CeleScoT

格物致知 识微通元

单细胞ATAC分析结题报告

content

服务电话:025-58165529
Email: singleron@singleronbio.com
(mailto:singleron@singleronbio.com)

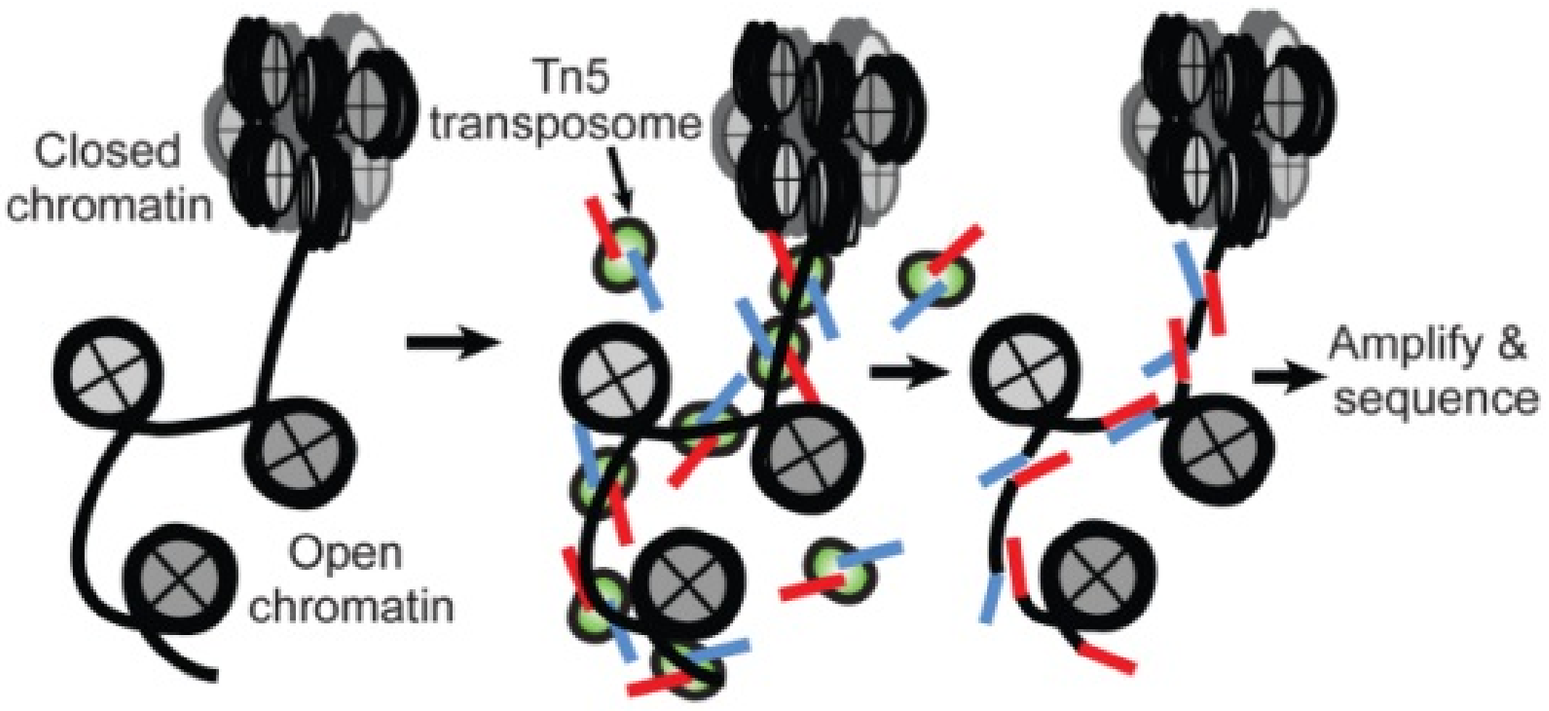
新格元单细胞ATAC-seq分析结题报告

合同信息	合同内容
项目编号	P23032302
项目名称	P23032302
样本名称	A140417
流程版本	Singlecell_ATAC_seqV1
结题日期	_____

1 前言

ATAC-seq，全称是Assay for transposase-accessible chromatin with high-throughput sequencing，是基于高通量测序对开放性染色质（open chromatin）进行研究的技术。而单细胞ATAC-seq技术，顾名思义就是在单细胞水平上的ATAC-seq技术，兼具单细胞技术的高分辨率及ATAC-seq的优势，是目前研究基因表观组学的热门技术。

ATAC-seq技术利用转座酶Tn5容易结合在开放染色质的特性，在切割获取DNA的同时加入接头，经过PCR扩增后即可进行高通量测序，获得全基因组范围内开放染色质的序列信息：



单细胞ATAC-seq的应用：

- 单细胞ATAC-seq分析可以解决细胞异质性的问题
- 单细胞ATAC-seq分析可以在细胞水平揭示开放染色质差异性
- 单细胞ATAC-seq分析可以预测转录因子结合位点，推测染色质调控活性。

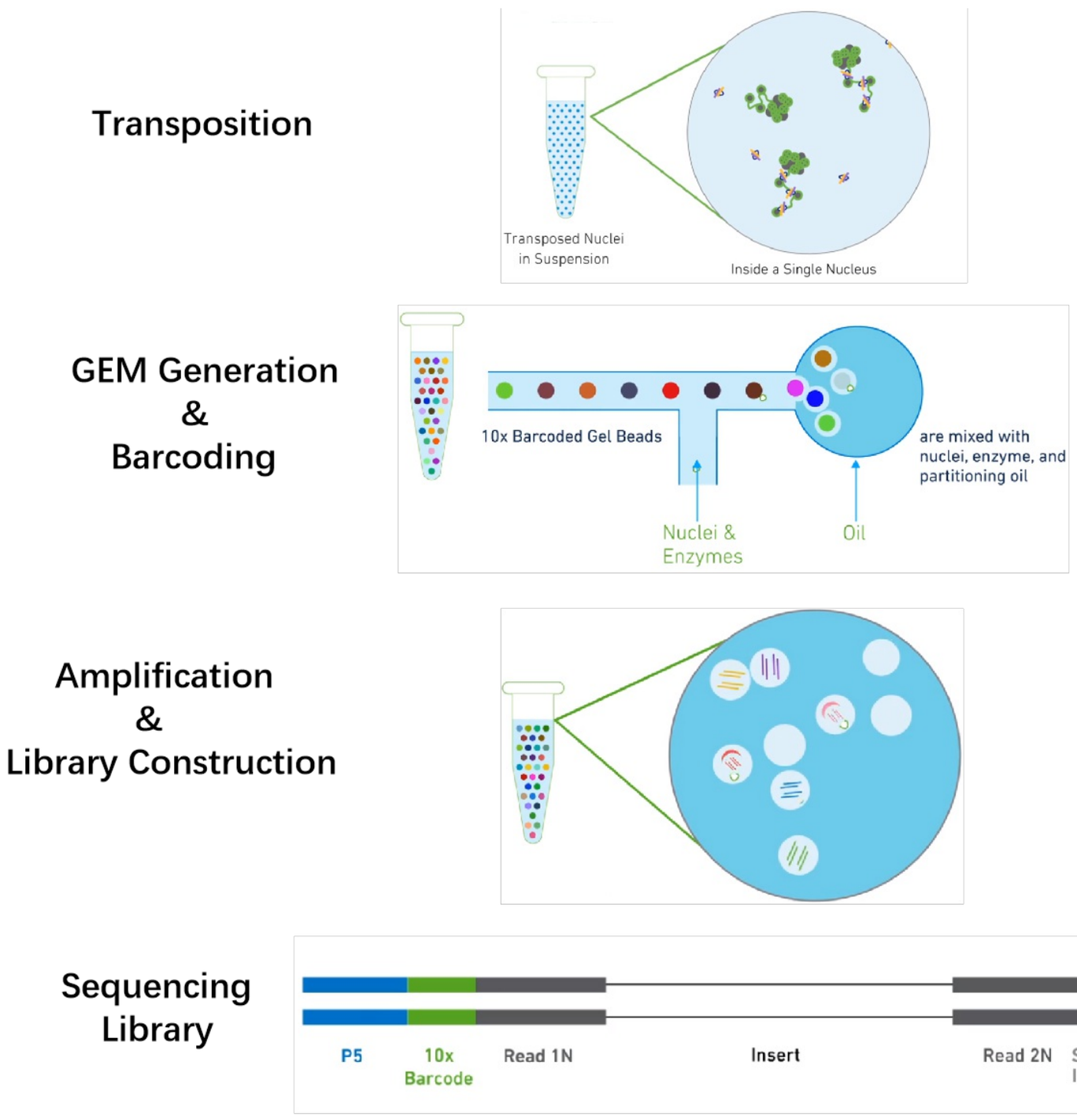
单细胞ATAC-seq分析目前已经应用于肿瘤学、干细胞、发育生物学、免疫学、神经生物学、生殖医学等研究领域。

