任何版权内容或商标(注册或未注册)的权利或许可。未经本单位书面同意, 任何人不得擅自使用、修改、复制、公开传播、更改、分发或发布本手册的程序或内容, 不得使用或利用设计技巧使用或占有本单位或本单位关联方的商标、标识或其他专有信息(包括图像、文本、网页设计或形式)。
- 4.此处的任何内容都无意于或应被理解为对此处列出或描述的任何产品的性能的任何保证、表达或暗示。适用于本文所列任何产品的任何和所有保证均载于购买该产品所附的适用销售条款和条件。深圳华大生命科学研究院不做任何保证, 并在此声明对本文所述任何第三方产品或协议的使用不做任何保证。

目录

第一章 简介	
1.1. 关于软件	1
1.2. 关于说明书	1
1.3. 术语和概念	1
第二章 文件格式	
2.1. BAM 文件	3
2.2. CID 对应 reads 数列表	3
2.3. 基因表达文件（可视化文件）	4
2.4. 基因表达矩阵文件	6
2.5. 图像金字塔	7
参考资料	8
联系我们	9

第一章 简介

1.1. 关于软件

STOmics分析流程软件包¹ (STOmics Analysis Workflow, SAW) 整合了多个STOmics基因表达分析工具, 用于将原位测序数据与空间位置信息结合, 实现空间重构。

- 将SAW下载链接 (Docker Hub) :<https://hub.docker.com/r/stomics/saw>
- SAW Github:<https://github.com/BGIResearch/SAW>

1.2. 关于说明书

本文档包含SAW分析过程中产生的关键文件的格式说明, 用于帮助用户从分析结果数据中提取有用信息。

1.3. 术语和概念


表 1-1 术语概念说明

缩写	全称	说明
SN	Serial Number	STOmics 基因表达芯片的唯一编号。
RIN	RNA Integrity Value	RNA 完整值, 用于测量样品的 RNA 的降解程度, 反映 RNA 完整性, 评估样品质量。其中 1 代表降解最严重, 10 代表最完整。空间转录组分析中, RIN 大于 7 为高质量样本, 可用于后续实验和分析。
CID	Coordinate ID	用于标识空间位置信息的人工合成核酸序列, 类似 barcode。
MID	Molecular ID	用于标识从样本组织捕获到的 mRNA 的核酸序列, 类似 umi。不同的 mRNA 序列片段可以来自相同的分子, 使用相同的 MID 标识。
DNB	DNA Nanoball	以单链环状 DNA 为模板, 经过滚环扩增 (Rolling-Circle Replication, RCR) 后的产物叫 DNA 纳米球 (DNA Nanoball), 是 STOmics 基因表达芯片上的最小捕获单元。
Bin	Bin	用于标识基因表达热图的分析单元大小, 一个 bin 表示一个固定大小的方形区域, 区域内 DNB 表达量累加, 区域间不重合, 数字表示方形边长。STOmics 基因表达芯片上每个 DNB 在基因表达热图上表现为一个 spot, 此时的分析单元为 Bin1, 即一个 spot 只包含一个 DNB 的数据。将相邻 N×N 个 DNB 数据合并, 在基因表达热图上以一个 spot 的形式展示, 此时分析单元为 Bin N, 即一个分析单元包含 N×N 个 DNB 区域的数据。如 Bin100 表示基因表达热图上一个分析单元包含 10,000 个 DNB 区域的数据。

第二章 文件格式

BAM² 格式为存储测序数据和参考基因组比对、注释结果的常用二进制格式。SAW **mapping** BAM 标签栏添加自定义标签记录 reads 的坐标、CID、和 MID 相关信息，**count** BAM 添加注释信息。标签说明见表 2-1。

标签	说明
Cx:i	CID 对应的空间位置 x 坐标。
Cy:i	CID 对应的空间位置 y 坐标。
UR:Z	原始 MID 以二进制编码后用十六进制打印的结果。
XF:Z	该序列比对到参考基因组的区域，有效值包括 0=EXONIC，1=INTRONIC，2=INTERGENIC，分别表示外显子区、内含子区和基因间区。
GE:Z	注释基因名称。
GS:Z	该序列比对到参考序列的正链（+）或负链（-）。
UB:Z	SAW count 工具校正后的 MID。