2 建库测序

新格元使用10X Genomics Chromium Single Cell ATAC技术进行单细胞文库制备，并进行文库检测，质量合格后方进行上机测序：

2.1 文库构建

单细胞ATAC-seq测序技术具体过程如下所示：



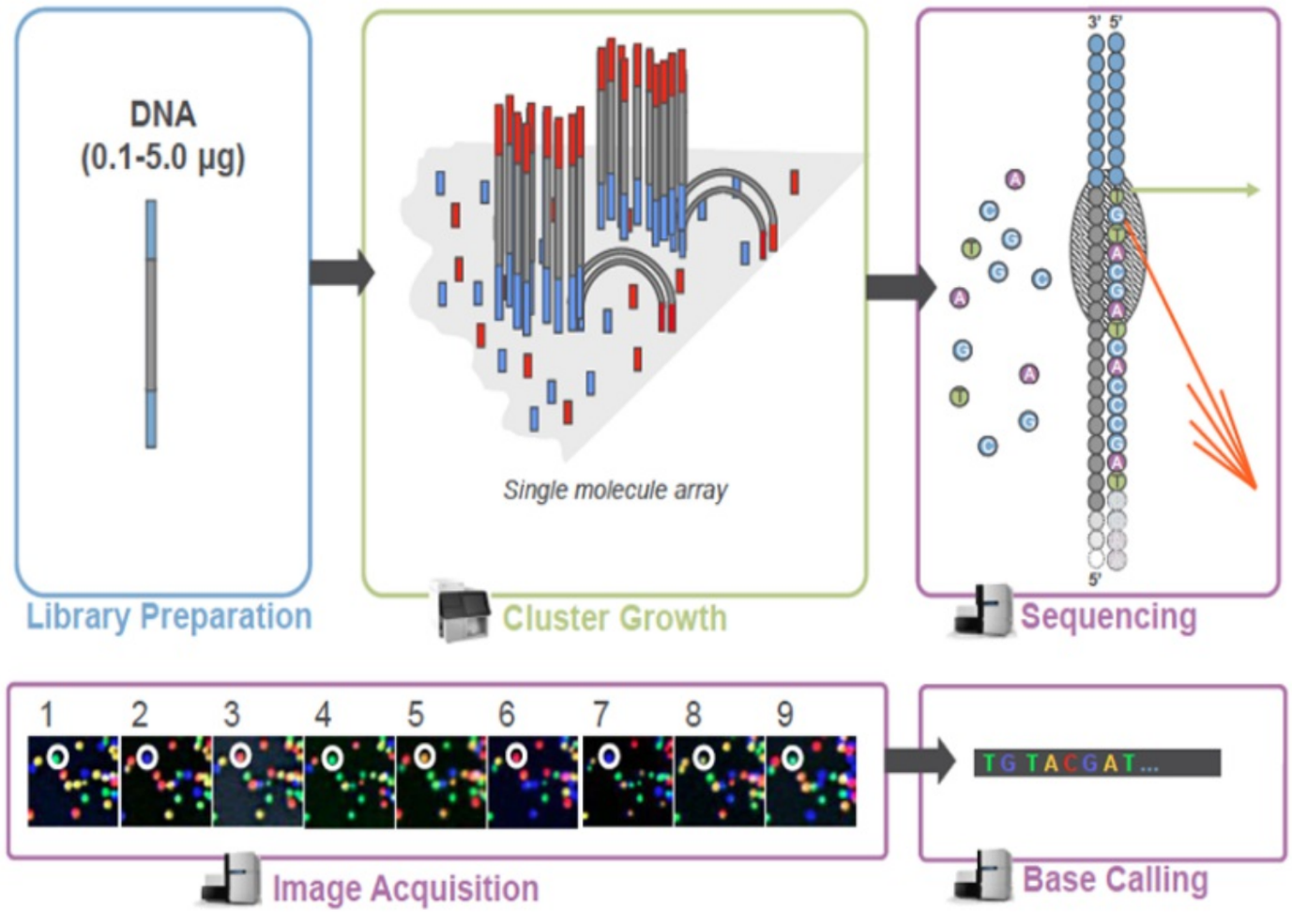
注：文库蓝色及黄色部分为测序接头P5和P7，10X barcode及sample index分别用于区分细胞及样本，Read 1N及Read2N则是测序引物结合区域。

2.2 文库质检

文库构建完成后，先使用Qubit2.0 Fluorometer进行初步定量，稀释文库至1.5ng/ul，随后使用Agilent 2100/4200 bioanalyzer对文库的insert size进行检测，insert size符合预期后，qRT-PCR对文库有效浓度进行准确定量（文库有效浓度高于2nM），以保证文库质量。

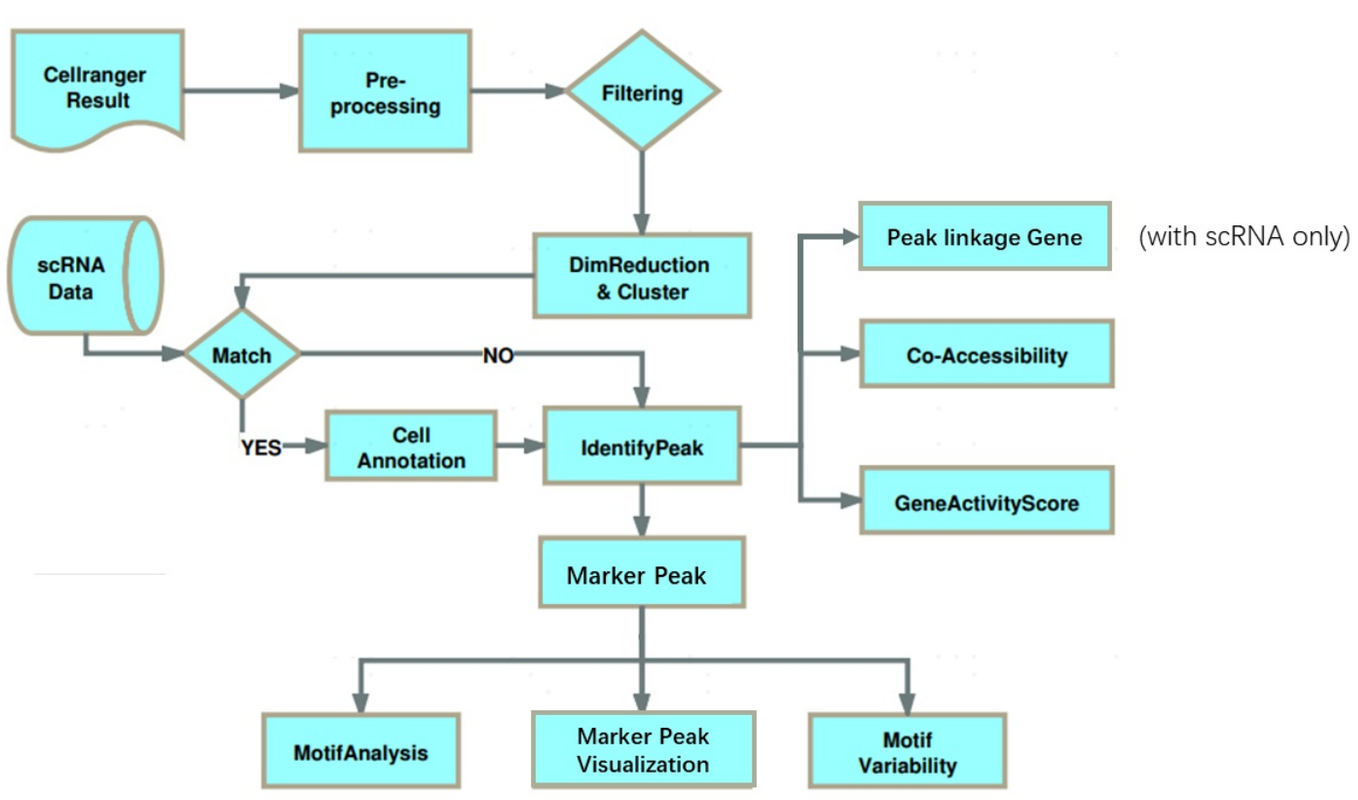
2.3 上机测序

库检合格后，把不同文库按照有效浓度及目标下机数据量的需求pooling后进行Illumina测序。测序的基本原理是边合成边测序（Sequencing by Synthesis）。在测序的flow cell中加入荧光标记的dNTP、DNA聚合酶以及接头引物进行扩增，在每一个测序簇延伸互补链时，每加入一个被荧光标记的dNTP就能释放出相对应的荧光，测序仪通过捕获荧光信号，并通过计算机软件将光信号转化为测序峰，从而获得待测片段的序列信息。测序过程如下图所示：



3 信息分析流程

对于单细胞ATAC-seq测序数据，新格元使用如下分析流程，此流程基于10X的cellranger-atac分析结果进行后续分析，因此无数据质控、比对等结果：



如流程图所示，根据cellranger的分析结果进行后续分析，主要进行数据过滤、细胞分群cluster、cluster间聚类分析、细胞分型降维展示、marker基因展示、peak calling、cluster差异可及性分析、共可及性分析、GeneActivityScore分析、motif鉴定、motif富集分析（GO富集）、cluster差异motif分析。

若有相对应的单细胞转录组数据，可基于scRNA分析进行注释：如使用seurat的FindTransferAnchors和TransferData函数，或使用数据库（新格元数据库singleron、公共数据库CellMarker/panglaodb等）完成注释。

3.1 数据质控分析

3.1.1 测序数据说明

fastq数据结构如下所示：

```
fastq_path
|-- Rports
|-- Stats
|-- test_sample_S1_L001_I1_001.fastq.gz
|-- test_sample_S1_L001_R1_001.fastq.gz
|-- test_sample_S1_L001_R2_001.fastq.gz
|-- test_sample_S1_L001_R3_001.fastq.gz
```

（1）I1: Dual index i7 read (optional)

（2）R1: Read 1

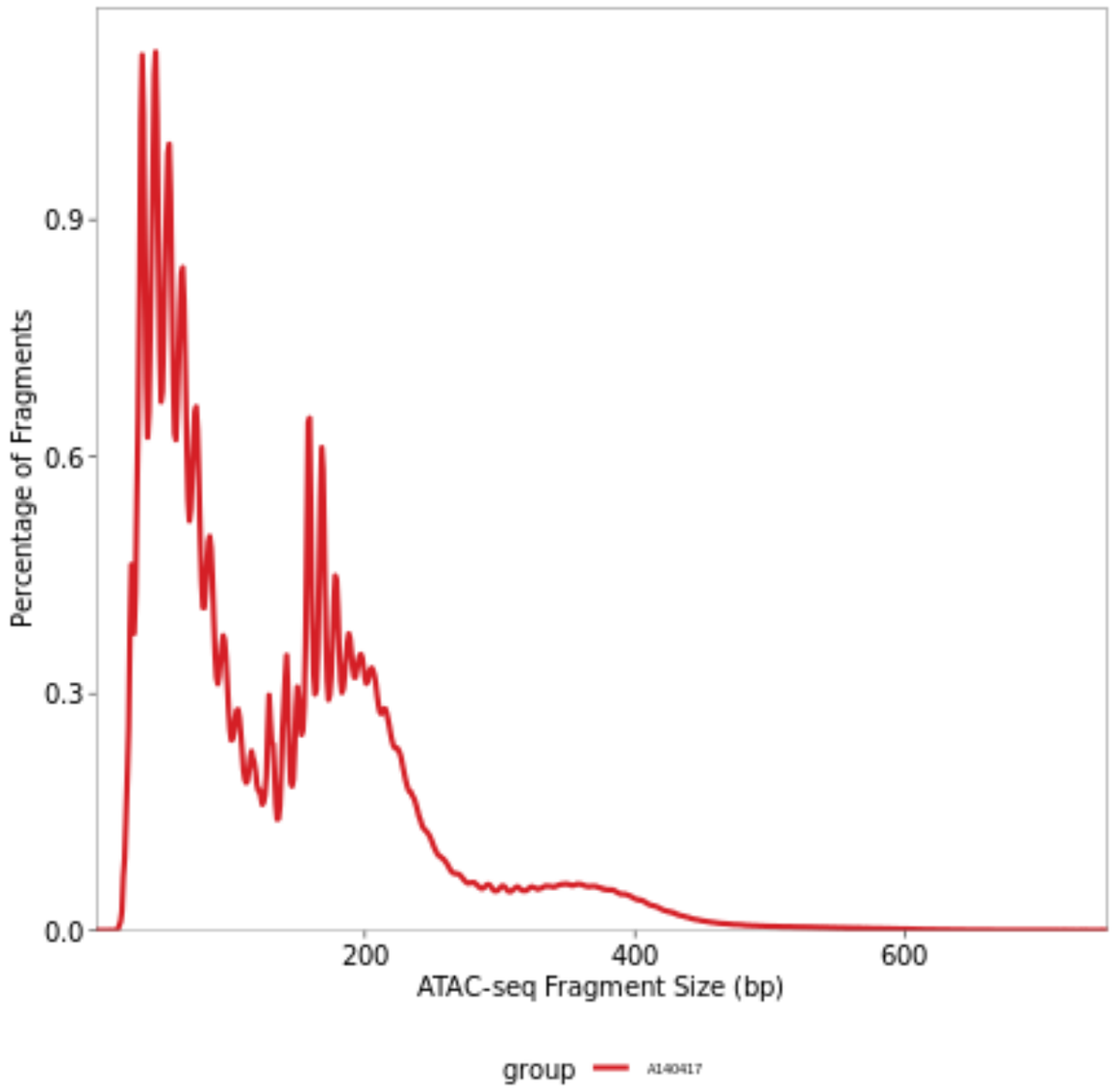
（3）R2: Dual index i5 read

（4）R3: Read 2

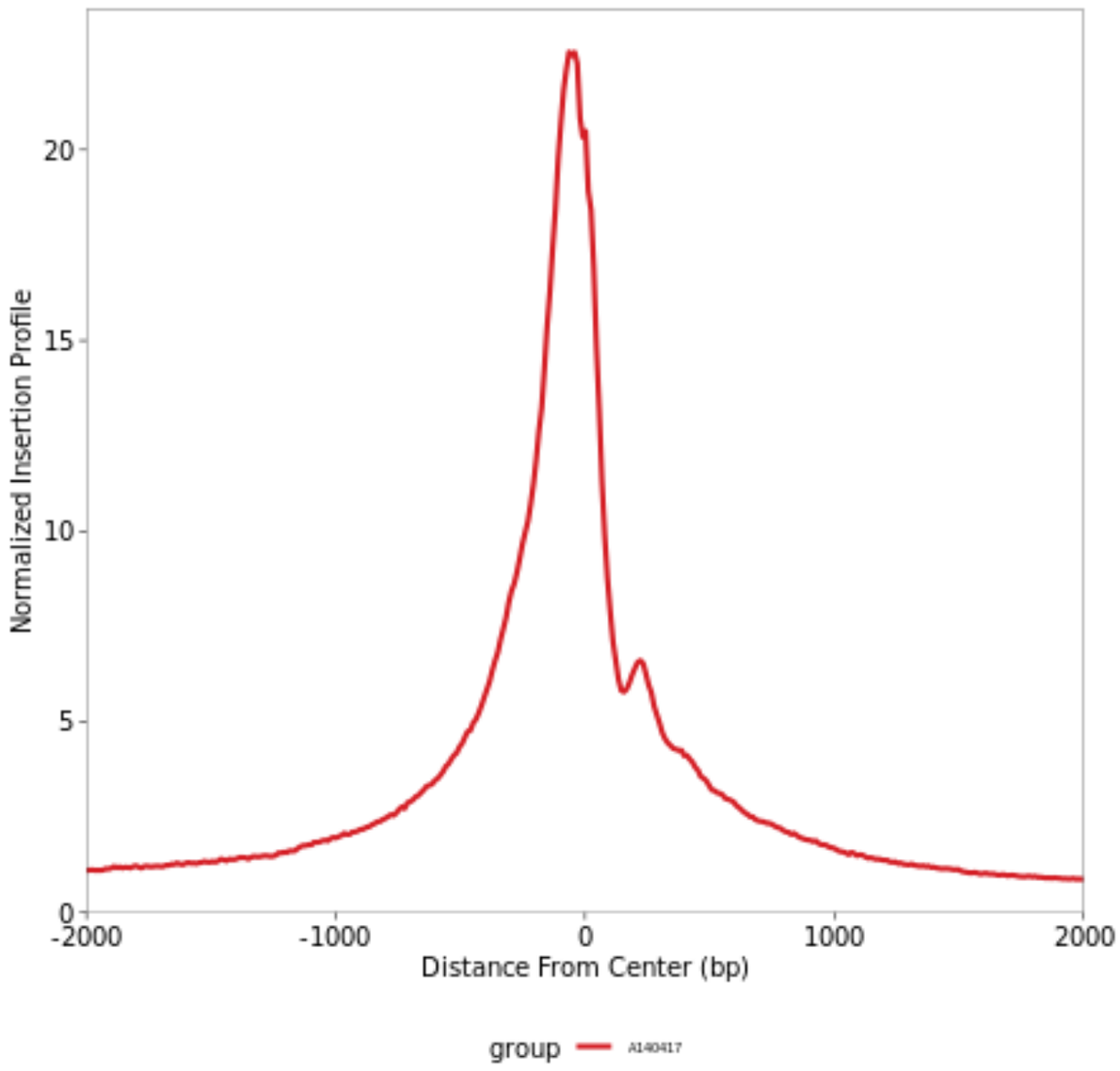
3.1.2 **Barcode** 质控分析

为了保证后续分析的准确性，需要对细胞及fragment进行过滤。首先，过滤去除在所有实验中都高度活跃的fragment，以此去除背景噪音；其次需要挑选高质量的细胞，默认使用cellranger过滤得到的细胞（通过fragment占peak区域的比例等参数选取真正的细胞），当然也可以不依据cellranger结果，根据两个指标（unique fragments数及在转录起始位点富集分数）去除指标数值过低的细胞。样本细胞的fragments分布，unique fragments数及TSS enrichment score的分布如下所示：