 E100026571L1C009R00301275185 16 1 3000095 255 26M121066N74M * 0 0

 GGCTTTTTTTTTTTTTTTTTTTTTTTTTTCTAAATATTGGGTTTTATTAGCACCATGATAACTGTATATTAATTTGCACT

 GACTGTCATAACAAAATAC

 G+:GFFGGFGFFGFFGFGGFFGFFFFFCFGFCFGGGFGGFGFFFFGGFGGFGFFFGGFFGFFFGFGFGFFGFFGFGFFFF

 GFFFFFFFFGGFFGGFFGEF

 NH:i:1 HI:i:1 AS:i:88 nM:i:0 Cx:i:4826 Cy:i:11598 UR:Z:6FA29

```
E100026571L1C002R00703943265    1040    1    3082766 255    11M132671N89M    *    0    0
CTGCTGCAGCTTTTTTTTTCTTTGAGATTTATTTTTATGCTATGTGTATGGGTATTTTGCCTGCATATATGTCTATGCACCATGT
GTGTGCAGTGCTTGAG
FFFFFECGFDCFGDGFEE@EEGIBFGGCGFFGACGFCGFFDGDGFFFFFEGCDFCGFFGG@FFF=EFFDGGGGGFDGFFFGG
GFGFFGGGFFGGGDFG
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
```

mapping 工具将 CID 还原回组织空间位置后生成 CID 对应 reads 数列表，此列表记录 CID 的坐标 (x, y) 和坐标对应的 reads 数。列表无表头，三列数据分别为 x 坐标、y 坐标、和 MID 数。

CID 对应 reads 数列表文件示例：

14195	16619	1
19945	14450	2
14548	9438	1

2.3. 基因表达文件（可视化文件）

基因表达文件（GEF）是一种为可视化展示时空组基因表达空间分布而设计的文件，其文件结构是一种包含多个 bin size 基因表达矩阵、有层级关系的结构。

每个 GEF 文件整理一组空间基因表达矩阵。该文件包含组（group）和数据集（dataset）两种主要对象类型。数据集是一种多维数组，而组是可以包含数据集和其他组的管理结构。

基因表达文件第一层可包括“geneExp”（必须），“wholeExp”（可选），和“stat”（可选）三个组。“geneExp”中包含一种或多种 bin size 下每个基因的表达数据；“wholeExp”中包含一种或多种 bin size 下每个坐标点（spot）的表达数据和基因个数；“stat”中包含基因名称、每个基因的总表达量、以及每个基因的富集程度打分。基因表达文件每个数据集集中的“Attributes”记录数据集属性信息。文件格式图示见下方链接，字段说明见表 2-2。SAW 流程运行生成三个 GEF 区别见表 2-3。

GEF 格式图示 :<https://www.processon.com/view/link/610cc49c7d9c087bbd1ab7ab#map>

表 2-2 基因表达文件字段说明

Attributes			
属性	数据类型	示例	说明
version	uint 32	1	基因表达文件格式版本号。
geftool_ver	uint 32	0,6,2	geftool 程序版本号, 此工具可单独使用处理 GEF 文件。
geneExp/binN/expression: “expression”数据集是一个记录每个基因在 bin N 分辨率下的坐标和 MID 数的一维数组。			
属性	数据类型	示例 (bin1)	说明
minX	uint 32	59820	Bin N 分辨率下 x 坐标最小值。
minY	uint 32	102086	Bin N 分辨率下 y 坐标最小值。
maxX	uint 32	73040	Bin N 分辨率下 x 坐标最大值。
maxY	uint 32	120539	Bin N 分辨率下 y 坐标最大值。
maxExp	uint 32	28	Bin N 分辨率下 MID 数最大值。
resolution	uint 32	500	相邻捕获点之间的物理距离。
数据类型： 复合型	数据类型	示例 (bin1)	说明
x	uint 32	71032	binN 下 x 坐标。
y	uint 32	103180	binN 下 y 坐标。
count	uint 8/uint 16/uint 32	1	binN 时坐标 (x, y) 的 MID 数。“count”字段的数据类型与“Attributes”中的“maxExp”一致。