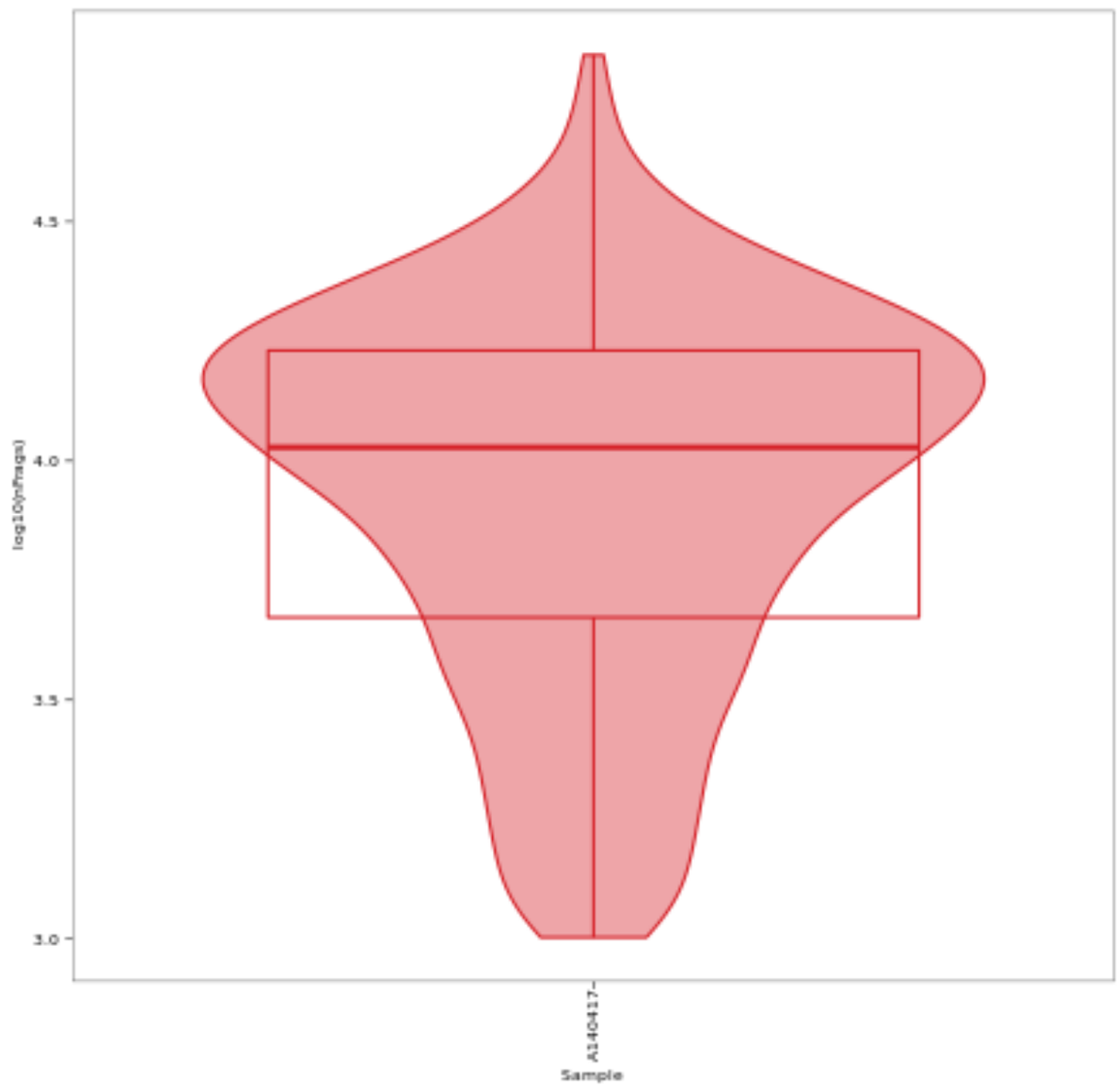
结果位于XX_ATAC_result/Plots/00.QCmetrics。



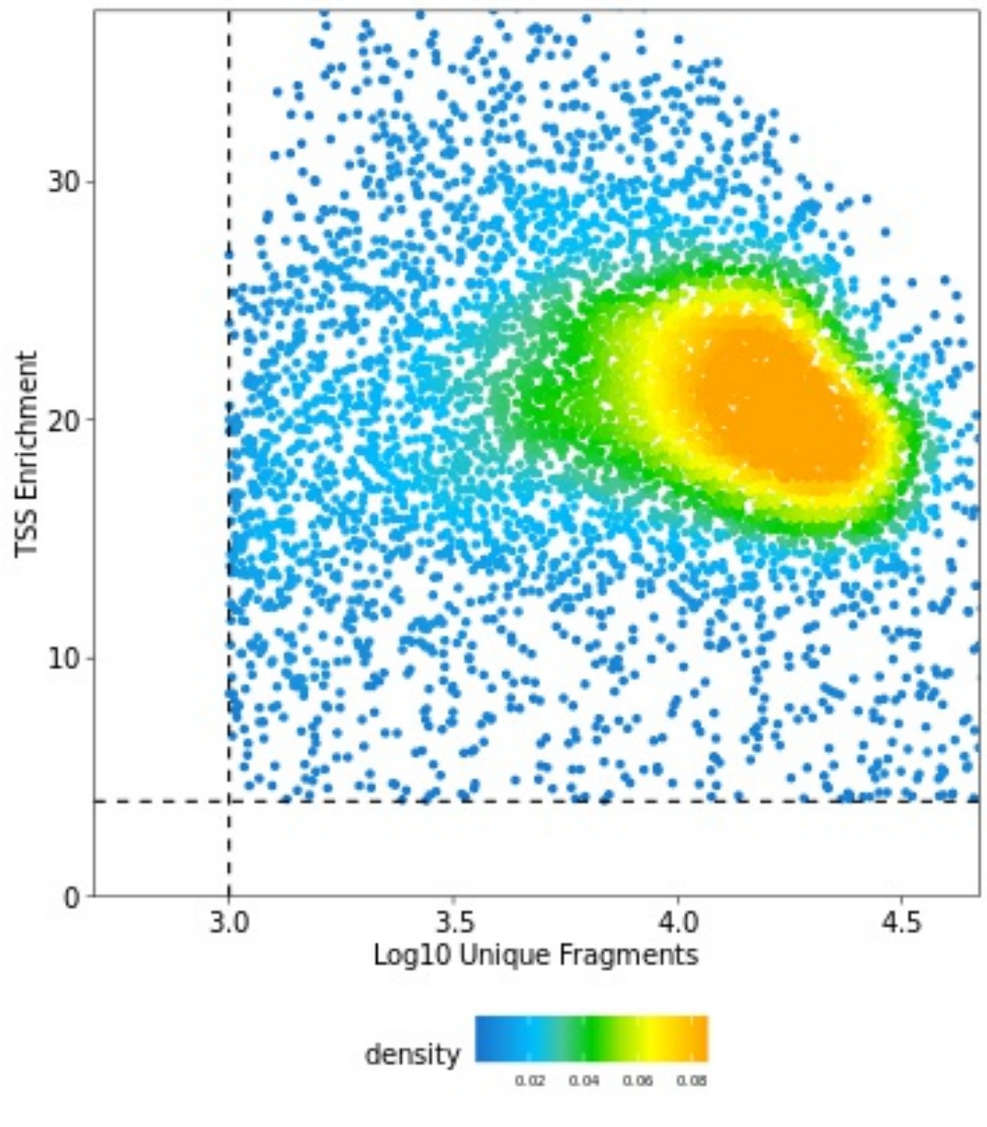
注：展示样本细胞不同长度的fragments分布情况。



注：展示样本细胞TSS富集情况。



注：展示细胞的两个指标数量分布情况。



注：展示细胞在这两个指标下的质量分布情况，颜色由蓝到黄细胞分布密度越大。

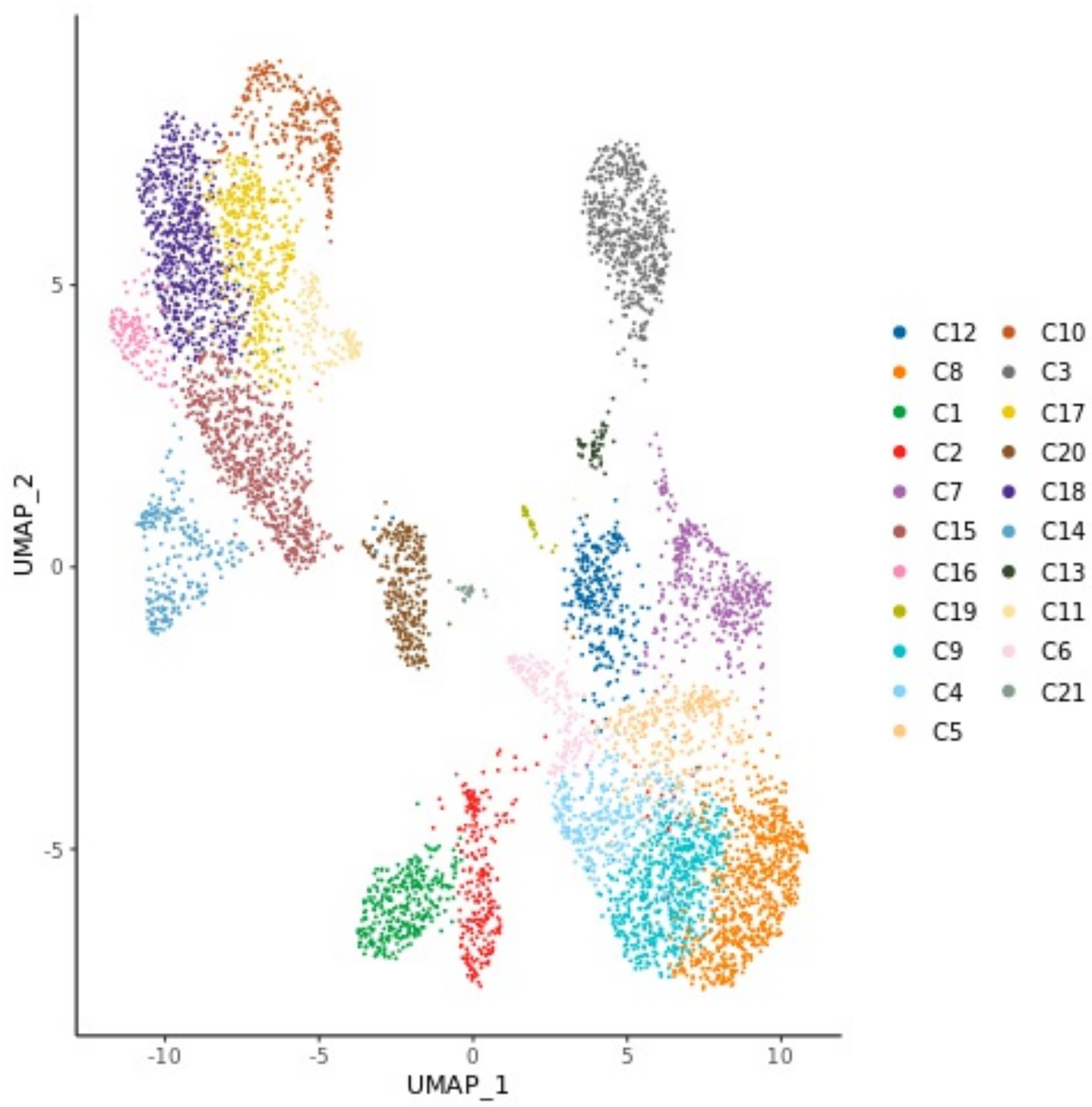
3.2 细胞分群

ArchR软件默认使用LSI算法进行降维，再通过Graph-based的方法进行细胞聚类分群。

结果位于XX_ATAC_result/Plots/01.DimReduction (如果有scRNA 添加注释则注释结果输出到02.Cellannotation文件夹)。

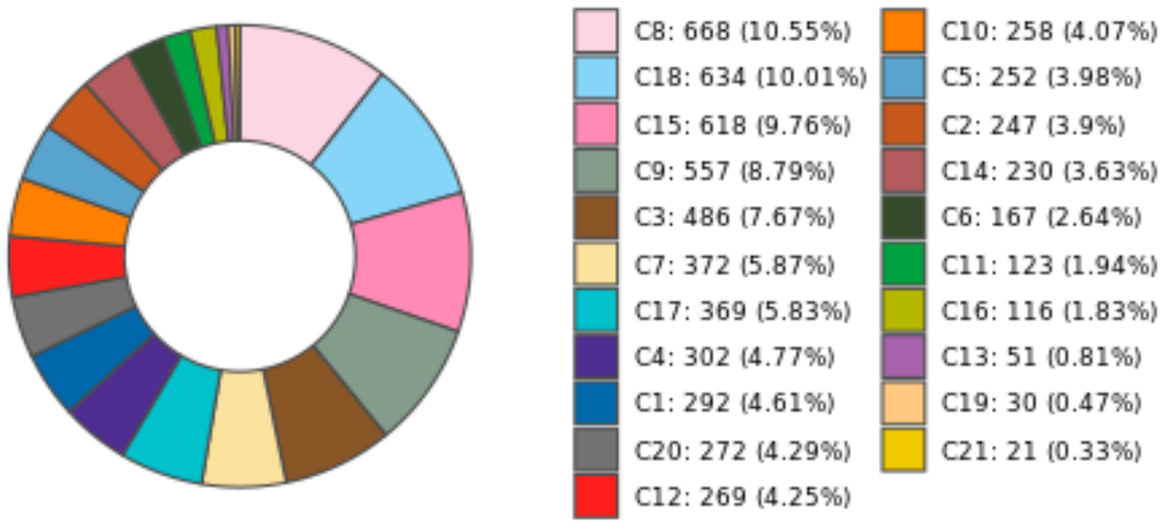
3.2.1 细胞分群图

为了进一步了解细胞之间的异质性，ArchR 采用潜在语义索引法（LSI）对细胞进行分群，再进行tSNE与UMAP降维分析用于可视化。报告中只展示UMAP的降维情况：



3.2.2 细胞cluster占比信息

流程会输出各个cluster的占比信息并进行可视化：



3.2.3 细胞cluster信息

流程会输出各个样本的分群信息以供后续分析需求，部分展示如下：

X	A140417
C12	269
C8	668
C1	292
C2	247

注：展示样本中cluster数量分布情况。

3.3 Peak calling

Peak指的是fragments富集的区域，即基因组上开放染色质富集的区域，这些区域在转录调控等方面可能发挥作用。ArchR调用MACS2软件进行peak calling，再将每个cluster得到的peak合并，生成cell X peak矩阵。Peak calling的结果以xls格式进行展示,结果如下，其它详见结果文件：

结果位于XX_ATAC_result/Plots/03.1.MacsCallPeak。

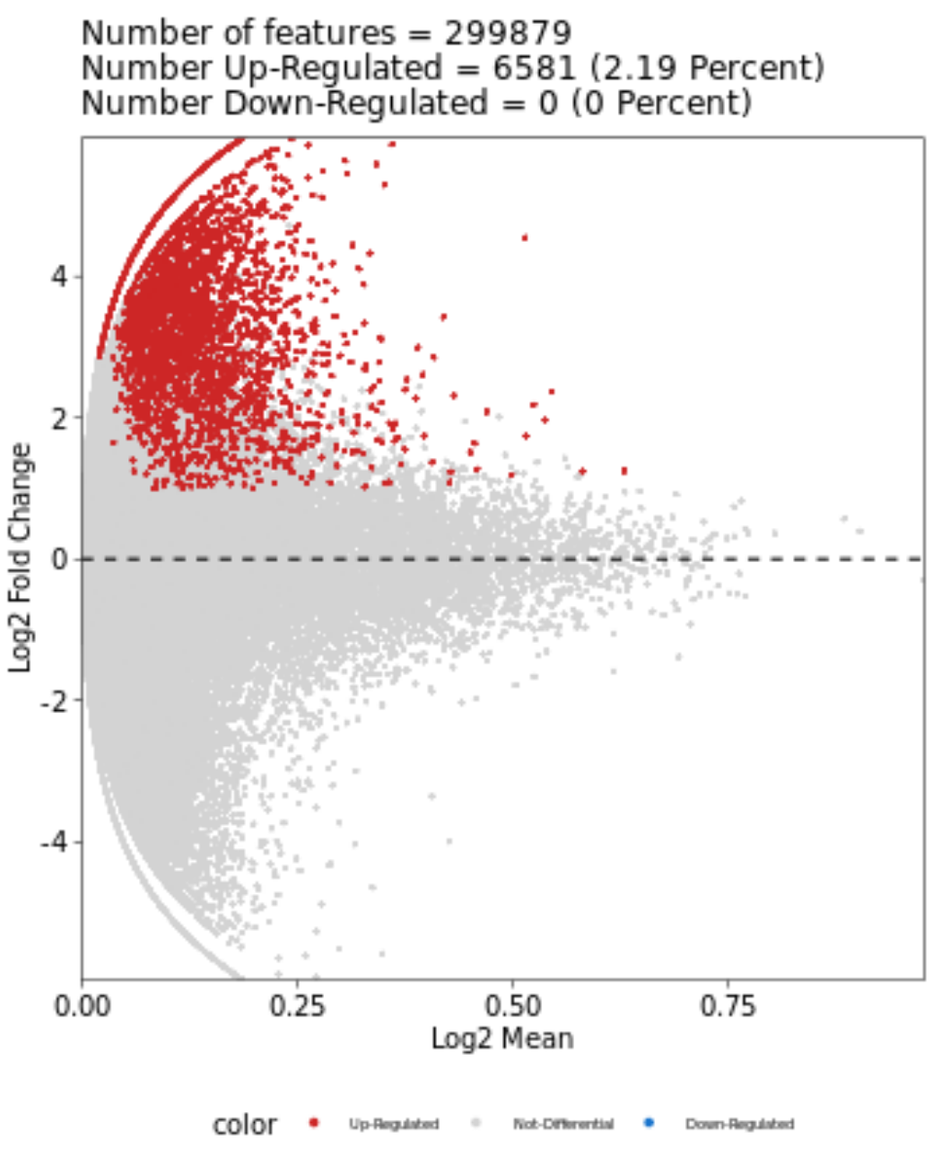
celltype	seqnames	start	end	width	distToGeneStart	nearestGene	peak
C12	chr1	9877356	9877856	501	31031	Mcmdc2	Intro
C12	chr1	16809302	16809802	501	121095	Ly96	Dista
C12	chr1	20933404	20933904	501	17971	Efhc1	Intro

（1）celltype: 细胞类型
（2）seqnames：开放区域染色体位置
（2）start：peak起始位点
（3）end：peak终止位点
（4）width: peak 长度
（5）distToGeneStart: 与基因起始位置的距离
（6）nearestGene：Peak 临近基因
（7）peakType：Peak 注释类型
（8）distToTSS：与基因 TSS