geneExp/binN/gene: “gene”数据集是一个记录基因名称,该基因在“expression”数据集中的起始行号, 以及该基因从起始行号开始占用的行数。			
数据类型: 复合型	数据类型	示例 (bin1)	说明
gene	S 32	b' Gm16045'	基因名称。
offset	uint 32	21	基因在“expression”数据集中的起始行号。 在示例中,“expression”数据集中的第 21 行开始为基因“Gm16045”的基因表达数据。
count	uint 32	2	行数。 在示例中,“expression”数据集中的第 21 和 22 行(共 2 行)为基因“Gm16045”的数据。
wholeExp/binN: 数据集“binN”是用于存储每个数据点的 MID 数和基因种类数的 2D 数组(矩阵)。			
属性	数据类型	示例 (bin1)	说明
number	uint 64	22879557	稠密矩阵中非零数据点个数。
minX	uint 32	59820	binN 下 x 坐标最小值。
lenX	uint 32	13221	x 范围长度。
minY	uint 32	102086	binN 下 y 坐标最小值。
lenY	uint 32	18454	y 范围长度。
maxMID	uint 32	2155	数据点中 MID 数最大值。
maxGene	uint 32	846	数据点中基因种类数最大值。
resolution	uint 32	500	相邻数据点间距 (nm)。
数据类型: 2D 数组 (X x Y), 复合型	数据类型	示例 (bin1)	说明
MIDcount	uint 32	1	数据点中的 MID 数。数据点坐标可通过 2D 矩阵的横纵轴索引加属性中的 x 和 y 最小值得到。
genecount	uint 16	1	数据点捕获的基因数。数据点坐标可通过属性信息和 2D 数组的索引得到。
stat/gene: “gene”数据集是一个存储每个基因 MID 数和空间富集程度打分 (E10) 的 1D 数组。数组由 MID 数降序排列。			
属性	数据类型	示例 (bin1)	说明
maxE10	float32	65.53	E10 打分最大值。
minE10	float32	0.	E10 打分最小值。
cutoff	float32	0.1	筛选用于计算 E10 的数据点的阈值。 示例中, 0.1 表示一个基因的全部数据点根据 MID 数排序后, MID 数排在前 10% 的数据点用于计算该基因的空间富集程度

数据类型： 复合型	数据类型	示例 (bin1)	说明
gene	S 32	b' Ptgds'	基因名称。
MIDcount	uint 32	229502	基因的 MID 数。
E10	float 32	65.53	基因的空间富集程度 (E10) 打分。

SAW 分析过程中生成三个 GEF 文件，其内容区别见表 2-3。

表 2-3 SAW 输出 GEF 文件说明

GEF 名称	SAW 流程步骤	示例	说明
SN.raw.gef	count	SS200000135TL_D1.raw.gef	count 生成原始 GEF 文件, 仅包含 bin size 为 1 的 geneExp 部分。矩阵中坐标原点矫正为 (0,0)。
SN.gef	tissueCut	SS200000135TL_D1.gef	tissueCut 生成完整 GEF 文件, 包含 bin1, 10, 20, 50, 100, 200, 500 的 geneExp 和 wholeExp 部分, 且包含 stat 组。矩阵中坐标原点矫正为 (0,0), offset 与 SN.raw.gef 一致。可视化读取展示矩阵。
SN.tissue.gef	tissueCut	SS200000135TL_D1.tissue.gef	tissueCut 生成的组织覆盖区域基因分布信息的 GEF 文件, 仅包含 bin size 为 1 的 geneExp 部分。矩阵中坐标与和 offset 皆与 SN.raw.gef 一致。

2.4. 基因表达矩阵文件

SAW 流程中生成的 GEF 文件可通过小工具转换成纯文本表格格式基因表达矩阵（GEM），基本格式为含表头的四列数据，分别为基因名称、x 坐标、y 坐标、和 MID 数。表达矩阵中原点位置校准至（0，0）。基因表达矩阵的表头部分包含六行以“#”开头的注释信息。表头字段说明见表 2-4。

表 2-4 基因表达矩阵表头字段说明

表头字段	数据类型	示例	说明
#FileFormat	string	GEMv0.1	基因表达矩阵文件格式版本号。
#SortedBy	string	None	基因表达矩阵排序策略, 有效值包括“geneID”, “x”, “y”, “MIDCount”, “None”。
#BinSize	uint 16	1	(见 1.3 术语和概念 Bin。)
#StereoChip	string	SS200000135TL_D1	STOmics 基因表达芯片 SN。
#OffsetX	uint 32	1	原点校准至 (0, 0) 前 x 坐标。
#OffsetY	uint 32	1	原点校准至 (0, 0) 前 y 坐标。
geneID	string	Cr2	基因名称。
x	uint 32	16809	x 坐标。
y	uint 32	8546	y 坐标。
MIDCount	uint 32	1	该行基因在坐标 (x, y) 处的 MID 数。

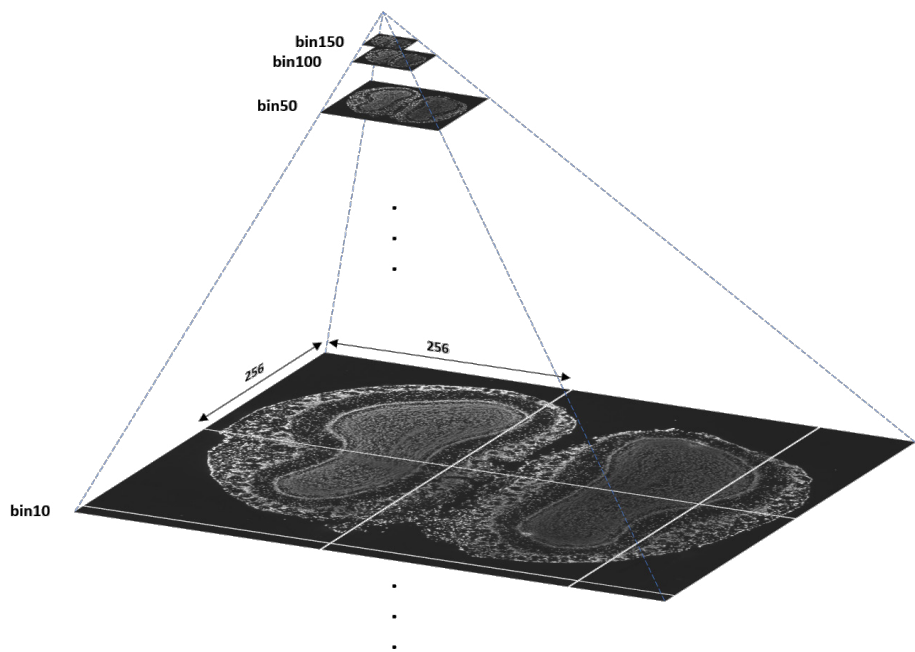
基因表达矩阵示例：

```
●●●
#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. 图像金字塔

图像金字塔模型是一种多分辨率层次模型，可根据需要以不同分辨率进行图像的存储与显示。所表示的图像范围不变的情况下，金字塔越靠近底层所表示的图像信息越详细，比例尺越大。对 **register** 工具处理后的配准图进行梯度降采样处理，得到多个图像数据以图像金字塔形式保存。每个分辨率层级中，将完整的组织配准图像切分为尺寸 256 x 256 像素的图像碎片进行保存，若在当前分辨率下的图片尺寸小于 256 x 256，则无需切割。文件名以 “.ssDNA.rpi” 或 “.conA.rpi” 结尾，分别对应染色细胞核和细胞膜两种染色方式。

图像金字塔结构示例：



参考资料

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>.

联系我们

深圳华大生命科学研究院

网址:<https://www.stomics.tech>

邮箱:support@stereomics.com