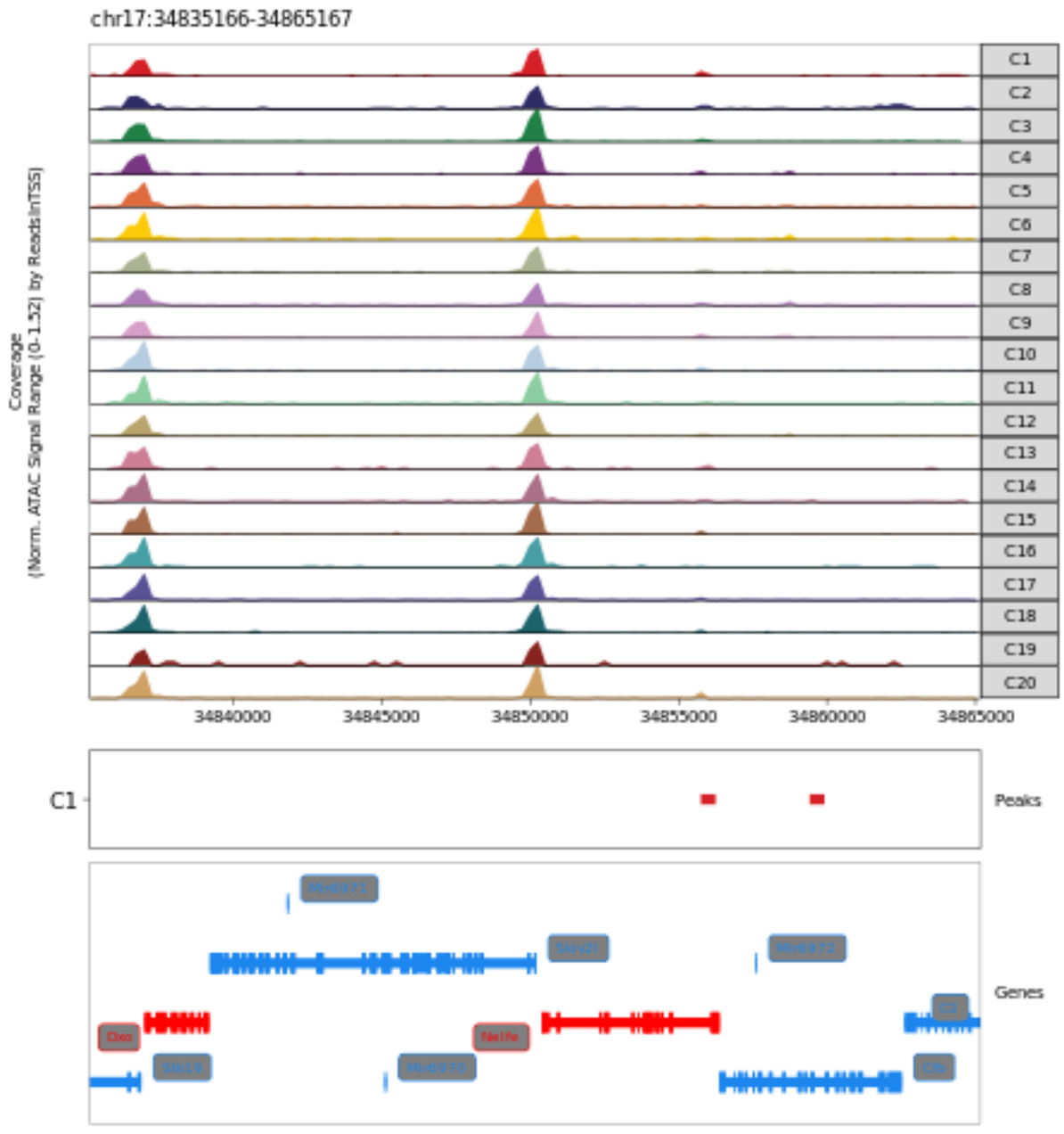
3.4 Marker peak分析结果展示

ArchR分析软件将细胞降维聚类后，利用得到的peak矩阵，找到各cluster特异的marker peak。

结果位于XX_ATAC_result/Plots/03.2.Markerpeak。



注：基于降维聚类的结果，展示样本中各cluster的marker peak的MA图分布情况。



注：展示各cluster中marker gene相关的peak在所有cluster的分布情况。

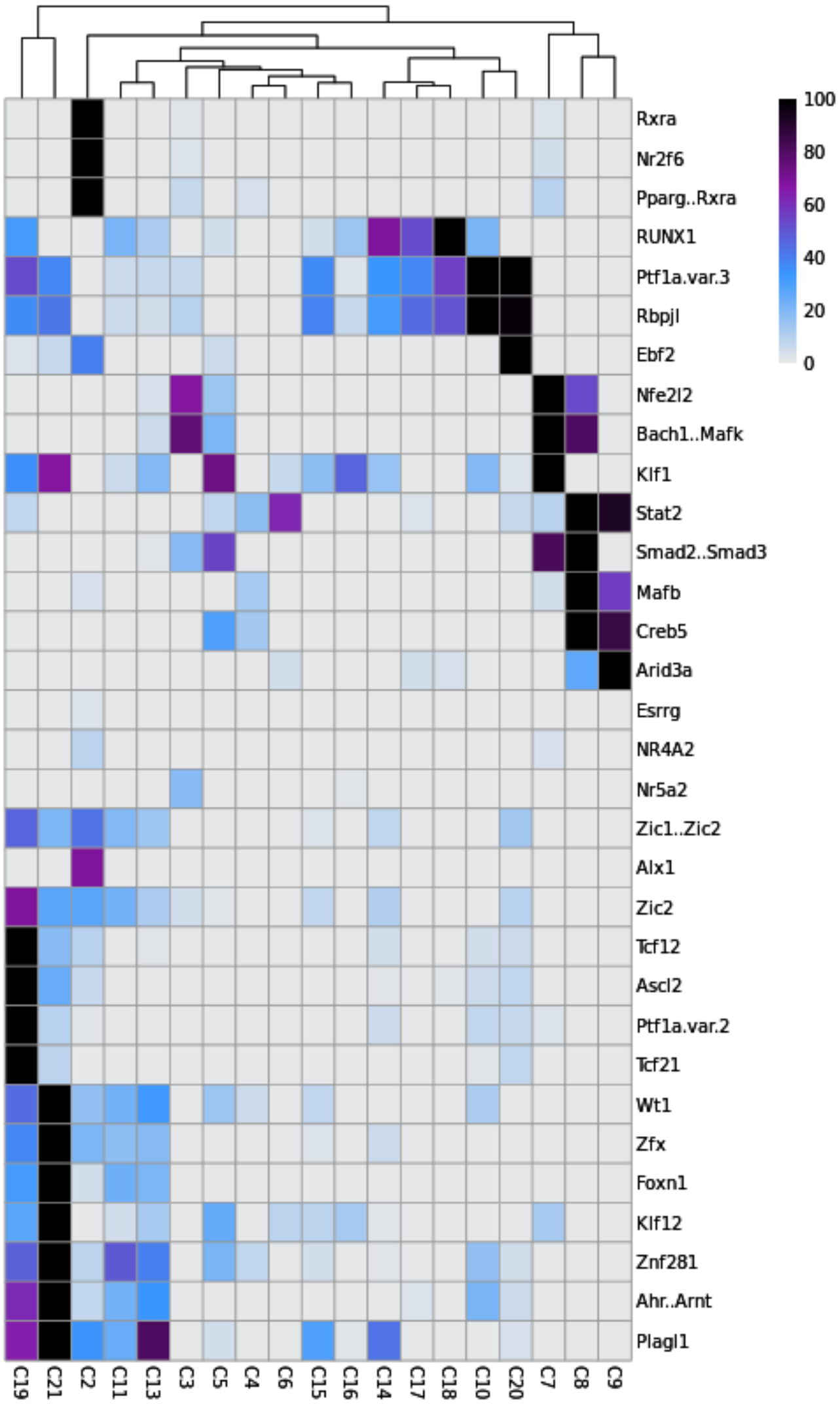
3.5 Motif分析

ArchR提供多种富集分析，一种是通过对各个cluster的marker peak进行motif富集分析，来预测活跃的染色质开放区段和转录因子特异结合的事件，例如，我们经常发现在cluster特异性的可及染色质区域中富集了关键的TFs; 一种是计算感兴趣的细胞或样本之间每个motif或注释的变异性，进一步实现对样本聚类，研究细胞或样本间的相似性，可及性与变异性的差异和motifs间的相关性。

3.5.1 MarkerPeak的Motif鉴定

基于对各个cluster的Marker Peak的分析，我们可以在各cluster中寻找在特异的的peak富集的motif。

结果位于XX_ATAC_result/Plots/04.MarkerpeakMotifEnrichment。

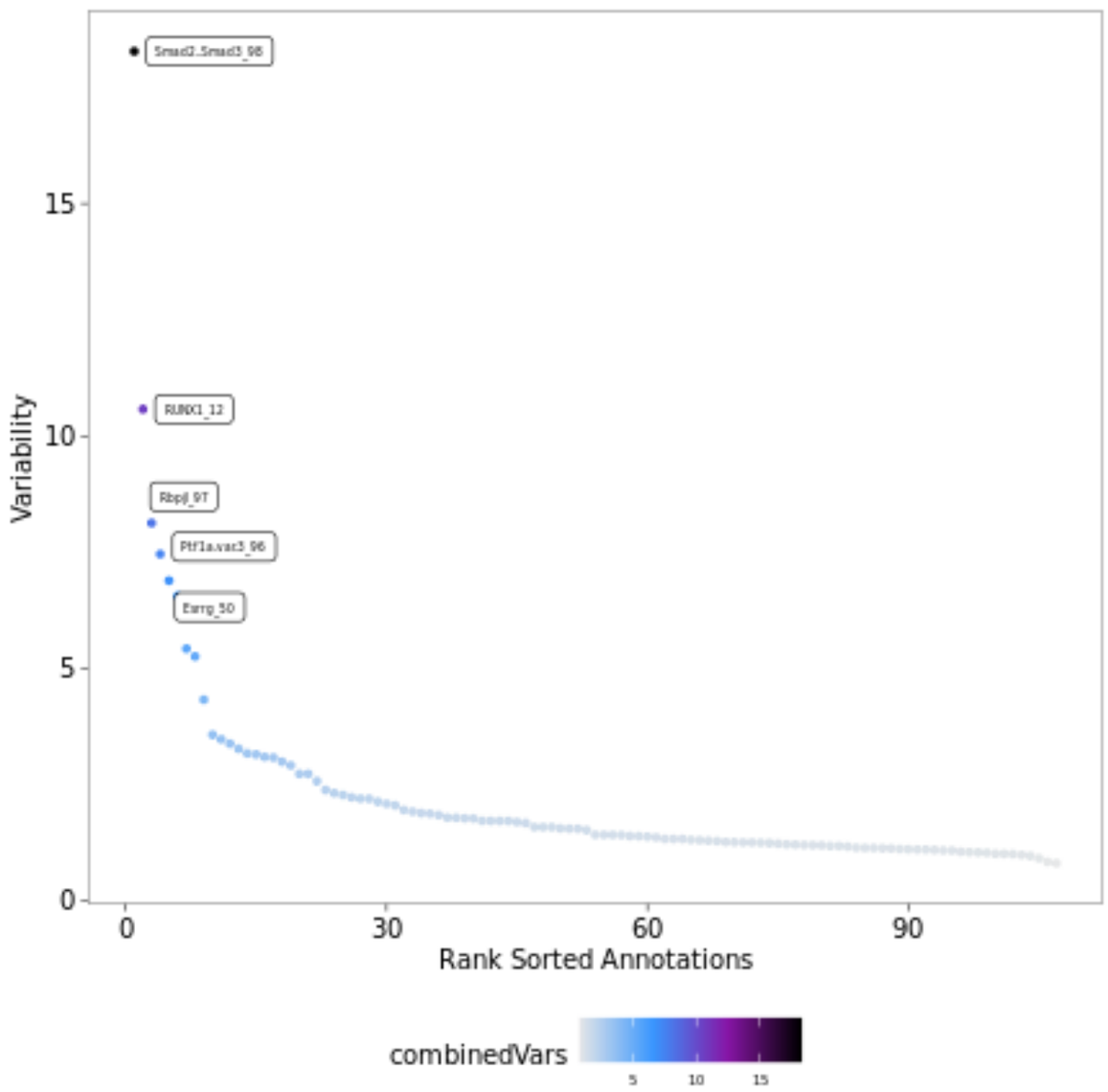


注：motif富集热图，根据它们的富集的显著性为它们着色。

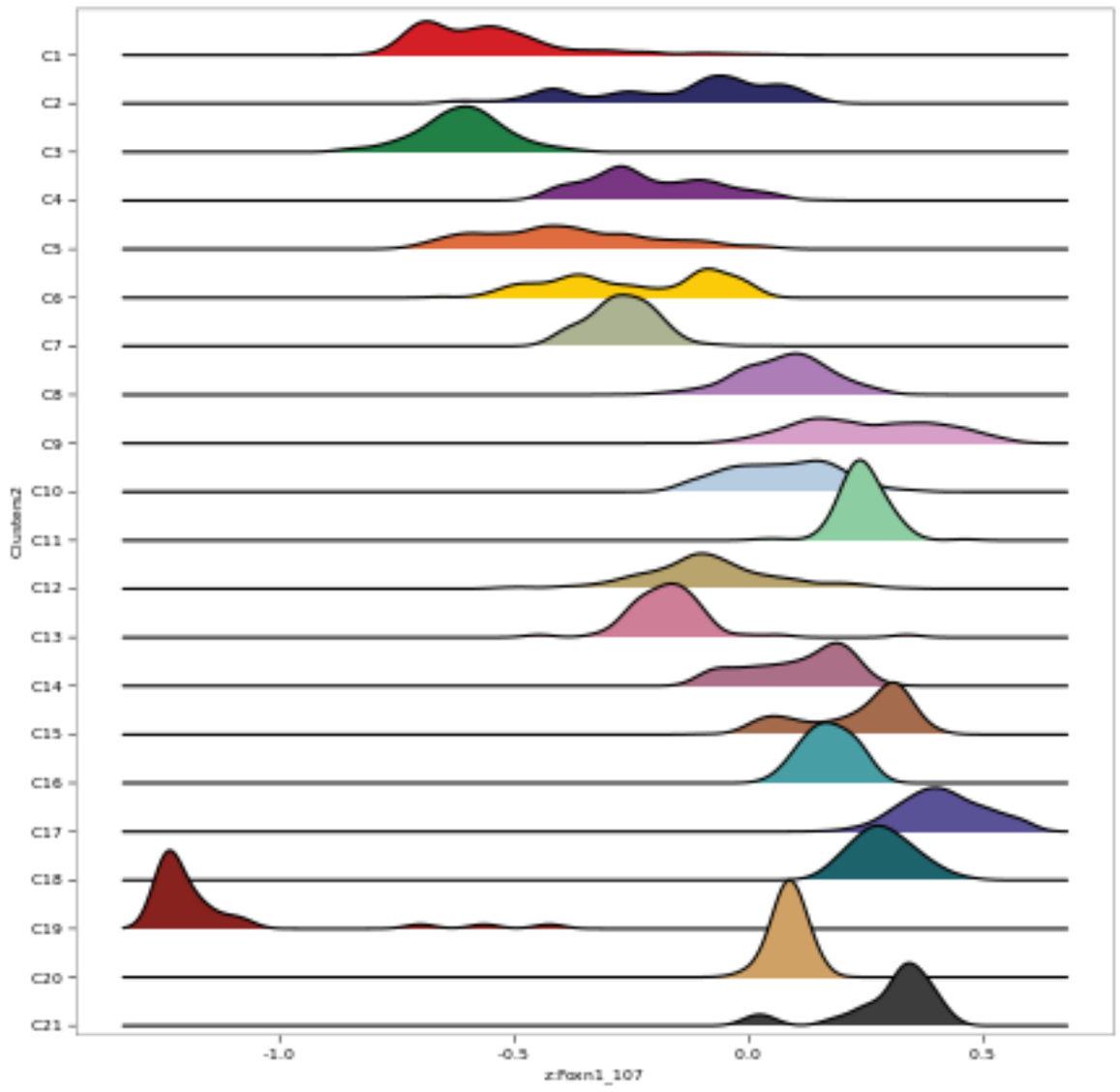
3.5.2 **ChromVAR Motif可变性分析**

以每个细胞为基础计算，考虑到偶Tn5转座酶的插入序列偏倚，在稀疏染色质可及性矩阵中预测富集的转录因子活性。

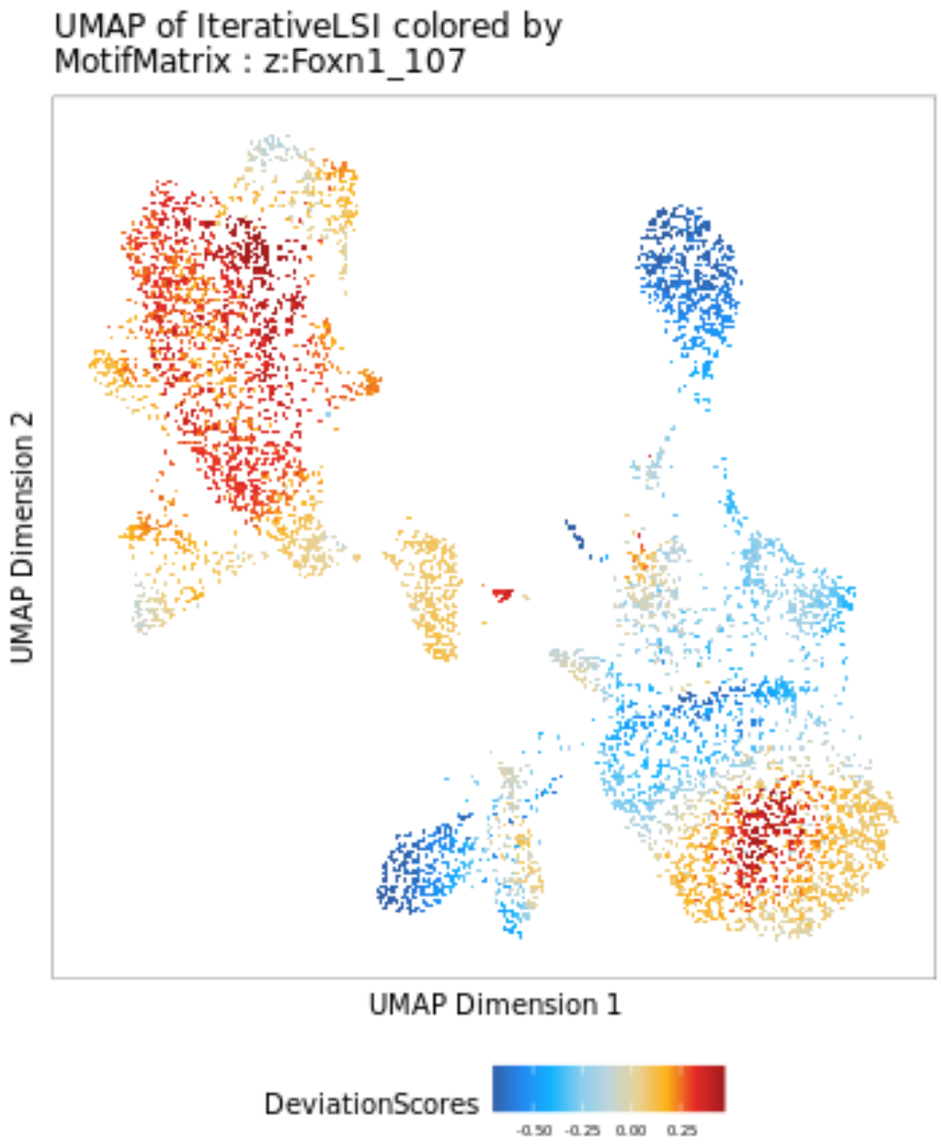
结果位于XX_ATAC_result/Plots/05.ChromVAR.Variable.motif。



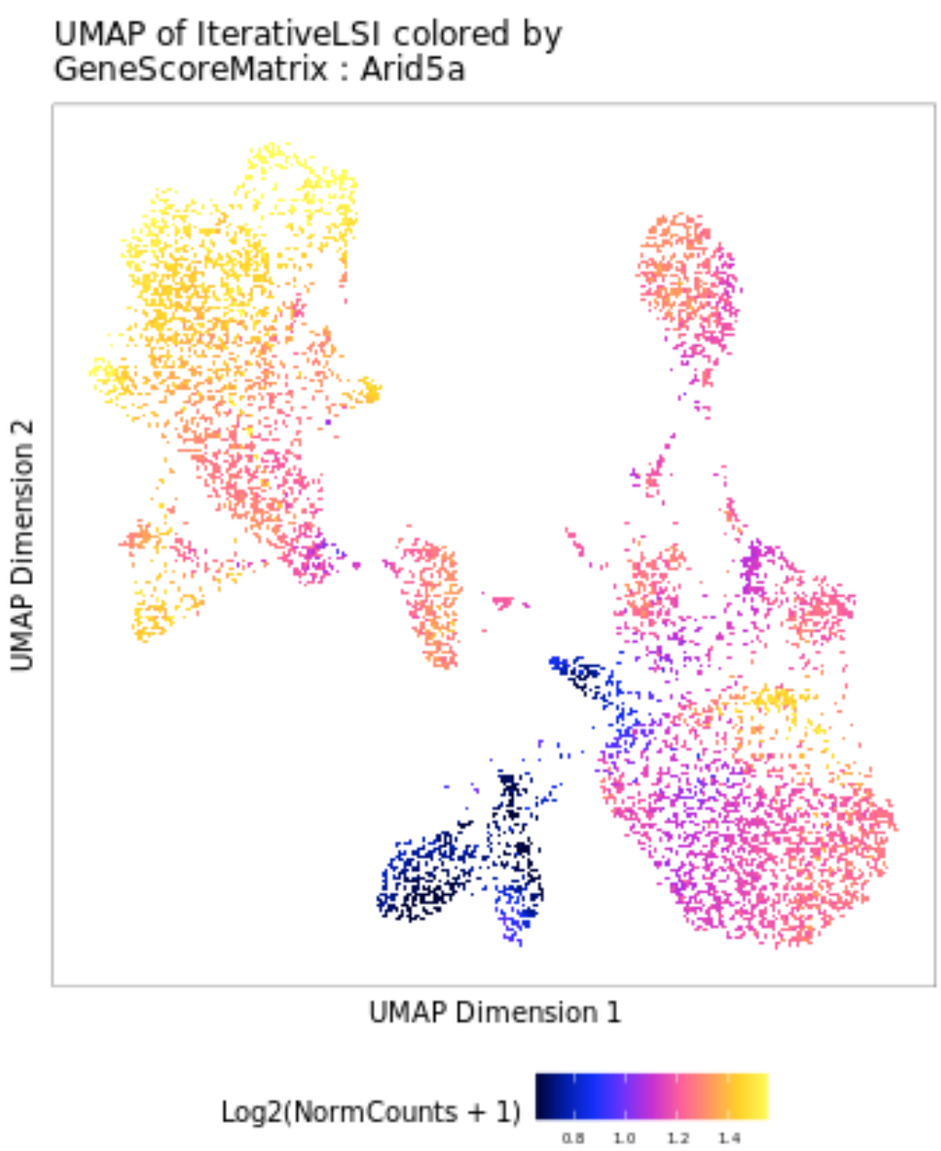
注：基于motif注释里每个细胞的偏差建立偏差矩阵，图中展示的偏差越大代表变异越明显。



注：从motif的偏差分析中挑选出变异性最大的motif,绘制出每个cluster的ChromVAR偏差分数的分布。



注：将最显著的motif的偏差z-score分布展示在UMAP中。



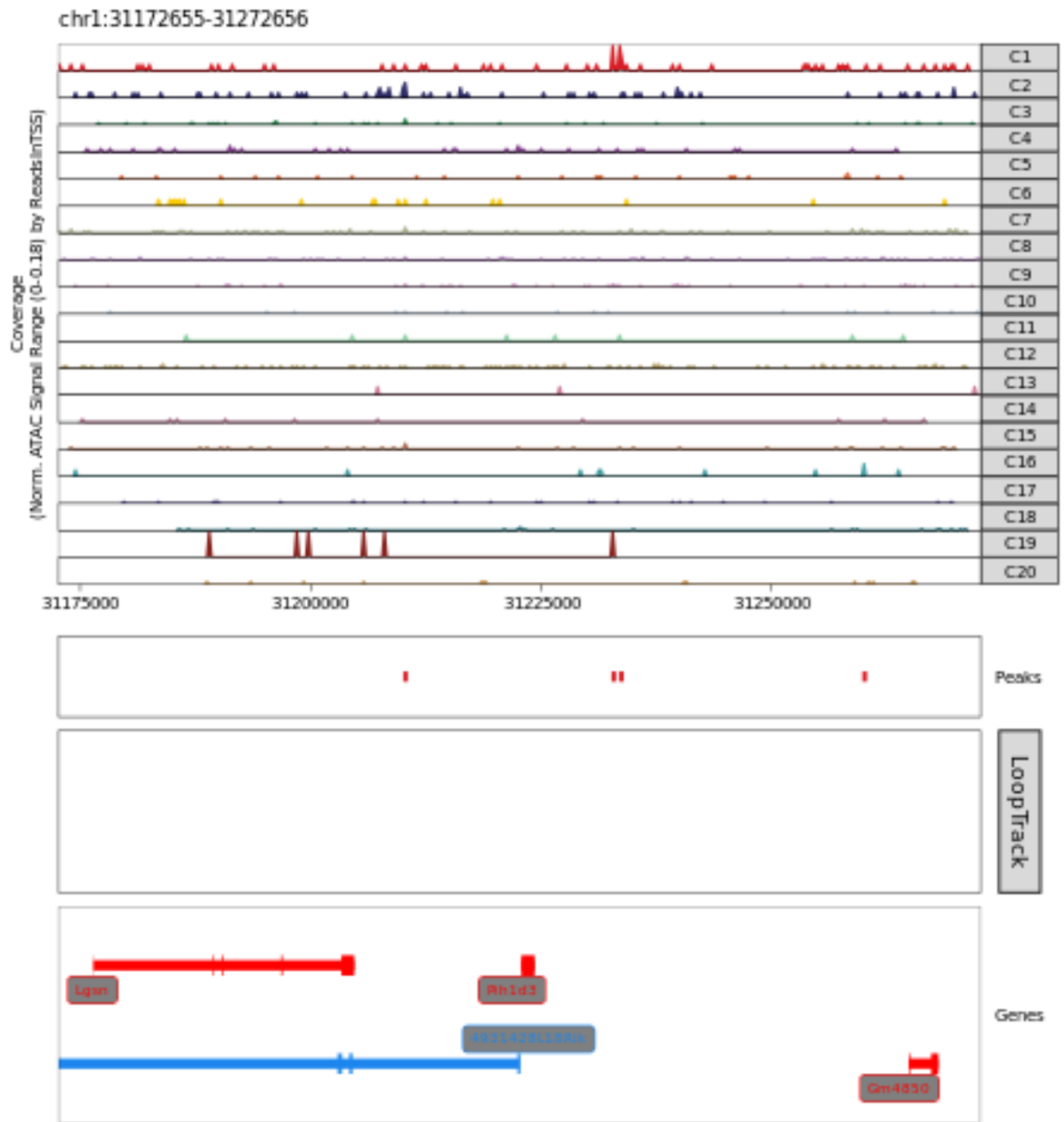
注：为了了解这些TF偏差z-score和推断的相应TF基因的gene score间的关系，将显著变异的TF的基因评分叠加在UMAP图上。

3.6 共可及性分析

共可及性是指跨多个单细胞的两个峰间的可及性相关性，及当A峰在单细胞内可接近的时候，B峰通常也可接近。有时会将特定细胞类型的峰标识为共可接近的，这是因为这些峰在特定的细胞类型中（相比较其他细

胞类型）展现出了较强的关联性，但这不一定意味着这些峰之间存在调控关系。

结果位于XX_ATAC_result/Plots/06.Co-accessibility。



注：browser tracks图展示了各cluster的共可及区域。

4 备注

4.1 结果文件解压方法

压缩格式	用户类型	方法
***.tar.gz形式的压缩文件**	Unix/Linux/Mac用户	使用tar -zxvf *.tar.gz命令
	Windows用户	使用解压缩软件如WinRAR、7-Zip等
***.gz形式的压缩文件**	Unix/Linux/Mac用户	使用gzip -d *.gz命令
	Windows用户	使用解压缩软件如WinRAR、7-Zip等
***.zip形式的压缩文件**	Unix/Linux/Mac用户	使用unzip *.zip命令
	Windows用户	使用解压缩软件如WinRAR、7-Zip等

4.2 结果文件格式说明

文件类型	文件描述	打开方式
***.fasta**	序列文件，fasta格式，一般为基因序列或者基因组序列。因文件一般较大，打开较为困难	unix/Linux/Mac用户使用less或more命令
		windows用户使用高级文本编辑器Editplus/Notepad++等
***.fq/fastq**	序列文件，fastq格式，一般为reads序列；因文件一般较大，打开较为困难	unix/Linux/Mac用户使用 less 或 more 命令
		windows用户使用高级文本编辑器Editplus/Notepad++等
***.xls*,.txt**	结果数据表格文件；文件以制表符Tab分隔	unix/Linux/Mac用户使用 less 或 more 命令
		windows用户使用高级文本编辑器Editplus/Notepad++ 等，也可以用Microsoft Excel打开
***.png**	结果图像文件；位图,无损压缩	unix/Linux/Mac用户使用display命令打开
		windows用户可以使用图片浏览器打开，如photoshop等。
***.pdf**	结果图像文件；矢量图，可以放大和缩小而不失真，方便用户查看和编辑处理，可使用Adobe Illustrator进行图片编辑，用于文章发表等	windows/Mac用户可以使用Adobe Reader/福昕阅读器/网页浏览器等打开
		unix/Linux用户使用evince命令打开

4.3 分析软件列表及版本

分析	软件	版本
单细胞ATAC分析软件	Seurat	4.0.4
	ArchR	1.0.1
	readr	2.0.1
	pheatmap	1.0.12
	tidyr	1.1.3
	dplyr	1.0.7
	chromVARmotifs	0.2.0
	ggplot2	3.3.5

其他分析软件	软件	版本
	R	4.0.5
	perl	5.26.2

5 参考文献

- 1.Buenrostro J D , Giresi P G , Zaba L C , et al. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics[J]. Nature Methods, 2013, 10(12):1213-1218.
- 2.Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. Cell, 2019, 177(7):1888-1902.e21.
- 3.Darren A C, Andrew J H, Delasa A, et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell, 23 August 2018, 174(5):1309-1324.
- 4.Rongxin Fang, Sebastian Preissl, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K. Shiau, Eran A. Mukamel, Yanxiao Zhang, M. Margarita Behrens, Joseph Ecker, Bing Ren. Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types. bioRxiv 615179. (SnapATAC)
- 5.Lafon S , Lee A B . Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning and data set parameterization[C]2006.
- 6.De l P J , Herbst B M , Hereman W , et al. An introduction to diffusion maps[J]. 2008.
- 7.Macosko, Basu, Satija et al., Spatial reconstruction of single-cell gene expression data. Nature Biotechnology volume 33, pages 495–502 (2015).(Seurat)
- 8.Andrew Butler, Efthymia Papalexi ,Rahul Satija et al.,Integrating single-cell transcriptomic data across different conditions, technologies, and species Nature Biotechnology volume 36, pages 411–420 (2018) (Seurat)
- 9.Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL (2014). The dynamics and regulators of cell fate decisions are revealed by pseudo-temporal ordering of single cells. Nature Biotechnology.
- 10.Zhang Y , Liu T , Meyer C A , et al. Model-based Analysis of ChIP-Seq (MACS)[J]. Genome biology, 2008, 9(9).
- 11.Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589
- 12.Etienne Becht, Charles-Antoine Dutertre, Immanuel W.H. Kwok, Lai Guan Ng, Florent Ginhoux, Evan W Newell.Evaluation of UMAP as an alternative to t-SNE for single-cell data. BioRxiv .(UMAP for single-cell data)
- 13.Leland McInnes, John Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426. (UMAP)
- 14.Jia Pan & Dinesh Manocha. Bi-Level Locality Sensitive Hashing for K-Nearest Neighbor Computation,2012.
- 15.Peter Hall,Byeong U.Park & Richard J. Samworth. Choice of Neighbor Order In Nearest Neighbor Classification,2008.
- 16.Schep A N , Wu B , Buenrostro J D , et al. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data[J]. Nature Methods, 2017.
- 17.Buenrostro, J.D., et al., Single-cell chromatin accessibility reveals principles of regulatory variation. Nature, 2015. 523(7561): p. 486-90.
- 18.Pliner H A , Packer J S , McFaline-Figueroa José L, et al. Cicero Predicts, cis -Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data[J]. Molecular Cell, 2018:S1097276518305471-.

- 19.Luo W, Brouwer C. Pathview: an R/Biocondutor package for pathway-based data integration and visualization. Bioinformatics, 2013, 29(14):1830-1831, doi: 10.1093/bioinformatics/btt285.
- 20.Luo W, Pant G, Bhavnasi YK, Blanchard SG, Brouwer C. Pathview Web: user friendly pathway visualization and data integration. Nucleic Acids Res, 2017, Web Server issue, doi: 10.1093/nar/gkx372.
- 21.Mclean C Y , Bristor D , Hiller M , et al. GREAT improves functional interpretation of cis-regulatory regions[J]. Nature Biotechnology, 2010, 28(5):495-501.
- 22.Granja, J.M., Corces, M.R., Pierce, S.E. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet 53, 403–411 (2021). https://doi.org/10.1038/s41588-021-00790-6 (https://doi.org/10.1038/s41588-021-00790-6). (ArchR)
- 23.Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods 14, 975–978 (2017). (chromVAR)